



An Investigation into the Generalisability of Fake News Detection Models

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy (Ph.D.)

of the

Department of Computer Science,
Brunel University London

by

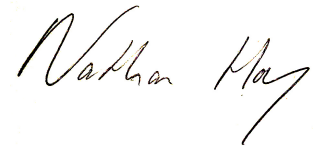
Nathaniel Hoy

Principal Supervisor:	Dr Theodora Koulouri
Supervisor:	Dr Stephen Swift
Submission Date:	2024-11-30

Declaration

I hereby declare that this thesis, titled ‘An Investigation into the Generalisability of Fake News Detection Models’ is the result of my own original research. It has been composed by me and has not been submitted, in whole or in part, for any degree or professional qualification at any other university or institution.

Where other sources of information have been used, they are cited in accordance with standard academic practices. All efforts have been made to ensure that the research presented herein adheres to the ethical and professional standards expected in the field and the University.

A handwritten signature in black ink, appearing to read 'Nathaniel Hoy', with a stylized, cursive script.

Nathaniel Hoy
November 2024

Publications

The research presented in this thesis has resulted in the following papers, which have either been published or are currently under submission for publication:

1. N. Hoy and T. Koulouri, “Exploring the Generalisability of Fake News Detection Models” *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, 2022, pp. 5731-5740 (**published**)
2. Hoy, N. and Koulouri, T., 2025. An exploration of features to improve the generalisability of fake news detection models. *Expert Systems with Applications*, p.126949. (**published**)
3. N. Hoy and T. Koulouri, “Machine Learning Approaches to Detect Fake News: A Systematic Review”, *IEEE Access*, 2025 (**under review**)

Abstract

Fake news has emerged as a significant societal challenge, influencing public discourse, spreading disinformation, and eroding trust in democratic institutions. While supervised machine learning has become the predominant approach to addressing this issue, existing methods often struggle with generalisability. These limitations stem from an overreliance on coarsely labelled datasets, which fail to capture nuanced distinctions between fake and real news, and the widespread use of token-based features, such as Bag-of-Words, TF-IDF, Word2Vec, and BERT. These features, while effective within specific datasets, are highly sensitive to dataset biases and source-specific patterns. Traditional evaluation techniques, such as hold-out testing and K-fold cross-validation, exacerbate this issue by assuming the data is representative, an assumption often invalid when models are tested against real-world data.

This thesis addresses these limitations by exploring strategies to enhance the generalisability of fake news detection models. It proposes the use of stylistic features, which focus on linguistic characteristics such as sentence structure, punctuation, readability, and persuasive language. These features are less reliant on specific word patterns and more robust to source biases. Additionally, the thesis introduces a novel set of ‘social-monetisation’ features to capture the economic motivations behind fake news. These include the presence of advertisements, social media share buttons and affiliate links. Together, these features offer a new perspective on detecting disinformation by focusing on the financial incentives driving its production.

To assess generalisability, the research combines K-fold cross-validation with external validation. In this approach, models are tested internally within each fold and externally on a manually labelled dataset after every fold. This dual framework ensures performance is rigorously evaluated under both experimental conditions and real-world scenarios. By combining these strategies, the research addresses the shortcomings of traditional methods, providing a robust understanding of generalisability.

Results demonstrate that token-based models, while effective within specific datasets, perform poorly in cross-dataset scenarios. In contrast, stylistic and social-monetisation features show greater resilience to dataset-specific biases and provide a more nuanced understanding of fake news characteristics. External validation further highlights the importance of evaluating models on diverse data to assess real-world performance.

This research advances fake news detection by identifying the limitations of current approaches, proposing robust feature sets, and advocating for rigorous evaluation methods. Specifically, it has made four key contributions: demonstrating

the advantages of stylistic features in improving fake news detection, introducing a novel category of features focused on social dissemination behaviors and economic incentives, developing a reduced and simplified feature set to enhance generalisability and efficiency, and establishing a novel evaluation framework for assessing model performance in this domain.

Acknowledgements

This thesis would not have been possible without the guidance, support, and encouragement of many individuals to whom I am deeply grateful.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Theodora Koulouri, for her invaluable guidance, patience, and expertise throughout my time at Brunel University. Her support has been pivotal not only during the course of this PhD but also during my undergraduate studies. Her mentorship has inspired me to achieve my best at every stage of my academic journey and her kindness has been invaluable during times of personal difficulty.

I would also like to extend my deepest thanks to my second supervisor, Dr. Stephen Swift, and Prof. Rob Macredie for their insightful feedback and guidance, particularly in relation to the research papers I have submitted for publication. Their expertise and encouragement have greatly enriched this work and strengthened its contributions. I would additionally like to extend my thanks to Dr. Mahir Arzoky and Dr. Zear Ibrahim for their steadfast support and invaluable advice, which have motivated me to persevere throughout this process.

I am also profoundly grateful to my family, whose unwavering love and support have sustained me throughout this challenging yet rewarding journey. My Mum's steadfast faith in me and her guidance during moments of self-doubt have been a pillar of strength, while my Dad's unique perspectives have helped shape and influence this research. My step-father, Darren, has also provided a constant source of positivity into my life, providing much-needed reassurance during difficult times. I am especially thankful to my younger brother, Stephen, whose own PhD journey has given me a close family member who truly understands the highs and lows of this process and has been a source of invaluable support throughout.

I am also deeply appreciative of countless colleagues and peers at Brunel University, whose camaraderie and collaboration have enriched my research experience. The stimulating discussions, shared challenges, and mutual support have been a vital part of this journey, and I am grateful to have been part of such a vibrant academic community. Having been a student at Brunel since 2016, I have had the privilege of growing both academically and personally within this supportive environment, which has played an integral role in shaping who I am today.

This thesis represents the culmination of years of effort, and it would not have been possible without the encouragement, guidance, and support of these remarkable individuals.

Thank you all for making this achievement possible.

*In loving memory of Nana, Papa, Grandma and Grandad,
whose love, wisdom, and guidance continue to inspire me.*

Your memory lives on in all that I do.

Contents

1	Introduction	17
1.1	Motivation	18
1.2	Current Approaches	19
1.3	Research Aim and Objectives	21
1.4	Research Questions	22
1.5	Research Approach and Roadmap	23
1.6	Chapter Summary	26
2	Background	27
2.1	Introduction	27
2.2	Evolution of News	27
2.3	Concept of Fake News	29
2.4	Impacts of Fake News Articles	30
2.5	Overview of Methods to Detect Fake News	32
2.5.1	Human-Based Approaches	32
2.5.2	Machine-Based Approaches	33
2.6	Chapter Summary	35
3	Machine Learning Approaches to Detect Fake News: A Systematic Review	37
3.1	Introduction	37
3.2	Motivation	38
3.3	Research Questions Addressed	39
3.4	Method	40
3.4.1	Search Process	40
3.4.2	Study Selection and Evaluation	42
3.4.3	Quality Assessment	43
3.4.4	Data Extraction	44
3.4.5	Threats to Validity	45
3.5	Results	46
3.5.1	Overview of Included Studies	47

3.5.2	Quality Assessment Results	48
3.5.3	Methods of Fake News Article Detection (RQ1)	50
3.5.4	Effectiveness of Current Methods (RQ2)	63
3.6	Discussion	71
3.7	Chapter Summary	75
4	Methodology	77
4.1	Introduction	77
4.2	Research Method	78
4.3	Data Collection	80
4.4	Pre-Processing	82
4.5	Feature Extraction	84
4.5.1	Bag of Words	84
4.5.2	Term-Frequency Inverse Document Frequency (TF-IDF)	85
4.5.3	Static Embeddings	86
4.5.4	Contextual Embeddings	87
4.5.5	Stylistic Features	89
4.6	Machine Learning Algorithms	89
4.6.1	Naive Bayes	90
4.6.2	Logistic Regression	91
4.6.3	Support Vector Machines (SVMs)	92
4.6.4	Decision Trees	93
4.6.5	Gradient Boosting	95
4.6.6	Random Forest	95
4.6.7	Feed-Forward Neural Networks (FFNNs)	96
4.6.8	Long-Term Short-Term Memory Networks (LSTMs)	98
4.7	Evaluation Methods	99
4.7.1	Holdout Testing	99
4.7.2	K-Fold Cross-Validation	100
4.7.3	External Validation	100
4.8	Evaluation Metrics	101
4.8.1	Accuracy	101
4.8.2	Precision	102
4.8.3	Recall	102
4.8.4	Specificity	103
4.8.5	F-1 Score	104
4.8.6	Other Metrics	104
4.9	Model Interpretability Techniques	105

4.9.1	Local Interpretable Model-Agnostic Explanations (LIME)	105
4.9.2	Permutation Feature Importance (PFI)	106
4.10	Chapter Summary	106
5	Study 1: Intra-Domain Generalisability	108
5.1	Introduction	108
5.2	Motivation	109
5.3	Research Questions Addressed	109
5.4	Methods	110
5.4.1	Datasets	111
5.4.2	Pre-Processing	115
5.4.3	Features	115
5.4.4	Algorithms	117
5.4.5	Evaluation	119
5.5	Results	120
5.5.1	Baseline K-fold Cross Validation	121
5.5.2	External Validation	128
5.5.3	Interpreting Models Trained on Token-Representations	136
5.6	Discussion	139
5.7	Chapter Summary	141
6	Study 2: Exploring Features for Generalisable Fake News Detection	143
6.1	Introduction	143
6.2	Motivation	144
6.3	Research Questions Addressed	145
6.4	Method	146
6.4.1	Datasets and Data Processing	147
6.4.2	Experiment 1 Features: Token-Representations	150
6.4.3	Experiment 2 Features: Stylistic and Proposed Social-Monetisation Features	151
6.4.4	Machine Learning Algorithms	156
6.4.5	Evaluation	157
6.5	Results	158
6.5.1	Experiment 1: Generalisability of Token-Representations	158
6.5.2	Experiment 2: Generalisability of Stylistic and Social-Monetisation Features	162
6.5.3	Analysis with Permutation Feature Importance	165

6.6	Discussion	171
6.7	Chapter Summary	173
7	Conclusions and Future Work	175
7.1	Introduction	175
7.2	Summary of Thesis	176
7.3	Research Objectives Revisited	182
7.4	Research Questions Revisited	184
7.5	Contributions	186
7.5.1	C-1 – Stylistic Features	186
7.5.2	C-2 – Social-Monetisation Features	188
7.5.3	C-3 – Simplified Feature Set	188
7.5.4	C-4 – Evaluation Approach for Fake News Detection Models	189
7.6	Limitations and Future Research	190
7.6.1	Features over Model Architectures	190
7.6.2	Expanding Beyond Textual Features	191
7.6.3	Reliance on Existing Datasets	192
7.7	Ethical Considerations	193
7.7.1	Model Accuracy and Bias	193
7.7.2	Transparency and Accountability	194
7.7.3	Freedom of Speech and Censorship	195
7.7.4	Risk of Misuse	195
7.7.5	Recommendations	196
A	Embedding Algorithms	198
A.1	Word2Vec Embedding Algorithm	198
A.2	BERT Embedding Algorithm	199
B	Stylistic Feature-Sets	200
B.1	Fernandez Feature-Set	200
B.2	Abonizio Feature-Set	201
B.3	LIWC	202
B.4	NELA Feature-Set	203

List of Figures

1.1	Thesis Roadmap	23
2.1	Fake News Article - 2020 U.S. Election (Hoft, 2020)	31
2.2	Fake News Article – PepsiCo (Moreno, 2016)	32
3.1	Study Selection Flowchart	43
3.2	Study Sources	47
3.3	Years of Publish	47
3.4	Frequency of Overall QA Scores	48
3.5	QA Results per Question	48
3.6	Top 15 Fake News Datasets (Reversed Order)	51
3.7	Comparison of Feature Types	53
3.8	Overview of Feature Usage	54
3.9	Percentage of Token-Representations by Year	55
3.10	Combinations of Content-Based Features	58
3.11	Overview of Social-Context Features	60
3.12	Overview of Fused-Feature Usage	61
3.13	Overview of Algorithm Usage	62
3.14	Feature Performance	65
3.15	Performance of Machine Learning Algorithms	67
4.1	Text Classification Process	78
4.2	Text-Preprocessing	83
4.3	Logistic Regression	91
4.4	SVM	92
4.5	Decision Tree	94
4.6	Neural Network	97
5.1	ISOT Dataset - Top 20 Common Words	112
5.2	Kaggle Fake or Real - Top 20 Common Words	113
5.3	Kaggle Fake News - Top 20 Common Words	113
5.4	FakeNewsNet - Top 20 Common Words	114

6.1	Study 2 - Overview	146
6.2	Articles per Source Prior to Extraction	148
6.3	Articles per Source Post-Extraction	149
6.4	Experiment 1 - Flowchart	151
6.5	Experiment 2 - Flowchart	152
6.6	NELA Feature Importance	167
6.7	External Validation Feature Importance	168

List of Tables

3.1	SLR - Thesis Research Questions Addressed	39
3.2	Total Papers Collected by Database	41
3.3	Quality Assessment Criteria	44
3.4	Data Extraction Fields	45
3.5	Dataset Details	52
3.6	Content-Based Feature Descriptors	56
3.7	Overview of Social-Context Features	59
3.8	Average Accuracy of Features	64
3.9	Average Accuracy of Machine Learning Algorithms	68
5.1	Study 1 - Thesis Research Questions Addressed	110
5.2	Fernandez Feature-Set	118
5.3	ISOT - K-Fold Results	124
5.4	Kaggle Fake or Real - K-Fold Results	125
5.5	Kaggle (Fake News) - K-Fold Results	126
5.6	FakeNewsNet - K-Fold Results	127
5.7	External Validation Across Datasets	128
5.8	External Validation - Features	131
5.9	External Validation - Algorithms	133
5.10	Frequency Distribution of Keywords Contributing to Classification . .	137
6.1	Study 2 - Thesis Research Questions Addressed	145
6.2	Dataset Summary	150
6.3	Token-Representations Baseline Results	159
6.4	Token-Representations Cross-Dataset Results	160
6.5	Stylistic Features & S-M Features Baseline Results	163
6.6	Stylistic Features & S-M Features Cross-Dataset Results	164
6.7	Relevant features to both datasets	169
6.8	Reduced Feature-Set Results	170
7.1	Key Contributions of the Thesis	186

B.1	Fernandez Feature-Set	200
B.2	Abonizio Feature-Set	201
B.3	LIWC	202
B.4	NELA Feature-Set	203

Acronyms

AI	Artificial Intelligence
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CV	Cross-Validation
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
F1	F1 Score
FN	False Negative / Fake News
FP	False Positive
FFNN	Feed-Forward Neural Network
GPU	Graphics Processing Unit
GNN	Graph Neural Network
HAN	Hierarchical Attention Network
LIME	Local Interpretable Model-Agnostic Explanations
LSTM	Long Short-Term Memory
ML	Machine Learning
NB	Naive Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
NN	Neural Network
FND	Fake News Detection
GPT	Generative Pre-trained Transformer
OHE	One-Hot Encoding
PFI	Permutation Feature Importance
RF	Random Forest
RNN	Recurrent Neural Network
SLR	Systematic Literature Review
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TP	True Positive

Chapter 1

Introduction

The rapid spread of ‘fake news’, defined as intentionally misleading or false information presented as legitimate news, across digital platforms has posed a significant challenge for society, influencing public opinion and undermining trust in reliable sources. Fake news detection has emerged as a critical area of research, with machine learning models commonly deployed to identify misleading content. However, while many of these models achieve high accuracy within specific datasets, they often struggle to maintain performance when exposed to new or varied data—a limitation that hinders their practical effectiveness in real-world applications. This thesis addresses this challenge by investigating the generalisability of machine learning models for fake news detection, focusing on their adaptability across diverse datasets and contexts.

To improve model generalisability, this research focuses on the features used by fake news detection models. In particular, it examines a range of feature sets—including token-based, stylistic, and novel social-monetisation features—that capture both the linguistic and contextual nuances of fake news. By analysing how these feature types influence intra-domain generalisability as well as applying them to real-world data, this thesis provides insights into enhancing the adaptability and reliability of fake news detection models in dynamic, practical environments.

This introductory Chapter is structured as follows: in Section 1.1, the motivation for this research is presented, introducing the key concepts and highlighting the importance of addressing the challenges associated with fake news detection in today’s digital environment. In Section 1.2, existing approaches and solutions are reviewed, providing an overview of current research in fake news detection which motivates the research reported in this thesis. Section 1.3 presents the research aim and objectives, outlining the intended contributions of this thesis to advance the robustness and adaptability of fake news detection models. Section 1.4 defines the research questions driving this study, focusing on the generalisability and effectiveness of various feature sets within fake news detection. To achieve these aims,

Section 1.5 provides an overview of the approach and thesis structure, offering a roadmap and Chapter-by-Chapter summary of the research process.

1.1 Motivation

The spread of fake news has become a pervasive issue in today's digital age, where information circulates rapidly. While fake news and disinformation has existed throughout history, their scale and influence have dramatically expanded with the rise of online platforms and social media. These digital channels allow for the rapid dissemination of both credible and deceptive information, enabling fake news to reach vast audiences quickly (Allcott and Gentzkow, 2017). As a result, the line between legitimate journalism and intentionally misleading news has become increasingly blurred, allowing disinformation to influence large audiences before it can be adequately addressed. The effects of this type of disinformation are both broad and significant. In democratic societies, fake news can manipulate public opinion, sway elections and undermine governance (Morgan, 2018). By shaping narratives that resonate emotionally or confirm biases, fake news can alter how people perceive political candidates, policies and events. This manipulation of public sentiment can not only influence voting behaviour but also weaken confidence in the electoral process itself, as individuals question the legitimacy of information shaping their choices (Bovet and Makse, 2019). Over time, this eroded trust impacts governance, as leaders face increasing scepticism and polarisation within the public, making it more challenging to implement policies and build consensus. As societies rely on informed citizen engagement, the unchecked spread of fake news threatens the foundations of accountable and representative governance.

More broadly, fake news also has wider societal impacts that extends beyond politics. The rapid spread of fake news can exacerbate social divisions and deepen existing conflicts within communities (Del Vicario et al., 2016). By exploiting controversial issues, fake news fosters an environment of mistrust and animosity, where individuals are less likely to engage in constructive dialogue or seek common ground. This polarisation can create echo chambers, reinforcing biases and limiting exposure to diverse perspectives (Törnberg, 2018). As a result, individuals may become increasingly isolated from views that differ from their own, making it difficult to bridge divides and weakening the sense of shared community. Over time, this fragmentation can disrupt social cohesion, making societies more vulnerable to conflicts and less resilient in times of crisis. The erosion of mutual understanding and respect risks creating an environment where empathy is diminished, and divisions are amplified, ultimately destabilising the foundations of civil society.

This societal impact can extend to areas such as healthcare, where disinforma-

tion and fake news about medical issues can lead to harmful consequences (Bratu, 2018). During public health crisis, such as pandemics, the spread of false information can create confusion and distrust in health authorities, undermining efforts to provide accurate guidance and implement effective measures (Do Nascimento et al., 2022). For instance, during the COVID-19 pandemic, disinformation about the virus, vaccines and treatments spread rapidly, often deterring individuals from following health guidelines or seeking necessary medical care (Rocha et al., 2023). This disinformation not only placed individuals at risk, but also strained healthcare systems as false beliefs about the virus and its prevention circulated widely.

Similarly, in sciences, the spread of fake new and disinformation can have detrimental effects on public understanding and trust in scientific research (Scheufele and Krause, 2019). False or misleading information about topics, such as climate change, genetic engineering, or renewable energy, can shape public opinion in ways that are disconnected from empirical evidence. This can hinder support for critical scientific initiatives, influence policy decisions, and delay action on urgent issues (Harper et al., 2020). For instance, disinformation about climate change has fuelled scepticism, slowing efforts to address environmental challenges and influencing political agendas around sustainability. Moreover, the prevalence of fake news in science gives rise to ‘alternative facts’, where individuals may reject well-supported scientific conclusions in favour of unsupported beliefs (Allchin, 2018). As a result, the gap between scientific consensus and public opinion widens, complicating efforts to mobilise communities around evidence-based solutions. This erosion of trust in science presents long-term challenges for societies that rely on scientific advancements to address complex problems, from environmental sustainability to medical innovation. Over time, as disinformation continues to distort perceptions of scientific knowledge, the credibility of researchers and institutions may be undermined, threatening the role of science as a foundation for informed decision-making.

In summary, the spread of fake news presents a fundamental threat to information integrity, public trust and social cohesion. By exploiting digital platforms’ immediacy and reach, disinformation has the power to manipulate opinions, amplify social divisions, and disrupt public well-being in critical contexts. As fake news continues to proliferate, the urgency to understand and mitigate its impact has never been greater.

1.2 Current Approaches

In an attempt to address the problem of the spread of fake news, a number of methods have been developed, which can be broadly divided into two categories: ‘human-based’ approaches and ‘machine-based’ approaches.

Traditional, human-based approaches to fake news detection, such as fact-checking by journalists or dedicated fact-checking organisations, play a crucial role in identifying and addressing disinformation (Vo and Lee, 2019). However, these methods are inherently limited in scalability; fact-checking is a labour-intensive and time-consuming process that often allows false information to spread widely before verification can take place. The rapid pace at which fake news circulates on digital platforms poses an additional challenge, as disinformation can reach millions within minutes, far outpacing the capacity of human fact-checkers to keep up (Karagiannis et al., 2020). Additionally, human-led detection efforts may introduce subjective biases, making it difficult to maintain consistency across diverse topics and sources.

These limitations underscore the need for machine-based approaches, which offer automated and scalable solutions for fake news detection. Machine learning (ML) and natural language processing (NLP) techniques have shown promise in automating this process by identifying linguistic patterns, contextual cues, and stylistic markers that distinguish fake news from credible information (Zhou and Zafarani, 2020). Automated systems can rapidly process vast quantities of data, making them suitable for handling the high volume and speed of information flow on social media and other digital platforms.

However, a key challenge persists: many current ML models struggle with generalisability; that is, while they may perform well on specific datasets, these models often fail to maintain accuracy and robustness when applied to new or unseen data (Gautam and Jerripathula, 2020; Blackledge and Atapour-Abarghouei, 2021; Janicka et al., 2019b). The issue of generalisability is compounded by the limited availability of diverse datasets and the frequent reliance on coarsely labelled data, where accuracy is assumed based on source credibility rather than content verification. Such limitations hinder a model’s ability to perform beyond the datasets on which they were trained on, restricting its capacity to handle the nuanced and evolving nature of fake news in real-world contexts. Addressing this gap motivates the research in this thesis, which seeks to explore new feature sets and evaluation approaches that can enhance the adaptability and reliability of fake news detection models across diverse, real-world contexts.

Furthermore, current research in fake news detection focuses largely on developing automated machine learning models that use token-based features, such as Bag-of-Words (BoW), TF-IDF, and embeddings like Word2Vec and BERT (Capuano et al., 2023). These token-based methods rely on word patterns and frequencies to classify fake news, and they often achieve high accuracy within the boundaries of specific datasets. By capturing basic linguistic structures, token-based models can identify commonalities in fake news language within isolated datasets. However, these methods often lack the flexibility needed to adapt to the varying tones, sen-

sationalism, and nuanced cues that fake news can exhibit across different contexts. Consequently, models trained with token-based features may perform well within their training datasets but often fail to generalise effectively outside of the datasets on which they were trained.

A significant gap in current research is the limited exploration of feature sets beyond token-based methods. While token-based approaches focus primarily on specific words and phrases, they overlook broader characteristics that may be crucial for identifying fake news. Features such as stylistic cues—e.g., tone, sentence structure, and readability—and social-monetisation attributes, such as advertisements, affiliate links, and social share prompts, offer valuable insights but have remained under-explored (Allcott and Gentzkow, 2017; Rehman et al., 2022; Ceylan et al., 2023). Stylistic features help capture the presentational elements of fake news, which can indicate attempts at deception through exaggerated or sensationalist writing. Other types of features, such as features that are linked to financial motivations behind spreading misleading information could also be explored. By incorporating these types of features, models could potentially develop a more adaptable and robust approach to fake news detection, capable of handling a wider range of disinformation strategies and presentation styles.

In addition to feature set limitations, current approaches often rely on traditional evaluation methods, such as holdout testing and K-fold cross-validation, which may overestimate model effectiveness (Cabitza et al., 2021). Since these methods assess performance on the same dataset used for training, they do not fully reflect how models perform on new, unseen data or across diverse topics and contexts. This can result in inflated performance metrics, as the models are not tested on data that truly challenges their adaptability. Exposing models to datasets that reflect real-world complexity, can lead to a more accurate assessment of model generalisability, ensuring that models are capable of performing effectively in diverse disinformation scenarios outside of controlled research settings.

By broadening the scope of features and introducing more realistic and robust evaluation methods, this thesis aims to address the gaps identified in current approaches, paving the way for more reliable and adaptable fake news detection systems.

1.3 Research Aim and Objectives

Motivated by the impact of fake news discussed in Section 1.1 and the challenges identified within current approaches outlined in Section 1.2, this thesis aims to investigate methods for improving the generalisability of fake news detection models. Specifically, it seeks to enhance the adaptability and effectiveness of machine learning

models in detecting fake news by exploring new approaches to feature selection and model evaluation.

The following objectives will support this research aim:

- **Objective 1.** Conduct a comprehensive literature review on fake news detection using machine learning, identifying current approaches, evaluating their effectiveness, as well as highlighting specific challenges and gaps related to model generalisability.
- **Objective 2.** Systematically test and compare the impact of different feature sets and ML algorithms on generalisability.
- **Objective 3.** Create a novel evaluation framework that combines training on widely available datasets with testing on manually labelled data, simulating real-world scenarios and enabling more accurate assessments of model performance.
- **Objective 4.** Propose and test novel feature sets specifically designed to improve generalisability, with a focus on features beyond text that capture the motivations for fake news creation and dissemination.
- **Objective 5.** Provide practical guidelines and recommendations for developing generalisable fake news detection models.

1.4 Research Questions

To address the aim and objectives of this thesis, the following set of research questions are framed:

- **RQ1.** What are the current methods to detect fake news?
- **RQ2.** How effective are current methods to detect fake news?
- **RQ3.** To what extent do existing fake news detection methods generalise across datasets?
- **RQ4.** What current features contribute to more generalisable models in the context of fake news detection?
- **RQ5.** How can novel features that extend beyond the text—such as social dissemination behaviours and economic incentives—enhance the generalisability of fake news detection models?

These research questions, together with the objectives outlined in Section 1.3, are designed to systematically investigate the factors influencing generalisability in fake news detection models. By examining both existing features and novel, non-textual elements, this thesis seeks to identify and address the limitations of current models, ultimately contributing to the development of more generalisable models.

1.5 Research Approach and Roadmap

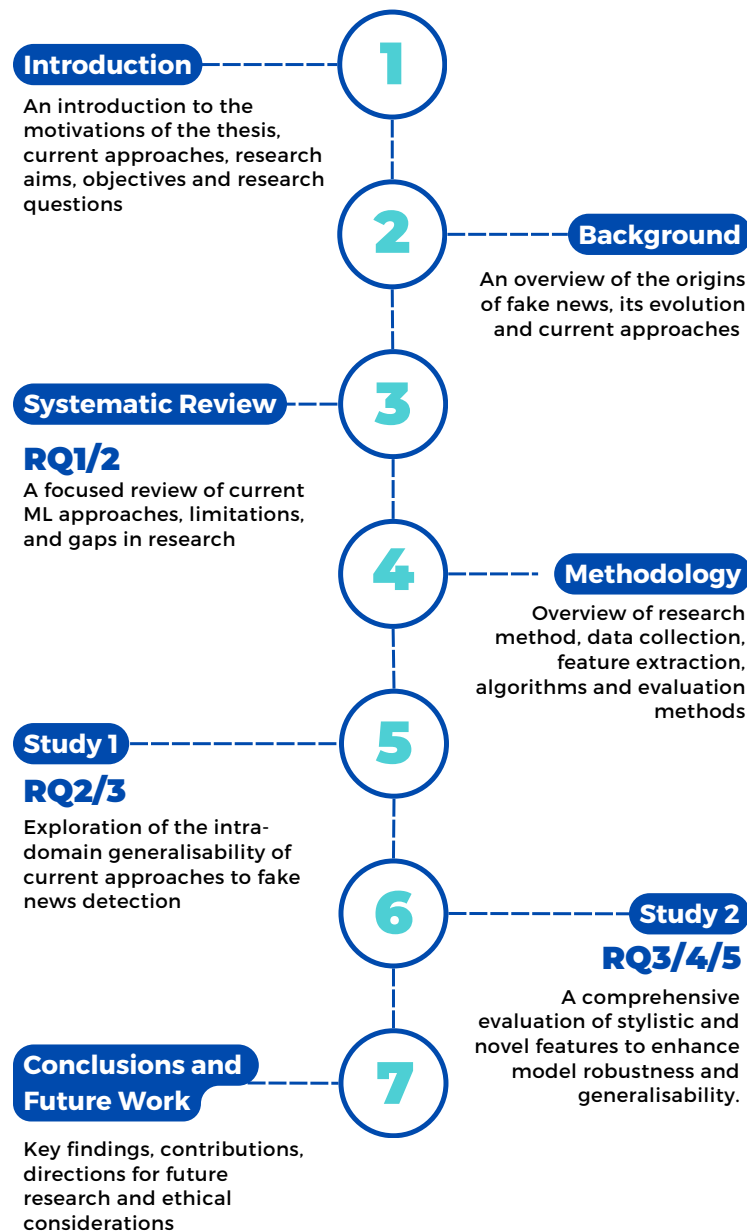


Figure 1.1: Thesis Roadmap

This thesis adopts an empirical research approach to investigate how different feature sets can improve the generalisability of fake news detection models. By addressing the limitations identified in Section 1.2, this approach focuses on evaluating the effectiveness of various feature types in enhancing model robustness and adaptability across diverse datasets and contexts. Through a combination of feature engineering, model training, and evaluation using a combination of K-fold cross-validation and external validation techniques, this research aims to contribute insights into building more adaptable and resilient fake news detection systems.

The thesis is structured as follows:

Chapter 2: Background

This chapter provides an overview of the historical context and evolution of fake news and disinformation, examining its impact on society and the unique challenges it poses. The chapter then transitions to an overview of current approaches to fake news detection, including both traditional human-led methods and automated machine-based techniques, highlighting their respective strengths and limitations. This examination of foundational concepts establishes the groundwork for the systematic review presented in the following chapter.

Chapter 3: Systematic Review on Machine Learning Approaches to Detect Fake News

Building on the background provided in Chapter 2, this chapter presents a focused investigation into machine learning approaches for fake news detection. Through a systematic review of existing studies, the chapter identifies the predominant reliance on token-based features and the generalisability limitations of current models. Key gaps in existing datasets, feature diversity and evaluation methods are highlighted, motivating the need for alternative feature sets and robust evaluation frameworks, which the thesis addresses in the subsequent empirical chapters.

Chapter 4: Methodology

This chapter outlines the research methodology, detailing the processes of data collection, pre-processing, feature extraction, and model selection. It describes the development and implementation of various feature sets and explains the experimental setup used to test model generalisability. The chapter also covers evaluation strategies, including holdout testing, K-fold cross-validation, and external validation (testing on different datasets than those used in testing).

Chapter 5: Study 1 - Intra-Domain Generalisability

Motivated by the findings from the systematic review in Chapter 3, this chapter investigates the fundamental question of intra-domain generalisability in fake news detection models. It assesses how token-based and stylistic features impact model robustness within a single domain, providing empirical insights into which features enhance adaptability and capture the nuanced patterns of fake news.

Chapter 6: Study 2 - Engineering Features for Generalisable Fake News Detection

Building on the findings of Chapter 5, this chapter explores the effectiveness of models trained on coarsely labelled datasets but tested on manually labelled data, providing a more realistic assessment of model generalisability. Additionally, it offers an in-depth exploration of stylistic features for fake news detection, expanding on their potential identified in the previous chapter to enhance model adaptability and robustness across diverse data contexts. This chapter also introduces novel ‘social-monetisation’ features, which capture economic incentives behind disinformation, further contributing to the development of more generalisable detection models.

Chapter 7: Conclusions and Future Work

The final chapter summarises the key findings, limitations, and contributions of the thesis. It also provides recommendations for future research directions, including further exploration of non-textual features, advanced model optimisation, and the integration of multimodal data to enhance fake news detection systems. Additionally, this chapter addresses ethical considerations for the field, advocating for transparency, fairness, and accountability in the development and deployment of fake news detection technologies, while highlighting the importance of avoiding censorship and ensuring the protection of free speech.

This roadmap provides a structured overview of the research journey, outlining the steps taken to address the research questions and objectives. Through this approach, the thesis aims to develop insights that contribute to building more generalisable and reliable fake news detection models capable of handling the complexities of disinformation in diverse, real-world contexts.

1.6 Chapter Summary

This introductory chapter defined the core concepts underpinning this research and provided an overview of the research work that will be reported in this thesis. It established the motivation of the research, which is the growing impact of fake news, and presented the current approaches in fake news detection, identifying wider issues and gaps. These issues and gaps motivated the research aim and objectives of the thesis, which focus on enhancing the generalisability of machine learning models for fake news detection, giving rise to specific research questions. Finally, the chapter outlined the approach of the work and structure of the thesis and provided a summary of the remaining chapters.

Chapter 2

Background

2.1 Introduction

This chapter explains concepts and issues relevant to the area of fake news in more detail and reviews, at a high level, the methods and techniques to identify fake news and prevent its spread. It starts with Section 2.2, which provides an overview of the development and significance of news, tracing its evolution from ancient times, print media and into the digital age. It examines how news has transitioned through various formats and technologies, and the role it plays in informing the public and shaping public opinion. Section 2.3 addresses the pressing issue of fake news, detailing its various forms and manifestations and how it has evolved in lockstep with new technologies. Section 2.4 explores its impact and how it undermines societal trust, influences public opinion, and contributes to disinformation. This exploration includes a thorough discussion of the social, political, economic, and psychological impacts of fake news. The chapter concludes with Section 2.5, which presents an overview of the diverse approaches used to address the problem of fake news, encompassing both human-driven methods and machine-based solutions, the latter primarily involving supervised machine learning techniques.

2.2 Evolution of News

News, defined as information regarding recent events or developments, has evolved significantly over time. Naturally, news under this definition has existed since humans began to communicate, relying on word-of-mouth. While the earliest forms of written communication were found in Mesopotamia in 3400-3300 BC, the first news publication, known as *Acta Diurna*, wasn't established until 59BC during the time of Julius Caesar in Ancient Rome (Eaman, 2021). Typically recorded on stone or metal and displayed in public spaces, this early form of daily communication was

created to inform the public of official decrees, significant political events and societal updates. Similarly in ancient China, reports known as ‘Diabo’, were circulated by government officials as early as the Han Dynasty (206BC – 220AD) (Zhao and Sun, 2018). These early forms of news were limited to a relatively small audience and were often controlled by those in power.

The communication of news did not see a sizable shift until the invention of the printing press by Johannes Gutenberg in Germany, 1440 (Briggs and Burke, 2009). This allowed for the mass production of books and leaflets, making written news more accessible to the wider public. By the early 17th century, the first true newspapers began to appear, with “Relation aller Fürnemmen und gedenckwürdigen Historien” printed by Johann Carlous in Strasbourg starting in 1605, often cited as one of the earliest examples (Weber, 2006). As a result, the 18th century saw an increased demand for information, leading to the distribution of newspapers across Europe and the Americas. During this time, the content and scope of newspapers expanded, covering a broader range of topics and reflecting diverse viewpoints. The invention of the steam-powered press in 1810 further allowed newspapers to be produced more quickly, acting as a catalyst for their widespread availability (Forrester, 2020).

In the 19th century, the concept of journalism began to solidify as a professional field. The rise of the penny press in the United States (that is, newspapers characterised by their low price of one cent), played a crucial role in this transition (Nerone, 1987). Such penny press papers as the New York Sun made newspapers affordable to the masses, often focussing on sensationalist stories (known as ‘yellow journalism’) to attract a wider audience as different publications competed for readership (Wiener, 2011). This competition saw the emergence of full-time journalists and newsrooms, giving rise to the notion of ‘newsworthiness’ (Udeze and Uzuegbunam, 2013). This is where a story’s value is determined by its potential to attract readers, based on factors such as timeliness, impact and novelty. This shift was a notable departure from previous newspapers that focussed on practical needs to share information, such as publishing weather reports and information about local events.

As society progressed into the 20th century, several technological advancements introduced further dimensions to news distribution. The advent of radio broadcasting in the 1920s allowed for the instantaneous transmission of news, giving rise to news programs providing real-time updates on current events, allowing listeners to receive information as it happened (Allan, 2004). By the 1950s, news dissemination saw another notable shift with the invention of the television. This allowed for the combination of auditory and visual elements, not only providing viewers with real-time updates but live footage of events, with the first 24-hour news channel,

CNN, established by 1980 (Napoli, 2020). The introduction of the internet going into the 21st century saw a further dramatic shift in how news was distributed and consumed (Leighton and Sagan, 2010). Online news platforms and social media have further increased the immediacy in which news is communicated, giving rise to the democratisation of news production (Jebril et al., 2015). This has meant anyone with an internet connection could publish and share content. While this democratisation has redistributed the influence of media away from those in power, offering a more diverse range of perspectives, it also presents a number of challenges.

In particular, social media has introduced new ways of interaction and engagement, allowing users to comment and share news stories to others within their respective networks. This environment has greatly increased the levels in which individuals engage with news, fostering more dynamic discourse. However, it has also given rise to the creation of ‘echo-chambers’, whereby individuals are primarily exposed to viewpoints that reinforce their own (Terren and Borge-Bravo, 2021). This has led to increased levels of confirmation bias and hostilities between different groups. The influence of algorithms on news feeds can further exacerbate this issue, negating the positive factors of the democratisation of news by limiting exposure to diverse perspectives (Forster and Wong, 2024). Such algorithms additionally present users with inflammatory content to increase engagement in favour of generating profit through advertising.

2.3 Concept of Fake News

A subset of such inflammatory content is **disinformation**, that is, false information created with the intent to deceive. This is in contrast with **misinformation**, which is inaccurate content spread without any intention to deceive. A subset of such disinformation is ‘fake news’, defined as intentionally misleading news created to generate profit through advertising or exert political influence (Allcott and Gentzkow, 2017). While the term ‘fake news’ was popularised during President Trump’s 2016 presidential election campaign (Sharma et al., 2019b), such disinformation has historically evolved alongside advancements in technology.

In ancient Rome, Octavian, the adopted son of Julius Caesar, launched a ‘fake news’ campaign against Mark Anthony to win public (and therefore political) support, accusing him of disrespecting Roman values due to his affair with Cleopatra. Octavian’s campaign involved the use of various forms of media, including speeches, poetry, and coinage (Watson, 2018). Upon the invention of the printing press, fake news began to spread more rapidly throughout Europe. During the Reformation period in the 16th century, the Catholic and Protestant church printed leaflets containing false claims in an attempt to smear the opposition (Maus, 2020). By the

19th century and advent of ‘yellow journalism’, false stories were written to generate profit (Campbell, 2019), with one such series of stories by the New York Sun which published a series of articles later known as ‘The Great Moon Hoax’, describing the lives of creatures living on the moon (Vida, 2012). In the 20th century, propaganda became a popular tool to deceive and manipulate public opinion, exemplified by extensive disinformation campaigns during both World Wars and the Cold War. In the Great War, The Times and The Daily Mail printed news articles claiming German forces were using the corpses of their own soldiers to boil down for fats, the goal being to paint the Germans as a barbaric people (Adena et al., 2015). During the Second World War, propaganda was spread on an even larger scale following the invention of the radio. The Nazis in particular, were adept in the use of radio propaganda to solidify their position and spread antisemitic rhetoric to justify their actions (Adena et al., 2015). The end of the 20th century, following the proliferation of the television and later the internet, saw an increase in satirical fake news, with shows such as The Daily Show and publications such as The Onion blending humour with social commentary (Brewer and Marquardt, 2007; Wenner, 2002).

Disinformation and ‘fake news’ have therefore existed throughout history with various motivations. From attempts to sway public opinion, generate profit through the publication of sensationalist stories and hoaxes, to satire. While the motivations to create fake news have remained largely unchanged, the digital age presents further advancements to how fake news can be presented. Manipulated videos or images can misrepresent reality; this includes photoshopping images, editing videos to alter their context, or creating deepfakes—hyper-realistic digital forgeries that depict people saying or doing things they never actually did (Cao et al., 2020). Misleading statistics can also be used to support false narratives by presenting facts out of context or selectively highlighting data to reinforce a particular viewpoint (Budak et al., 2024). Sensationalist and misleading headlines, known as ‘clickbait’, headlines, can further encourage users to consume and share content without verifying its accuracy (Chen et al., 2015) – further exacerbated by the nature of social media algorithms. All these tactics of disinformation can culminate in the form of fake news articles, making them one of the most dangerous forms of disinformation media today.

2.4 Impacts of Fake News Articles

Fake news articles can have significant impacts on society. Aside from swaying political opinions (and therefore elections), such articles can have widespread impacts on healthcare, the economy and public safety.

A recent example of such disinformation can be found in the 2020 U.S. Pres-



Figure 2.1: Fake News Article - 2020 U.S. Election (Hoft, 2020)

idential Election, where it was falsely claimed that the election was ‘stolen’ from the incumbent, President Trump. While such rhetoric originated from Trump himself, the media, including producers of false news articles, amplified this notion such that a significant portion of the population came to believe that the results of the election were illegitimate (Calvillo et al., 2023). This in turn, led to significant polarisation, culminating on the 6th January, 2021, where rioters stormed the Capitol in an attempt to overturn the results of the election. These events underscore the significance of fake news in promoting civil unrest and deepening polarisation.

In terms of impacts on healthcare, numerous examples can be found from the COVID-19 pandemic. A number of articles promoted the idea that the virus was a hoax, that it was intentionally released as a bioweapon and that treatments such as inducing bleach were effective (Dharawat et al., 2022). Such disinformation led to higher rates of hospitalisations, on already strained health services, due to people following unverified medical advice. The psychological impact on the public was also significant, with disinformation exacerbating anxiety and mistrust in health authorities, leading to lower rates of vaccine uptake (van der Linden et al., 2020). These events further highlight the impact that fake news articles can have on public health and safety.

In terms of the economy, fake news articles can have a significant impact on businesses. Examples of this can be found in Clarke et al. (2018) and Kogan et al. (2022), which demonstrated that articles posted on the website, Seeking Alpha, artificially inflated stock prices. This can have a damaging impact on investors, who buy such stocks at the inflated price until manipulators sell-off, resulting in a significant drop in value. Fake news articles can also have significant impacts on individual businesses. Such an example can be found in Trump’s presidential election bid in 2016, where comments by PepsiCo CEO Indra Nooyi were misrepresented, leading to calls for a boycott against the Pepsi brand (Figure 2.2). Such actions can cause severe reputational damage and negatively affect revenue.

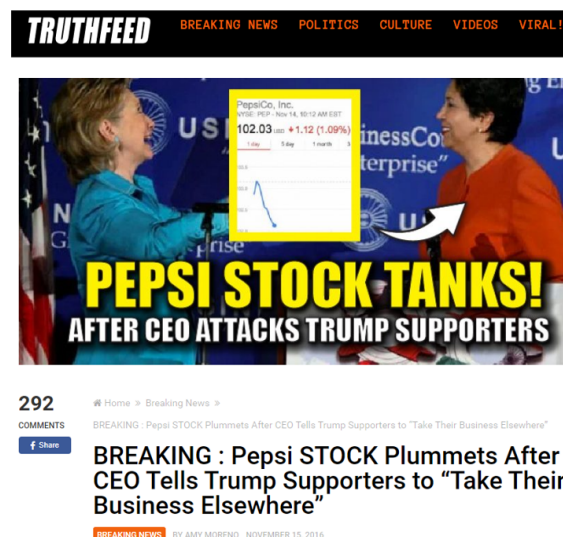


Figure 2.2: Fake News Article – PepsiCo (Moreno, 2016)

2.5 Overview of Methods to Detect Fake News

The previous section demonstrated that fake news articles present a significant threat to society, in terms of influencing political discourse, threatening public health and destabilising businesses. As such, a number of methods have been developed in an attempt to combat this threat. These can broadly be divided into two categories: human-based approaches and ‘machine-based’ approaches.

2.5.1 Human-Based Approaches

Human-based approaches, as the name suggests, typically rely on a manual approach to fake news detection. One of the primary human-based approaches is manual fact-checking. This is typically carried out by teams of researchers, who attempt to verify claims by cross-referencing with reliable sources. Organisations that carry out this work include FullFact, Snopes and PolitiFact (Vo and Lee, 2019). Such organisations typically publish their findings, while citing evidence to support their views on whether a particular piece of content is accurate or misleading. While this method can be considered reliable, it is resource intensive and time consuming (Karagiannis et al., 2020).

An extension of this approach is community reporting. This relies on individuals on platforms such as Facebook and X (formerly known as Twitter) to flag potentially misleading media. Such media is then reviewed by moderators, or is sent to third-party fact-checkers, to determine whether such content is misleading or harmful (Wu, 2023). This aids in addressing a weakness of purely manual fact-checking which struggles to scale over the vast amount of content that is created on a daily basis. Community reporting allows for quicker identification of such content and

therefore the speed in which such content can be addressed. However, while such an approach does increase efficiencies in the identification of such content, there are a number of weaknesses. These include the accuracy of human reports which may be influenced by bias or personal beliefs. Without guidance from external sources, humans typically only predict fake news accurate 64% of the time, based on an article’s content (Snijders et al., 2023). Such inaccuracies can negate the improved efficiencies associated with community reporting, as moderators must assess large quantities of content that has been falsely flagged.

2.5.2 Machine-Based Approaches

The weaknesses in human-based approaches necessitate a more efficient and robust method for detecting disinformation, which leverages technology. While some alternative techniques, such as knowledge graphs (Ciampaglia et al., 2015), have been explored for their ability to represent relationships between entities and infer credibility, the field is predominantly driven by supervised machine learning algorithms (Zhou and Zafarani, 2020). These algorithms excel in identifying patterns within fake news and detecting it before it can spread, making them the cornerstone of computational approaches to combat disinformation.

This automated approach to fake news detection begins with data collection, which typically involves one of three approaches: using established datasets, creating custom datasets, or a combination of both these methods. Established datasets, often sourced from platforms such as Kaggle or from other institutions, provide a ready-made foundation with labelled articles categorised as fake or real. Alternatively, custom datasets are created by collecting new data directly from sources such as news websites or social media feeds. This approach allows for the inclusion of more recent or niche topics, ensuring the dataset is tailored to specific research needs. Finally, a combined approach leverages the strengths of both methods by supplementing established datasets with additional data. This can improve the diversity and relevance of the data, as researchers can ensure coverage of both foundational patterns in fake news and emerging trends.

Following data collection, the process moves to pre-processing, where raw text is cleaned and standardised to make it suitable for analysis. This stage often includes removing unwanted characters, URLs, and HTML tags, converting the text to lowercase to maintain consistency, and tokenizing (or splitting) the text into individual words or phrases (Bird et al., 2009). Additional steps, such as removing common stop words (like “the” or “and”) and stemming words to their root form, help reduce noise, making it easier for the model to focus on patterns in the text that are likely to indicate fake news. By ensuring the data is clean and formatted consistently,

pre-processing improves the reliability of the features that will be extracted in the next stage.

In the feature extraction stage, pre-processed text is transformed into numerical representations that machine learning models can interpret. Basic approaches, like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), count word occurrences to create a numerical profile, though these methods capture little beyond word frequency (Tabassum and Patil, 2020). More advanced techniques, such as word embeddings like Word2Vec and GloVe, represent words as fixed vectors based on their overall semantic relationships in a large corpus, enabling models to identify nuanced language patterns that differentiate fake news from real news (Selva Birunda and Kanniga Devi, 2021). However, these embeddings have limitations, such as their inability to handle out-of-vocabulary (OOV) words or distinguish between words that are morphologically similar but semantically different. Contextual models like BERT address these weaknesses by generating dynamic embeddings that adapt to the context of surrounding text, allowing for more accurate detection of complex patterns and relationships often associated with disinformation (Devlin, 2018). In addition to token-based features, stylistic indicators (e.g., sentiment, tone, and phrasing) provide further clues, as fake news often employs emotionally charged or sensational language (Lagutina et al., 2019). Non-textual features, such as social media engagement metrics (likes, shares, and comments) or visual elements in multimedia, can also be valuable, as they reflect user interaction patterns or imagery often associated with misleading content (Shu et al., 2017).

With features extracted, the next stage is model training, where the machine learning model learns to recognise patterns associated with fake news. During training, the model is provided with labelled data, enabling it to adjust its internal parameters and “learn” from the features. Basic algorithms like Logistic Regression, Naïve Bayes, and Support Vector Machines (SVMs) offer reliable baselines, while ensemble methods like Random Forests and Gradient Boosting can capture more complex interactions (Varshney and Wadhwani, 2023). For even greater sophistication, deep learning models - such as Long Short-Term Memory (LSTM) networks - are often used. These models excel at understanding complex language structures and contextual information, which can significantly improve detection accuracy (Padalko et al., 2023). Deep learning models are particularly useful for large-scale fake news detection, as they can adaptively learn the nuanced patterns in language, though they require substantial data and computing resources.

Finally, the evaluation phase assesses the model’s ability to accurately classify fake and real news using techniques like holdout testing and K-fold cross-validation (Resnik and Lin, 2010). In holdout testing, the dataset is split into separate training and testing sets, allowing the model to be evaluated on data it has not seen

during training, providing a straightforward measure of its performance. K-fold cross-validation further strengthens evaluation by dividing the dataset into K subsets, or folds, and iteratively training the model on K-1 folds while testing on the remaining fold. This process is repeated K times, with each fold serving as the test set once, and the average results across folds provide a more comprehensive estimate of the model's robustness. Through these techniques, key evaluation metrics can be produced, including accuracy, precision, recall, and F1-score, provide deeper insights into the model's performance, allowing for a clear understanding of its strengths in identifying fake news and areas where it may need improvement.

In addition to these standard techniques, external validation is crucial for assessing a model's ability to generalise beyond its original dataset (Cabitza et al., 2021). This involves testing the model on an entirely separate dataset that was not used during training, providing insight into how well the model performs when applied to new data sources, topics, or styles of fake news. External validation is particularly valuable in the fake news detection domain, as it highlights whether the model can handle variations in language, format, and content that are common in real-world scenarios.

2.6 Chapter Summary

This chapter provided an overview of the background concepts essential to understanding the challenges and approaches involved in fake news detection. It began by examining the evolution of fake news, tracing its roots from historical instances of disinformation to its current, rapidly evolving forms in the digital age. As social media and online news sources have proliferated, so too has the ease with which fake news can be created and spread, highlighting the pressing need for effective detection methods. The concept of fake news was defined to clarify its characteristics, including its intentional spread of false or misleading information, often designed to influence public opinion, provoke emotions, or generate financial gain.

The chapter also explored the impact of fake news articles in the modern era, demonstrating how disinformation has had significant consequences on political discourse, public health, and societal trust in media. High-profile cases in elections and public health crises were highlighted as examples of how fake news can shape public perception and behaviour with far-reaching effects.

An overview of methods to detect fake news was provided, focusing on two main categories: human-based and machine-based approaches. Human-based approaches, such as manual fact-checking by organizations like PolitiFact and Snopes, were discussed for their reliability but limited scalability. Community-based reporting on social media, which enables users to flag potentially misleading content, was also cov-

ered, as it allows for quicker identification of suspect articles despite potential biases in reporting. Machine-based approaches were introduced as an efficient alternative, emphasising the scalability of with machine learning techniques for automated fake news detection emerging as the most predominant approach. Traditional machine learning methods, as well as more advanced deep learning models, were discussed for their ability to identify patterns in language, style, and context that distinguish fake news from legitimate information.

Chapter 3

Machine Learning Approaches to Detect Fake News: A Systematic Review

3.1 Introduction

Building on the foundation laid in the previous chapter, which explored the evolution, concept, and societal impact of fake news as well as the broad categories of detection methods, this chapter presents a systematic review of machine learning approaches to detect fake news. With the limitations of human-based approaches in mind, this chapter focuses on the potential of machine learning (ML) to provide a scalable and robust solution to disinformation. The review is structured to provide a comprehensive overview of the range of datasets, features, and machine learning algorithms employed in fake news detection.

Section 3.2 of this review provides the motivation for conducting this systematic review, highlighting the limitations of existing reviews that may often focus on high-level descriptions or narrow comparisons without systematically assessing the effectiveness of different methods across various types of fake news. By addressing these gaps, this review aims to offer a more nuanced understanding of the strengths, limitations, and applicability of current machine learning techniques in detecting fake news. Section 3.3 then covers the research questions guiding this review. First introduced in Chapter 1, these research questions aim to identify the existing approaches to fake news detection and evaluate how effective they are. Section 3.4 outlines the methodology used for conducting the systematic review, detailing the process of selecting and analysing relevant studies to ensure a thorough and unbiased synthesis of findings. Section 3.5 presents the results, summarising the key findings on the various machine learning methods, datasets, and features employed, as well

as insights into their effectiveness and generalisability.

3.2 Motivation

As discussed in Chapter 2, fake news manifests in various forms and spreads through different means of distribution, leading to a correspondingly diverse landscape of detection methods. With numerous ML and NLP techniques now applied to tackle disinformation, each approach has varying suitability and effectiveness, depending on the type of fake news and the distribution channels involved. Determining which techniques work best under different conditions is essential for guiding future research and industry practices.

Although there are many literature reviews that provide overviews of the different methods of detecting fake news in general (Sharma et al., 2019a; Elhadad et al., 2019; Lahlou et al., 2019; Kaliyar and Singh, 2019; Parikh and Atrey, 2018; Bondielli and Marcelloni, 2019; Hassan and Meziane, 2019; Pierri and Ceri, 2019; Guo et al., 2021; Rana et al., 2018; Manzoor et al., 2019; Klyuev, 2018). Much fewer literature reviews provide insight into the effectiveness of certain methods. However, because these reviews are often not aimed to be systematic or have a different scope, they report the results of a limited number of papers or state the limitations of certain approaches but do not provide an in-depth comparison of the techniques used (Zanettou et al., 2019; Zhang and Ghorbani, 2020; Hirlekar and Kumar, 2020; Zhou et al., 2019; Vishwakarma and Jain, 2020). The reviews by Sharma and Sharma (2019); Mahid et al. (2018); Sharma et al. (2019a) provide insight to the limitations of particular approaches, with the review by George et al. (2020) offering a comprehensive comparison of approaches in terms of their effectiveness; however, due to the limited number of studies included in the comparison, the results, while valuable, are not conclusive. Mahid et al. (2018) cites hybrid approaches as being more effective than other approaches and this is somewhat supported by the review by Manzoor et al. (2019), which states that “the analysis of fake news content is not sufficient to establish an effective and reliable detection system” and that other aspects of fake news including author and user analysis as well as social context should also be explored.

Moreover, literature reviews may often include and compare studies that address different ‘types’ of fake news (rumours, clickbait, social media posts, etc.), as previously mentioned in Chapter 2. It could be argued that these forms of fake news have different characteristics, such that different approaches may be more effective. As such, more focused investigations – primary studies and literature reviews – are needed to assess the suitability of approaches, or combination of approaches, for different types of fake news.

Given these gaps, there is a need for focused, systematic reviews that not only catalogue the available techniques but also analyse their effectiveness across various types and sources of fake news. This review aims to address this need by providing a focused investigation into approaches for detecting fake news articles, specifically those designed to deceive for purposes of political influence or generating profit through advertising, rather than clickbait or satirical content. By doing so, it seeks to offer valuable insights into the effectiveness of various detection methods, guiding researchers and practitioners toward more targeted and impactful strategies for combating disinformation in an ever-evolving digital landscape.

3.3 Research Questions Addressed

This section outlines the thesis research questions, introduced in Section 1.4, addressed by this review. Specifically, this review addresses RQ1, which investigates the methods currently used to detect fake news, and RQ2, which evaluates the effectiveness of these methods:

Table 3.1: SLR - Thesis Research Questions Addressed

RQ	Description
RQ1	What are the current methods to detect fake news?
RQ2	How effective are current methods to detect fake news?
RQ3	To what extent do existing fake news detection methods generalise across datasets?
RQ4	What current features contribute to more generalisable models in the context of fake news detection?
RQ5	How can novel features that extend beyond the text—such as social dissemination behaviours and economic incentives—enhance the generalisability of fake news detection models?

To ensure a thorough and structured examination, each research question is broken down into the following sub-questions:

RQ1. What are the current methods to detect fake news?

- **1.1.** What datasets are used in developing fake news detection models?
- **1.2.** What features are used to detect fake news?
- **1.3.** What machine learning algorithms are used?

RQ2. How effective are existing methods to detect fake news?

- **2.1.** What groups of features are most effective for fake news detection?
- **2.2.** What machine learning algorithms are most effective for fake news detection?
- **2.3.** How generalisable are current approaches to fake news detection?

RQ1 is intentionally broad to ensure that all methods of detecting fake news are captured. It seeks to identify the various approaches used in fake news detection by examining the datasets employed in the development of detection models (RQ1.1), the features utilized to differentiate fake news from real news (RQ1.2), and the machine learning algorithms that are applied to these tasks (RQ1.3). These sub-questions aim to provide a comprehensive overview of the key components involved in current fake news detection methods.

RQ2 addresses how effective these features and machine learning algorithms are at addressing the fake news problem by primarily comparing accuracy metrics. Due to the number of variables between different papers (including the datasets used and the differing implementations of ML methods and NLP techniques), multiple analyses are performed to ensure that more reliable comparisons can be made.

While RQ2.1 and RQ2.2 specifically focus on evaluating the performance of individual algorithms and approaches in isolation, RQ2.3 takes a broader perspective by examining the literature to assess how generalisable these approaches are across different contexts. The goal of RQ2.3 is to determine whether the approaches being studied can be generalised beyond their original contexts, which is crucial for developing reliable and scalable solutions that are usable in ‘real-world’ conditions.

3.4 Method

This review follows the guidelines from systematic literature reviews as described by Kitchenham (2004). To better manage the review, the tool ‘Parsifal’¹ was used. Adhering to Kitchenham’s guidelines, this tool allows researchers to import studies, specify exclusion criteria and write comments regarding reasons for exclusion. It also includes features for carrying out Quality Assessments and Data Extraction.

3.4.1 Search Process

As the term “fake news” gained significant popularity during the 2016 US election, as mentioned in Section 2.3, articles published in the period between 1st January 2016

¹<https://parsif.al/>

and 31st December 2023 were collected. The search process was largely automated by searching databases including IEEEExplore, ACM, ScienceDirect and Scopus.

Derived from the research questions, the chosen search string for this systematic review encompasses a comprehensive range of terms related to fake news detection, ensuring the inclusion of relevant literature across various domains and disciplines. The inclusion of terms such as “Fake” “Disinformation” “False” “Unverified” “Inaccurate” and “Rumour/s” captures different facets of disinformation, acknowledging its diverse forms and manifestations. Additionally, terms like “News” “Article/s” “Media” and “Information” broaden the scope to include different types of content disseminated through various channels. Finally, incorporating terms like “Detect” “Detection” and “Classification” focuses on literature specifically related to the identification and categorisation of fake news, providing a targeted approach to retrieving relevant studies. By combining these terms logically with Boolean operators (AND/OR), the search string aims to yield a comprehensive dataset for analysis, ensuring that no relevant literature is overlooked in the systematic review process. The resulting search string is as follows:

(“Fake” OR “Disinformation” OR “False” OR “Unverified” OR “Inaccurate” OR “Rumour/s”) AND (“News” OR “Article/s” OR “Media” OR “Information”) AND (“Detect” OR “Detection” OR “Classification”)

Due to fake news being a topic that spans research areas outside of Computer Science, searches were limited, where possible, to peer-reviewed Computer Science journals and conferences, given the focus of this review on machine-based approaches to fake news detection. Following collection, the majority of duplicates were removed automatically through Parsifal. Duplicates that were not captured by Parsifal were excluded manually upon content review. This stage of data collection is presented in Table 3.2.

Table 3.2: Total Papers Collected by Database

Database	Number of Papers
IEEE Xplore	796
ACM	343
Scopus	975
ScienceDirect	108
Total	2222
After All Duplicates Removed	1307

3.4.2 Study Selection and Evaluation

Following the collection of papers, a set of exclusion and inclusion criteria was defined in order to filter out papers that were not relevant to the study or did not align with the definitions and research questions of this review:

Inclusion Criteria

- IC1. Computer Science Papers
- IC2. Date = 2016–2023
- IC3. Language = English
- IC4. Primary Studies
- IC5. Relevant to research questions

The inclusion criteria are typical for systematic reviews, whereby studies must be relevant to the research questions and subject area as well as be primary studies. The date range was selected because academic interest in fake news gained significant traction after the 2016 Presidential Election (as discussed in Section 2.3). Studies were also required to be written in English, such that the authors could understand the content. For studies to be included in the review, they were required to satisfy all the above criteria.

Exclusion Criteria

- EC1. Does not focus on news articles.
- EC2. Does not address detection of fake news articles.
- EC3. Does not present any results for the detection of fake news articles.
- EC4. Does not focus on detection of fake news written in English
- EC5. Focuses on single, unprecedented events (e.g. COVID-19)

The exclusion criteria expand on IC5 by stating what is required for papers to be relevant to the research questions. EC1 excludes studies that do not focus on news articles, thereby narrowing the scope to literature directly related to the detection of fake news in written, long-form, journalistic content. EC2 further refines the selection by excluding studies that do not address the detection of fake news articles specifically, ensuring that only research directly relevant to the review’s objectives is considered. EC3 ensures that only studies presenting results for the detection of fake

news articles are included, enhancing the robustness of the review by focusing on empirical evidence. EC4 serves to exclude studies that do not focus on the detection of fake news written in English, enabling a more targeted analysis of literature relevant to English-speaking contexts. Finally, EC5 excludes studies that focus on single, unprecedented events, ensuring that the review encompasses a diverse range of contexts and scenarios, thus enhancing its applicability and generalisability. Together, these criteria contribute to the rigor and relevance of the systematic review by ensuring that only studies meeting specific criteria are included for analysis.

Study Selection

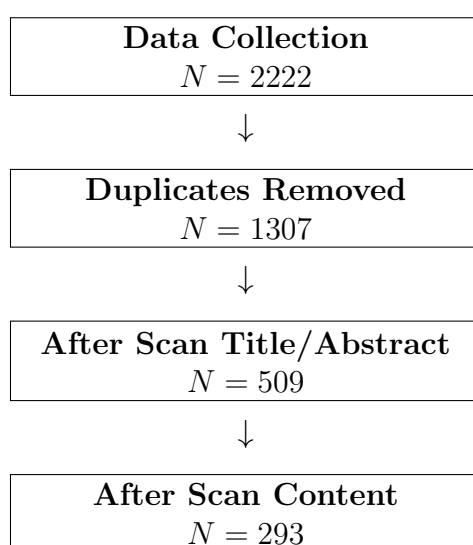


Figure 3.1: Study Selection Flowchart

The study selection process followed a systematic approach to ensure that only the most relevant studies were included in the review. As shown in Figure 3.1, the initial data collection phase yielded 2222 studies. Following this, duplicates were removed, reducing the total number to 1307. A further scan of titles and abstracts, along with the application of the defined inclusion/exclusion criteria, narrowed the pool down to 509 studies. After applying the inclusion/exclusion criteria during a detailed scan of the full content, 293 studies remained, which were deemed appropriate for inclusion in the final review.

3.4.3 Quality Assessment

Following the study selection phase, a Quality Assessment was carried out on the included studies. In systematic reviews, the Quality Assessment can have two purposes: it can be used as a means to either exclude studies, or to support data synthesis Yang et al. (2021). In this study, the Quality Assessment was used to

Table 3.3: Quality Assessment Criteria

Question	Justification
QA1. Does the paper provide an adequate definition or explanation of 'Fake News'?	The term "fake news" is often used loosely, or as an umbrella term to refer to different types of intentionally or unintentionally misleading online content (e.g., rumours, satirical content, social media posts, etc.). As such, it is important that research papers define or describe the "fake news" content being addressed.
QA2. Does the paper disclose/provide access to the dataset used (if any)?	For purposes of reproducibility and evidence of adherence to scientific method.
QA3. If applicable, were the attributes of the dataset used adequately described?	Disclosure of features used is beneficial to the reader so inferences may be derived on what features are best for future research.
QA4. Did the performance metrics used provide a reliable evaluation of the performance of the models?	Disclosure of metrics and the results of the evaluation allow readers to assess the true effectiveness of the model. A single metric might not provide a complete picture of the model's performance. For example, accuracy alone can be misleading, especially in cases of class imbalance. If a combination of metrics was used (e.g., precision, recall, AUC-ROC), this would offer a more comprehensive evaluation, contributing to the reliability of the assessment.
QA5. Did the discussion critically interpret the results?	Ensures the discussion section of the reviewed studies provides a thorough and critical interpretation of the obtained results, facilitating the understanding of their significance and implications.

support data synthesis and analysis. This enables the review to capture the current state of fake news research and identify areas of improvement more accurately. The criteria of the Quality Assessment along with an explanation and motivation for each criterion can be found in Table 3.3. The criteria were formulated as questions, and answers to these questions were restricted to "Yes", "Partially" and "No", each with a numerical score of 1, 0.5 and 0 given, respectively.

3.4.4 Data Extraction

The data extraction phase serves to collect data to address the research questions. This was organised by means of a spreadsheet exported from Parsifal where each row contained the selected papers. Appended to this list of papers, a number of attributes were added in relation to the research questions. These are summarised in Table 3.4. Some fields pertaining to the details of the publication and the authorship were automatically collected through the export process. The remaining fields were filled in manually. The two major groups of data that were manually collected were

Table 3.4: Data Extraction Fields

Data Extraction Fields
Title
Year of Publication
Authors
Source
Journal/Conference
Dataset Used
Features Used
Token-Representation Groups
Stylistic Feature Groups
Social Feature Groups
Other Feature Groups
Machine Learning Algorithm(s)
Accuracy
F1
Precision
Recall
AUC

as follows: the method of detection including the dataset, features and algorithms used, addressing RQ1, as well as the performance including metrics such as F1-score, accuracy, precision, recall and Area Under the ROC Curve (AUC), addressing RQ2.

During data extraction, some papers did not directly give the figures for some fields. Where possible, these fields were populated by deriving results from other data collected (for example, F1 score may have been derived using the precision and recall, or through a confusion matrix). In cases where studies included results from other papers to be used as a baseline, only the primary results were included in the data extraction to avoid duplication. An exception to this is where a paper repeated another’s method and produced new results through that method. In cases where several results were presented for the same method, with the independent variable not being the method, only the average score was included.

One prominent issue during the data extraction phase were the variations of the same basic algorithm; an example of this is NuSVM and Linear SVM or variations in the different Gradient Boosting algorithms such as AdaBoost or XGBoost. These were recorded as they were presented by the selected papers but were also grouped by the algorithms from which they were derived in order to provide a high-level overview.

3.4.5 Threats to Validity

As is common in systematic literature reviews, there are a number of threats to validity which may introduce bias into the outcomes of the review. These include

publication bias and errors in data collection, study exclusion and data extraction. To mitigate against these threats, the following counter-measures were implemented. In terms of publication bias, whereby studies are more likely to select positive results over negative ones, this is mitigated through the Quality Assessment which attempts to ascertain whether studies discuss their results with limitations. In regards to this review, as the aim is to report the efficacy of different methods in the field, rather than present new results of its own, there is also no motivation from this review to only include studies that report positive results—similar to other SLRs identified by Kitchenham et al. (2010). To mitigate against omitting studies based on the search criteria, a broad search string was used as discussed in Section 3.4.2. It could be argued that the date range used could be expanded to studies that were published before 2016; however, as discussed in Section 2.3, fake news only became popularised from this year onwards. This decision is further justified by the results presented later in this Chapter, in Section 3.5.1, which showed a steep increase in publications addressing fake news starting from 2017. Regarding errors in study exclusion and data extraction, where studies may have been incorrectly excluded or the data extraction erroneous, this was mitigated through a review by a secondary researcher. In the initial stages of study selection based on title and abstract, this was carried out by a single author with the sole purpose to only exclude studies that were undoubtedly out of scope (erring on the side of inclusion for any title/abstract that was deemed doubtful). During the selection by content stage, a random sample of papers was taken and reviewed by the secondary author. This approach appears to be the most popular for SLRs, as demonstrated by Carver et al. (2013), although there is no standard amount of papers to use for this random sample. It was agreed that a significant but manageable number of papers should be undertaken for review by a secondary author, in this case 20%, with an agreement threshold of 90%. This percentage of papers and agreement between the two researchers was also used for a different random sample in the data extraction stage to ensure data was extracted accurately.

3.5 Results

In the following sections, the results of the study are discussed. Initially, an overview is provided of the included studies in Section 3.5.1. Section 3.5.2 provides the results of the Quality Assessment. In Section 3.5.3, results relating to RQ1 are presented about the methods used in the studies – including choice of datasets, features and machine learning algorithms. Finally, Section 3.5.4 describes the results relating to RQ2 which focuses on the effectiveness of these approaches. The results conclude by outlining the studies that focus on generalisability of current approaches (RQ2.3).

3.5.1 Overview of Included Studies

This study identified 293 papers from 2016–2023 that were relevant to the research questions. 97 of these were from journals and the remaining 196 were from conference proceedings. Figure 3.2 displays where these studies were found, and Figure 3.3 displays the year in which the studies were published. As can be seen from Figure 3.3, most selected studies were from later years in the defined range, with no studies being identified in 2016. This supports the decision to keep the selection range between 2016 to 2023. The steep slope in Figure 3.3 may also indicate that it is a relatively new but also rapidly growing area of interest.

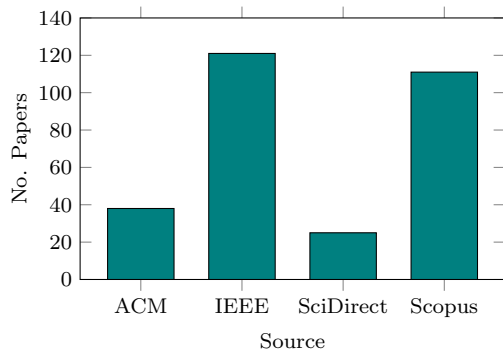


Figure 3.2: Study Sources

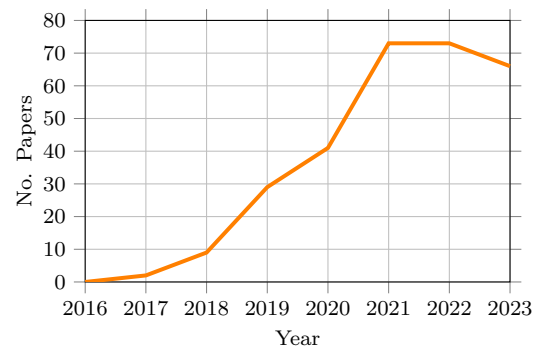


Figure 3.3: Years of Publish

Upon reviewing the most popular journals and conferences that contributed to this body of work, several key findings emerge. Among the journals, *Multimedia Tools and Applications* and *IEEE Access* were the most prolific, each contributing seven papers to the research landscape. Following closely are *Expert Systems with Applications*, which published six relevant papers, and *Procedia Computer Science*, which contributed four.

In terms of conference contributions, the *IEEE International Conference on Big Data* led with eight papers, demonstrating its prominence in the field. This was followed by the *International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, and the *International Conference on Computing Communication and Networking Technologies (ICCCNT)*, each contributing four papers. The *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* and the *ACM International Conference on Information and Knowledge Management (CIKM)* also each added four papers to the overall count, underscoring the significant role these conferences play in disseminating research findings in this area.

3.5.2 Quality Assessment Results

A Quality Assessment (QA) of the studies was performed principally to assist in data synthesis as well as to provide insights for future research Kitchenham (2004). Five quality assessment questions were derived and can be found in Section 3.4.3.

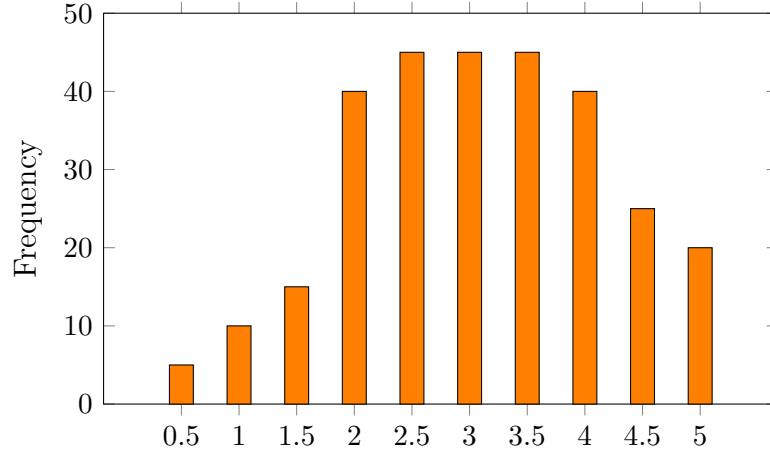


Figure 3.4: Frequency of Overall QA Scores

The distribution of scores presented in Figure 3.4 demonstrate that the majority of studies achieved moderate to high-quality ratings, with the highest frequency observed for scores between 2.5 and 4.0. This indicates that most studies met key quality criteria but also highlights areas for improvement in achieving methodological robustness. Lower scores (0.5–1.5) were less frequent, suggesting that only a small subset of studies lacked significant aspects of quality assurance. Conversely, few studies achieved the highest score of 5.0, indicating room for improvement even among the better-rated works.

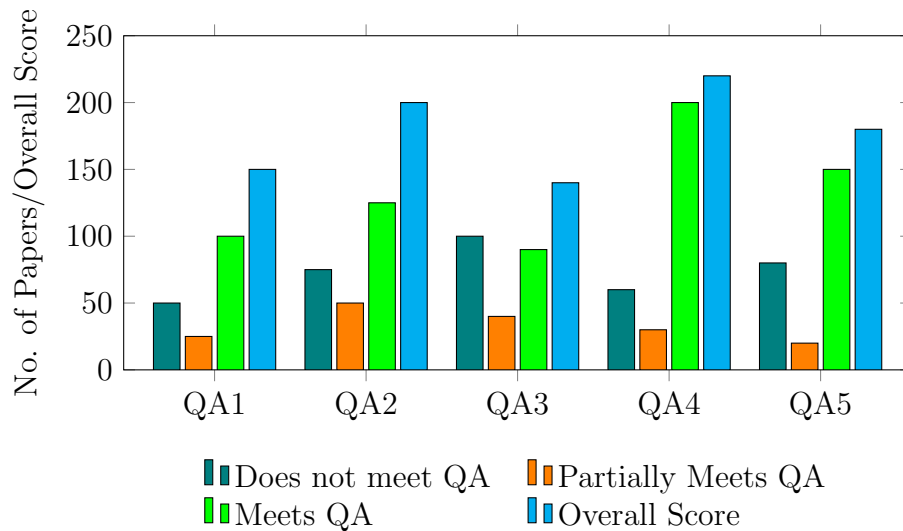


Figure 3.5: QA Results per Question

QA1: Definition of Fake News

Relating to QA1, it was noted that 31% of the studies did not provide any definition of “Fake News”; rather, these papers discussed the current state of fake news and opportunities in research before moving onto their own approach for solving the problem. 17% of studies were marked as “partially” in answering QA1 and were generally studies that alluded to what fake news is, without providing an explicit definition. An example of this can be found in Nath et al. (2021), which, in the introduction, explains the impact of fake news which gives the reader some insight to what fake news is but without providing an explicit definition. Lack of clarity in the definitions used in a study may be seen as problematic. As described in Section 3.2, the study of fake news is an emerging field with no agreed definition of what fake news is. This means that there are deviations in how the fake news problem is being understood and, in turn, being approached and solved.

QA2: Disclosure and Access to Datasets

Relating to QA2, 34% of studies did at least partially disclose what dataset was used, typically by citing a previous study that has used a dataset while omitting a direct reference to the dataset, or, by describing a dataset on Kaggle² without explicitly citing which dataset. A further 43% disclosed the dataset fully with a direct citation to the dataset used. On the other hand, the 23% of the studies marked as not disclosing the dataset at all were typically studies where a custom dataset was used, which was created by the authors. These studies would largely describe how the dataset was produced, typically through web-scraping and labelling based on where they were scraped from but would not provide access to the dataset. Disclosing the dataset used could help create performance benchmarks, support transparency and discourage concerns around bias.

QA3: Dataset Contents

In relation to QA3, 45% of studies did not adequately describe the contents of the dataset that was used, particularly in studies that presented models which only trained on textual features. This meant that it was unclear what aspects of a news article were used; for example, whether the headline, author and publication date were used in training. As many of the models are not easily explainable, knowing the contents of the dataset used to train the model could provide some transparency into how a model differentiates between different types of news in the dataset.

²<https://kaggle.com/>

QA4: Evaluation of Methods

The majority of selected papers (73%) provided a robust set of evaluation metrics as part of QA5, including accuracy, precision, recall, F1-score, and ROC-AUC. Alternatively, papers including the confusion matrices were also considered as fulfilling this criteria. This comprehensive approach allows for a nuanced assessment of model performance, addressing various aspects and facilitating better comparison across studies. In contrast, 12% of the papers only partially met this criterion, typically reported two metrics, such as accuracy and F1-score. Finally, 15% of papers did not provide a robust set of metrics, relying solely on accuracy. This is problematic as it can omit details on how well a model performs in predicting different classes. Providing a robust set of metrics helps create detailed performance benchmarks, enhances transparency, and mitigates concerns about bias, leading to more reliable and credible research findings.

QA5: Discussion of Results

Extending on QA5, 48% of papers provided an adequate discussion of results, highlighting study weaknesses, areas for improvement, and comparisons with the literature. These discussions often included detailed analyses of performance, error sources, methodology limitations, and future work proposals. This thoroughness helps understand the study's impact and fosters transparency and advancement. 7% of studies partially met this criterion by providing comparison tables with existing literature but lacked in-depth analysis. These papers missed opportunities to contextualise findings, discuss performance nuances, or explore implications, leaving readers with an incomplete understanding. 45% of papers failed to provide an acceptable discussion, merely enumerating results without analysis or context. It was noted that most of these studies were conference papers, suggesting that page restrictions likely limited the amount of critical discussion.

3.5.3 Methods of Fake News Article Detection (RQ1)

This section presents the findings related to RQ1 on methods for detecting fake news articles. The first sub-section addresses RQ1.1 and explores the datasets used in developing fake news detection models. This is followed by an examination of the features employed to train fake news detection models (RQ1.2) and the machine learning algorithms applied in these models (RQ1.3). The analysis covers a comprehensive review of the current literature, highlighting the diversity in datasets, the range of features used in training models, and the various machine learning algorithms implemented. The results provide a detailed overview of the state-of-the-art

in fake news detection methods, offering insights into the strengths and limitations of different approaches.

Datasets (RQ1.1)

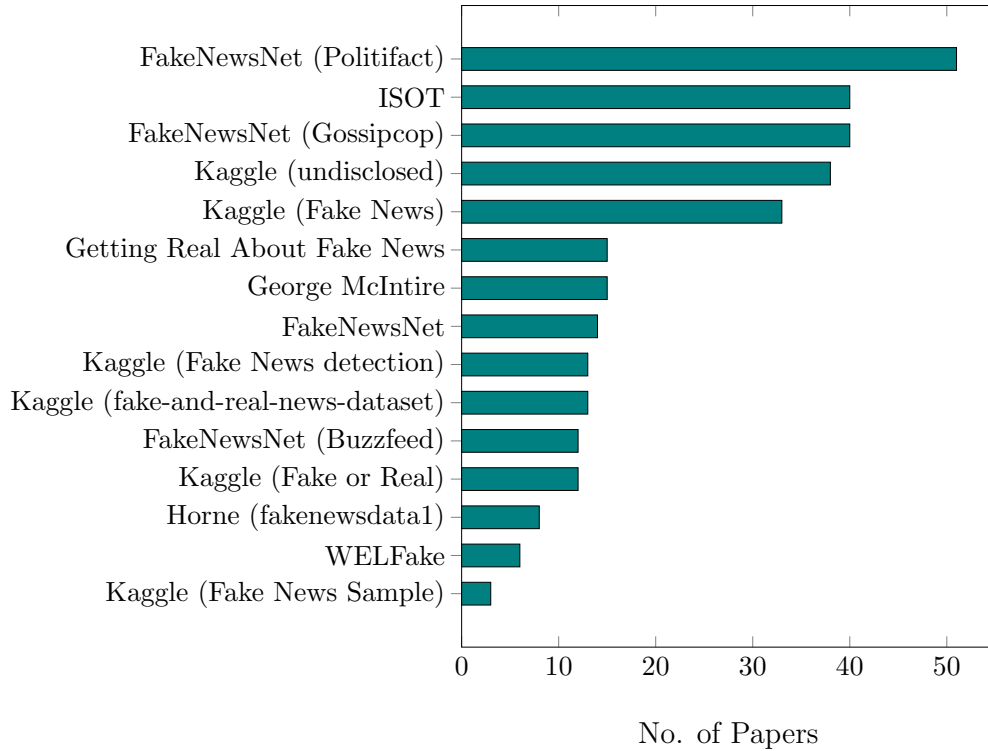


Figure 3.6: Top 15 Fake News Datasets (Reversed Order)

Figure 3.6 outlines the top 15 most popular established datasets, accounting for 70% of the total datasets used across all studies captured by this review. 17.5% of studies either did not disclose the dataset used, gathered their own dataset, or combined a number of datasets together to use in training. Among these, 19 studies used combined datasets, while 24 studies employed custom datasets. Approximately 40 other datasets make up the remaining 12.5% of datasets used in the collected studies, as such these less popular datasets have been excluded from this section of the review.

As can be seen from Figure 3.6, variations of the FakeNewsNet datasets are among the most popular in the literature. There are two versions of this dataset, one that includes Politifact and GossipCop segments and an old version that includes Politifact and BuzzFeed segments. Different studies use different combinations of these segments, hence why they have been separated in Figure 3.6, often preferring to train models on different segments individually. Given the Politifact segment exists in both versions of the dataset, it is logical that this segment is the most popular, found in 11% of the studies collected by this review. Of the other FakeNewsNet

Table 3.5: Dataset Details

Name	Real	Fake	Description
FakeNewsNet (Politifact) (Shu et al., 2019b)	624	432	Articles labelled as “Fake” or “True” by Politifact. Attributes: ID, URL, Title, Tweet ID
ISOT (Ahmed et al., 2018)	21,417	23,481	Real news articles collected from Reuters and fake news articles collected from sites listed as unreliable by Politifact and Facebook. Generally covers political news from 2016-2017. Attributes: Article text, Title, Date, Topic, Label
FakeNewsNet (GossipCop) (Shu et al., 2019b)	16,187	5,323	Articles labelled as “Fake” or “True” by GossipCop. Attributes: ID, URL, Title, Tweet ID
Kaggle (undisclosed)	N/A	N/A	Datasets extracted from Kaggle, but not specifically disclosed in the literature
Kaggle (Fake News) (Lifferth, 2018)	10,413	10,387	A fake news challenge dataset hosted on Kaggle. Data broadly covers the 2016 US Presidential Election. Attributes: ID, Title, Author, Text, Label
Getting Real About Fake News (M. Risdal, n.d.)	0	12,999	Hosted on Kaggle, this dataset only contains fake articles labelled by BS Detector. Covers the 2016 US Presidential election. Typically combined with other datasets for the ‘real’ class. Attributes: Author, Publish Date, Title, Article Text, Date Crawled, Site URL, Country.
George McIntire (McIntire, 2017)	5,279	5,279	Fake news articles collected from Kaggle (likely, the Getting Real About Fake News dataset – although this is unclear) with real news collected from All Sides. Broadly covers US news from 2015-2016, including the election. Attributes: Title, Article Text, Label
FakeNewsNet (Shu et al., 2019b)	16,811	5,755	Studies that utilise both sections of the FakeNewsNet dataset (typically, the Politifact and GossipCop segments) Attributes: ID, URL, Title, Tweet ID
Kaggle (Fake News detection) (R. Jain, n.d.)	1,872	2,137	Dataset hosted on Kaggle, no information provided on the date range collected, contents or labelling strategy. Attributes: URL, Headline, Body, Label
Kaggle (fake-and-real-news-dataset) (C. Bisailon, n.d.)	20,826	17,903	Dataset hosted on Kaggle, no information provided on the date range collected, contents or labelling strategy. Attributes: Title, Text, Subject, Date
FakeNewsNet (Buzzfeed) (Shu et al., 2019b)	91	91	An older version of the FakeNewsNet repository, covering news from the 2016 US Election. Attributes: ID, Title, Text, URL, Top Image, Movies, Authors, Source URL, Publish Date, Movies, Images
Kaggle (Fake or Real) (Jillani, n.d.)	3,128	3,128	Dataset hosted on Kaggle, no information provided on the date range collected, contents or labelling strategy. Attributes: Title, Text, Subject, Date
Horne (fakenewsdata1) (Horne and Adali, 2017)	128	123	Incorporates 2 datasets, “Buzzfeed” and “Random Political News”. Random Political News was gathered from Business Insider’s “Most Trusted” list and Zimdars 2016 Fake news list. Buzzfeed originally collected and labelled by journalist Craig Silverman. Attributes: Text, Label
WELFake (Verma et al., 2021)	37,106	35,028	An aggregation of four datasets (Kaggle (undisclosed), McIntire, Reuters, BuzzFeed Political). Attributes: Title, Text, Label
Kaggle (Fake News Sample) (Pontes, n.d.)	N/A	N/A	Dataset hosted on Kaggle, no information provided on the date range collected or contents. Not labelled binarily, includes clickbait, satire and fake labels. Appears to focus on news from 2018. Attributes: ID, Domain, Type, URL, Content, Title

segments, the GossipCop segment is the next most popular (found in 8% of studies) followed by the older BuzzFeed segment (found in 2.7% of studies).

It is notable that a significant number of studies use datasets from Kaggle, but do not disclose specifically which datasets from Kaggle are used. Failing to cite a dataset presents a number of issues including limiting the reproducibility of studies and the amount of scrutiny that can be performed on a study (for example, if the underlying dataset is inherently biased). While Kaggle datasets are clearly popular in the literature, care must be taken when using these community datasets as they often lack documentation and are not necessarily verified for quality. Of the Kaggle datasets that are cited specifically, the “Kaggle (Fake News)” (Lifferth, 2018) dataset, used as part of a Kaggle competition, is the most popular. Other datasets of note include the ISOT dataset (Ahmed et al., 2018), found in 9% of the studies of this review. These datasets, along with others counted in the Top 15, are outlined in more detail in Table 3.5.

Features (RQ1.2)

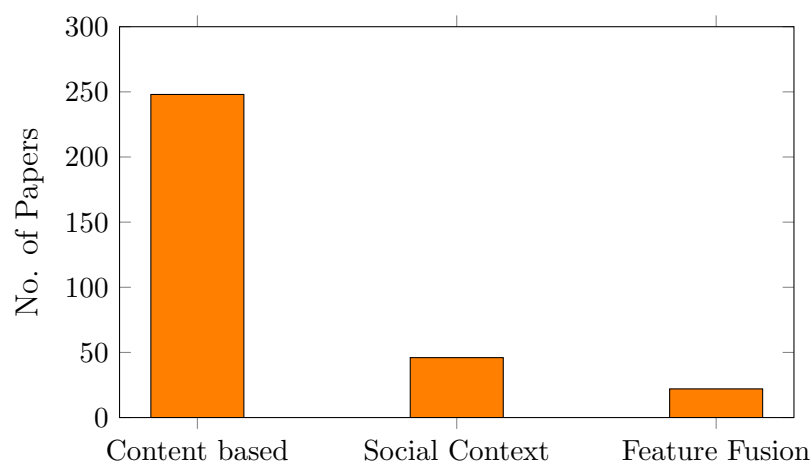


Figure 3.7: Comparison of Feature Types

In this section, we examine the different types of features utilised in fake news detection models. Xie et al. (2020a) offers a broad categorisation of three types of approach depending on the features used. These features are as follows:

- **Content Based:** Features derived from the main body of the article, including textual features and visual features.
- **Social-Context Based:** Features derived from user profiles, social media post and propagation paths.
- **Feature-Fusion:** Features that combine the first two categories.

Figure 3.7 clearly shows that content-based features are the most prevalent, featuring in 248 studies. This prevalence is likely due to two factors: most datasets in the literature focus primarily on textual features (as observed in Table 3.5), and many studies use experiments on textual features as a baseline for comparison when incorporating social-context or fused features. Social context features follow but are significantly less favoured, likely due to the limited number of datasets that include these features and the increasing restrictions on social network access for research purposes. The least popular approach are fused-features that combine both content-based and social-context features.

Content-Based Features

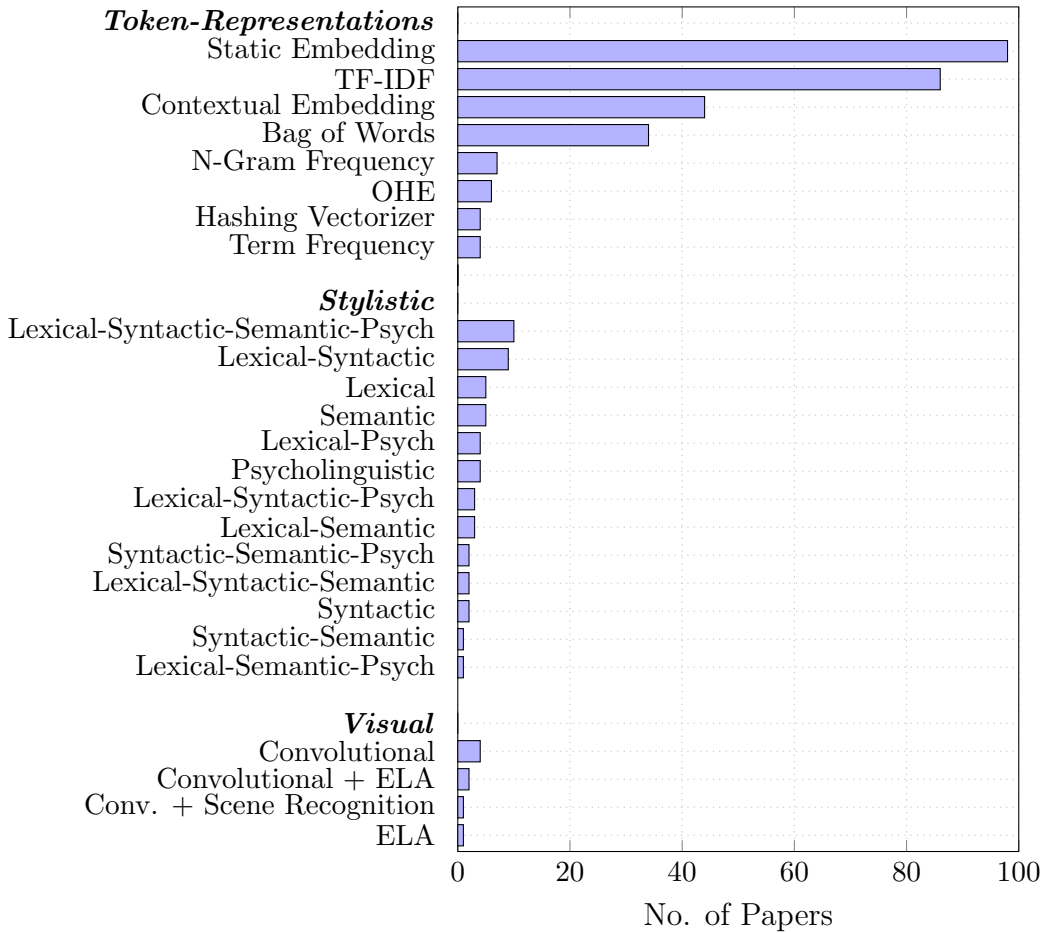


Figure 3.8: Overview of Feature Usage

Focussing on content-based features, we can broadly divide these into the following categories:

- **Token-Representations:** refer to numerical representations of words, such as token-occurrence analysis (TOA) based approaches (such as Bag-of-Words

or TF-IDF), static embeddings (such as Word2Vec and FastText) and contextual embeddings (such as ELMO and BERT).

- **Stylistic Features:** pertain to the unique characteristics and patterns in the writing style and structure of a text. These features help analyse how the text is constructed and its stylistic elements.
- **Visual Features:** refer to the specific attributes or characteristics extracted from images that algorithms use to understand and classify visual data.

Table 3.6 outlines the specific features captured by this review categorised under each of these types, providing a comprehensive overview of the different features utilised in fake news detection. Figure 3.8 highlights the prominence of these individual features used in isolation in the studies reviewed, with token-representations and stylistic features appearing in 205 and 38 studies respectively. This preference towards textual features is indicative of the current focus on text-based datasets as outlined in Section 3.5.3. The use of token-representations in nearly all studies covered by this review, also underscores their frequent use in comparison experiments when evaluating novel approaches.

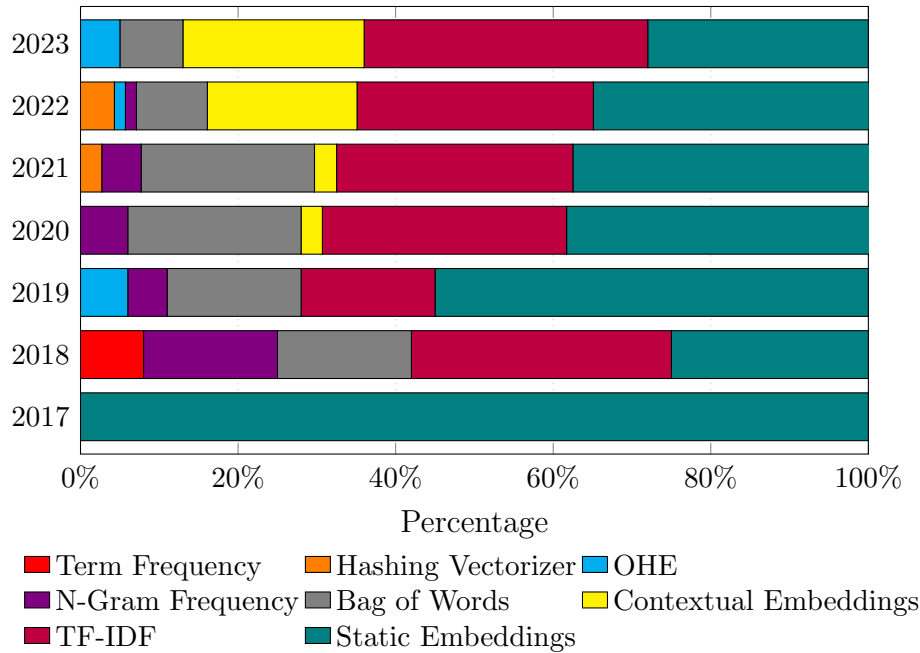


Figure 3.9: Percentage of Token-Representations by Year

Regarding token-representations, static embeddings are the most frequently used. Of these static embeddings, Word2Vec, Glove and embedding layers from frameworks such as Keras are the most popular, featured in 22, 39 and 27 studies respectively. Simpler token-occurrence analysis (TOA) methods, such as Bag-of-Words and TF-IDF, are frequently employed as well with 34 and 86 studies respectively

Table 3.6: Content-Based Feature Descriptors

Category	Feature	Description
Token-Representations	Bag-of-Words	Represents text by counting the frequency of each word in the document, disregarding word order and context.
	Term Frequency (TF)	A metric that quantifies how often a particular word or term appears in a document, typically normalised by the total number of words in the document to provide a relative measure of its importance within that specific text.
	TF-IDF	Computes the importance of a word in a document relative to its frequency in the entire corpus, balancing common and rare words.
	Static Embeddings	Static word embeddings like Word2Vec or GloVe, where each word is represented by a fixed vector regardless of its context.
	Contextual Embeddings	Dynamic word embeddings generated by models like ELMo or BERT, where word representations vary depending on the surrounding context.
	One-Hot-Encoding	Represents each word as a binary vector with a length equal to the vocabulary size, where only the index of the word is set to 1, and all other entries are 0.
	N-Gram Frequency	Captures sequences of N words or characters (e.g., bigrams, trigrams) to include some context and word order information in the text.
Stylistic	Hashing Vectorizer	Uses a hashing function to convert text into a fixed-size vector, efficiently handling large vocabularies by hashing word tokens into indices.
	Lexical	Focuses on features related to word choice and vocabulary usage, such as the frequency of different types of words. While these features can include token-occurrence analysis methods like Bag-of-Words and TF-IDF, in this study, we have chosen to classify lexical features separately from these token-representations, focussing on lexical features that are distinct from token-representations.
	Syntactic	Pertains to the structure and arrangement of words and phrases in sentences, including sentence complexity, grammar patterns, and syntactic structures.
	Semantic	Relates to the meaning and interpretation of words and sentences, including conceptual similarity, meaning relationships, and semantic roles. While these features can include embeddings such as Word2Vec and BERT, in this study, we have chosen to classify semantic features separately from these embeddings, focussing on semantic features that are distinct from these embeddings.
Visual	Psycholinguistic	Examines language from the perspective of cognitive and psychological processes, such as language processing, linguistic cues related to emotional states, and cognitive load.
	Convolutional	Uses Convolutional Neural Networks (CNNs) to analyse and extract features from images, often using pre-trained models such as ResNet or VGGNet.
	ELA	Detects image manipulation by analysing differences between the original and edited versions of an image, highlighting inconsistencies.
	Scene-Recognition	Identifies and categorizes scenes or contextual elements in images, such as places, weather conditions, or seasons. For example, recognising an image as “rainy” or “sunny” or classifying a scene as “beach” or “mountain.”

utilising these approaches. Advanced contextual embeddings are also notable, with 46 studies incorporating them. Among these, embeddings based on BERT-like models are the most widely used, likely due to their open-source nature compared to embeddings like GPT. The higher apparent popularity of static embeddings can be attributed to their availability before 2018, which is relevant given this review’s coverage of studies from 2016 to 2023. Figure 3.9 provides evidence of this, showing the percentage of studies using various token representations by year. Static embeddings feature prominently throughout this period, whereas contextual embeddings do not feature until 2020. After their introduction in fake news detection literature, the use of contextual embeddings grows significantly in 2022 and 2023, accounting for a similar percentage of studies as static embeddings. Therefore, the perceived prominence of static embeddings in Figure 3.8 is likely influenced by their earlier introduction compared to more recent contextual embeddings.

The lower popularity of stylistic features relative to token-representations may be attributed to two factors. Firstly, as token-representations such as BoW and TF-IDF can be classed as lexical features and embeddings such as Word2Vec and BERT classed as semantic features, it could be argued that lexical and semantic features are more prominent due to the widespread use of such token-representations. However, for this study, we focus specifically on stylistic features that are distinct from these token-representation techniques in order to offer a more comprehensive analysis of the differing types of features used in the literature. Secondly, stylistic features are less frequently used in comparison experiments. Such features are more time-consuming to reproduce in comparison to established token-representation approaches where pre-trained models and libraries to produce such features are readily available.

Observing Figure 3.8, we can see that various combinations of stylistic features are utilised in the literature. The most common combination incorporates all four groups of stylistic features denoted in Table 4, observed in 10 studies. Among these, five use Linguistic Inquiry Word Count (LIWC) (Ahmad et al., 2020; Spezzano et al., 2021; Shu et al., 2019b; Gôlo et al., 2021; Rai et al., 2022), a text analysis tool that quantifies emotional, cognitive, and structural components of language by categorising words into predefined psychological and linguistic categories. Following this combination is “Lexical-Syntactic”, appearing in 9 studies (Gravanis et al., 2019; Reddy et al., 2020; Sheikhi, 2021; Abeynayake et al., 2022; Aluri et al., 2022; Sverdrup-Thygeson and Haddow, 2021; Castillo et al., 2021; Kumar Jain et al., 2020; Seddari et al., 2022). This combination often captures the frequency of different types of words (such as nouns, adjectives, and adverbs), as well as variations in punctuation and sentence complexity. Following these combinations, lexical and semantic features used individually are the next most popular, likely due to their

use in ablation experiments where different groups of stylistic features are included and excluded to observe the resulting effects on models. Syntactic features are used alone in only 2 studies (Castillo et al., 2021; Uppal et al., 2020), suggesting they may not be considered effective for accurate fake news detection when used in isolation. Similarly, psycho-linguistic features are not used exclusively. Overall, most studies combine multiple groups of stylistic features, with lexical and syntactic features being the most prominent across these combinations.

In regard to visual features, these are the least utilised, primarily because only a small number of datasets include images associated with news articles, or the source URL of the articles such that images may be extracted independently. Among these visual features, convolutional features – often extracted using pre-trained models like ResNet and VGGNet (Athira et al., 2022; Mangal and Sharma, 2020) – are particularly favoured, likely because of the convenience of applying these pre-trained models, similar to textual vectors such as TF-IDF and BERT. Additionally, Error Level Analysis (ELA) is also seen to be relevant feature, due to its ability to reveal inconsistencies and manipulations in digital images, which are often used in conjunction with fake news to deceive readers Meel and Vishwakarma (2021).

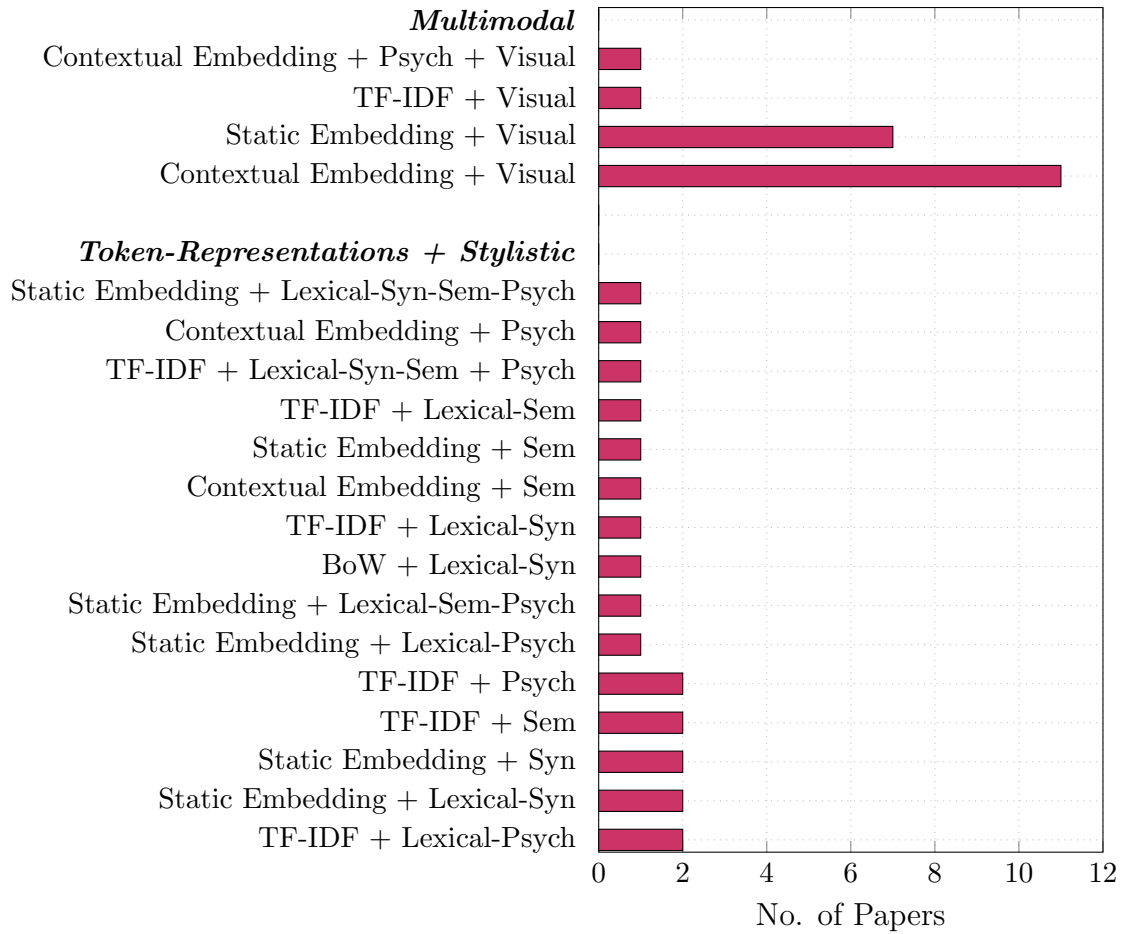


Figure 3.10: Combinations of Content-Based Features

Where Figure 3.8 outlines the use of features used in isolation, Figure 3.10 outlines the number of studies that use varying combinations of these content-based feature groups. In terms of studies that combine textual and visual features (also known as multimodal approaches), 10 studies utilise contextual embeddings (Cui et al., 2019; Masciari et al., 2020; Singhal et al., 2021; A et al., 2022; Zhou et al., 2023; Liang, 2023; Giachanou et al., 2020; Madhusudhan et al., 2020; Guo et al., 2023; Xiong et al., 2023) with 7 studies combining static embeddings (Mangal and Sharma, 2020; Ferreira et al., 2022a; Raj and Meel, 2021; Babar et al., 2024; Zhang et al., 2022; Nadeem et al., 2023b; Cui et al., 2019). Comparing this to the visual features used in isolation in Figure 3.8, this suggests that visual features alone are perhaps not considered effective for accurate fake news detection, with studies generally preferring to combine visual features with textual features. In terms of combinations of token-representations and stylistic features, static embeddings and TF-IDF are more frequently combined with a variety of stylistic features.

Social-Context Features

Regarding social-context features, Shu et al. (2017) offers a broad categorisation of these features. Table 3.7 outlines how these groups are considered in this study.

Table 3.7: Overview of Social-Context Features

Category	Description
User	These features represent the characteristics of users interacting with news on social media. These may encompass features aimed at assessing the user’s credibility, demographic or activity metrics.
Post	These features analyse the posts containing news, rather than evaluating the news articles themselves. Similar textual features may be extracted from these posts, encompassing token-representations, stylistic or visual features. Alternatively, metrics such as likes and re-tweets may be analysed.
Network	Users form different networks on social media based on interests, topics, and relationships, making network-based features valuable for detecting fake news. These features are extracted by constructing various networks, such as stance networks, co-occurrence networks, friendship networks, and diffusion networks, to represent network patterns and apply metrics like degree and clustering coefficient, or by learning latent node embedding features.

Of the three groups of social-context features, network and user-based features are the most prominent, appearing in 8 (Davoudi et al., 2021; Soga et al., 2023; Jeong et al., 2022; Zhou and Zafarani, 2019; Kaur, 2023; Wu and Wang, 2021; Wu, 2023; Davoudi et al., 2022) and 6 studies respectively (Xie et al., 2020b; Cui et al.,

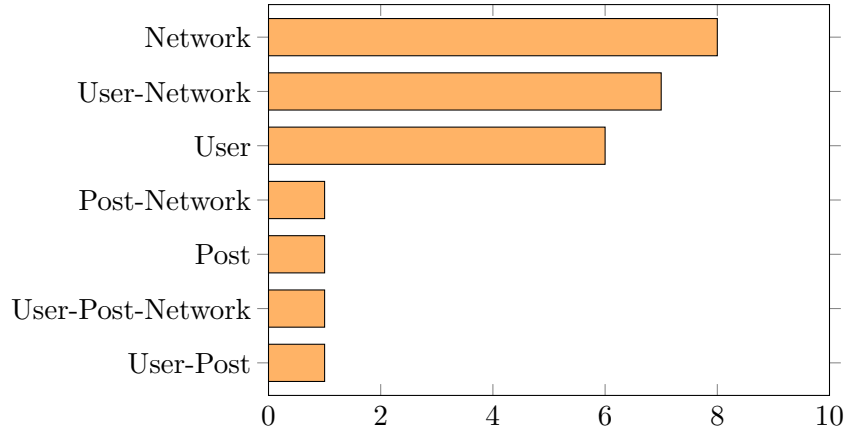


Figure 3.11: Overview of Social-Context Features

2019; Chowdhury et al., 2020; Shu et al., 2019b; Mahmud et al., 2022; Wu et al., 2023), with a combination of these two groups occurring in 7 studies (Freire and Goldschmidt, 2019; Shu et al., 2019b; Kaliyar et al., 2021; Qureshi et al., 2022; Su et al., 2023; Saikia et al., 2022; Tschatschek et al., 2018). This indicates a strong preference for incorporating network-based and user-related features in fake news detection, with comparatively fewer studies focusing solely on post features or their combinations with other feature types. This emphasis is perhaps due to the studies’ focus on fake news articles, rather than fake news occurring in social media posts. This suggests that the demographics of users sharing these articles and their propagation through social networks are considered more relevant in the detection of fake news articles.

Fused Features

Figure 3.12 provides a high-level overview of the studies that employ “fused-features”. As can be seen from this figure, token-representations are the most commonly combined with social-features. This is expected given the popularity of token-representations in content-based approaches. Similar to content-based approaches, static embeddings are the most prominently used in conjunction with social-context features. Reflecting the findings of social-context features used exclusively, user and network features are the most commonly used throughout. Only four studies leverages all groups of features (token-representations, stylistic, social and visual) (Ferreira et al., 2022b; Nadeem et al., 2023a; Cui et al., 2019; Hlaing and Kham, 2020).

Machine Learning Algorithms (RQ1.3)

This section provides an overview of the various machine learning algorithms used in the field of fake news detection. These algorithms range from more traditional

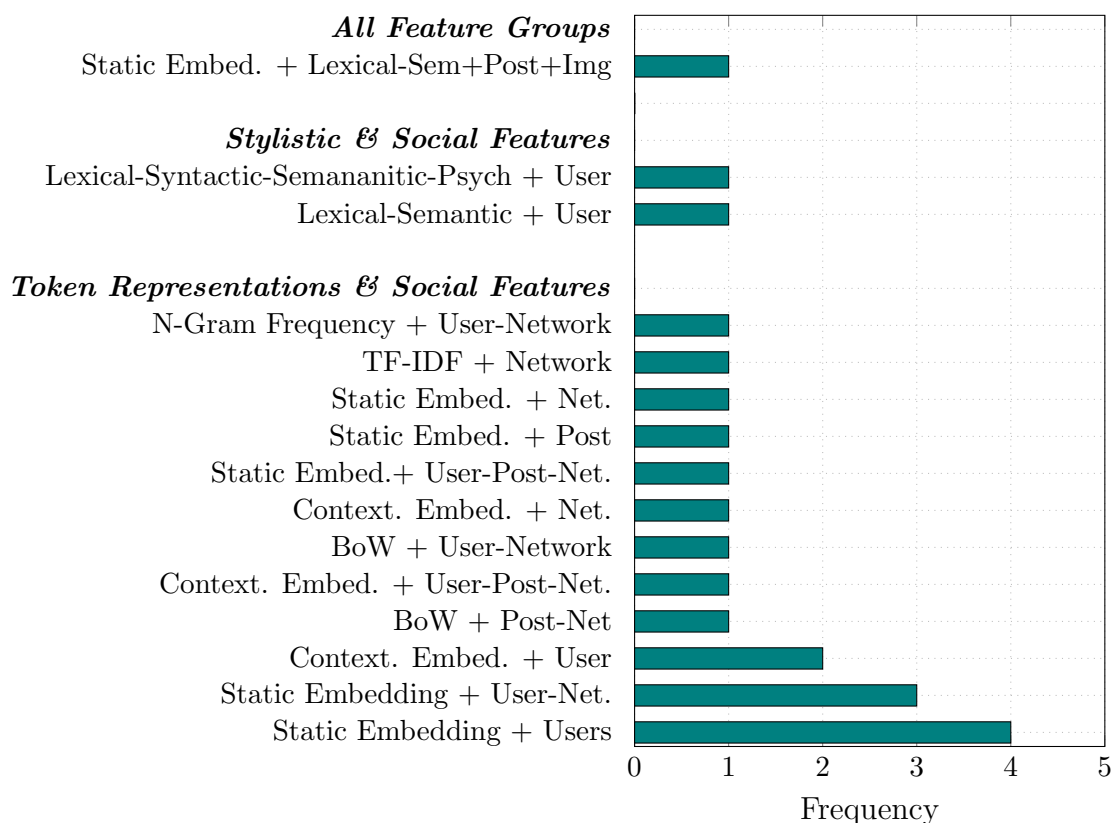


Figure 3.12: Overview of Fused-Feature Usage

approaches such as Logistic Regression and Decision Trees, to more advanced models such as Long Short-Term Memory Networks (LSTMs) and Transformers.

Figure 3.13 shows the frequency of these algorithms across the papers collected in this review. In terms of more classical machine learning algorithms, Logistic Regression, SVMs and Naïve Bayes are among the most favoured for the fake news detection task, each observed in over 80 studies collected by this review. As fake news detection is often considered a binary classification task, whereby news articles are typically labelled as ‘true’ or ‘fake’, these algorithms and their effectiveness in handling binary outcomes make them well suited for this task. Additionally, their simplicity also sees these algorithms frequently used in comparison experiments, facilitating analysis between these simpler algorithms and more complex, novel models proposed in research. Furthermore, given the relatively small size of current datasets in this domain, these algorithms may be favoured due to their ability to generalise with smaller amounts of data. As such, these algorithms are often seen paired with simpler feature representations such as Bag-of-Words and TF-IDF, as well as stylistic features. Decision Trees, while seeing less use than these algorithms, may also be used for these reasons. Similar to Logistic Regression and Naïve Bayes, Decision Trees are often considered more interpretable, which is important in this domain given the high degree of similarity between real news and true news. This allows re-

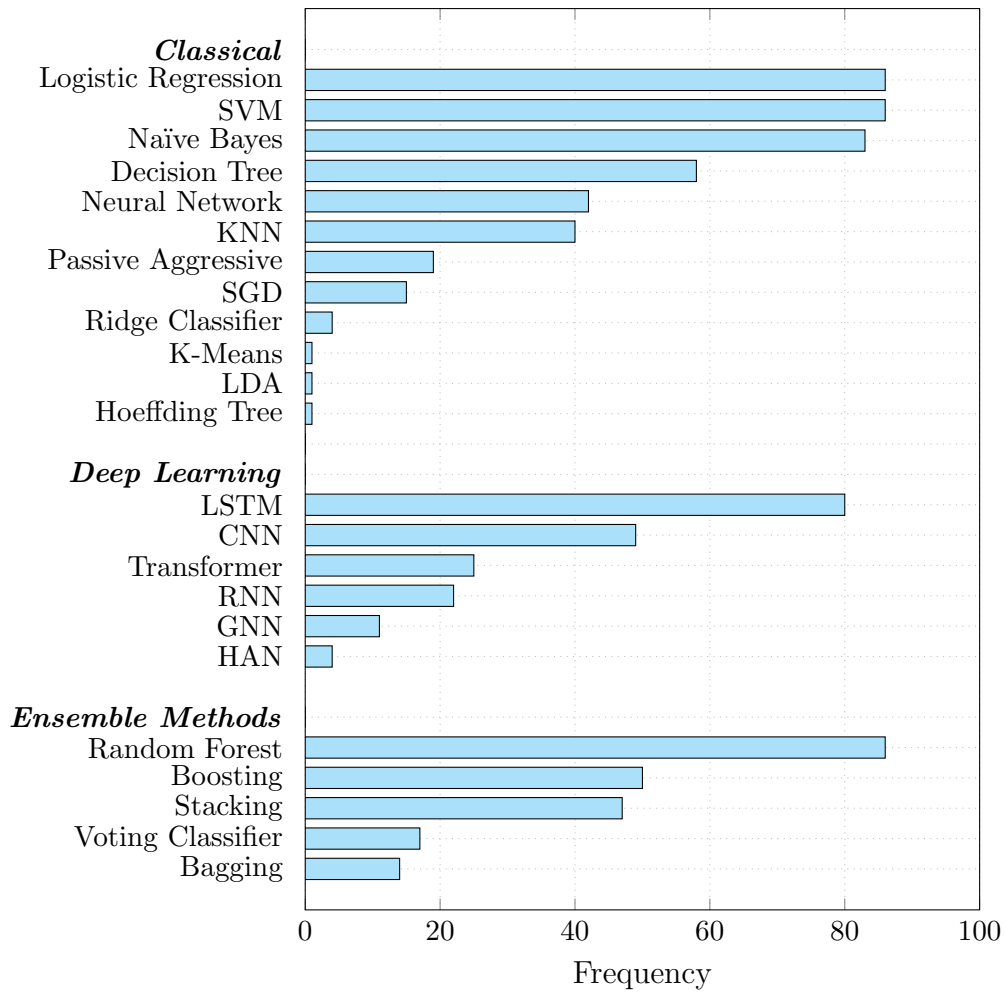


Figure 3.13: Overview of Algorithm Usage

searchers to understand and explain the decision-making process, providing valuable insights into the distinguishing features of fake news.

Classical algorithms are in contrast to deep learning algorithms, which are computationally expensive to train and are less interpretable. However, such algorithms are able to capture complex relationships, which is potentially valuable in the challenging task of distinguishing fake news from real news. Of these deep learning algorithms, LSTMs are the most favoured, likely due to their efficacy on sequential data such as text. LSTMs are typically employed alongside static embeddings like Word2Vec or contextual embeddings such as BERT, which enhance their ability to represent semantic and contextual nuances in text, further improving their performance in fake news detection tasks. While LSTMs are equally popular to the more simplistic algorithms previously mentioned, they are not typically used for comparative analysis, instead their use is aligned with more complex, novel models proposed in the literature. Interestingly, CNNs see more use than RNNs, perhaps as a result of the popularity of LSTMs which address some of the weaknesses in RNNs, such as the vanishing gradient problem. Additionally, RNNs struggle with long sequence

lengths, such as those seen in news articles, potentially making them inappropriate for the fake news detection task. Similarly, Transformers are used less frequently due to their later introduction in 2018, with their increasing popularity likely reflecting trends in contextual embeddings noted in Figure 3.9.

In terms of ensemble methods, that is, methods that combine multiple models, Random Forest is the most popular with a similar level of popularity to Naïve Bayes, Logistic Regression and SVMs. This popularity may be attributed to similar reasons to those for classical machine learning algorithms. This includes the relative computational efficiency of Random Forest compared to other ensemble methods (particularly those reliant on deep learning) as well as its interpretability in terms of providing feature-importance scores. Boosting algorithms follow Random Forest in Figure 3.13, encompassing algorithms such as Gradient Boosting, AdaBoost and XGBoost. Of these algorithms, Gradient Boosting is the most prevalent, appearing in 21 studies with AdaBoost and XGBoost appearing 15 and 14 times respectively. Similar to Random Forest, such algorithms are relatively simple to implement, potentially explaining their high popularity. This is in contrast to stacked models, observed in 37 studies which are more complex. Such models are often observed in ‘multimodal’ and ‘feature-fusion’ approaches, that incorporate different types of features, such as textual and visual data. Reflecting on the low availability of such features in current datasets, this potentially explains why stacked models see less use in the literature overall.

3.5.4 Effectiveness of Current Methods (RQ2)

This section presents the findings related to RQ2 on the effectiveness of current methods for detecting fake news articles. The first sub-section addresses RQ2.1 and explores the performance of the various groups of features outlined in Section 3.5.3. This is followed by an examination of the machine learning algorithms used in classification (RQ1.2). During the data extraction phase of this review, it was noted that the most frequently used metrics for measuring performance were accuracy, F-score, precision and recall. Accuracy was the most commonly used metric in the reviewed studies, and, as such, it will be used as the primary metric in this analysis, providing a high-level overview of performance. Additionally, we consider the number of datasets on which features and algorithms have been trained, as this factor helps inform the generalisability and robustness of the methods employed.

During the data extraction process, it was observed that approximately 98% of studies rely on K -fold cross-validation or hold-out testing for evaluation. Hold-out testing involves dividing the dataset into separate training and testing sets to evaluate model performance. K -fold cross-validation, on the other hand, involves

dividing the dataset into K subsets, training the model on $K-1$ subsets, and testing it on the remaining subset, repeating this process K times to ensure a comprehensive evaluation. When multiple results were presented for a model with differing hyperparameters, the best result was selected. In instances where multiple results were provided with different train/test splits for the same model, the average was taken.

Performance of Features (RQ2.1)

Table 3.8 and Figure 3.14 provide an overview of the performance of various groups of features observed in the literature. Due to the extensive use of token-representations in the literature, it was decided to include the specific features within this group individually. In contrast, features that are represented in fewer studies are consolidated into their high-level groups.

Table 3.8: Average Accuracy of Features

Group	Feature	No. Datasets	Mean Acc.
Token-Representations	Term Frequency	3	0.92
	Hashing	3	0.79
	One Hot Encoding	3	0.86
	N-Gram	3	0.79
	Bag of Words	11	0.88
	Contextual Embedding	15	0.85
	TF-IDF	19	0.84
	Static Embedding	21	0.86
Other Groups	Stylistic & Social Features	2	0.94
	Token-Rep. & Social	5	0.83
	Social	4	0.83
	Multimodal	8	0.86
	Token-Rep. & Stylistic	9	0.86
	Visual	4	0.85
	Stylistic	12	0.78

Observing Table 3.8, it can be determined that the majority of features used in the literature achieve $\sim 80\%$ accuracy or better, indicating a high level of effectiveness across different types of features. In terms of mean accuracy, the most performant of these groups of features is Term Frequency and the combination of Stylistic and Social features. However, these features are tested on a relatively small number of datasets, therefore it is premature to conclude that these features are the most

performant. More testing on these features is therefore necessary to confirm their robustness.

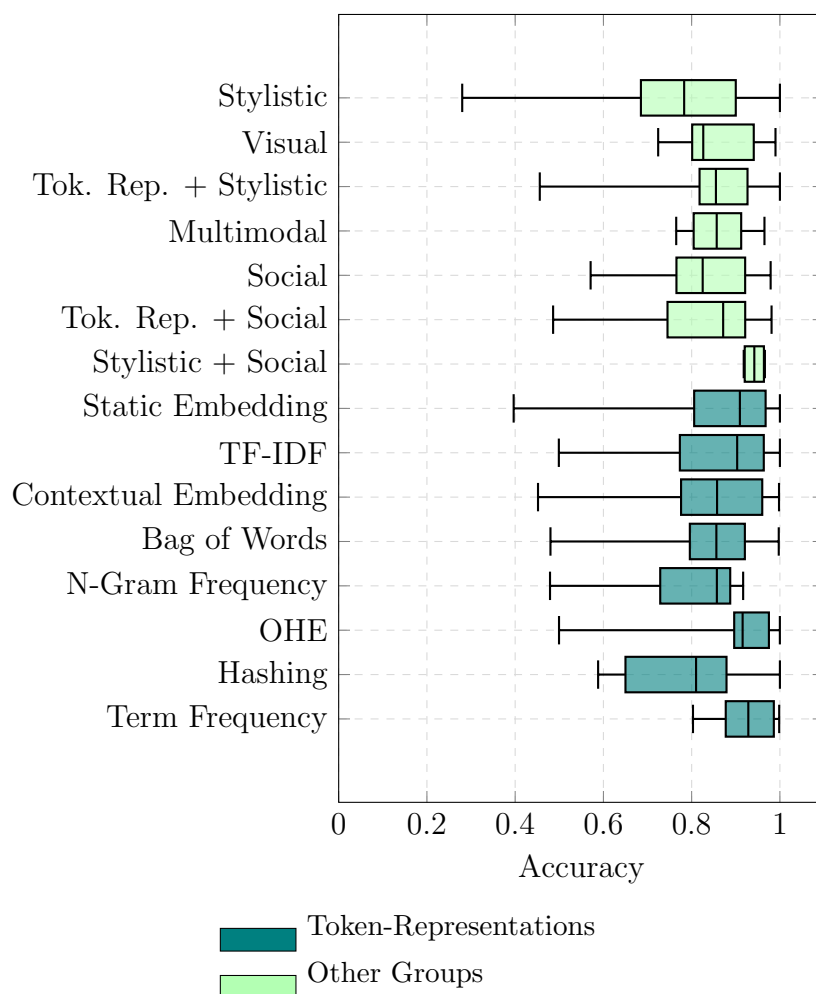


Figure 3.14: Feature Performance

Features such as token-representations that are observed in more studies and tested on a larger number of datasets see a larger degree of variation. In terms of static embeddings, observed in approximately a third of studies collected in this review, the mean accuracy is 86%. Considering the larger number of different datasets and studies that use these embeddings (with Word2Vec, GloVe and FastText being the most popular), we can have greater confidence in the robustness and effectiveness of these features. In regard to contextual embeddings, while appearing in fewer studies, their application across 15 datasets and mean accuracy of 85% also indicates that these embeddings are effective. Simpler means of token-representation also see a high-level of performance, with TF-IDF and Bag-of-Words achieving mean accuracies of 84% and 88% respectively. Interestingly, these results indicate similar or better performance than contextual embeddings, indicating that these simpler approaches still have value for the task of fake news detection. Furthermore, given the performance features, they may be favoured over more complex embeddings owing

to their computational efficiency and interpretability.

In terms of other groups of features, or those that combine various feature groups, stylistic features are the most prevalent in the literature tested across 12 different datasets. While less effective than token representations at a mean accuracy of 78%, it is important to note that there is large degree of variation in the different types of stylistic feature used, as observed in Section 3.5.3. This is in contrast to token representations, which are more standardised and therefore may tend to exhibit more consistent performance. Combining token representations and stylistic features exhibit an improved performance, achieving a mean accuracy of 86%, suggesting this combination leverages the strengths of both sets of features. Similarly, multimodal approaches also result in an average accuracy of 86%, albeit on a smaller number of datasets. Studies incorporating social features also perform well, with mean accuracy ranging from 83% with social features alone and social features incorporating token-representations to 94% with social features combined with stylistic features. However, as noted in Section 3.5.3, fewer studies utilise these features. Consequently, more research is needed to fully understand the effectiveness of these approaches and to confirm their robustness across a wider range of datasets, however, this may be a challenge as social media companies become more restrictive in the data available for collection. Overall, these findings highlight the potential benefits of incorporating additional features beyond textual data, suggesting that a more comprehensive feature set could enhance fake news detection models.

Performance of Machine Learning Algorithms (RQ2.2)

Table 3.9 and Figure 3.15 provide an overview of the performance of the various machine learning algorithms observed in the literature for the task of fake news detection. While the previous section summarised some of the features into their higher-level groups, all the groups of algorithms have been enumerated in this section.

Comparing Figure 3.15 containing the performance of machine learning algorithms and Figure 3.14 the performance of features, it appears there is less variation in the median performance of the machine learning algorithms. The relatively stable performance of machine learning algorithms indicates that these models tend to perform consistently across different datasets when given varying types of features. This consistency suggests that the algorithmic methods themselves are robust and capable of leveraging the information provided by the features effectively. In contrast, the greater variability in the performance of features highlights that the choice of features has a significant impact on the outcomes of fake news detection. The varying performance levels of different features emphasise that certain types of features may be more or less effective depending on how well they capture relevant

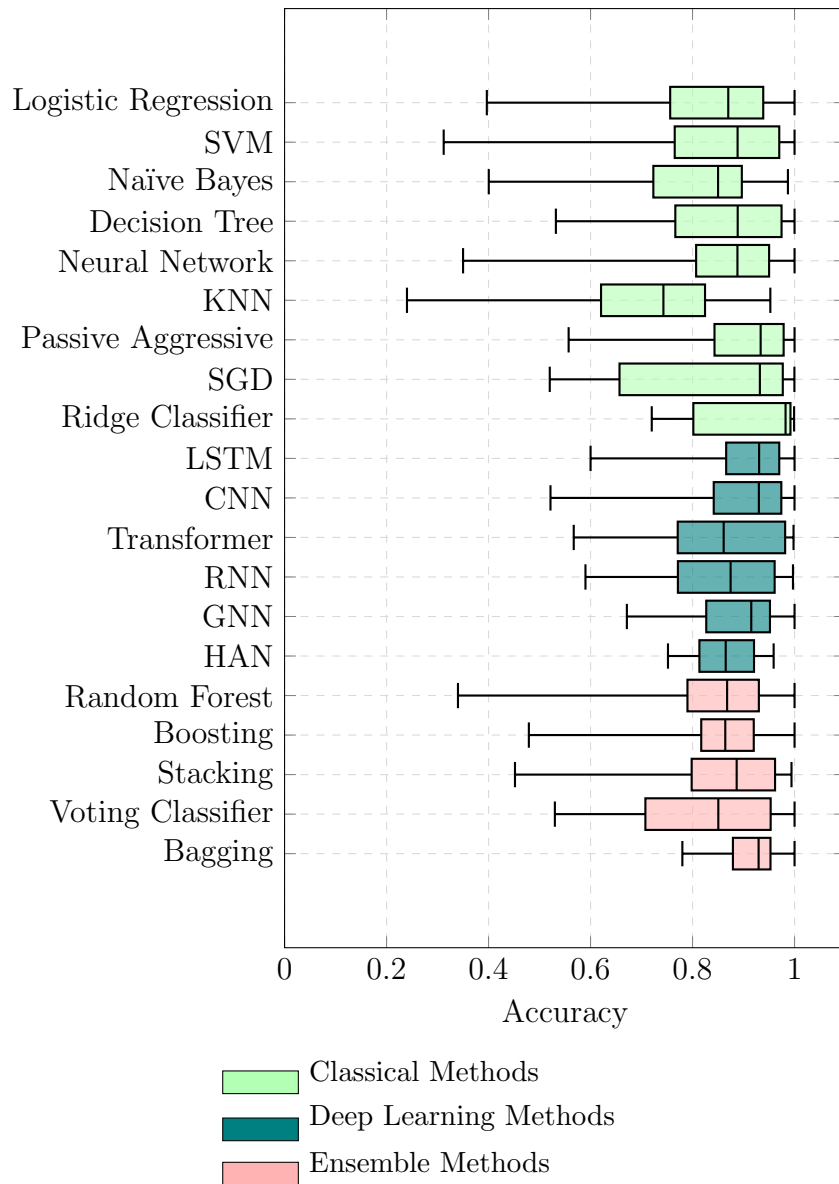


Figure 3.15: Performance of Machine Learning Algorithms

information for distinguishing between fake and real news. As such, it can be argued that the choice of features is more important than the choice of algorithm for the task of fake news detection.

While the overall performance of these algorithms is more consistent when compared with features, there still remain a number of variations between these algorithms. Observing classical machine learning algorithms, we can determine K-Nearest Neighbour (KNN) is the worst performing across all the algorithms collected in this review with a mean accuracy of 72%. Given its relatively poor performance and the number of datasets it has been applied to, it can be argued that KNN is less suitable for the fake news detection task. Its reliance on distance metrics, which may not be effective in high-dimensional textual data, likely contributes to its lower performance. Logistic Regression and Naïve Bayes, while more effective than KNN,

Table 3.9: Average Accuracy of Machine Learning Algorithms

Group	Algorithm	No. Datasets	Mean Acc.
Classical	Logistic Regression	16	0.83
	SVM	20	0.85
	Naïve Bayes	20	0.81
	Decision Tree	17	0.85
	Neural Network	15	0.85
	KNN	16	0.72
	Passive Aggressive	12	0.9
	SGD	9	0.81
	Ridge Classifier	5	0.9
Deep Learning	LSTM	19	0.91
	CNN	17	0.89
	Transformer	11	0.86
	RNN	11	0.85
	GNN	10	0.89
	HAN	6	0.87
Ensemble Methods	Random Forest	19	0.84
	Boosting	19	0.86
	Stacking	16	0.86
	Voting Classifier	7	0.82
	Bagging	7	0.92

also exhibit certain limitations. Logistic Regression achieves a mean accuracy of 83% across 16 datasets, while Naïve Bayes achieves a mean accuracy of 81% across 20 datasets. Both models are relatively simple and fast, but their performance can be hampered by their linear assumptions (in the case of Logistic Regression) and the assumption of feature independence (in the case of Naïve Bayes). Support Vector Machines (SVMs), Decision Trees, and Neural Networks show more promising results, each achieving a mean accuracy of 85%. SVMs, tested on 20 datasets, are particularly strong in high-dimensional spaces and are effective when there is a clear margin of separation between classes. Decision Trees, tested on 17 datasets, are easy to interpret and can handle both categorical and numerical data well. Neural Networks, tested on 15 datasets, are highly effective at capturing complex patterns in data but can require significant computational resources and fine-tuning. Passive Aggressive Models and Ridge Classifiers stand out with high mean accuracies of 90%, although they are tested on fewer datasets (12 for Passive Aggressive and 5 for

Ridge Classifier). Passive Aggressive Models are well-suited for online learning and adapt quickly to new data, which may be advantageous in adapting models inline with changing news landscapes. In contrast, Ridge Classifiers excel at managing multi-collinearity and regularization, offering robustness against overfitting.

Turning to deep learning algorithms, LSTM networks achieve the highest mean accuracy of 91% across 19 datasets. As mentioned in Section 3.5.3, LSTMs are particularly well-suited for long sequence data, making them effective for text classification tasks like fake news detection. CNNs and Graph Neural Networks (GNNs) also perform well, with mean accuracies of 89% across 17 and 10 datasets respectively. RNNs, while less popular in the literature overall, see similar mean accuracy to Transformers at 85% and 86% respectively. Owing to the high computational demands of Transformers, it is common for studies to use a pre-trained model and fine-tune it to fake news detection task. Consequently, while Transformers have the advantage of capturing intricate patterns and contextual nuances in data, their effectiveness in some studies may be limited by the choice of pre-trained models rather than the full potential of Transformer architectures. This limitation may be further exacerbated by the relatively small size of current datasets, which restrict the ability of Transformers to effectively fine-tune and adapt to specific tasks.

In contrast, ensemble methods generally show robust performance, leveraging the strengths of multiple models to improve accuracy. Bagging, tested on 7 datasets, achieves the highest mean accuracy of 92%, indicating its effectiveness in reducing variance and improving stability. Boosting and Stacking, both with mean accuracies of 86% across 19 and 16 datasets respectively, also perform well, benefiting from their approaches to reducing bias and combining multiple models. Random Forests, tested on 19 datasets, show a mean accuracy of 0.84, highlighting their robustness and ability to handle diverse data types. Voting Classifiers, with a mean accuracy of 0.82 across 7 datasets, aggregate predictions from multiple models to enhance overall performance.

Generalisability of Fake News Detection Models (RQ2.3)

During the data extraction process of this systematic review, it was noted that the vast majority of studies solely relied on holdout testing or K-fold cross-validation. Only four studies performed additional testing, such as external validation, where models were evaluated on entirely independent datasets not used during the training phase. This approach provides a more rigorous test of model generalisability, defined as the ability of a model to perform well outside the dataset on which it was trained, and robustness. The studies that provided evidence that their models generalise are discussed below.

Horne et al. (2020) explored the generalisability of a model over time. Using the NELA-GT dataset, their findings indicate that as time progresses, the classification performance for both unreliable and hyper-partisan news classification gradually degrades. However, this degradation occurs slower than expected, illustrating that hand-crafted, content-based features, such as writing style, are fairly robust to changes in the news cycle. They also show that this small degradation can be mitigated using online learning, where the predictive model is updated as new data becomes available. Additionally, they examine the impact of adversarial content manipulation by malicious news producers, testing three types of attacks based on changes in the input space and data availability. Their results show that static models are susceptible to content manipulation attacks, but online models can recover from such attacks.

In contrast to this, Gautam and Jerripathula (2020) investigated the cross-domain generalisability of two distinct models by examining how well they performed across different news topics, specifically celebrity and political news. They found a significant drop in accuracy when testing models between these topics, with a 39% accuracy drop for the political model tested on celebrity news and an 8% drop for the celebrity model tested on political news. It is important to note, however, that the smaller size of the datasets (490 articles each) limits the reliability of these findings on larger corpora. The celebrity model's performance was less affected by cross-dataset testing, but this may be due to its lower initial accuracy of 78% compared to the political model's 95%, resulting in less potential for a substantial decline in performance.

Similarly, a study by Blackledge and Atapour-Abarghouei (2021) also explored how well models generalise across different topics by testing across two datasets: the ISOT dataset and the Combined Corpus (CC) dataset. Most of the data contained in the ISOT dataset is political in nature whereas the Combined Corpus covers additional topics such as healthcare, sports and entertainment. Additionally, these datasets are significantly larger than the datasets used in Gautam and Jerripathula (2020) at 44,898 and 79,548 rows respectively. This experiment therefore is perhaps more representative of generalisability across topics. It was found that the hold-out test performance was high at over 90% for each dataset. When testing across datasets however, a drop in accuracy was observed of approximately 25% on the model trained on the ISOT dataset and tested on the CC dataset. A less significant drop was found between the model trained on the CC dataset and tested on the ISOT dataset of around 15%. This further supports the finding that models do not generalise well across topics. The less significant drop in accuracy between the CC dataset model and the ISOT dataset could be attributed to the fact that both datasets contain political news whereas the ISOT dataset does not cover all the

topics contained in the CC dataset. It is also possible that there is a degree of duplicity between the two datasets as the CC dataset combines data from other datasets which may, in fact, include the ISOT dataset.

Janicka et al. (2019a) also found similar results in cross-domain generalisability across four datasets and points to the issue of generalisability arising, in part, due to the state of current datasets used in the literature. They advocate for the development and utilisation of more diverse datasets that better represent the wide range of fake news scenarios encountered in real-world contexts. Additionally, they emphasise the importance of employing more robust labelling strategies to ensure that datasets accurately capture the nuanced characteristics of fake and real news. Currently, nearly all datasets in the literature rely on a coarse labelling strategy, whereby articles are labelled by their publisher as a proxy for accuracy. Janicka et al. (2019a) highlight that this reliance on publisher-based labelling introduces significant biases, as it assumes that all content from a particular publisher can be uniformly classified as either fake or real. This coarse labelling fails to account for the subtleties within individual articles, such as instances where ‘credible’ publishers may inadvertently (or intentionally) publish misleading information or where traditionally unreliable sources may produce accurate content. Consequently, this approach can lead to misleading evaluations of fake news detection models, as the models may appear more effective than they truly are when tested against such oversimplified datasets.

These studies highlight that despite the broadly positive results of current approaches to fake news detection in Section 3.5.4, as well as their performance over time, there are weaknesses in terms of the generalisability of current approaches that require further investigation. Specifically, the existing approaches often perform well in controlled settings such as in holdout-testing or cross-validation but may struggle when applied to diverse or novel contexts.

3.6 Discussion

Fake news detection is a relatively new field, as can be seen from the substantial increase in publications from 2016 onwards, as presented in Section 3.5.1. This rapid growth in interest has led to a wide range of approaches aimed at addressing the issue, primarily leveraging machine learning (ML) and natural language processing (NLP) techniques. Many of these methods are developed using datasets from platforms such as Kaggle or by researchers in the field, yet there remains no standardised set of approaches or established baseline datasets. As the field continues to evolve, it is important to explore both the methods used and their effectiveness in detecting fake news. To this end, this study investigated two key research questions: What

methods are available for detecting fake news (RQ1); and how effective are these methods (RQ2). The investigation included a quality assessment and analysis of the literature. It provided insight by addressing each of the research questions and sub-questions. In addition, it revealed some fundamental, wider issues within the field of fake news detection. The findings relating to the research questions and these wider issues are discussed below.

In regard to RQ1.1, this study identified a number of datasets used within the field. Among them, the most popular are those already established in the literature, particularly those hosted on the Kaggle platform. Additionally, custom and hybrid datasets, which are either tailored for specific studies or combine multiple sources, are also widely used. Regardless of their category, the datasets commonly adopt a coarse labelling strategy, where the publisher is used as a proxy to classify articles as “fake” or “real.” This is likely due to the significant manual effort to label articles individually. Despite taking this more efficient approach, established datasets in the literature are often relatively small in the field, with an average size of ~10,500 articles per class.

The majority of these datasets are also textual datasets, which is reflected in the features used for fake news detection (RQ1.2). The features used are overwhelmingly content-based, with token-representations such as Bag-of-Words, TF-IDF, static embeddings and contextual embeddings being the most commonly used. Similarly, a large variety of stylistic features are also applied, which focus on analysing the writing style, structure, and linguistic nuances of articles. Studies using social-context, visual features and fused features are significantly less popular, reflecting the datasets in the literature whereby only a couple of datasets, such as FakeNews-Net, include attributes that make the extraction of these features possible. Given social media companies’ increased restrictions on researcher access to their data, it is likely that content-based features will remain the most predominant in future research. As such, the research landscape in fake news detection will likely continue to prioritise content-based approaches, as these are the most accessible and readily applicable across various datasets. While content-based features relying on textual features provide a solid foundation for detecting deceptive information, their focus on textual content alone might limit the ability to capture other dimensions of fake news. As such, it may be beneficial to investigate other features beyond the text, such as visual features, to enhance the performance and generalisability of fake news detection models.

In terms of the performance of these features (RQ2.1), content-based features such as Bag-of-Words, TF-IDF, and embeddings (both static and contextual) consistently demonstrate strong performance, with average accuracies ranging from 84% to 88% across various datasets. Stylistic features, while less commonly used

and showing greater variability in their performance, often enhance model accuracy when combined with content-based approaches. The variability in performance for stylistic features is likely due to the diverse range of stylistic elements analysed. Additionally, combining token-representations with visual and social-context features has been observed to improve performance compared to using these features in isolation. In regard to visual features in particular, this provides evidence that the use of additional features external to the text can provide an enhancement to fake news detection models in the absence of social-context features going forward.

Regarding machine learning algorithms (RQ1.3), the selection of algorithms demonstrates a large variety in approaches to the fake news detection task, with classical machine learning algorithms such as Logistic Regression, SVMs, and Naïve Bayes being widely used for their simplicity and often employed as baselines to compare to more complex models. While deep learning algorithms are less commonly used overall, it was noted that LSTMs are also particularly favoured for this task as well as ensemble methods such as Random Forest and Gradient Boosting. In terms of the performance of these algorithms (RQ2.2), the results generally indicate that there is less variability between algorithms in comparison to features, suggesting that the features used have a more significant impact on performance. Although deep learning algorithms like LSTMs tend to perform slightly better than classical machine learning methods, the overall impact of feature selection appears to be more influential in determining the effectiveness of fake news detection systems.

Finally, it was found that the evaluation of the models typically relied on holdout testing or k -fold cross-validation (RQ2.3). This approach raises questions regarding the generalisability of these models, simply put, whether they are effective beyond the dataset on which they have been trained. This argument is revisited in the discussion of wider issues below.

Section 2.3 of Chapter 2 highlighted the varying definitions of what can be classified as ‘fake news’, as clickbait, rumours, satire or verifiably false articles have invariably been referred to as fake news in the literature (Bondielli and Marcelloni, 2019). This was also observed during the study selection process, in which several studies define their focus to be fake news but, on closer inspection, they were found to deal exclusively with clickbait articles. Due to these varying definitions, it has been argued that implementation of these models will lead to AI bias concerns and arguments that it will also undermine democracy and infringe free speech (Rainie et al., 2017). Furthermore, there is no consensus on whether to optimise models for better recall (capturing all instances of true news) or better specificity (capturing all instances of fake news). This lack of agreement on optimisation benchmarks complicates the development of effective models. Models optimised for high recall might capture more instances of true news but risk including false positives, while

those optimised for high specificity might miss some true news but ensure that fake news is accurately identified. Given these issues, it could be argued that it may be impossible to create a model that satisfies everyone’s definition of fake news across different topics. However, that does not exclude such models from being applicable to certain situations and remain useful, for example, for social media companies which are under increasing pressure to police their platforms.

If models are to be applied in real-world scenarios, they must be accurate, robust, and generalisable. A key factor in achieving this is the size and quality of the datasets used. However, as observed in 3.5 in Section 3.5.3, the most popular datasets in the field are relatively small, with a combined average size of approximately ~10,500 records per class. This limitation poses significant challenges for training machine learning models, as small datasets may not adequately capture the complexity and diversity of fake news encountered in real-world scenarios. Additionally, small datasets increase the risk of overfitting, where models perform well on the training data but fail to generalise effectively to unseen data. While the results from this review indicate that the mean accuracy of models is around 85%, this performance might not fully reflect the challenges that models will face when applied to more diverse and larger datasets. The relatively high accuracy observed could be partly due to the models being tailored to the specific characteristics of these limited datasets rather than truly generalising to new, varied examples of fake news. Section 3.5.4 partially demonstrates this, showing that cross-domain generalisability is a significant issue. However, it is also crucial to assess how well models generalise within the same domain, as holdout testing and K-fold cross-validation may not provide sufficient insight into a model’s robustness. These methods often assume that the variations within the training and validation sets are representative of the entire domain. However, this assumption may fail to account for the nuances and variability inherent within the same domain. Consequently, there is a need for more robust evaluation strategies to accurately assess a model’s ability to generalise effectively in its intended context. While results in the literature may appear promising, they should be interpreted with caution, as limitations in the underlying datasets can obscure the true challenges of fake news detection in broader and more complex scenarios.

These issues are compounded by the limitations of existing annotation approaches. Manual annotation, which provides high-quality labelled data, is labour-intensive and expensive, resulting in smaller datasets. Automated annotation enables the creation of larger datasets but often sacrifices label accuracy. For example, Kaggle’s “Getting Real About Fake News” dataset, containing only 13,000 articles, is labelled using the “BS Detector,” which identifies unreliable articles based solely on domain names. The accuracy of the BS Detector is not well-documented, raising concerns

that such automated approaches may perpetuate a domino effect, where poorly labelled data leads to suboptimal models. To develop high-quality datasets at scale, there is a critical need for ongoing investment in manual labelling or the advancement of more sophisticated annotation algorithms. Unsupervised methods may provide a promising avenue for streamlining the data labelling process by offering initial labels, which can then be refined through manual validation, thus balancing scalability with accuracy. However, care must be taken to ensure that the initial labels provided by such models are not taken as definitive. Instead, these labels should be treated as a starting point, subject to rigorous evaluation and adjustment, to avoid propagating errors that could compromise the reliability of downstream models. By improving both the size and quality of datasets, researchers can better equip models to perform effectively in real-world applications.

This review of the literature highlights that despite the broadly positive results of current approaches to fake news detection in Section 3.5.4, as well as their performance over time, there are significant gaps in understanding their generalisability. While cross-domain testing has revealed models' struggles to adapt across distinct topics like politics and entertainment, a more fundamental question remains unexplored: intra-domain generalisability. If models cannot reliably detect fake news within a single domain (e.g., different sources of political news), they are even less likely to perform effectively across domains or in real-world applications. This limitation is particularly critical as practical applications require models to handle both topic diversity and content variation within a domain. Current approaches often perform well in controlled settings such as holdout-testing or cross-validation but may struggle with even subtle variations in writing style, source bias, or temporal context within their intended domain. Understanding these intra-domain limitations is therefore crucial as a first step toward developing more robust and adaptable fake news detection systems. By identifying specific weaknesses in data, feature representations, or model architectures within a controlled, single-domain context, approaches for improving both intra- and cross-domain generalisability can be developed.

3.7 Chapter Summary

This chapter provided a systematic review of machine learning approaches to fake news detection, focusing on the datasets, features, and algorithms used in the field, as well as their effectiveness. The review highlighted that most commonly used datasets in the literature are not manually annotated, with many studies relying on community datasets from platforms like Kaggle, which often lack reliability. Additionally, the relatively small size of many datasets poses challenges for training

robust machine learning models, as these datasets may not adequately capture the complexity and diversity of fake news encountered in real-world scenarios.

In terms of features, token representations such as Bag-of-Words, TF-IDF, and embeddings (static and contextual) emerged as the most frequently used approaches, reflecting the predominance of textual datasets in the field. While these content-based features have demonstrated strong performance, the review noted that their sole reliance on textual elements limits their ability to capture other dimensions of fake news, such as social or visual context. Stylistic features, though less commonly used, were shown to enhance performance when combined with content-based approaches. However, features leveraging social and visual contexts were less explored due to limited dataset availability.

The review also highlighted the wide range of machine learning algorithms applied to the task, including classical algorithms like Logistic Regression, Support Vector Machines (SVMs), and Naïve Bayes, which are often used as benchmarks. Ensemble methods such as Random Forest and Gradient Boosting, as well as deep learning approaches like LSTMs, were also employed, with deep learning models showing slight performance advantages. However, it was observed that the choice of features generally had a greater impact on model performance than the selection of algorithms.

Finally, the chapter emphasised the significant challenges associated with evaluating fake news detection models. Most studies rely on holdout testing or k-fold cross-validation, which may not fully capture the generalisability of models to unseen data. While cross-domain testing has revealed the difficulty of adapting models across different topics, the more fundamental task of intra-domain generalisability remains unexplored. Addressing these limitations is essential for developing more robust and adaptable fake news detection systems capable of handling the complexities of real-world applications.

Chapter 4

Methodology

4.1 Introduction

The two previous Chapters indicated that machine learning approaches are the predominant method for detecting fake news, largely leveraging token-based and stylistic features. While these approaches have demonstrated strong performance under holdout and cross-validation conditions, they often face significant challenges when applied to new or unseen data not included in the training dataset. This lack of generalisability limits their practical applicability, particularly in the dynamic and evolving context of fake news detection. The variability in language, style, and context across datasets can reduce the effectiveness of these models, highlighting the need for methodologies that enhance robustness and adaptability. This chapter focuses on addressing these challenges by presenting a comprehensive framework for developing and evaluating machine learning models aimed at identifying the reasons behind poor generalisability and finding solutions towards improvement.

To this end, the chapter outlines a systematic methodology for assessing and improving intra-domain generalisability. The approach involves curating diverse datasets, employing a range of pre-processing and feature extraction techniques, experimenting with multiple machine learning algorithms, and evaluating models using robust validation strategies. Additionally, interpretability methods are applied to provide insights into the decision-making process of the models, ensuring the findings are both actionable and transparent.

The chapter is structured as follows. Section 4.2 outlines the research method, providing an overview of the experimental design and approach to investigating intra-domain generalisability. Section 4.3 focuses on data collection, detailing the different methods of data collection in the domain. Section 4.4 describes the pre-processing steps applied to the data to ensure consistency and prepare it for analysis. Section 4.5 explores the feature extraction techniques used to represent textual and

stylistic information. Section 4.6 introduces the machine learning algorithms employed for classification. Section 4.7 explains the methods used for evaluating model performance, with Section 4.8 detailing the metrics used to measure effectiveness. Finally, Section 4.9 discusses model interpretability techniques, highlighting how these methods provide transparency and insights into the factors influencing predictions.

4.2 Research Method

This section provides a high-level overview of the research method that shall be used in addressing the research questions outlined in Section 1.4. This research will adopt the experimental method, which aims to test hypotheses by systematically manipulating variables within controlled settings to determine causal relationships (Kamiri and Mariga, 2021).

In the context of text classification tasks, this is applied by making adjustments to different aspects of the text classification process. The review by Kadhim (2019) provides an overview of the text classification process, outlined in Figure 4.1. The steps within this process and techniques as applied in this thesis are outlined in Sections 4.3–4.9.

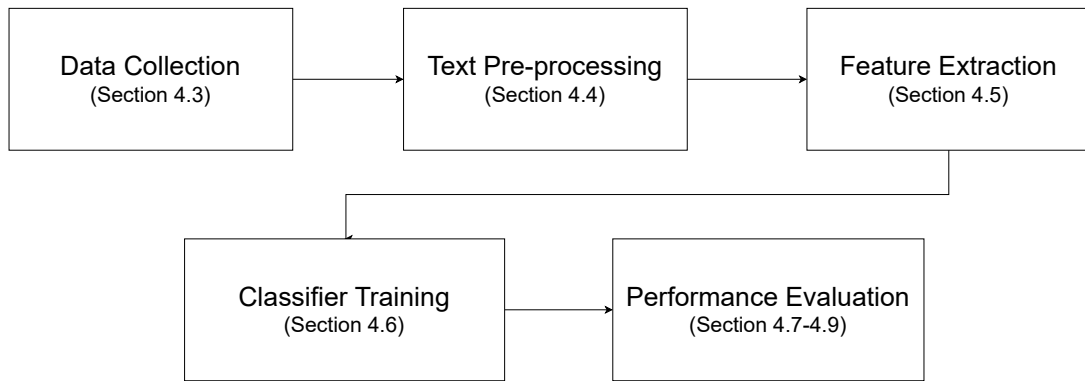


Figure 4.1: Text Classification Process

The process depicted in the Figure 4.1 provides a structured approach to machine learning text classification. It begins with Data Collection, which involves gathering textual data relevant to the classification task. Next, the data undergoes pre-processing, where noise is removed, and the text is standardized for analysis, including steps like tokenization, stopword removal, and stemming or lemmatization (Vijayarani et al., 2015). During feature extraction, the pre-processed text is transformed into numerical representations using techniques such as Bag-of-Words, TF-IDF, or word embeddings, which are essential for input into machine learning models (Pintas et al., 2021). This is followed by classifier training, where algorithms

are applied to learn patterns in the data and build predictive models (Li et al., 2022). Finally, performance evaluation assesses the model’s effectiveness using metrics like accuracy, precision, recall, and F1-score, ensuring the classifier meets the requirements for the intended application (Zhou et al., 2021). This pipeline ensures a rigorous and methodical approach to developing reliable text classification models.

When applying the experimental method to this process, researchers may investigate using different datasets to assess how model performance varies across different types of news articles, such as those observed in Section 3.5.4 which evaluated the performance of models across different news topics. Alternatively, in terms of text-preprocessing, researchers may observe the effects of different tokenisation techniques (such as breaking the text down into n-grams or to the sub-word level), or, varying the degree to which certain textual elements are removed from the text (such as punctuation, numbers or stop-words). In regard to feature-extraction, researchers may observe the effects of using different techniques, such as Bag-of-Words and TF-IDF or more advanced embeddings such as BERT. Extending this, researchers may look to apply dimensionality reduction techniques such as Principal Component Analysis (PCA) to observe the effects of simplifying the feature-space. Finally, researchers may look to experiment with different machine learning algorithms and their hyperparameters.

A key advantage of applying the experimental method in this way is its ability to offer precise insights into how different components of the text classification process affect model performance. By systematically varying datasets, preprocessing techniques, feature extraction methods, and classification algorithms, researchers can identify which factors contribute most effectively to improving model performance (Bouthillier and Varoquaux, 2020). This structured approach allows for a thorough evaluation of each element’s impact, leading to well-informed decisions on optimising models for better performance. Furthermore, the experimental method enables the replication of experiments, which enhances the reliability of the results. By adhering to controlled conditions and standardised metrics, researchers can validate findings and ensure that they are consistent across different trials and setups. This also facilitates systematic reviews, such as the one in Chapter 3, by providing a basis for comparing and synthesising results from various experiments, contributing to a more robust understanding of effective practices in fake news detection.

However, while the experimental method applied in this context provides several benefits, it also has disadvantages. For example, it can be resource-intensive, requiring significant computational power to train and test machine learning algorithms. Specifically training, especially with large datasets or deep learning algorithms, demands substantial processing resources. Using rigorous evaluation methods, such as K-fold cross-validation, further adds to this demand, requiring models to be re-

peatedly trained and tested K -times with different segments of the dataset (Gorriz et al., 2024). Furthermore, experiments done in this manner may not fully capture the complexities of real-world scenarios, potentially limiting the applicability of the results. This may be a result of poorly labelled datasets or overfitting, where models fit too closely to the training data, affecting their performance on new, unseen datasets.

Despite the demands on computational resources, this approach remains fundamental in addressing the research questions outlined in Section 1.4. Producing robust and comprehensive results is a key aspiration of this thesis, and the structured framework provided by this method is essential for systematically exploring and validating the effectiveness of various techniques and models. This rigorous approach is intended to offer valuable insights that drive advancements in the field of fake news detection, ensuring that the findings are both reliable and applicable to real-world contexts. The following sections shall elaborate on each step of this process in more detail and provide justification for the specific tools and techniques chosen.

4.3 Data Collection

This section outlines the data collection phase. Typically, studies in the literature will take one of three approaches to data collection.

The first and most commonly used approach is to use existing datasets, such as those hosted on Kaggle or through research institutions. Section 3.5.3 of Chapter 3 lists a number of these datasets, accounting for approximately 70% of the datasets used in the literature. Using such datasets is advantageous, as the time-intensive work of web-scraping and initial data gathering has already been completed. These datasets also often undergo a degree of pre-processing, excluding records that contain erroneous data as well as labelling. A further advantage to using such datasets is their reusability, enabling researchers to directly compare their findings to others using the same dataset.

However, while such datasets are convenient and readily available, there are a number of issues with these datasets. First, the methodology behind the data collection process may not always be transparent, particularly for datasets hosted on platforms such as Kaggle, where many datasets popular in the literature do not provide detailed information on the collection process (Hutchinson et al., 2021). This lack of transparency can make it challenging to understand the origins of the data and assess its suitability for specific research objectives. Furthermore, this lack of transparency may also make it more difficult to discern whether there are certain biases within the data which may invalidate research findings, particularly in the

context of classification. Without clear insights into the data collection and labelling process, researchers may not be able to identify or correct these biases, which could affect the accuracy of classification results. This can lead to misleading conclusions and undermine the reliability of fake news detection models, potentially impacting their effectiveness and generalisability in real-world applications.

An alternative approach to using already established datasets is to create custom datasets from the ground up (Roh et al., 2019). This can be particularly advantageous in gathering current or topic-specific data that may be relevant in answering specific research questions. While excluded from the systematic review in Chapter 3, studies that may favour this approach include those that focus on specific events such as the COVID-19 pandemic, reflecting the latest news around such events. This approach enables the inclusion of recent and relevant examples, which may improve model’s applicability in these specific circumstances. Furthermore, creating a custom dataset also offers the opportunity to employ a more refined labelling strategy, in contrast to most established datasets which typically employ a coarse labelling strategy based on news publisher reliability.

While this approach addresses some of the disadvantages of already established datasets, it also presents a number of distinct challenges. As previously noted, such an approach can be time-consuming and resource intensive (L’heureux et al., 2017). For instance, news publishers often do not have APIs to facilitate the extraction of data. This necessitates the use of web-scraping libraries, such as BeautifulSoup, to extract data from webpages. However, this approach to data extraction can present issues. Differing news publishers have different structures to their websites and therefore adjustments may have to be made to web-scraping scripts to account for these differences. This variability can make it challenging to standardise the data-collection process and ensure that the extracted data is both comprehensive and accurate. Furthermore, websites and APIs may additionally have protections in place to prevent the extraction of data or the abuse of API access. These may include CAPTCHAs and rate-limiting, which can hinder the scraping process and require additional strategies to bypass. Such challenges may encourage limiting extraction to a fewer number of sources which may result in a narrower dataset that does not represent the variability of news in the real-world.

The third approach involves hybrids of the above two approaches, typically by combining two or more pre-built datasets together or blending established datasets with additional, scraped data (Ahmad et al., 2020; Reddy et al., 2020; Kumar et al., 2023; Uppal et al., 2020). Owing to the limited size of current datasets, as explored in Section 3.5.3, this approach can enhance existing datasets by including current events. Alternatively, researchers may look to broaden datasets to include a larger range of topics. This approach therefore allows researchers to benefit from the pre-

processing and standardisation of established datasets while incorporating news, context-specific data to improve the overall applicability of the dataset. However, while this hybrid approach allows researchers to benefit from the advantages of the two prior approaches it can also introduce the disadvantages. These can include the difficulties in scraping news data, as well as the poor transparency of some established datasets, thus making it challenging to ensure data consistency.

Owing to the high popularity of currently established datasets and given the focus of this thesis on exploring current approaches to fake news detection, this thesis will also use established datasets. Utilising these datasets allows the thesis to benefit from the significant effort already invested in data preparation while also facilitating comparability with existing research. While these datasets often lack transparency in their data collection methodologies, this is an aspect that will be actively explored in this thesis. Specifically, RQ2 and RQ3 investigate how effective and generalisable current approaches are for fake news detection. This aims to examine whether the lack of transparency and potential biases in these widely used datasets could lead to misleading results or affect models' robustness and generalisability. This exploration is crucial as it will provide insights into the reliability of current approaches in fake news detection and help identify potential areas for improvement in dataset creation and utilisation. In this thesis, the specific datasets to be used in the experimental work will be detailed in the empirical chapters, Chapters 5 and 6, ensuring a comprehensive understanding of the data sources relevant to the experiments undertaken.

4.4 Pre-Processing

This section outlines the next steps in the text-classification process, pre-processing the raw textual data and the extraction of features from this data. Pre-processing is necessary to clean the text and remove unwanted noise, which helps ensure that data that is used as input to machine learning models is consistent. After this pre-processing stage, feature engineering and extraction is necessary to translate textual data into numerical inputs that models can process.

The first stage of the pre-processing phase typically begins with tokenisation of the data, where text is broken down into individual units. These are typically at the word-level, but text may also be broken down into the sub-word or character level. Alternatively, the text may be broken down into n-grams, which capture sequences of adjacent words or characters. This step is necessary before further processing as it converts words into an array of strings that can be further processed in subsequent stages (Vijayarani et al., 2015). Following tokenisation, other common pre-processing steps include the removal of stop words. These are words

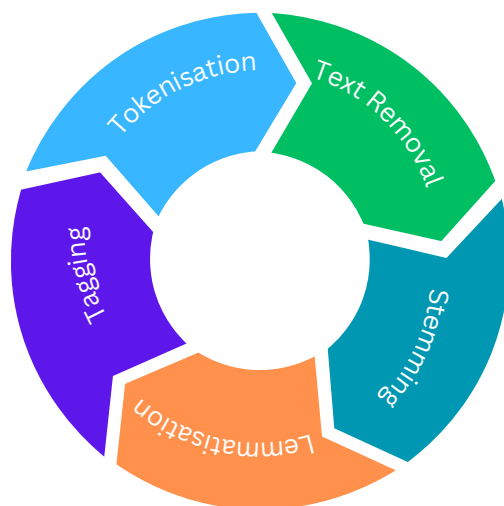


Figure 4.2: Text-Preprocessing

that are commonly used but do not contribute significant meaning, such as connectives like ‘and’, ‘or’ and ‘but’ or other conjunctions and prepositions. This helps remove unwanted noise from text, leaving words that are more relevant to the overall meaning of the text (Hickman et al., 2022). Other pre-processing steps may involve the removal of punctuation, eliminating symbols that are unnecessary for analysis. Furthermore, words may undergo stemming or lemmatisation, processes that reduce them to their root forms to minimise vocabulary size and focus on the core meanings of words. Additionally, part-of-speech (POS) tagging may be applied to identify the grammatical roles of words within the text, which can enhance downstream analysis by providing syntactic context (Chai, 2023).

It’s important to note that depending on the feature extraction approach, some or all of these steps may not be carried out. For example, static embeddings such as Word2Vec and contextual embeddings such as BERT often require less pre-processing (Albalawi et al., 2021). Static embeddings are typically pre-trained on large corpora and have a pre-defined vocabularies, negating the need for extensive processing or normalisation. Words that do not fall into these models’ vocabularies are either mapped to a placeholder token or ignored, allowing embeddings to still perform effectively with slightly noisy data. Similarly, contextual embeddings are designed to understand and adapt to various textual contexts and can often process text with minimal pre-processing. This inherent capability allows these models to handle diverse linguistic features and variations without the need for extensive pre-processing, thus simplifying the overall text preparation process.

In contrast, simpler approaches to token-representation such as Bag-of-Words or TF-IDF typically require more rigorous pre-processing (Pimpalkar and Raj, 2020). These approaches inherently rely on transforming text into a matrix of token counts or term frequencies and therefore can be significantly affected by noise (such as those

presented by stop-words or irrelevant punctuation). The presence of such noise can distort the feature-matrix, leading to less meaningful representations of the text. Cleaning the data and reducing noise therefore results in feature vectors that are more representative of the underlying content, thereby enhancing the performance of text classification models that rely on the token-occurrence analysis approaches.

Chapters 5 and 6 of this thesis, which detail the experiments, shall outline the specific pre-processing steps taken for each dataset and model. While this overview has covered common techniques such as tokenisation, stop word removal, and punctuation stripping, the precise pre-processing procedures applied in each study will be detailed in their respective chapters. These steps will be tailored to meet the specific requirements of the models and datasets used, with simpler approaches like Bag-of-Words and TF-IDF potentially requiring more extensive cleaning compared to advanced models such as static or contextual embeddings.

4.5 Feature Extraction

Following pre-processing, feature engineering and extraction transforms textual data into a format that can be used by machine learning models. The SLR reported in the previous chapter identified the features that are prevalent in the literature - a choice which is often dictated by the datasets used, which are generally composed of text-based data. The following subsections describe these different types of feature representations and discuss the characteristics, strengths and limitations of each technique.

4.5.1 Bag of Words

The Bag of Words (BoW) approach is a foundational technique in text classification and natural language processing. This method transforms text into fixed-length vectors based on word frequencies, without considering the order of words or their contextual relationships (Zhang et al., 2010). In practice, BoW represents each document as a vector where each element corresponds to the frequency of a specific word from the entire vocabulary. For example, if the vocabulary consists of the words: ['dog', 'barked', 'at', 'the', 'moon'], a document containing the sentence "The dog barked at the moon" would be represented by a vector reflecting the count of each word's occurrence: [1, 1, 1, 2, 1]. If another document contains the sentence "The dog sat by the fire" the vector representation based on the same vocabulary would be [1, 0, 0, 2, 0], with zeros indicating the absence of the words 'barked', 'at', and 'moon'. This vectorisation approach highlights the presence or absence of words but does not preserve their sequence or relationships.

A significant aspect of BoW is its disregard for word order. It treats the document as a collection of individual words, ignoring their arrangement and any syntactic structure (Qader et al., 2019). Consequently, sentences like “The dog ate my thesis” and “My thesis ate the dog” would yield identical vectors, as BoW only accounts for the frequency of words rather than their specific placement in the sentence. This limitation means that BoW does not capture the nuances of meaning that can arise from different word sequences. BoW also overlooks the contextual meaning of words, as it considers each word in isolation without regard to the surrounding words that might influence its meaning. For example, the word “bank” could refer to a financial institution or the side of a river, but BoW treats these instances as the same word without distinguishing the context. This lack of contextual awareness can reduce the effectiveness of BoW in tasks requiring a deeper understanding of text.

Despite these limitations, BoW remains popular due to its computational efficiency and ease of implementation (Wu et al., 2010). It simplifies text data into numerical vectors that can be easily processed by machine learning algorithms, making it a practical choice for many text classification tasks. In the context of fake news detection, the systematic review conducted in Chapter 3 found that BoW has been utilised across 11 datasets in 34 studies, achieving an average accuracy of 88%. This strong performance highlights BoW’s effectiveness specifically for fake news detection and supports its continued exploration within this thesis. While BoW has limitations, such as its inability to capture nuanced contextual meanings, its practical benefits and high accuracy justify its inclusion. Examining BoW alongside more advanced techniques will provide insights into the generalisability of current approaches to fake news detection and their effectiveness across different contexts.

4.5.2 Term-Frequency Inverse Document Frequency (TF-IDF)

Similar to BoW, TF-IDF transforms words into numerical vectors, but it improves upon BoW by capturing the relative importance of a word within a document in comparison to the entire corpus (Zhao et al., 2018). TF-IDF achieves this by combining two key components: term frequency and inverse document frequency.

Term frequency measures how often a word appears in a particular document, similar to the BoW approach. This aspect reflects the significance of a word within that specific text. Inverse document frequency adjusts the term frequency by accounting for how common or rare the word is across all documents in the corpus. Specifically, it assigns higher weights to words that appear infrequently in the corpus, making them more significant, and lower weights to words that are common across many documents.

This approach helps address a notable limitation of BoW: the lack of differentiation between frequent but less informative words and rare, potentially more important terms (Zhang et al., 2018). For example, common words like “and”, “the” or “of” appear frequently in most documents but do not provide meaningful distinctions between them. TF-IDF reduces the weight of these common words, thereby minimising their impact on the vector representation. Conversely, words that are rare or unique to certain documents receive higher weights, enhancing their ability to differentiate between documents. By adjusting for word frequency in relation to the entire corpus, TF-IDF provides a more refined representation of text than BoW. It emphasises terms that are likely to carry significant meaning and enhances the differentiation between documents.

Although TF-IDF represents an improvement over BoW by addressing the frequency of words and their relative importance, it still retains some limitations shared with BoW, such as disregarding word order and lacking contextual nuance (Zhao et al., 2018). Nonetheless, TF-IDF’s capability to highlight the significance of less common words based on their rarity and frequency makes it a valuable advancement over the BoW approach, making it effective for text classification and information retrieval.

The systematic review in Chapter 3 underscores the relevance of TF-IDF in fake news detection, having been applied in 86 studies with an average accuracy of 84% across 19 diverse datasets. This performance highlights TF-IDF’s effectiveness and justifies its role in investigating the generalisability of current fake news detection methods. Although TF-IDF shares some limitations with BoW, such as disregarding word order and lacking contextual nuance, its enhanced capability to prioritise significant terms makes it a valuable tool for further research, particularly in examining the generalisability of existing approaches in fake news detection.

4.5.3 Static Embeddings

Static embeddings represent words as dense vectors in a continuous vector space, capturing their meanings based on their usage in a large text corpus. Techniques like Word2Vec generate these embeddings by training shallow neural networks to predict a target word from its surrounding context (Mikolov, 2013). This results in a vector space where words with similar meanings, such as “king” and “queen”, are positioned close to each other, reflecting their related meanings (Yilmaz and Toklu, 2020). Compared to traditional methods like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), which focus on word frequencies and term importance, static embeddings offer a more nuanced representation by capturing semantic relationships. For example, “cat” and “kitten” would have similar vectors,

showing their related meanings.

While static embeddings offer a significant improvement over token-occurrence analysis approaches such as BoW and TF-IDF, they also have their own distinct limitations. One major issue is their context-independent nature. As each word is represented by a single, fixed vector that does not change based on how the word is used in different contexts, words that are morphologically the same but semantically different, such as “bank” in the context of a financial institution versus “bank” as the side of a river, receive the same embedding (Huang et al., 2012). This can lead to misunderstandings or inaccuracies when the context changes but the word form remains the same. Additionally, static embeddings also struggle with handling out-of-vocabulary (OOV) words. Since embeddings are created based on a fixed vocabulary learned during training, any word not present in this vocabulary cannot be accurately represented (Kwon et al., 2021). This limitation poses challenges when encountering new or rare words that were not part of the original training data, as these OOV words lack associated vectors and thus cannot be effectively incorporated into the model. Consequently, the inability to dynamically adapt to new words further limits the flexibility and applicability of static embeddings in various text processing tasks.

Despite these limitations, the systematic review in Chapter 3 highlights the efficacy of static embeddings in the context of fake news detection, having been utilised in 86 studies with an average accuracy of 86% across 21 datasets. This strong performance underscores the effectiveness of static embeddings in capturing meaningful patterns and relationships in textual data. The use of static embeddings across this large number of datasets also emphasises their potential in producing generalisable models that perform across various datasets and contexts. This performance justifies their inclusion in assessing the generalisability of current approaches to fake news detection.

4.5.4 Contextual Embeddings

Contextual embeddings address some of the limitations of static embeddings by incorporating the surrounding context of words into their representations. Unlike static embeddings, which assign a single fixed vector to each word regardless of its usage, contextual embeddings adapt based on the context in which a word appears. This means that the representation of a word can change depending on the surrounding words in a sentence. Models such as Bidirectional Encoder Representations from Transformers (BERT) achieve this by training a Transformer on tasks such as Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) (Devlin, 2018). MLM trains the model to predict missing words within a sentence,

which helps the model understand the relationships between words in context, while NSP trains the model to predict whether one sentence logically follows another, enhancing the model’s grasp of sentence structure and relationships. Together, MLM and NSP enable BERT to generate embeddings that reflect both the local context (relationships between nearby words) and the global context (relationships between sentences). This ability to produce embeddings that dynamically adjust based on context significantly improves the model’s performance on various natural language processing tasks. This allows models such as BERT to produce nuanced representations of words (Miaschi and Dell’Orletta, 2020). For example, in contrast to static embeddings, the word “bank” will have different embeddings in the sentences “I went to the bank to withdraw money” and “I sat on the bank of the river”, reflecting its different meanings based on context. This dynamic approach allows contextual embeddings to capture both the semantic and syntactic properties of words more accurately than static embeddings.

While contextual embeddings offer advanced language representations by adapting to the surrounding context of words, they require substantial computational resources for both training and inference due to the complexity of the models. To address this challenge, pre-trained models are typically used. These models are trained on large, diverse datasets and then adapted for specific tasks, allowing users to benefit from rich, context-aware embeddings without needing to invest in the extensive computational resources required for training from scratch. However, while these pre-trained models mitigate the need for substantial computational resources, these models can be influenced by biases in the training data. As such models learn from a large and varied corpora, they may inadvertently perpetuate biases, which can affect the fairness and accuracy of their outputs (Srinivasan et al., 2024). Additionally, although contextual embeddings capture nuanced meanings more effectively than static embeddings, they may struggle with highly specialised or domain-specific language, potentially limiting their effectiveness in certain contexts.

Although contextual embeddings were first introduced for fake news detection in 2020, with 44 studies incorporating them as noted in the systematic review in Chapter 3, they have already demonstrated impressive performance with an average accuracy of 85% across 15 datasets. In contrast, static embeddings have been used since 2016, reflecting a more established history in this area. The strong performance of contextual embeddings underscores their potential for capturing nuanced meanings and relationships within text. This capability addresses some limitations of static embeddings and justifies further exploration into their generalisability in the task of fake news detection.

4.5.5 Stylistic Features

Stylistic features provide an alternative approach to text representation, focusing on the manner in which text is written rather than its specific content. Unlike token-based methods, which aim to capture the meaning of individual words or phrases, stylistic features delve into the structural, lexical, and expressive properties of text (Verma and Srinivasan, 2019). These features encompass various dimensions, including lexical characteristics such as word frequency, sentence length, and vocabulary richness, which provide insights into the text’s complexity and formality. Syntactic features, on the other hand, analyse elements like part-of-speech distribution and sentence structure to identify patterns indicative of different writing styles or genres.

In addition to lexical and syntactic features, some approaches incorporate deeper analyses through semantic and psycholinguistic features. Tools such as Linguistic Inquiry and Word Count (LIWC) are widely employed to capture psychological and social dimensions, analysing aspects like emotional tone, cognitive processes, and interpersonal dynamics reflected in the text (Tausczik and Pennebaker, 2010). These features are particularly valuable in tasks such as sentiment analysis, deception detection, and authorship attribution, where understanding how something is written is as critical as understanding what is written. For instance, sentence complexity and variability in punctuation might serve as indicators of deceptive writing, while patterns of emotional expression can offer clues about the writer’s intent.

While stylistic features do not directly model the specific words or phrases within a text, they allow machine learning models to focus on overarching patterns and nuances that are less tied to the vocabulary of a particular dataset. This makes them particularly robust for applications where generalisation across datasets or domains is required (Holmes et al., 2023). However, as stylistic features focus on the form and style of writing, they may not always capture the deeper semantic relationships or domain-specific knowledge that token-based methods can provide. Nonetheless, their ability to capture subtle cues in writing style enhances the interpretability and effectiveness of machine learning models in tasks where the nuances of text expression play a crucial role.

4.6 Machine Learning Algorithms

The next step in the text classification pipeline is classifier training, where machine learning algorithms are used to identify patterns in the extracted features. The selection of algorithms is guided by the findings of the systematic review, which highlighted methods commonly applied in fake news detection. Each algorithm brings distinct advantages: Naïve Bayes offers simplicity and efficiency, SVM and Logistic

Regression provide robust performance in high-dimensional spaces, and tree-based models like Random Forests and Gradient Boosting handle non-linear relationships effectively. More advanced models, such as Neural Networks and LSTMs, capture complex patterns and sequential dependencies but require greater computational resources. The following subsections outline the strengths and limitations of these algorithms, ensuring their suitability for the study’s focus on accuracy and generalisability.

4.6.1 Naïve Bayes

Naïve Bayes is a probabilistic classifier grounded in Bayes’ Theorem, which provides a mathematical framework for updating the probability of a hypothesis as new evidence is introduced. The “naïve” aspect of this classifier stems from its assumption that all features are conditionally independent of each other given the class label (Webb et al., 2010). This means that, when making predictions, Naïve Bayes treats each feature as if it contributes to the outcome independently of any other feature. While this assumption simplifies the model, it is rarely true in real-world datasets where features often interact or are correlated. Despite this simplifying assumption, which can be a significant limitation in some contexts, Naïve Bayes has consistently demonstrated strong performance in text classification tasks (Jiang et al., 2011). The systematic review in Chapter 3 offers evidence of this strong performance, with Naïve Bayes achieving a mean accuracy of 81% across 20 datasets in the fake news detection task.

One of the key strengths of Naïve Bayes is its simplicity. The model is easy to understand and interpret, as the probability estimates generated by the model provide clear insights into the decision-making process, allowing users to understand why certain predictions are made (Schneider, 2005). This is particularly advantageous in the context of fake news detection, where the ability to understand what features are most relevant in classifying a document as ‘real’ or ‘fake’ can inform future work, as well as inform understanding of what terms are mostly associated with real and fake news. Furthermore, Naïve Bayes’ is robust to irrelevant features, which is particularly advantageous in fake news detection where text can include residual noise from web-scraping (Witten et al., 2005). This robustness comes from the model’s ability to naturally down-weight the influence of less informative features, allowing it to focus on those that are more significant for the classification task.

However, it’s important to recognise the limitations of Naïve Bayes. The assumption of conditional independence, which simplifies the model, can lead to inaccuracies when features are actually interdependent, as is often the case in complex text datasets (Ting and Zheng, 2003). For instance, in fake news detection, the rela-

tionship between certain words or phrases may play a critical role in determining the credibility of a news article. If these interactions are overlooked, Naïve Bayes may underperform in scenarios where understanding the nuanced relationships between features is crucial for accurate predictions. Despite these drawbacks, Naïve Bayes remains a valuable tool in fake news detection, especially when its assumptions align reasonably well with the dataset characteristics. Its computational efficiency makes it a popular baseline algorithm for comparison, and this likely contributes to its popularity in the literature where it features in 83 studies.

4.6.2 Logistic Regression

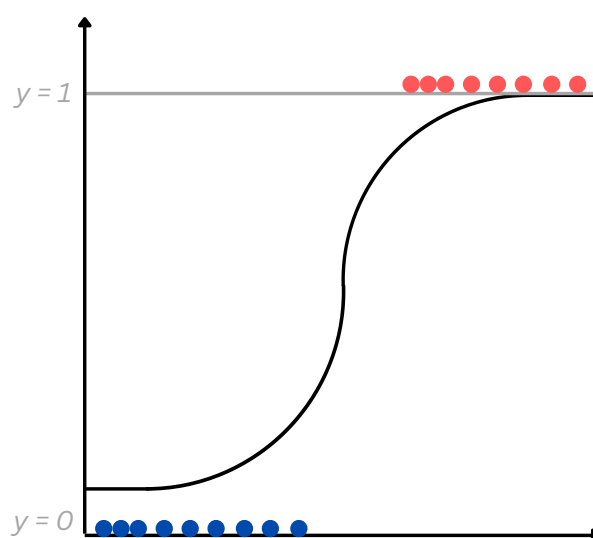


Figure 4.3: Logistic Regression

Logistic Regression is a widely used statistical model for binary classification tasks, grounded in the principles of logistic function. Unlike linear regression, which predicts continuous outcomes, Logistic Regression predicts probabilities for discrete classes by applying the logistic function to a linear combination of the input features. This function maps predicted values to a probability range between 0 and 1, making it particularly suitable for tasks like fake news detection where the goal is to classify text into ‘real’ or ‘fake’ categories. The systematic review in Chapter 3 demonstrates the efficacy of Logistic Regression in this context, showing that it achieves competitive performance in the fake news detection task with a mean accuracy of 83% across 16 datasets.

Similar to Naïve Bayes, one of the key strengths of Logistic Regression is its interpretability (Slack et al., 2019). The model coefficients reveal the strength and direction of the relationship between each feature and the probability of the outcome, allowing users to understand how different features influence the classification

decision. Like Naïve Bayes, Logistic Regression is also computationally efficient and does not require extensive parameter tuning, making it a popular baseline model. This is reflected in its presence in 86 studies within the literature. Additionally, Logistic Regression performs well with smaller datasets and is less prone to overfitting compared to more complex models. As highlighted in the systematic review, this feature is particularly advantageous for current fake news detection datasets, which typically average around 10,500 articles per class.

However, Logistic Regression, while widely used and effective in many contexts, has its own set of limitations (Nick and Campbell, 2007). It assumes a linear relationship between the features and the log-odds of the outcome, which may not capture complex, non-linear relationships in the data. In the context of fake news detection, where interactions between features can be intricate and multi-faceted, this assumption may restrict the model's performance. Moreover, Logistic Regression can struggle with high-dimensional data if not regularised appropriately, potentially leading to issues with overfitting. Despite these limitations, Logistic Regression remains a robust tool in fake news detection, valued for its simplicity, efficiency, and ability to provide meaningful insights into feature importance. Its presence in numerous studies highlights its ongoing relevance and utility in the field.

4.6.3 Support Vector Machines (SVMs)

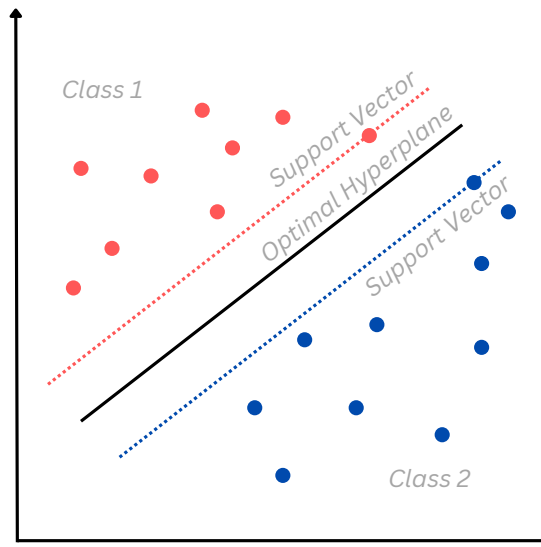


Figure 4.4: SVM

Support Vector Machines (SVMs) are advanced classifiers designed to find the optimal hyperplane that maximizes the margin between different classes in a high-dimensional space. Unlike Logistic Regression, which predicts probabilities based on a linear combination of features, SVMs focus on identifying the boundary that best

separates classes by considering the most critical data points—the support vectors (Bhavsar and Panchal, 2012). This approach allows SVMs to handle sparse and high-dimensional data effectively, such as text data used in fake news detection.

Unlike Naïve Bayes, which operates under the assumption of conditional independence among features, SVMs are adept at capturing complex, non-linear relationships through the use of the kernel trick (Schölkopf, 2000). This technique enables SVMs to function in high-dimensional or even infinite-dimensional spaces, allowing them to model intricate class boundaries. Such a capability is particularly valuable for tasks like fake news detection, where text features often interact in subtle and non-linear ways. The systematic review in Chapter 3 highlights this advantage, demonstrating that SVMs achieve a mean accuracy of 85% across 20 datasets. This performance slightly surpasses that of Naïve Bayes and Logistic Regression, emphasising the effectiveness of SVMs in managing the complexities of text data compared to the more linear assumptions of Naïve Bayes.

However, unlike Naïve Bayes and Logistic Regression, SVMs tend to be less interpretable. While Naïve Bayes provides insights based on feature independence and Logistic Regression offers clear understanding through model coefficients, SVMs prioritise finding the optimal boundary between classes using support vectors and the kernel trick. This focus on complex, high-dimensional spaces can obscure the direct relationship between features and classification outcomes, making it more challenging to interpret how specific features influence the final decision (Siddique et al., 2024). Furthermore, SVMs can be computationally intensive, especially with large datasets or complex kernels, which may require significant processing power and memory (Nandan et al., 2014). Selecting the appropriate kernel and tuning hyperparameters can also be complex and time-consuming. Despite these challenges, SVMs remain a popular tool in fake news detection, as evidenced by their presence in 86 studies within the SLR. Their slightly superior performance, through managing high-dimensional data and modelling non-linear relationships, underscores their value in the field. This effectiveness and ongoing relevance are reflected in their extensive usage across numerous studies, highlighting their robust performance in detecting fake news.

4.6.4 Decision Trees

Decision Trees are a powerful and versatile classification tool that operates by recursively splitting the feature space into distinct regions based on feature values. This process generates a tree-like model of decisions, where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label (De Ville, 2013). This method allows Decision

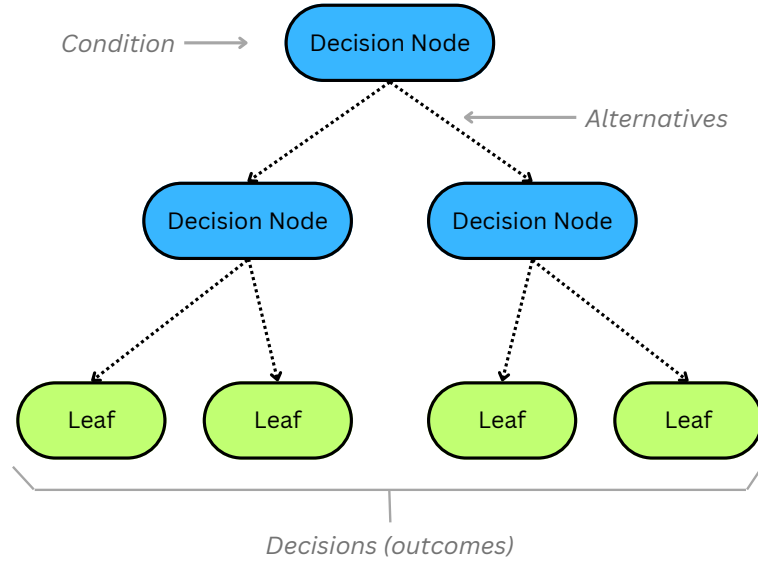


Figure 4.5: Decision Tree

Trees to handle a mix of numerical and categorical features effectively, making them particularly suited for complex tasks like fake news detection.

A key advantage of Decision Trees is their ability to model complex feature interactions without requiring any feature transformations or making assumptions about the relationships between variables. This flexibility allows them to capture intricate patterns within the data, which is particularly useful when working with the nuanced and varied text data involved in fake news detection. Like Naïve Bayes and Logistic Regression, Decision Trees offer interpretability by providing a clear, visual representation of the decision-making process (Slack et al., 2019). Coupled with their computational efficiency, this makes them a popular choice for fake news detection, as demonstrated by their presence in 58 studies from the SLR. Their average accuracy of 85% across 17 datasets in this review further supports their standing as an effective algorithm in the field.

While Decision Trees provide several advantages in the context of fake news detection, they are prone to overfitting, especially when the tree becomes very deep and complex (Bramer, 2007). This overfitting occurs because the model may capture noise and outliers in the training data rather than the underlying patterns, leading to reduced generalisation to new data. Additionally, as the tree grows, it can become less interpretable, with the decision-making process becoming more opaque. While they provide clear decision rules, the complexity of deeper trees can make it difficult to trace how specific features influence the classification. Despite these challenges, Decision Trees continue to be a valuable tool in fake news detection. Their ability to handle diverse feature types and interactions, combined with their robust performance across numerous studies, underscores their practical utility and

ongoing relevance in the field.

4.6.5 Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds a strong classifier by combining the outputs of several weaker classifiers, typically decision trees. Unlike traditional decision trees, which operate independently, Gradient Boosting works iteratively, where each new tree corrects the errors of the previous ones (Natekin and Knoll, 2013). This approach allows Gradient Boosting to model complex patterns in the data by sequentially reducing the residual errors, making it highly effective for tasks such as fake news detection, where subtle patterns in text need to be captured.

One of the key strengths of Gradient Boosting is its ability to handle non-linear relationships and interactions between features (Kalusivalingam et al., 2022). By focusing on the hardest-to-predict examples in each iteration, Gradient Boosting can capture complex dependencies within the data. This is especially valuable in fake news detection, where textual features may interact in intricate ways. Moreover, the flexibility of Gradient Boosting allows it to adapt well to various types of datasets, achieving high accuracy even in challenging classification tasks.

While Gradient Boosting is highly effective, achieving 86% accuracy on average in fake news detection, it comes with some trade-offs, particularly in terms of interpretability and computational efficiency. Unlike simpler models such as Decision Trees or Logistic Regression, Gradient Boosting models are harder to interpret, as they involve many sequential decision trees working together (Welchowski et al., 2022). This can make it difficult to understand how individual features contribute to the final prediction, which is a consideration when transparency is important in applications like fake news detection. Additionally, Gradient Boosting can be computationally intensive, especially as the number of iterations or trees increases, leading to longer training times and higher resource consumption. Despite these challenges, Gradient Boosting remains a highly popular algorithm due to its strong predictive performance. The systematic literature review (SLR) highlights its widespread use in fake news detection, featuring in 50 studies and consistently achieving high accuracy rates across different datasets.

4.6.6 Random Forest

Similar to Gradient Boosting, Random Forest is an ensemble learning method that leverages multiple decision trees to improve predictive performance. However, unlike Gradient Boosting, which builds trees sequentially to correct the errors of previous ones, Random Forest constructs trees independently in parallel. By averaging or taking the majority vote of the trees' predictions, Random Forest reduces overfitting

and increases robustness, making it a strong candidate for tasks like fake news detection (Biau and Scornet, 2016).

One of the major advantages of Random Forest is its ability to handle large amounts of data with high-dimensional feature spaces, such as text data in fake news detection. By averaging the predictions from multiple trees, Random Forest is capable of capturing complex patterns while maintaining stability in its predictions. This makes it particularly useful when dealing with noisy or unbalanced datasets, where individual models might struggle to perform consistently. Additionally, Random Forest is less sensitive to outliers and irrelevant features, as the random sampling of features reduces the likelihood of these elements having an undue influence on the final model.

Furthermore, Random Forest is relatively interpretable compared to more complex models like Gradient Boosting or Support Vector Machines. While not as transparent as a single decision tree, Random Forest allows for some insights into feature importance by evaluating the contribution of each feature across the entire ensemble (Haddouchi and Berrado, 2019). This interpretability, combined with its flexibility, makes Random Forest a valuable tool in fake news detection, where understanding the key indicators of disinformation is crucial. Furthermore, Random Forest is computationally efficient in comparison to more resource-intensive algorithms, as each tree can be built and evaluated independently, allowing for parallel processing.

Despite its advantages, Random Forest has limitations, particularly in terms of computational efficiency. Training hundreds or thousands of trees can be resource-intensive, both in terms of memory and processing power, especially with very large datasets (Biau and Scornet, 2016). Additionally, while Random Forest offers improved accuracy over single decision trees, it may not perform as well as gradient boosting when subtle relationships between features need to be captured. Nonetheless, its strong performance in the field is highlighted by the fact that Random Forest was featured in 86 studies in the SLR, achieving an average accuracy of 84% across 16 datasets, which demonstrates its robustness and continued relevance in fake news detection tasks.

4.6.7 Feed-Forward Neural Networks (FFNNs)

Feed-forward neural networks (FFNNs) are a class of artificial neural networks where information flows in one direction—from input nodes, through hidden layers, to output nodes—without any feedback loops. Unlike traditional algorithms like decision trees or Naïve Bayes, FFNNs are highly flexible and can model complex, non-linear relationships in data, making them well-suited for tasks like fake news detection. By

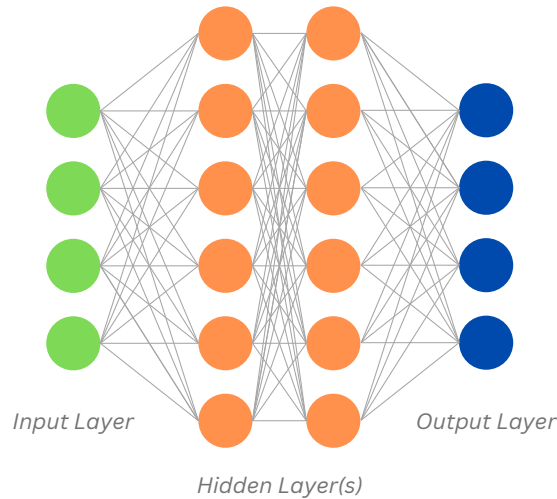


Figure 4.6: Neural Network

learning hierarchical feature representations, FFNNs can capture intricate patterns in textual data that might be missed by simpler models (Surdeanu and Valenzuela-Escárcega, 2024).

One of the primary strengths of FFNNs is their ability to handle large, high-dimensional datasets, particularly when text is represented using advanced techniques like word embeddings or TF-IDF. This flexibility enables FFNNs to excel in fake news detection, where subtle semantic nuances between real and fake news articles are crucial. FFNNs also offer a high degree of customisability, as their architecture (e.g., number of layers, neurons per layer) can be tailored to the specific characteristics of the dataset.

However, unlike Decision Trees or Logistic Regression, which provide clear insights into feature importance, FFNNs function more like a “black box” (Benítez et al., 1997). It is difficult to trace how individual input features influence the final classification decision, which can be a drawback when transparency is required. Moreover, training neural networks can be computationally expensive, especially when dealing with large datasets or deep architectures. FFNNs require careful tuning of hyperparameters such as learning rate, activation functions, and the number of hidden layers, and they are prone to overfitting if regularisation techniques like dropout are not applied.

Despite these challenges, FFNNs remain a powerful tool in fake news detection due to their ability to capture complex patterns in text data. The systematic review highlights that FFNNs were featured in 42 studies, achieving a competitive mean accuracy of 85% across 15 datasets. This performance reflects their ability to generalise well across diverse datasets, particularly when other models might struggle with the non-linear intricacies of text. Although they are less popular compared to

SVMs, decision trees, and Logistic Regression, their strong performance and ability to capture complex patterns underscore their value in the field. Their continued relevance is reflected in their robust performance, which highlights their potential as a powerful tool for detecting fake news.

4.6.8 Long-Term Short-Term Memory Networks (LSTMs)

Long Short-Term Memory networks (LSTMs) are a specialised type of recurrent neural network (RNN) designed to handle sequential data and capture long-term dependencies within it (Sherstinsky, 2020). Unlike feed-forward neural networks (FFNNs), which process data in a fixed direction, LSTMs are adept at learning from sequences of text, making them particularly well-suited for tasks like fake news detection where context and word order are crucial.

One of the primary strengths of LSTMs is their ability to remember and utilise information over long sequences due to their unique architecture, which includes gates that regulate the flow of information. This allows LSTMs to capture intricate patterns and dependencies in text data, such as the context of a news article or the sequence of words that may signal whether content is real or fake. This temporal aspect is crucial in fake news detection, where understanding the context and progression of information can significantly impact classification accuracy.

However, LSTMs face certain challenges compared to more traditional models like SVMs, decision trees, and Logistic Regression. Training LSTMs requires significant processing power and memory, especially with large datasets or long sequences, which can make them less accessible for some applications (Sen and Raghunathan, 2018). Additionally, similar to SVMs and FFNNs, LSTMs can be difficult to interpret. Unlike Decision Trees, which provide a clear visual representation of decision-making, or Logistic Regression, which offers insights through model coefficients, LSTMs operate as a complex “black box,” making it challenging to understand how they arrive at specific predictions.

Despite these limitations, LSTMs have proven their effectiveness in fake news detection. The SLR indicated that LSTMs were featured in 80 studies, achieving an average accuracy of 91% across 19 datasets. Their ability to handle sequential dependencies and capture nuanced patterns in text data underscores their value in the field. The strong performance and continued relevance of LSTMs highlight their potential as a powerful tool for detecting fake news, particularly in applications where understanding context and sequence is crucial.

The diverse array of algorithms explored in this section—Naïve Bayes, Logistic Regression, Support Vector Machines, Decision Trees, Gradient Boosting, Ran-

dom Forests, Feed-Forward Neural Networks, and Long Short-Term Memory networks—each brings distinct strengths and weaknesses to the task of fake news detection. Traditional models like Naïve Bayes and Logistic Regression offer simplicity and interpretability, while advanced techniques such as SVMs and ensemble methods like Random Forests and Gradient Boosting excel in handling complex interactions and high-dimensional data. Neural network approaches, including Feed-Forward and Long Short-Term Memory networks, are adept at capturing intricate patterns and sequential dependencies in text. The prevalence of these algorithms and their demonstrated effectiveness, as evidenced by the systematic literature review, justifies their inclusion and analysis in this thesis. Their varied computational demands and interpretability challenges underscore their collective importance in enhancing the accuracy and robustness of fake news detection systems. The ongoing evolution and application of these methods reflect their critical role in advancing the field.

While this section has offered an overview of these algorithms and their rationale for inclusion in this thesis, the specific implementations will be detailed in the empirical chapters, Chapters 5 and 6. These chapters will provide a comprehensive explanation of how each algorithm was applied to the datasets, along with the associated results and analysis.

4.7 Evaluation Methods

The final step of the text-classification process involves evaluating the machine learning models that have been trained. This evaluation is critical for assessing the models' performance and ensuring their effectiveness on new data. The remaining sections of this chapter will detail the methods used for model evaluation, including how models are tested and validated, the metrics used to measure their performance, and the techniques employed to interpret and understand their predictions

4.7.1 Holdout Testing

Holdout testing is a basic yet essential evaluation technique used to assess the performance of machine learning models. In this method, the dataset is randomly divided into two separate subsets: a training set and a test set. The training set is used to build and tune the model, while the test set, which remains unseen during the training process, is used to evaluate the model's performance. The primary advantage of holdout testing is its simplicity and ease of implementation. It allows for a quick assessment of how well the model generalises to new, unseen data. However, the results can be sensitive to the specific partitioning of the data. If the split is not representative of the overall dataset, the performance metrics obtained might

not accurately reflect the model’s ability to handle real-world data. Additionally, the performance estimate can vary depending on the random seed used for the data split.

4.7.2 K-Fold Cross-Validation

K-Fold Cross-Validation attempts to address the limitations of holdout testing by using multiple splits of the data. In this approach, the dataset is divided into K equally sized folds or subsets. The model undergoes training and evaluation K times, each time with a different fold reserved as the test set while the remaining $K-1$ folds are used for training. This means that each data point is used for both training and testing, providing a comprehensive view of model performance. The results from each of the K folds are aggregated, typically by averaging, to produce a final performance metric. This method reduces the variance associated with a single train-test split and offers a more robust estimate of model performance. K-Fold Cross-Validation is particularly useful in situations where the dataset is limited, as it maximises the use of available data. However, it can be computationally expensive, especially with large datasets or complex models, due to the repeated training and evaluation processes.

4.7.3 External Validation

External validation refers to the process of testing a model on an independent dataset that was not involved in the training or validation stages. This dataset is typically sourced from different domains or distributions than the original data used in training, as observed in Section 3.5.4 of Chapter 3, providing a rigorous test of the model’s ability to generalise. This helps confirm that the model’s performance is not constrained to the dataset in which it was trained on, which may not necessarily represent the broader range of scenarios the model could encounter in real-world conditions.

While this approach to testing can offer a unique perspective on model performance, one limitation is that this method of testing can sometimes lead to an overestimation or underestimation of a model’s performance, depending on how closely the external dataset aligns with the model’s original training data. If the external dataset is too similar, it may not provide a meaningful test of generalisability. Conversely, if the external dataset is too different, the model’s performance may drop significantly, which could be more a reflection of domain shift rather than the model’s general competence. Moreover, external validation is typically performed only once, meaning the evaluation is based on a single snapshot of performance. This can be problematic if the external dataset itself has biases or does not capture

the full range of variability present in the target domain, leading to conclusions that might not hold across other unseen data. Finally, external validation does not always provide insights into why a model might perform poorly on new data. While it can indicate issues like overfitting, it does not inherently offer a mechanism for diagnosing or addressing such issues. Therefore, other techniques must be used in conjunction with external validation to provide insight into potential reasons why a model may generalise poorly.

Given the thesis’s focus on producing robust results, K-fold cross-validation and external validation will serve as the primary methods for evaluating model performance. Holdout testing will be employed during the development phase to ensure that models are functioning correctly.

4.8 Evaluation Metrics

This section outlines the metrics used in evaluating text classification models for the fake news detection task. Typically, studies in the literature rely primarily on four metrics: accuracy, precision, recall and F1-score. Each of these metrics provides different insights into the performance of a model, helping to assess its effectiveness in distinguishing between genuine and fake news.

4.8.1 Accuracy

Accuracy is one of the most straightforward metrics and is calculated as the ratio of correctly predicted instances to the total number of instances in the dataset. This provides a general measure of how well the model performs across all classes, the formula for accuracy is as follows:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

While accuracy is useful for an overall assessment, it can be misleading in cases of class imbalance, where one class is significantly more prevalent than the other. For example, if a dataset has 90% true news and only 10% fake news, a model that always predicts “true” would still achieve a high accuracy of 90%. This could mask poor performance in detecting the minority class (fake news), which is often of greater interest in such tasks. In practice, the vast majority of studies in the literature use balanced datasets for training and testing models, negating this particular issue.

However, accuracy has other limitations. It does not capture the performance of each class individually or provide insight into the types of errors a model makes. For instance, accuracy does not differentiate between false positives and false negatives, nor does it reflect the trade-offs between different types of errors. As a result, relying solely on accuracy can provide an incomplete picture of a model's effectiveness, particularly in scenarios where distinguishing between classes is critical.

4.8.2 Precision

Precision, also known as positive predictive value, measures the proportion of true positive predictions among all positive predictions made by the model. This provides a measure of how accurate the model is when it identifies a case as positive, the formula for precision is as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

In the context of fake news detection, precision provides a measure of the reliability of a model's predictions of the 'true' class, as fake news detection literature typically designate the positive class as 'true news' and the negative class as 'fake news.' High precision indicates that when the model predicts news as 'true', it is likely to be correct, minimising the risk of falsely labelling fake news as true. This is particularly important in applications where the consequences of misclassifying false information as true could be severe, such as on social media platforms or in news outlets, where maintaining credibility is essential. However, focusing solely on precision without considering other metrics, like recall, could result in a model that identifies only a few true news articles, missing many others. This trade-off between precision and recall is why it's essential to consider multiple metrics when evaluating the effectiveness of a fake news detection model.

4.8.3 Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive cases that the model successfully identifies out of all actual positive cases. The formula for recall is as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is a key metric for assessing a model’s ability to identify all actual instances of the positive class (in this context, typically ‘true news’) in fake news detection. A high recall score indicates that the model is proficient at capturing most true news articles, thereby reducing the risk of failing to identify genuine news. This capability is crucial in scenarios where prioritising the detection of all instances of true news is more important than minimizing false positives. However, a focus on high recall can sometimes lead to lower precision, meaning that more fake news articles might be misclassified as true news. Therefore, balancing recall with precision is essential to ensure the model is both thorough in identifying true news and accurate in distinguishing it from fake news.

4.8.4 Specificity

Specificity, also referred to as the true negative rate, measures the proportion of true negative cases that the model correctly identifies out of all actual negative cases. The formula for specificity is as follows:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Specificity is a critical metric for evaluating a model’s ability to correctly identify instances of fake news in fake news detection. A high specificity score indicates that the model is effective at minimising the number of false positives (i.e., incorrectly classifying fake news as true). This capability is particularly important in scenarios where ensuring that fake news is not falsely accepted as true is critical to mitigating the spread of disinformation and its associated consequences. However, focusing solely on specificity may come at the cost of lower recall for true news, where some genuine news articles might be misclassified as fake. Balancing specificity with recall ensures that the model is both precise in detecting fake news and comprehensive in identifying true news.

4.8.5 F-1 Score

The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful when dealing with imbalanced datasets, where focusing on just one metric (precision or recall) could be misleading. The formula for F1-Score is:

$$\text{F-1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F-1 Score provides a holistic measure of a model's performance by integrating both precision and recall, reflecting its ability to accurately identify true news and capture all relevant true news articles. A high F1-Score signifies a balanced performance, where the model effectively manages both precision and recall, which is crucial when the impact of missing true news or misclassifying fake news is equally critical. Despite its usefulness, the F1-Score combines precision and recall into a single metric, which can mask specific weaknesses in either area. Therefore, it is often analysed alongside these individual metrics to offer a more comprehensive evaluation of the model's effectiveness.

4.8.6 Other Metrics

While accuracy, precision, recall, and F1-score are commonly utilised in evaluating model performance, other metrics that are sometimes reported in the literature include AUC-ROC and confusion matrices. AUC-ROC evaluates the model's ability to distinguish between classes by plotting the true positive rate against the false positive rate across different thresholds. A higher AUC-ROC value indicates better performance in differentiating between fake and genuine news, regardless of the classification threshold. Confusion matrices offer a detailed breakdown of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives. This provides a more granular view of the model's performance, helping to identify specific areas where it might be over or under-predicting certain classes.

Despite the value of these additional metrics, they are less frequently reported, potentially due to practical constraints such as the limited space available in conference papers which made up the majority of papers accepted in the systematic review. This often leads to a focus on the more commonly used metrics like accuracy, precision, recall, and F1-score.

Given that it is typical to define 'true news' as the positive class and 'fake news' as the negative class, other metrics such as specificity may also be beneficial. Specificity measures the model's ability to correctly identify negative cases, helping to ensure that fake news is accurately detected without falsely labelling true news. Incorporating a broader set of metrics, including specificity and others like AUC-ROC, could provide a more comprehensive evaluation of model performance, particularly in the complex task of fake news detection.

4.9 Model Interpretability Techniques

As mentioned in Section 4.7.3, while external validation is useful in assessing a model's ability to generalise to data outside the dataset in which it was trained, it does not provide insights into how the model arrives at its predictions. Therefore, to fully understand and trust a model's decisions, interpretability techniques should be employed alongside external validation. These techniques offer a window into the inner workings of machine learning models, making it easier to understand and explain their predictions.

4.9.1 Local Interpretable Model-Agnostic Explanations (LIME)

Local Interpretable Model-Agnostic Explanations (LIME) is a technique designed to explain individual predictions made by a machine learning model. LIME works by generating a series of perturbed versions of the input data and observing how the model's predictions change. It then fits a simpler, interpretable model—often a linear model—that approximates the complex model's behaviour in the vicinity of the specific prediction being explained. This localised approach allows users to understand why the model made a particular prediction, even if the overall model is a black box. LIME's model-agnostic nature means it can be applied to any type of machine learning model, making it a versatile tool for interpreting predictions in various applications, including in the field of text classification (Biecek and Burzykowski, 2021).

In the context of a binary classification problem such as fake news detection, LIME returns an array of tuples containing a word and a number indicating whether the word had an impact in the model classifying one way or another. Within the scope of this thesis, words that carry a negative score mean they contributed to a 'fake' classification. Words that carry a positive score mean they contributed to a 'real' classification. This allows frequency distributions of these words to be analysed to identify which words or features are most influential in the model's

decision-making process for classifying news as either fake or real. By examining these frequency distributions, researchers can gain insights into the characteristics and patterns that the model relies on to make its predictions.

4.9.2 Permutation Feature Importance (PFI)

Permutation Feature Importance (PFI) is another key interpretability technique that helps in understanding which features are most influential in a model's decision-making process. PFI works by randomly shuffling the values of a feature in the dataset and measuring the impact on the model's performance. If the model's performance significantly drops when a feature is shuffled, it indicates that the feature is important for making accurate predictions (François et al., 2006). PFI is valuable for identifying the factors that contribute to the model's decisions, particularly in complex tasks like fake news detection. However, it can be computationally intensive, as it requires multiple evaluations of the model. Like LIME, PFI is model-agnostic, allowing it to be used with a wide range of machine learning models.

Using interpretability techniques like LIME and PFI in conjunction with external validation provides a comprehensive approach to model evaluation. This combination not only ensures that the model generalises well to new data but also that its predictions can be understood and trusted, which is crucial in applications where the stakes are high, such as in the fight against disinformation.

4.10 Chapter Summary

This chapter has provided an overview of the methodology and techniques used to address these research questions. This overview began by outlining the experimental research method and how this is typically applied in the context of text classification tasks. It outlined the advantages and disadvantages of this approach, concluding that its systematic and robust approach would be crucial in comprehensively addressing the aforementioned research questions.

Following this, each step of the text classification process was described in the context of the fake news detection task, outlining the specific techniques chosen to address the research questions and their justifications. This began by outlining the data collection process and the typical techniques used to collect data for the fake news detection task. Owing to this thesis focussing on current approaches to fake news detection, it was determined that already established datasets would be the primary data used throughout.

The chapter then moved on to pre-processing, emphasising the importance of cleaning and transforming raw data to improve model performance. Various tech-

niques, such as tokenisation, stop-word removal, and lemmatisation, were discussed as key steps in preparing the data for effective analysis. The specific pre-processing steps taken will be outlined in the relevant studies. Following this, the chapter explored feature extraction, focusing on how the relevant characteristics of the text are identified and transformed into features that can be used by machine learning algorithms, using techniques such as Bag-of-words, TF-IDF and embeddings or calculating stylistic features.

The chapter then provided an overview of the machine learning algorithms applied to the fake news detection task. A range of algorithms, including traditional models such as Logistic Regression, Decision Trees, and Support Vector Machines, as well as more advanced models like Neural Networks and LSTMs, were discussed. Similar to features, the choice of algorithms was justified based on their popularity identified in the systematic review in Chapter 3. The strengths and weaknesses of each algorithm were also examined, with particular emphasis on their suitability for the specific characteristics of the task.

The chapter then outlined the evaluation methods used to assess model performance, including holdout testing, K-fold cross-validation, and external validation. These methods were chosen to ensure that the models were rigorously tested for their ability to generalise to new, unseen data. Metrics such as accuracy, precision, recall, and F1-score were discussed as the primary tools for measuring the effectiveness of the models. The importance of considering multiple metrics was emphasised to provide a more comprehensive assessment of model performance, especially in tasks like fake news detection, where class imbalance and other challenges can skew results. Finally, the chapter concluded with a discussion of model interpretability techniques, such as Local Interpretable Model-Agnostic Explanations (LIME) and Permutation Feature Importance (PFI). These techniques were highlighted as essential for understanding the decision-making processes of the models, ensuring that their predictions could be trusted and effectively applied in real-world scenarios.

Chapter 5

Study 1: Intra-Domain Generalisability

5.1 Introduction

The systematic literature review reported in Chapter 3 identified limitations associated with the evaluation of fake news detection models. It found that most studies rely on holdout testing or K-fold cross-validation, which may not fully capture the generalisability of models to unseen data. While this review primarily highlighted issues related to cross-domain generalisability—where models often struggle to adapt across distinct and unrelated news topics—it also pointed to the largely unexplored area, and more fundamental issue, of intra-domain generalisability. This form of generalisability refers to a model’s ability to perform effectively within a consistent domain or topic, such as political news, where even subtle shifts in language, tone, or subject matter can influence performance. Focusing on intra-domain generalisability addresses the question of how well models can adapt within the same category of content, rather than across entirely different domains.

To investigate intra-domain generalisability, this chapter adopts a structured experimental approach, following the methodology outlined in Chapter 4. By examining a specific news domain, such as political news, the study evaluates whether a model trained on one subset of data within that domain can maintain its effectiveness when applied to other related subsets. To enhance the robustness of this evaluation, the chapter utilises external validation, testing models on datasets other than those used in training to simulate real-world conditions. This rigorous approach allows for a more comprehensive assessment of a model’s intra-domain adaptability, as it measures performance not only on training data but also on unseen datasets within the same topic.

The chapter is organised as follows: it begins with presenting the motivation be-

hind the study (Section 5.2). Next, the research questions (Section 5.3) are outlined, setting the study’s focus on model performance and adaptability within the same domain. The methods section (Section 5.4) describes the experimental design, detailing essential processes such as text classification, data preprocessing, feature selection, and the algorithms used, building on the methodologies discussed in Chapter 4. Section 5.5 then presents a comprehensive analysis of model performance across different datasets, using various performance metrics to assess generalisability. Finally, the chapter concludes with a discussion (Section 5.6) that synthesises the findings, identifying trends, limitations, and implications for the broader research questions.

5.2 Motivation

The motivation of this study stems from questions that emerged through the systematic literature review in Chapter 3 around the generalisability of fake news detection models. Real-world applications demand models that can adapt not only to new domains but also to diverse content within a single domain. The inability to generalise within a consistent domain, such as political news, highlights fundamental weaknesses in current models and calls into question their reliability when applied across broader, more varied contexts.

By focusing on the more foundational question of intra-domain generalisability, this study aims to clarify the underlying reasons for poor model adaptability. Investigating whether models can perform consistently within a single topic area allows for a more controlled analysis of the factors that contribute to generalisability. This approach provides a basis for identifying specific limitations within the data, feature representations, or algorithms that may inhibit robust performance. A clearer understanding of these factors in the context of intra-domain generalisability may also inform strategies to enhance models, providing a stepping stone for improved cross-domain generalisability.

This study therefore positions intra-domain generalisability as a critical first step in assessing and improving fake news detection models. By rigorously testing models within a single domain, this research aims to uncover insights into model behaviour and performance that are crucial for advancing fake news detection capabilities in real-world applications, where models must be adaptable, reliable, and resilient to varying content.

5.3 Research Questions Addressed

This section outlines the thesis research questions addressed in this study, focusing particularly on RQ2 and RQ3. RQ2 evaluates the effectiveness of existing methods

for detecting fake news, while RQ3 assesses the generalisability of these methods across different datasets. These questions, introduced in Section 1.4, frame the study’s focus on enhancing the robustness and adaptability of fake news detection models.

Table 5.1: Study 1 - Thesis Research Questions Addressed

RQ	Description
RQ1	What are the current methods to detect fake news?
RQ2	How effective are current methods to detect fake news?
RQ3	To what extent do existing fake news detection methods generalise across datasets?
RQ4	What current features contribute to more generalisable models in the context of fake news detection?
RQ5	How can novel features that extend beyond the text—such as social dissemination behaviours and economic incentives—enhance the generalisability of fake news detection models?

While the systematic review in Chapter 3 partially addressed RQ2 by summarising current methods and their effectiveness, this study further investigates these questions through empirical analysis, aiming to identify strategies for improving model adaptability across datasets.

5.4 Methods

This section outlines the methodological approach used to evaluate intra-domain generalisability in fake news detection models. Building upon the framework introduced in the previous chapter, the section begins by describing the datasets selected for analysis, followed by the preprocessing steps and feature extraction techniques applied to transform the data for model training. It then provides an overview of the algorithms employed in the study, each chosen based on their relevance and demonstrated effectiveness in fake news detection.

The evaluation is structured around three key experiments. The first experiment performs stratified cross-validation (SCV), providing a baseline for comparison across different algorithms and feature sets by ensuring balanced class distributions within each fold. The second experiment focuses on external validation, where models trained on one dataset are tested on entirely new datasets to assess their intra-domain generalisability. This step is crucial for determining how well the models transfer their decision boundaries to similar but unseen data. The third experiment

employs LIME (Local Interpretable Model-Agnostic Explanations) to interpret the model predictions, specifically aiming to identify the most influential keywords that determine whether an article is classified as real or fake. This final experiment adds a layer of interpretability, offering insights into the decision-making processes of the models.

5.4.1 Datasets

As this study focuses on intra-domain generalisability, the choice of datasets is critical to ensure that the models are trained and evaluated on data that represents news within a specific domain. Additionally, as fake news has been defined in several ways and can take different forms—ranging from entirely fabricated stories to misleading headlines or partially true articles—it is important that all the datasets used in this study contain the same type of fake news. This ensures consistency in model training and evaluation, and it prevents variability that could arise from differing definitions or categories of disinformation.

Section 3.5.3 of the systematic review in Chapter 3 identified a number of popular datasets used throughout the literature. From this list, it was determined that a number of datasets focused specifically on fake news surrounding the 2016 U.S. Presidential Election. As such, the most popular of these datasets were selected to explore intra-domain generalisability of fake news detection models. In this section, we will provide an in-depth overview of the datasets chosen for this study, discussing their key characteristics, size, and the types of disinformation they contain. To ensure that all the datasets largely focus on the same domain, a word-frequency analysis was conducted. This analysis helped verify that the most commonly occurring words (other than stop words) and themes across the datasets align with the political and election-related context of the 2016 U.S. Presidential Election. By examining the frequency of specific terms and topics, we confirm that the datasets represent a cohesive and consistent domain, further strengthening the study’s focus on intra-domain generalisability.

ISOT Dataset

The ISOT dataset, developed by the Information Security and Object Technology (ISOT) research lab at the University of Victoria, is a widely used dataset for fake news detection, featuring in 40 studies collected by the systematic review in Section 3.5.3. This dataset is highly relevant to the present study as it includes a collection of news articles from the 2016 U.S. Presidential Election, aligning well with the focus on intra-domain generalisability within this specific political context.

The ISOT dataset consists of two main categories: real news and fake news.

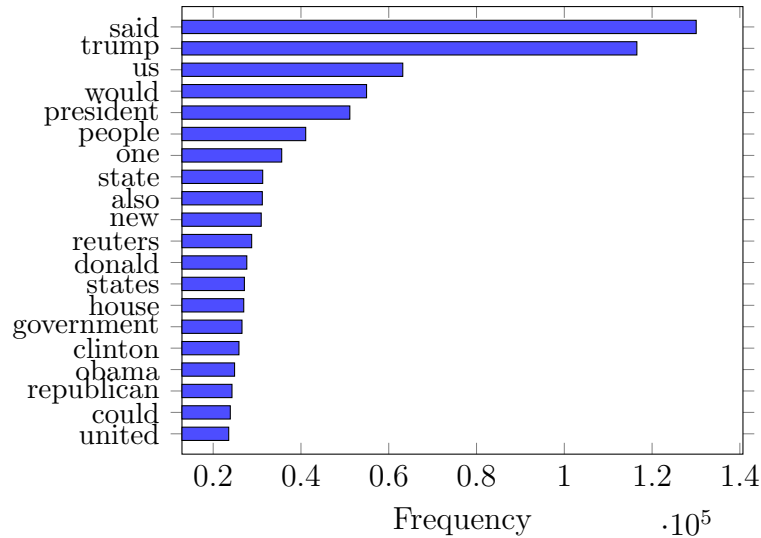


Figure 5.1: ISOT Dataset - Top 20 Common Words

The real news articles were sourced from Reuters, a well-established and credible news organisation, while the fake news articles were collected from various unreliable sources that were known for publishing fabricated or misleading content during the election period. The dataset contains a total of 44,898 news articles, with 21,417 real and 23,481 fake news samples, making it a well-balanced dataset in terms of class distribution (Ahmed et al., 2017). The word-frequency analysis in Figure 5.1 confirms that this dataset focusses on political news, particularly the 2016 Presidential Election. The most frequent words, such as “president” and “government”, and names like “Clinton” “Obama”, “Donald” and “Trump” reflect the political context of the content.

Kaggle Fake or Real Dataset

Not to be confused with the ‘Fake and Real’ Kaggle dataset, this dataset is relatively new according to its publish date on Kaggle but lacks transparency regarding the methods of data collection, which raises concerns about its reliability. Unlike the ISOT dataset, which provides detailed information about its sources, this dataset is contained in a single CSV file that includes the article title, full text, and a label of either ‘FAKE’ or ‘TRUE’. The dataset is evenly split, with 3,128 fake news articles and 3,128 real news articles, ensuring class balance.

During the research phase for this study, it was noted that this dataset bears a resemblance to the KDNuggets dataset hosted on GitHub, suggesting that it may be a repackaged version of an existing dataset rather than a new collection. This overlap highlights the importance of verifying the origin and uniqueness of datasets used in fake news detection research, particularly when evaluating intra-domain generalisability. Despite these concerns, the dataset remains popular, as demonstrated

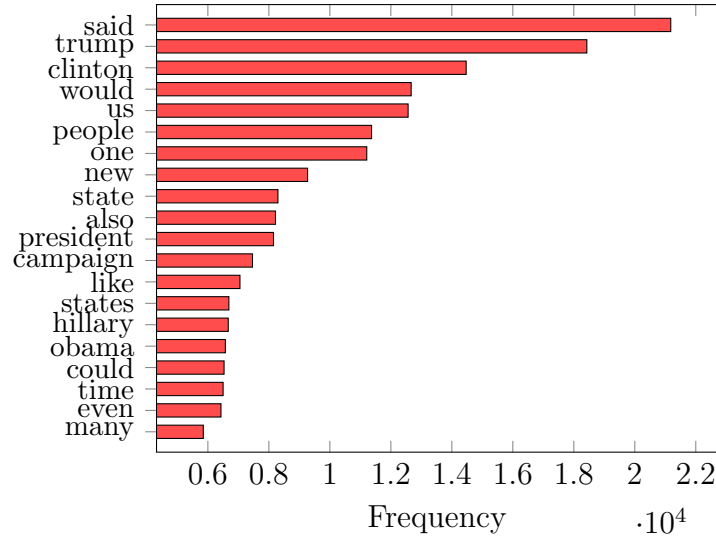


Figure 5.2: Kaggle Fake or Real - Top 20 Common Words

by its position as one of the top 15 datasets in the systematic review. Similar to the ISOT dataset, the word-frequency analysis conducted on this dataset revealed that the most common words, such as “Trump”, “Clinton”, “president”, and other political terms, closely align with the focus on political disinformation. This confirms that the dataset is appropriate for use in this study alongside the ISOT dataset.

Kaggle Fake News Dataset

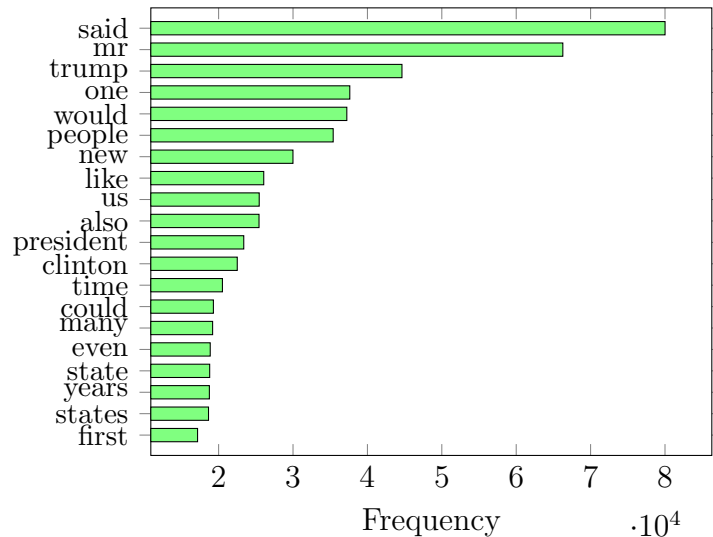


Figure 5.3: Kaggle Fake News - Top 20 Common Words

The Kaggle (Fake News) dataset is among the most popular Kaggle datasets in the literature, seeing use in 33 studies collected in the systematic review in Chapter 3. Similar to the previous Kaggle dataset, despite its popularity there is a lack of transparency regarding the methods of data collection. The dataset contains

five fields, including id, title, author, text and label with 10,413 articles allocated to the real news class and 10,387 allocated to the fake news class. Similar to the previous two datasets, the word-frequency analysis confirms the focus of this data on political news, with terms such as ‘Trump’, ‘Clinton’ and ‘president’ featuring frequently across the dataset. This alignment confirms its choice as a dataset for use in this study on intra-domain generalisability alongside the previous two datasets.

FakeNewsNet Dataset

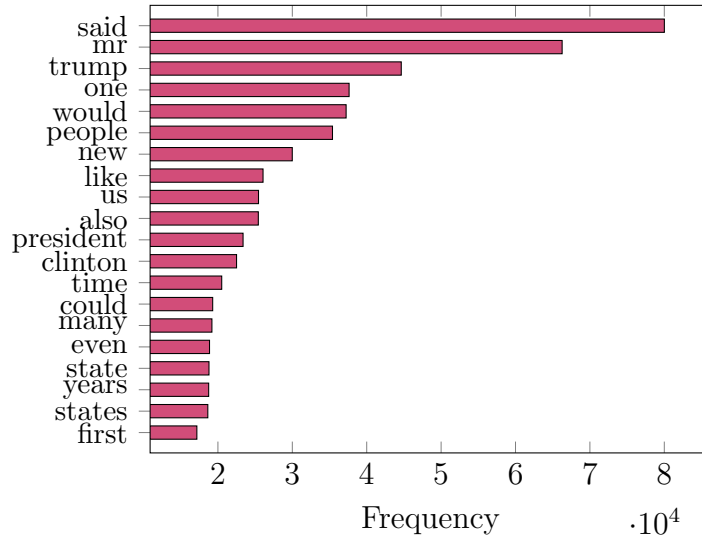


Figure 5.4: FakeNewsNet - Top 20 Common Words

The systematic review in Chapter 3 found that the FakeNewsNet dataset is among the most popular datasets in the literature, featuring in over 50 studies collected by the review. Similar to the ISOT dataset, FakeNewsNet discloses its data collection approach, sourcing articles from sites such as Politifact. While there is a more recent version of this dataset containing articles sourced from both Politifact and GossipCop (which incorporates celebrity news), it was determined that an older version of this dataset containing articles from BuzzFeed be used, owing to its stronger focus on the 2016 Presidential Election. In contrast to other datasets, the authors also provide code which allows you to collect social and spatiotemporal features from Twitter. However, as no other datasets provide such features, for consistency, only textual features were used for this dataset. The dataset is split into two parts, each containing news from Politifact and BuzzFeed, which were then combined. This resulted in an overall dataset of 522 articles with 211 labelled as ‘fake’ and 211 labelled as ‘true’ (Shu et al., 2017, 2019b). Similar to the previous datasets, the word-frequency analysis in Figure 5.4 confirms this dataset’s alignment with US political news, with terms such as ‘Trump’, ‘Clinton’ and ‘president’ featuring frequently.

5.4.2 Pre-Processing

Section 4.4 of Chapter 4 outlined several common pre-processing steps performed before feature extraction. In this experiment, following the data collection outlined in the previous section, the text was first converted to lowercase to prevent the model from treating identical words differently due to capitalisation. A lemmatisation step was then applied to reduce linguistic noise by simplifying words to their root form, enhancing consistency across the text data. Following this, additional noise such as punctuation, numbers, URLs, Twitter handles, extra whitespace, and stop words were removed, as these elements typically offer little meaningful contribution to the model’s predictive capabilities.

However, exceptions to this pre-processing pipeline were necessary for models utilising Word2Vec, BERT, and stylistic features. Word2Vec requires detailed contextual information to create accurate word embeddings, and BERT (Bidirectional Encoder Representations from Transformers) is highly sensitive to the input structure, as it uses the full sentence context. Therefore, only light cleaning was performed for these models to retain important contextual cues. For Word2Vec and BERT, the pre-processing was limited to converting text to lowercase, spell-checking, and removing URLs and Twitter handles, while other elements were left intact to preserve the original context. Similarly, stylistic features, which rely on statistical properties of the original text, required no pre-processing. By keeping the text unaltered for these models, the stylistic and contextual integrity necessary for accurate feature extraction and embedding creation was maintained.

5.4.3 Features

The previous Chapter outlined the different features that shall be utilised in this thesis. As found by the systematic review, these features are among the most popular for content-based fake news detection. This section provides a brief overview of these methods and how each method was implemented in the context of this study, detailing the tools and parameters used for processing article text.

Bag-of-Words

Bag-of-Words is a term frequency-based approach that converts text into a fixed-length vector by counting how many times each word appears. As this is purely a frequency-based approach, context and word order is not considered. This means that any information on the meaning of the text is lost. In the case of these experiments, SKLearn’s Count Vectorizer was utilised to create the Bag-of-Words representations. The key parameter used was `max_features=10,000`, which limits the vocabulary size to the top 10,000 most frequent terms across the dataset. This was

done to reduce the dimensionality of the feature space, balancing between retaining informative words and ensuring computational efficiency.

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) improves the Bag-of-Words approach by weighting words based on how often they appear in a document relative to how common they are across the entire dataset. This ensures that common words like “the” or “is” are down-weighted, while rarer, more informative terms are emphasised. In this experiment, SKLearn’s `TfidfVectorizer` was used to generate the TF-IDF representations. While default parameters were largely used in this experiment, the vocabulary was capped at 10,000 terms with the `max_features` parameter, similar to the Bag-of-Words approach. The analyser was set to ‘word’, meaning the vectorizer generated word-level features.

Word2Vec

Word2Vec is a type of static embedding (as discussed in Section 4.5.3 of Chapter 4), popular in the literature for its ability to represent words as vectors in a high-dimensional space. This approach assigns each word a unique vector based on its context, allowing similar words to be positioned close together while dissimilar words are further apart. While Word2Vec captures some semantic meaning by considering contextual information, it does not account for the order of words, making it context-independent. This limitation means that it assigns the same embedding to words that are morphologically identical but semantically different, such as “left” used as the past tense of “to leave” and “left” indicating direction.

In these experiments, Gensim was used to apply pre-trained Word2Vec embeddings. This library facilitated the transformation of textual data into vector representations efficiently, utilizing a pre-trained Word2Vec model trained on the Google News corpus. By leveraging this pre-trained model, the Word2Vec embeddings captured rich semantic relationships between words, enhancing the performance of the models without requiring additional training on the experimental data. The algorithm for extracting these embeddings can be found in Appendix A.1.

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a contextual embedding model (as discussed in Section 4.5.4 of Chapter 4) known for its ability to capture deep semantic and syntactic relationships in text. Unlike static embeddings such as Word2Vec, BERT considers the entire context of a word by analysing both its left and right surroundings, making it context-dependent. This enables BERT to

generate different embeddings for words with identical forms but varying meanings, such as the word “left” used as a direction versus “left” as the past tense of “leave.” The model’s deep bidirectional nature makes it especially powerful for capturing nuanced language patterns and word relationships.

In these experiments, the Transformers library by HuggingFace was used to implement a pre-trained BERT-based uncased model. This version of BERT does not differentiate between uppercase and lowercase text, which simplifies processing and reduces the complexity of training. The model was employed to generate contextual embeddings for each token in the input text, capturing both semantic and syntactic relationships at the sentence level (see Appendix A.2). This allowed the experiment to leverage BERT’s superior ability to model context and meaning compared to static embeddings.

Stylistic Features

As discussed in Section 4.5.5 of Chapter 4, stylistic features are typically statistical in nature or derived from the text such as with sentiment analysis. This experiment uses the set of 34 linguistic features (‘Linguistic Dimensions’ and ‘Punctuation Cues’) which was identified as producing the best performance in fake news classification in a series of experiments by Fernandez and Devaraj (2019). After these features were collected, they were combined to form a 34-dimensional vector that was then used for training on each model. A summary of these features collected from each document in the respective datasets is presented in Table 5.2.

5.4.4 Algorithms

Building on the findings from Chapter 3, a selection of well-established machine learning algorithms, recognised for their effectiveness in fake news detection, were employed to investigate the intra-domain generalisability of fake news detection models. This section presents a brief overview of the algorithms utilised, outlining their fundamental principles and method of implementation in this study.

The traditional machine learning algorithms—Naïve Bayes, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines (SVMs)—were implemented using scikit-learn. Naïve Bayes, a probabilistic model based on Bayes’ theorem, assumes feature independence and is computationally efficient, making it a useful baseline for text classification tasks such as fake news detection. Logistic Regression, a linear model, estimates the probability of a binary outcome based on input features and has demonstrated strong performance in text classification tasks. Decision Trees, a non-linear algorithm, split data hierarchically based on feature thresholds to classify instances, offering interpretability and sim-

Table 5.2: Fernandez Feature-Set

Feature	Description
Word Count	Total number of words
Syllables Count	Total number of syllables
Sentence Count	Total number of sentences
Word/Sent	Total words / Total Sentences
Long Words Count	Number of words with >6 characters
All Caps Count	Number of words in all caps
Unique Words Count	Number of unique words
Personal Pronouns %	% of words such as ‘I, we, she, him’
First Person Singular %	% of words such as ‘I, me’
First Person Plural %	% of words such as ‘we, us’
Second Person %	% of words such as ‘you, your’
Third Person Singular %	% of words such as ‘she, he, her, him’
Third Person Plural %	% of words such as ‘they, them’
Impersonal Pronouns %	% of words such as ‘it, that, anything’
Articles %	% of words such as ‘a, an, the’
Prepositions %	% of words such as ‘below, all, much’
Auxiliary Verbs %	% of words such as ‘have, did, are’
Common Adverbs %	% of words such as ‘just, usually, even’
Conjunctions %	% of words such as ‘until, so, and, but’
Negations %	% of words such as ‘no, never, not’
Common Verbs %	% of words such as ‘run, walk, swim’
Common Adjectives %	% of words such as ‘big, small, silly’
Comparisons %	% of words such as ‘better, greater, larger’
Concrete Figures %	% of words that represent real numbers
Punctuation Count	Total number of punctuation marks per document
Full Stop Count	Total number of full-stops
Commas Count	Total number of commas
Colons Count	Total number of colons
Semi-Colons Count	Total number of semi-colons
Question Marks Count	Total number of question marks
Exclamation Marks Count	Total number of exclamation marks
Dashes Count	Total number of dashes
Apostrophe Count	Total number of apostrophes
Brackets Count	Total number of brackets ‘()’

plicity. Random Forests, an ensemble method, combine multiple decision trees to enhance accuracy and reduce overfitting, striking a balance between performance and generalisability. Gradient Boosting, another ensemble method, builds models sequentially by correcting errors from previous iterations, enabling it to capture intricate relationships in data. SVMs, which are effective for high-dimensional data, classify instances by finding the optimal hyperplane that separates classes, offering robustness in handling complex datasets. All of these models were implemented using scikit-learn with default hyperparameters, ensuring consistent and fair comparisons across algorithms.

The Neural Network and LSTM models were both implemented using PyTorch. Neural Networks are effective for capturing complex, non-linear patterns, making them a suitable choice for binary classification tasks such as fake news detection. In this study, the Neural Network was designed with a single hidden layer containing 10 neurons, prioritising simplicity and computational efficiency. This minimalist architecture served as a baseline model to evaluate the performance of a straightforward approach on the classification task. To enhance generalisability and prevent overfitting, early stopping was employed to halt training when validation performance ceased to improve, ensuring that the model avoided overtraining while capturing essential patterns in the data.

The LSTM model, also implemented using PyTorch, was designed to leverage the sequential nature of textual data. Unlike the previous algorithms, the LSTM was applied only to the Word2Vec and BERT embeddings, as these representations inherently capture contextual and semantic relationships between words. Bag-of-Words and TF-IDF features were excluded from LSTM input, as they lack sequential information and treat text as unordered sets of terms, making them unsuitable for sequence-based models. The LSTM in this study consisted of an LSTM layer with 128 hidden units, enabling it to model complex temporal dependencies and contextual relationships within the text. A 40% dropout rate was applied to the LSTM's output to mitigate overfitting, while a fully connected layer mapped the final hidden state to two output classes (fake and real news).

5.4.5 Evaluation

This section outlines the evaluation methods employed to assess model performance and investigate the intra-domain generalisability of the fake news detection models. A combination of stratified K-Fold cross-validation (SCV) and external validation was used to ensure comprehensive and reliable evaluation. The Local Interpretable Model Explainer (LIME) was used to provide insight into the keywords used in discriminating articles between 'real' and 'fake'.

In the first experiment, SCV was performed to provide a baseline for comparing all combinations of features, machine learning models, and datasets. In this experiment, the dataset was split into $K=5$ folds, ensuring an 80/20 train-test split while maintaining an even class distribution across both the training and validation sets. This method was chosen specifically to avoid overfitting and ensure that each fold represented the overall data distribution. The process was repeated five times, with each fold serving as the validation set once, and the final performance metrics were averaged across all folds.

To assess the intra-domain generalisability of the models, the second experiment used external validation. In this experiment, each model was trained on all the data from one dataset and tested on the remaining datasets that had not been used in training. For models trained using Bag-of-Words and TF-IDF features, the vectorizers from this training phase were reused to transform the text from the other datasets, ensuring consistency in the feature extraction process. In contrast, for models using Word2Vec, BERT, or stylistic features, reusing vectorizers was not necessary as these methods were pre-trained, providing consistent representations across different datasets. In the case of stylistic features, the same approach for calculating these features was applied across all datasets.

In addition to these experiments, the third experiment employed LIME (Local Interpretable Model-Agnostic Explanations) to analyse the predictions made by the models. LIME functions by locally approximating the model’s decision boundary, enabling a thorough understanding of how particular features affect individual predictions. By pinpointing the features that significantly influence a specific classification, LIME sheds light on the model’s underlying mechanics.

The primary objective of using LIME in this context was to reveal the words and phrases considered most critical in determining whether an article is categorised as ‘real’ or ‘fake.’ This facilitated the creation of a frequency distribution that highlights which words were most frequently associated with each classification. This analysis aimed to investigate potential reasons for the models’ difficulties in generalising across datasets, as discussed in Section 3.5.4 of Chapter 3, which highlighted the challenges of cross-domain generalisability. By analysing the most influential words, it became possible to detect variations in linguistic patterns that differ across datasets, thereby highlighting the potential barriers to achieving strong generalisation.

5.5 Results

This section presents the findings from the experiments conducted to evaluate the intra-domain generalisability of the fake news detection models. The results are di-

vided into several subsections. The first subsection (5.5.1) outlines the results of the baseline experiment, which used stratified cross-validation. SCV was employed to create a benchmark for evaluating model performance, ensuring that the data distribution was consistent and that each algorithm was fairly assessed. These baseline results offer an initial view of how well the models performed within their original datasets.

Subsection 5.5.2 focuses on external validation, where models trained on one dataset were tested against entirely different datasets to assess their ability to generalise. This experiment aimed to evaluate the robustness of the models when applied to unfamiliar data, revealing how effectively the models' learned decision boundaries could be transferred to new domains of fake news detection. By comparing performance across different datasets, this section identifies key strengths and weaknesses in the models' generalisability.

Finally, the third subsection (5.5.3) shifts focus to the interpretability of the models, exploring how LIME (Local Interpretable Model-Agnostic Explanations) was used to explain model predictions. Rather than solely examining predictive accuracy, this section emphasises the importance of understanding the underlying factors that contribute to a model's decision. The goal is to uncover which features are most influential in classifying articles and to explore how these features vary across datasets, providing insights into the patterns that drive model performance in fake news detection.

5.5.1 Baseline K-fold Cross Validation

This experiment evaluated the baseline performance of the fake news detection models using stratified K-fold cross-validation, with K set to 5. This method ensured balanced class distribution across each of the five folds, providing a reliable assessment of the models' ability to handle in-domain data. The results from each fold were averaged to yield final performance metrics, establishing a benchmark for evaluating model effectiveness within the same dataset. This baseline experiment lays the groundwork for future comparisons through external validation, where the emphasis will transition to assessing the models' ability to generalise across various datasets.

Tables 5.3-5.6 provides a detailed breakdown of the results, which align closely with those reported in prior research, particularly for the ISOT, Kaggle Fake or Real, and Kaggle (Fake News) datasets. Across these three datasets, all models demonstrated strong performance when evaluated on an unseen portion of the datasets on which they were trained on, with accuracy scores generally ranging between 86% and 94%. These results confirm that the selected models are well-suited for han-

dling balanced, in-domain data and that the feature extraction methods, such as Bag-of-Words, TF-IDF, and Word2Vec, are effective for distinguishing between real and fake news in a controlled setting.

However the results for the FakeNewsNet dataset, despite aligning with the results from the literature, were notably weaker. This is likely due to its relatively small size compared to the other datasets used in this study. With fewer instances to train on, the model may have struggled to learn effectively, leading to a higher risk of overfitting and less robust generalisation capabilities. In machine learning, larger datasets typically provide more varied examples, enabling models to better understand the underlying patterns and nuances present in the data. The limited sample size in the FakeNewsNet dataset may have hindered the model's ability to generalise effectively, resulting in less accurate predictions. Additionally, this outcome may be attributed to the exclusive focus on textual features in this experiment. The literature often emphasises that, in addition to the text, social context features play a crucial role in this dataset. By omitting these social features in favour of purely textual ones, the model's capacity to capture the full spectrum of patterns commonly associated with fake news in the FakeNewsNet data was likely limited, contributing to the overall poorer performance.

Moreover, the Support Vector Machines (SVM) models trained on stylistic features showed underwhelming performance relative to other machine learning algorithms, which may be attributed to the hyperparameter settings. Given that uniform hyperparameters were applied across different feature types, it is possible that the default hyperparameters were more suited to token-representations (like TF-IDF or Word2Vec) and were not optimal for capturing stylistic features. This suggests that more tailored hyperparameter tuning could enhance the model's ability to pick up on stylistic or rhetorical patterns unique to fake news.

Similarly, the neural network trained on stylist features also exhibited a slight decline in performance relative to other feature-sets and models. While neural networks generally excel with token-representations due to their ability to capture non-linear relationships, the early stopping criterion applied to prevent overfitting may have inadvertently hindered the model's ability to learn from the stylistic features. The use of early stopping, while preventing overtraining, might have caused the neural network to terminate before it had adequately captured the intricacies of the linguistic patterns necessary for distinguishing between fake and real news. Consequently, the model may not have trained for enough epochs to fully leverage stylistic features.

Naive Bayes demonstrated underwhelming performance in certain cases, particularly when trained on Word2Vec, BERT, and stylistic features for the Kaggle Fake News dataset. This underperformance can be attributed to the algorithm's assump-

tion of feature independence, which does not align well with the dense, context-aware embeddings produced by Word2Vec and BERT or the complex, interdependent nature of stylistic features. While Naive Bayes remains effective with simpler, discrete feature sets like Bag-of-Words or TF-IDF, its application to more advanced and interdependent features underscores the need for careful consideration of feature-algorithm compatibility in fake news detection tasks.

Table 5.3: ISOT - K-Fold Results

Feature	Model	Acc.	Prec.	Rec.	Spec.	F1
Count	Naïve Bayes	0.87	0.85	0.90	0.84	0.87
	Logistic Regression	0.99	0.99	0.99	0.99	0.99
	Decision Tree	0.99	0.99	0.99	0.99	0.99
	Random Forest	0.98	0.98	0.97	0.98	0.97
	Gradient Boosting	0.97	0.97	0.97	0.97	0.97
	SVM	0.99	0.99	0.99	0.99	0.99
	Neural Network	0.99	0.99	0.98	0.99	0.99
TF-IDF	Naïve Bayes	0.89	0.89	0.88	0.90	0.88
	Logistic Regression	0.99	0.99	0.99	0.99	0.99
	Decision Tree	0.99	0.99	0.98	0.99	0.99
	Random Forest	0.99	0.99	0.99	0.99	0.99
	Gradient Boosting	0.99	0.99	0.99	0.99	0.99
	SVM	0.99	0.99	0.99	0.99	0.99
	Neural Network	0.99	0.99	0.98	0.99	0.99
Word2Vec	Naïve Bayes	0.88	0.85	0.92	0.84	0.88
	Logistic Regression	0.98	0.98	0.98	0.98	0.98
	Decision Tree	0.90	0.91	0.88	0.92	0.89
	Random Forest	0.99	0.99	0.99	0.99	0.99
	Gradient Boosting	0.99	0.99	0.99	0.99	0.99
	SVM	0.99	0.99	0.99	0.99	0.99
	Neural Network	0.99	0.99	0.99	0.99	0.99
	LSTM	0.98	0.98	0.98	0.98	0.98
BERT	Naïve Bayes	0.93	0.89	0.96	0.89	0.93
	Logistic Regression	0.96	0.96	0.97	0.96	0.96
	Decision Tree	0.93	0.94	0.91	0.95	0.93
	Random Forest	0.96	0.95	0.96	0.96	0.95
	Gradient Boosting	0.95	0.95	0.95	0.95	0.95
	SVM	0.98	0.98	0.98	0.98	0.98
	Neural Network	0.97	0.97	0.97	0.97	0.97
	LSTM	0.90	0.91	0.85	0.93	0.88
Stylistic Features	Naïve Bayes	0.70	0.62	0.93	0.48	0.75
	Logistic Regression	0.88	0.85	0.90	0.86	0.88
	Decision Tree	0.93	0.93	0.93	0.94	0.93
	Random Forest	0.96	0.97	0.95	0.97	0.96
	Gradient Boosting	0.95	0.94	0.95	0.94	0.95
	SVM	0.75	0.71	0.80	0.69	0.75
	Neural Network	0.84	0.89	0.79	0.94	0.94

Table 5.4: Kaggle Fake or Real - K-Fold Results

Feature	Model	Acc.	Prec.	Rec.	Spec.	F1
Count	Naïve Bayes	0.83	0.78	0.93	0.74	0.85
	Logistic Regression	0.95	0.95	0.93	0.96	0.94
	Decision Tree	0.79	0.80	0.78	0.80	0.79
	Random Forest	0.88	0.91	0.81	0.94	0.85
	Gradient Boosting	0.89	0.90	0.83	0.93	0.87
	SVM	0.94	0.95	0.91	0.96	0.93
	Neural Network	0.92	0.94	0.88	0.96	0.91
TF-IDF	Naïve Bayes	0.87	0.85	0.90	0.84	0.87
	Logistic Regression	0.96	0.95	0.95	0.96	0.95
	Decision Tree	0.80	0.79	0.80	0.79	0.80
	Random Forest	0.94	0.95	0.90	0.97	0.93
	Gradient Boosting	0.94	0.92	0.94	0.94	0.93
	SVM	0.92	0.90	0.92	0.92	0.91
	Neural Network	0.95	0.96	0.93	0.97	0.94
Word2Vec	Naïve Bayes	0.68	0.63	0.86	0.49	0.73
	Logistic Regression	0.95	0.95	0.93	0.96	0.94
	Decision Tree	0.77	0.77	0.77	0.77	0.77
	Random Forest	0.93	0.96	0.89	0.97	0.92
	Gradient Boosting	0.94	0.92	0.93	0.94	0.93
	SVM	0.96	0.96	0.95	0.97	0.96
	Neural Network	0.96	0.96	0.95	0.97	0.96
	LSTM	0.90	0.92	0.87	0.93	0.89
BERT	Naïve Bayes	0.80	0.77	0.84	0.75	0.81
	Logistic Regression	0.88	0.88	0.82	0.92	0.85
	Decision Tree	0.80	0.80	0.81	0.79	0.80
	Random Forest	0.88	0.89	0.81	0.92	0.85
	Gradient Boosting	0.88	0.88	0.83	0.91	0.85
	SVM	0.91	0.91	0.88	0.93	0.89
	Neural Network	0.87	0.87	0.84	0.90	0.85
	LSTM	0.93	0.92	0.95	0.91	0.93
Stylistic Features	Naïve Bayes	0.61	0.57	0.89	0.33	0.70
	Logistic Regression	0.92	0.93	0.87	0.95	0.90
	Decision Tree	0.77	0.77	0.77	0.77	0.77
	Random Forest	0.97	0.99	0.95	0.99	0.97
	Gradient Boosting	0.97	0.99	0.94	0.99	0.96
	SVM	0.77	0.74	0.71	0.81	0.72
	Neural Network	0.95	0.94	0.94	0.98	0.96

Table 5.5: Kaggle (Fake News) - K-Fold Results

Feature	Model	Acc.	Prec.	Rec.	Spec.	F1
Count	Naïve Bayes	0.88	0.82	0.93	0.84	0.87
	Logistic Regression	0.95	0.95	0.93	0.96	0.94
	Decision Tree	0.89	0.87	0.88	0.90	0.87
	Random Forest	0.88	0.91	0.81	0.94	0.85
	Gradient Boosting	0.89	0.90	0.83	0.93	0.87
	SVM	0.94	0.95	0.91	0.96	0.93
	Neural Network	0.92	0.94	0.88	0.96	0.91
TF-IDF	Naïve Bayes	0.89	0.88	0.85	0.91	0.87
	Logistic Regression	0.96	0.95	0.95	0.96	0.95
	Decision Tree	0.89	0.87	0.87	0.90	0.87
	Random Forest	0.94	0.95	0.90	0.97	0.93
	Gradient Boosting	0.94	0.92	0.94	0.94	0.93
	SVM	0.92	0.90	0.92	0.92	0.91
	Neural Network	0.95	0.96	0.93	0.97	0.94
Word2Vec	Naïve Bayes	0.66	0.77	0.32	0.93	0.45
	Logistic Regression	0.95	0.95	0.93	0.96	0.94
	Decision Tree	0.77	0.73	0.73	0.79	0.73
	Random Forest	0.93	0.96	0.89	0.97	0.92
	Gradient Boosting	0.94	0.92	0.93	0.94	0.93
	SVM	0.96	0.96	0.95	0.97	0.96
	Neural Network	0.96	0.96	0.95	0.97	0.96
	LSTM	0.90	0.91	0.85	0.93	0.88
BERT	Naïve Bayes	0.72	0.80	0.47	0.91	0.59
	Logistic Regression	0.88	0.88	0.82	0.92	0.85
	Decision Tree	0.76	0.73	0.73	0.79	0.73
	Random Forest	0.88	0.89	0.81	0.92	0.85
	Gradient Boosting	0.88	0.88	0.83	0.91	0.85
	SVM	0.91	0.91	0.88	0.93	0.89
	Neural Network	0.87	0.87	0.84	0.90	0.85
	LSTM	0.96	0.96	0.94	0.97	0.95
Stylistic Features	Naïve Bayes	0.94	0.96	0.90	0.97	0.93
	Logistic Regression	0.92	0.93	0.87	0.95	0.90
	Decision Tree	0.95	0.95	0.95	0.96	0.95
	Random Forest	0.97	0.99	0.95	0.99	0.97
	Gradient Boosting	0.97	0.99	0.94	0.99	0.96
	SVM	0.77	0.74	0.71	0.81	0.72
	Neural Network	0.95	0.94	0.94	0.98	0.96

Table 5.6: FakeNewsNet - K-Fold Results

Feature	Model	Acc.	Prec.	Rec.	Spec.	F1
Bag of Words	Naïve Bayes	0.52	0.52	0.53	0.51	0.52
	Logistic Regression	0.56	0.57	0.57	0.55	0.56
	Decision Tree	0.50	0.48	0.28	0.71	0.33
	Random Forest	0.56	0.58	0.58	0.55	0.57
	Gradient Boosting	0.52	0.52	0.59	0.44	0.54
	SVM	0.52	0.61	0.42	0.60	0.42
	Neural Network	0.50	0.29	0.41	0.58	0.50
TF-IDF	Naïve Bayes	0.53	0.52	0.58	0.47	0.55
	Logistic Regression	0.57	0.59	0.60	0.55	0.59
	Decision Tree	0.48	0.47	0.32	0.64	0.37
	Random Forest	0.59	0.62	0.62	0.56	0.61
	Gradient Boosting	0.53	0.53	0.63	0.43	0.57
	SVM	0.56	0.54	0.81	0.32	0.65
	Neural Network	0.51	0.46	0.41	0.61	0.43
Word2Vec	Naïve Bayes	0.53	0.53	0.73	0.34	0.61
	Logistic Regression	0.61	0.65	0.59	0.62	0.61
	Decision Tree	0.48	0.47	0.37	0.60	0.41
	Random Forest	0.58	0.60	0.64	0.52	0.61
	Gradient Boosting	0.57	0.58	0.62	0.52	0.59
	SVM	0.61	0.65	0.62	0.60	0.62
	Neural Network	0.50	0.39	0.19	0.81	0.31
	LSTM	0.56	0.54	0.63	0.49	0.56
BERT	Naïve Bayes	0.61	0.62	0.68	0.54	0.64
	Logistic Regression	0.58	0.59	0.60	0.56	0.59
	Decision Tree	0.51	0.50	0.33	0.68	0.38
	Random Forest	0.61	0.63	0.67	0.56	0.64
	Gradient Boosting	0.59	0.59	0.68	0.50	0.63
	SVM	0.62	0.65	0.68	0.56	0.65
	Neural Network	0.54	0.44	0.59	0.49	0.62
	LSTM	0.55	0.53	0.50	0.59	0.50
Stylistic Features	Naïve Bayes	0.55	0.56	0.38	0.71	0.44
	Logistic Regression	0.57	0.61	0.49	0.65	0.54
	Decision Tree	0.53	0.56	0.42	0.64	0.46
	Random Forest	0.57	0.57	0.65	0.48	0.60
	Gradient Boosting	0.56	0.56	0.64	0.48	0.60
	SVM	0.51	0.64	0.20	0.83	0.27
	Neural Network	0.51	0.51	0.65	0.46	0.55

5.5.2 External Validation

This section presents the results of the external validation experiments, designed to assess the generalisability of the fake news detection models across different datasets. Unlike the baseline experiment, where models were evaluated on the same dataset they were trained on using SCV, external validation focuses on testing models with entirely new data to evaluate how well they can generalise to unseen instances. This analysis was broken down into three sections. The first section provides an overview of the drops in accuracy observed when models were tested on datasets they were not trained on. The second section focuses on the average performance of the models across different feature-set, offering insight into which features generalised better. Finally, the third section assesses the average performance of various algorithms, highlighting which algorithms were more effective in maintaining accuracy when applied to new, unseen data. Together, these analyses offer a comprehensive view of how well models and feature sets generalised across datasets.

Generalisability by Dataset

Table 5.7: External Validation Across Datasets

Training Dataset	X-Dataset	Acc.	Prec.	Rec.	Spec.	F1
ISOT	FNN	0.53	0.55	0.23	0.83	0.32
	Kaggle (Fake News)	0.43	0.29	0.17	0.63	0.19
	Kaggle Fake or Real	0.59	0.68	0.37	0.81	0.45
Kaggle Fake or Real	FNN	0.56	0.56	0.56	0.56	0.56
	ISOT	0.65	0.65	0.61	0.69	0.61
	Kaggle (Fake News)	0.31	0.15	0.14	0.44	0.14
Kaggle (Fake News)	FNN	0.47	0.46	0.61	0.32	0.52
	ISOT	0.35	0.31	0.40	0.30	0.34
	Kaggle Fake or Real	0.30	0.33	0.48	0.11	0.39
FakeNewsNet	ISOT	0.61	0.58	0.65	0.58	0.60
	Kaggle (Fake News)	0.43	0.35	0.38	0.47	0.35
	Kaggle Fake or Real	0.61	0.62	0.60	0.62	0.60

Table 5.7 provide an averaged summary of the performance of several different models and feature-sets across the four datasets used in this study. By averaging these results, this analysis provides a broad perspective on how well models and features generalise when trained on one dataset and externally validated on the remaining datasets not used in training. As can be seen from these four tables, models suffer a significant drop in performance when compared to their baseline

results in Section 5.5.1. The analysis of the results highlights several key patterns regarding the generalisation capabilities of models trained on specific datasets when evaluated on external datasets.

When using ISOT as the training dataset, the models exhibit mixed performance when tested across the three remaining datasets not used in training. For instance, when tested against the Kaggle Fake or Real dataset, models trained on the ISOT dataset produce an average accuracy of 0.59, indicating a moderate ability to generalise from the ISOT dataset. However, models tested against this dataset are conservative in making predictions for the ‘true news’ class, as indicated by the high precision and low recall of the model. This pattern suggests that while the models are accurate when they predict true news, they tend to miss a significant number of actual true news instances when tested against the Kaggle Fake or Real dataset. A similar trend can be observed in models tested against the FakeNewsNet and Kaggle (Fake News) datasets, however performance against these datasets is worse, particular in the case of the Kaggle (Fake News) dataset.

In terms of models trained on the Kaggle (Fake News) dataset, the worst external validation performance is observed compared to models trained on other datasets. In all cases, average model accuracy is less than 0.5 when tested against the three datasets not used in training. However, compared to models trained on the ISOT dataset, higher F1-scores are observed indicating that models trained on this dataset have a slightly better balance between precision and recall when identifying instances of true news. Despite this, models trained using this dataset produce noticeably lower specificity relative to recall, particularly when tested against the Kaggle Fake or Real dataset. This indicates that models trained on the Kaggle (Fake News) dataset struggle to classify instances of fake news.

Interestingly, when observing models trained on the Kaggle Fake or Real dataset the opposite is true, as models trained on this dataset and tested against the Kaggle (Fake News) dataset have notably lower recall compared to specificity. Additionally, the lowest average accuracies are achieved when testing between these datasets compared to the other two datasets. This suggests a fundamental distinction between the articles in these two datasets, further underscoring the issues of coarsely labelled datasets in the literature. In contrast, when observing the performance of these models tested against the ISOT dataset, the highest accuracy is achieved compared to all other external validation tests conducted in this experiment. This finding suggests that the ISOT dataset shares more similarities with the Kaggle Fake or Real dataset, allowing the models to generalise better in this context. Given that all these datasets broadly cover the same topic and time period (as outlined in Section 5.4.1), this discrepancy in generalisability between datasets points to underlying differences in content structure, labelling practices, or editorial biases. These factors

could impact how models trained on one dataset perform when applied to another, further emphasising the need for more consistent dataset curation.

Observing the performance of models trained on the FakeNewsNet dataset and tested on the remaining datasets, further notable patterns emerge. Despite being the smallest dataset in this study and producing the worst baseline performance observed in Section 5.5.1, models trained on this dataset actually show improved average accuracy under external validation testing conditions. Moreover, in terms of precision, recall and specificity, models trained on this dataset manage to strike a better balance compared to those trained on other datasets. A likely explanation for this is that, compared to models trained on other datasets, those trained on FakeNewsNet struggle to overfit to the nuances of this particular dataset (as indicated by its poor baseline performance). As a result, these models are better positioned to generalise to other datasets.

Generalisability by Features

This section of the analysis focuses on the performance of the different groups of features used in this study. The performance of these different features—Bag-of-Words, TF-IDF, Word2Vec, BERT, and stylistic features—is presented in Table 5.8. Overall, the mean accuracies across these different sets of features are relatively comparable, with TF-IDF exhibiting the lowest mean accuracy at 0.47, while Stylistic Features show the highest mean accuracy at 0.52. However, a large degree of variation exists between all metrics which are explored below.

For Bag-of-Words, the performance metrics exhibit moderate variability across datasets, with accuracy ranging from 0.17 to 0.72 and a mean of 0.48. The mean precision (0.46) and recall (0.41) indicate that while the model can identify some patterns, it struggles to consistently distinguish between fake and true news. The interquartile range (IQR) for recall (0.37) and specificity (0.40) highlights notable fluctuations in performance, reflecting variability across different datasets. Additionally, the standard deviation values—0.19 for precision and 0.26 for specificity—suggest that Bag-of-Words fails to generalise effectively across datasets. These results indicate that while Bag-of-Words may perform reasonably well on some datasets, its inability to consistently capture the nuances in the data limits its utility for generalisable fake news detection. Similarly, TF-IDF shows a performance pattern similar to Bag-of-Words, with accuracy ranging from 0.18 to 0.72 and a mean of 0.47. However, TF-IDF achieves a slightly higher recall (mean = 0.43) compared to Bag-of-Words, though this comes at a marginal expense of precision (mean = 0.44). This trade-off suggests that TF-IDF may capture more relevant features, but it still struggles to consistently balance precision and recall. The standard deviations for recall (0.26) and specificity (0.26), along with IQR values, indicate even similar

Table 5.8: External Validation - Features

Bag-of-Words					
Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.17	0.06	0.00	0.03	0.01
Max	0.72	0.81	0.90	1.00	0.72
Mean	0.48	0.46	0.41	0.54	0.40
Median	0.52	0.49	0.43	0.56	0.41
IQR	0.19	0.28	0.37	0.40	0.31
Std. Dev.	0.13	0.19	0.23	0.26	0.20

TF-IDF					
Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.18	0.00	0.00	0.00	0.03
Max	0.72	0.79	0.99	1.00	0.73
Mean	0.47	0.44	0.43	0.51	0.41
Median	0.50	0.48	0.42	0.52	0.46
IQR	0.22	0.32	0.40	0.37	0.34
Std. Dev.	0.15	0.20	0.26	0.26	0.21

Word2Vec					
Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.18	0.10	0.06	0.05	0.08
Max	0.72	0.74	0.97	0.89	0.75
Mean	0.48	0.44	0.42	0.53	0.41
Median	0.50	0.49	0.37	0.54	0.41
IQR	0.22	0.33	0.35	0.22	0.35
Std. Dev.	0.15	0.19	0.24	0.20	0.20

BERT					
Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.15	0.09	0.04	0.03	0.06
Max	0.84	0.86	0.92	0.96	0.82
Mean	0.49	0.46	0.42	0.54	0.43
Median	0.51	0.50	0.45	0.61	0.46
IQR	0.30	0.41	0.35	0.36	0.35
Std. Dev.	0.18	0.22	0.23	0.25	0.21

Stylistic Features					
Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.31	0.17	0.09	0.02	0.13
Max	0.62	1.00	1.00	1.00	0.69
Mean	0.52	0.51	0.50	0.53	0.45
Median	0.53	0.52	0.43	0.60	0.46
IQR	0.05	0.09	0.49	0.45	0.32
Std. Dev.	0.07	0.14	0.31	0.29	0.17

degree of variation in performance across datasets compared to Bag-of-Words.

Word2Vec demonstrates comparable overall performance to both Bag-of-Words and TF-IDF, with accuracy ranging from 0.18 to 0.72 and a mean of 0.48. Its strength lies in the balance between precision (mean = 0.44) and recall (mean = 0.42), indicating that Word2Vec performs more consistently across datasets in identifying real and fake news. The IQR for precision (0.33) and recall (0.35) shows variability in performance similar to that of other token-based methods. However, the slightly narrower standard deviation for specificity (0.20) suggests greater stability when distinguishing between true news and fake news. Overall, Word2Vec demonstrates slightly better reliability as a feature extraction method compared to BoW and TF-IDF, though it does not achieve the peak performance seen in more advanced techniques like BERT.

In contrast, BERT exhibits a broad range of performance across datasets, with

accuracy varying from 0.15 to 0.84 and a mean of 0.49. This highlights BERT's potential for strong performance under optimal conditions, though it also reveals significant variability. Precision (mean = 0.46) and recall (mean = 0.42) are relatively balanced, but the interquartile ranges for precision (0.41) and recall (0.35), alongside the standard deviations (0.22 for precision and 0.23 for recall), indicate a sensitivity to dataset characteristics. The interquartile range for specificity (0.36) further underscores the fluctuations in performance across different datasets. While BERT demonstrates the capability to capture complex patterns and achieve strong results under certain conditions, its inconsistent performance suggests limited reliability for generalisation across varied datasets.

Stylistic features, compared to token-based methods, demonstrate more consistent performance with a mean accuracy of 0.52 and a notably lower standard deviation of 0.07, suggesting they maintain stable accuracy across different datasets. Precision also exhibits improved consistency, with a mean of 0.53 and a standard deviation of 0.16, which points to these features' reliable performance in correctly identifying true news compared to the more variable token-based methods. However, recall (mean = 0.48) and specificity (mean = 0.55) show greater variability, marked by standard deviations of 0.30 and 0.29, respectively. These figures are accompanied by interquartile ranges (IQR) of 0.44 for recall and 0.41 for specificity, underscoring significant performance fluctuations across datasets. This suggests that while stylistic features generally provide stable accuracy and precision, their effectiveness in distinguishing true from fake news can vary considerably depending on the specifics of the dataset. Despite these challenges, stylistic features maintain a better balance between recall and specificity than some token-based methods. This balance indicates a greater potential for stylistic features to effectively handle both classes of news. The overall stability in accuracy and precision, combined with moderate variability in recall and specificity, suggests that stylistic features warrant further investigation, particularly when considering the wide range of stylistic features not tested in this study.

Generalisability by Algorithm

This section analyses the performance of various machine learning algorithms utilised in the study, including Naïve Bayes, Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, SVMs, Neural Networks, and LSTMs. As in the previous section, Table 5.9 provide summary statistics of the performance metrics employed in this research. Overall, the mean accuracies across these algorithms are more consistent when compared to the differences in mean accuracy between feature sets, with the mean accuracies for algorithms ranging narrowly from 0.48 to 0.49, with the exception of LSTM which had a mean accuracy of 0.45. Similar to the analysis with

feature-sets, there is some degree of variation between the remaining performance metrics which are explored in the following paragraphs.

Table 5.9: External Validation - Algorithms

Naïve Bayes						Logistic Regression					
Stat	Acc.	Prec.	Rec.	Spec.	F1	Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.18	0.10	0.09	0.07	0.10	Min	0.21	0.07	0.04	0.03	0.06
Max	0.77	0.77	1.00	0.89	0.77	Max	0.74	0.81	0.98	0.96	0.74
Mean	0.49	0.46	0.48	0.50	0.45	Mean	0.49	0.46	0.44	0.53	0.43
Median	0.51	0.49	0.45	0.52	0.46	Median	0.52	0.52	0.43	0.54	0.41
IQR	0.12	0.19	0.33	0.27	0.29	IQR	0.24	0.34	0.44	0.33	0.37
Std. Dev.	0.11	0.14	0.24	0.21	0.17	Std. Dev.	0.16	0.21	0.26	0.25	0.21
Decision Tree						Random Forest					
Stat	Acc.	Prec.	Rec.	Spec.	F1	Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.26	0.15	0.05	0.02	0.09	Min	0.17	0.08	0.04	0.02	0.06
Max	0.67	0.78	1.00	0.94	0.67	Max	0.75	1.00	1.00	1.00	0.74
Mean	0.49	0.47	0.41	0.56	0.41	Mean	0.49	0.47	0.38	0.58	0.39
Median	0.52	0.50	0.38	0.60	0.40	Median	0.53	0.55	0.32	0.63	0.37
IQR	0.12	0.20	0.27	0.28	0.21	IQR	0.23	0.40	0.40	0.32	0.37
Std. Dev.	0.10	0.14	0.21	0.24	0.15	Std. Dev.	0.15	0.22	0.25	0.25	0.21
Gradient Boosting						SVM					
Stat	Acc.	Prec.	Rec.	Spec.	F1	Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.17	0.00	0.00	0.05	0.01	Min	0.18	0.05	0.03	0.03	0.04
Max	0.72	0.99	1.00	1.00	0.73	Max	0.79	0.81	0.99	0.95	0.80
Mean	0.49	0.47	0.41	0.57	0.41	Mean	0.49	0.46	0.46	0.51	0.44
Median	0.52	0.51	0.42	0.59	0.42	Median	0.52	0.51	0.45	0.49	0.46
IQR	0.18	0.32	0.45	0.33	0.37	IQR	0.26	0.30	0.38	0.39	0.29
Std. Dev.	0.14	0.21	0.26	0.27	0.21	Std. Dev.	0.16	0.20	0.25	0.24	0.21
Neural Network						LSTM					
Stat	Acc.	Prec.	Rec.	Spec.	F1	Stat	Acc.	Prec.	Rec.	Spec.	F1
Min	0.20	0.06	0.05	0.00	0.05	Min	0.15	0.11	0.04	0.03	0.06
Max	0.84	0.86	0.99	0.89	0.82	Max	0.64	0.71	0.92	0.96	0.71
Mean	0.48	0.45	0.47	0.47	0.44	Mean	0.45	0.43	0.40	0.50	0.38
Median	0.50	0.49	0.42	0.54	0.47	Median	0.51	0.47	0.43	0.52	0.40
IQR	0.22	0.35	0.44	0.44	0.36	IQR	0.20	0.27	0.45	0.34	0.36
Std. Dev.	0.16	0.21	0.28	0.28	0.21	Std. Dev.	0.15	0.18	0.25	0.27	0.20

Naïve Bayes demonstrates a mean accuracy of 0.49, which is comparable to other algorithms in this study. Precision (mean = 0.46) and recall (mean = 0.48) are relatively balanced, indicating that Naïve Bayes performs similarly in identifying both true and fake news cases. However, its recall variability, as indicated by a standard deviation of 0.24, suggests that the algorithm's ability to capture instances

of true news is more sensitive to dataset characteristics than its precision. Specificity (mean = 0.50) aligns closely with recall, reflecting an even-handed approach to both classes. The interquartile range (IQR) for recall (0.33) and specificity (0.27) highlights variability across datasets, suggesting that while Naïve Bayes offers a straightforward and computationally efficient approach, its generalisability remains limited when applied to diverse datasets.

Logistic Regression also demonstrates a mean accuracy of 0.49, which is consistent with the performance of Naïve Bayes. Precision (mean = 0.46) and recall (mean = 0.44) are similarly balanced, indicating that Logistic Regression performs slightly better at capturing true news cases compared to Naïve Bayes. Specificity (mean = 0.53) is slightly higher, reflecting an improved ability to correctly identify fake news. However, the variability in recall, as indicated by a standard deviation of 0.26, suggests that Logistic Regression's performance is more influenced by dataset characteristics, especially in its ability to identify true news. The interquartile range (IQR) for recall (0.44) and specificity (0.33) further highlights this variability. While Logistic Regression provides a simple and interpretable model, its sensitivity to different datasets limits its consistency, though it strikes a slightly better balance between recall and specificity compared to Naïve Bayes.

Decision Trees also achieve a mean accuracy of 0.49, comparable to both Naïve Bayes and Logistic Regression. Precision (mean = 0.47) and recall (mean = 0.41) indicate that Decision Trees are slightly more effective at identifying fake news compared to Logistic Regression, as reflected by their higher mean specificity (0.56). However, the standard deviation of 0.21 for recall suggests variability in their ability to correctly classify true news across datasets. Similarly, the interquartile range (IQR) for specificity (0.28) and recall (0.27) highlights moderate fluctuations in performance. While Decision Trees provide an interpretable and non-linear approach to classification, their sensitivity to dataset characteristics can lead to inconsistent performance. Nevertheless, the algorithm's higher specificity makes it particularly well-suited for detecting fake news, albeit with a trade-off in recall compared to Logistic Regression.

Random Forest demonstrates a mean accuracy of 0.49, which is consistent with the other algorithms analysed so far. Precision (mean = 0.47) is comparable to that of Decision Trees, but Random Forest achieves a slightly higher specificity (mean = 0.58), indicating a stronger ability to correctly identify fake news. However, this comes at the cost of recall (mean = 0.38), which is the lowest among the algorithms, suggesting that Random Forest is less effective at capturing instances of true news. The interquartile range (IQR) for both recall (0.40) and specificity (0.32) highlights moderate variability in its performance across datasets. Similarly, the standard deviation of 0.25 for specificity indicates fluctuations in its ability to detect fake news

reliably. Overall, Random Forest demonstrates a trade-off, favouring specificity at the expense of recall, making it particularly effective for fake news detection but less balanced in identifying true news cases compared to Logistic Regression or Decision Trees.

Gradient Boosting achieves a mean accuracy of 0.49, aligning closely with the other algorithms in this study. Precision (mean = 0.47) and recall (mean = 0.41) suggest that Gradient Boosting provides a balanced performance, similar to Random Forest. However, its slightly higher specificity (mean = 0.57) indicates a marginally better ability to correctly identify fake news cases. The interquartile range (IQR) for recall (0.45) and specificity (0.33) highlights greater variability in performance compared to Random Forest, particularly in its ability to capture true news cases. Additionally, the standard deviation for recall (0.26) and specificity (0.27) underscores the sensitivity of Gradient Boosting to dataset characteristics. While Gradient Boosting demonstrates comparable mean performance metrics to Random Forest, its increased variability suggests that its results are more dataset-dependent, potentially limiting its generalisability under diverse conditions.

The Neural Network demonstrates a mean accuracy of 0.48, slightly lower than some of the other algorithms analysed. Precision (mean = 0.45) and recall (mean = 0.47) are relatively balanced, indicating that the Neural Network performs similarly in identifying both true and fake news cases. However, its specificity (mean = 0.47) is notably lower than that of algorithms like Random Forest or Gradient Boosting, suggesting that it struggles more with correctly identifying fake news. The interquartile range (IQR) for recall (0.44) and specificity (0.44) highlights considerable variability in its performance, further emphasised by the standard deviations of 0.28 for both recall and specificity. This variability reflects the Neural Network's sensitivity to dataset characteristics. While the Neural Network demonstrates the capacity for strong performance under optimal conditions, its higher variability and lower specificity compared to other algorithms suggest that it is less consistent and generalisable across diverse datasets.

Support Vector Machines (SVMs) offer a more balanced and reliable performance among the algorithms tested, with a mean accuracy of 0.49 that is consistent with other models in this analysis. Precision (mean = 0.46) and recall (mean = 0.46) are well-aligned, reflecting the algorithm's ability to evenly classify both true and fake news cases. Specificity (mean = 0.51) is slightly higher, indicating a small bias toward correctly identifying fake news. The interquartile range (IQR) for recall (0.38) and specificity (0.39), along with standard deviations of 0.25 and 0.24, respectively, suggest moderate variability but more stability compared to Neural Networks or Gradient Boosting. This balance across metrics potentially makes SVMs a more consistent and dependable choice for fake news detection.

Long Short-Term Memory (LSTM) networks demonstrate a mean accuracy of 0.45, slightly lower than other algorithms in this analysis. Precision (mean = 0.43) and recall (mean = 0.40) are relatively balanced, but both metrics fall below those of simpler models like Logistic Regression or SVMs. Specificity (mean = 0.50) aligns closely with recall, indicating that LSTMs perform similarly in identifying both true and fake news cases. However, the interquartile range (IQR) for recall (0.45) and specificity (0.34), along with standard deviations of 0.25 and 0.27 respectively, highlight substantial variability in performance across datasets. This variability can be partially attributed to the relatively small amount of data available, which limits the LSTM’s ability to effectively learn and generalise complex sequential patterns. While LSTMs are theoretically well-suited for capturing temporal dependencies in text, their higher sensitivity to data limitations and variability in this study underscores the challenges of applying such models with constrained dataset sizes.

Overall, the performance of the algorithms tested in this study shows a narrow range of mean accuracy, with all models performing similarly, ranging between 0.45 and 0.49. While there are differences in precision, recall, and specificity, the variations between algorithms are less significant when compared to the differences observed across feature sets and datasets. This indicates that the choice of algorithm has less influence on generalisability than the features used and the characteristics of the datasets themselves.

5.5.3 Interpreting Models Trained on Token-Representations

Section 5.5.2 clearly demonstrated the poor generalisability of machine learning models trained on a variety of token representations and stylistic features. To further investigate the reasons behind this poor generalisability, this section extends the analysis of token representations by leveraging the LIME package, as detailed in Section 4.9. The LimeTextExplainer submodule was used to generate two lists of words: a list of words that were most fundamental to classifying a document as ‘fake’, and a list of words that were most fundamental to classifying a document as ‘real’ (Ribeiro et al., 2016). As LIME relies on an SKLearn pipeline that includes an SKLearn vectorizer and machine learning model, a Logistic Regression model was trained using TF-IDF features to be used as part of the pipeline. Training is performed for each dataset and lists are generated using unseen documents from each dataset. Due to the computational complexity of LIME, this analysis used a random sample of 100 unseen documents from each dataset. Next, a frequency distribution of the top 15 keywords that increase the likelihood of a model classifying ‘real’ or ‘fake’ for each dataset was produced. The FakeNewsNet dataset was excluded from this analysis, owing to its extremely poor baseline performance, as identified by the

experiment detailed in Section 5.4.5. Table 5.10 below shows the ranked list of 15 keywords that contributed to the classification for each datasets.

Table 5.10: Frequency Distribution of Keywords Contributing to Classification

ISOT				Kaggle Fake or Real				Kaggle (Fake News)			
Real		Fake		Real		Fake		Real		Fake	
Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq
Reuters	52	trump	32	president	20	2016	17	Clinton	14	Mr	26
Washington	19	just	20	state	19	Hillary	13	Hillary	11	president	20
Wednesday	13	image	11	Obama	13	October	13	2016	11	Ms	16
Trumps	11	Obama	11	house	12	election	11	October	8	twitter	15
Tuesday	11	Hillary	9	told	10	Russia	7	war	8	follow	11
minister	11	don	8	says	10	FBI	6	share	8	com	7
house	6	like	8	sanders	9	article	6	election	5	united	7
Friday	6	Could	7	campaign	8	just	6	Obama	4	new	6
Government	5	GOP	6	white	8	email	5	LA	4	news	5
Thursday	5	Doesn't	6	debate	8	war	5	source	4	Breitbart	5
election	5	Black	5	republican	7	world	4	Aleppo	4	Sunday	5
court	4	Americans	4	senate	7	Comey	3	November	4	percent	5
EU	4	Right	4	voters	6	share	3	FBI	4	York	4
month	4	Video	4	Islamic	6	daily	3	UK	3	Trumps	4

The frequency distribution presented in Table 5.10 generated by LIME for the fake and real classes across the three datasets (ISOT, Kaggle Fake or Real, and Kaggle Competition) reveals key patterns that provide insight into how the models differentiate between fake and real news articles. The distributions indicate that token-based features heavily rely on contextually salient words and reflect the linguistic patterns inherent to the datasets, potentially emphasising the limitations of these approaches in achieving generalisability.

For the ISOT dataset, the real class is dominated by terms associated with reputable news organisations and standard journalistic language, such as “Reuters”, “Washington”, and “Wednesday”. Conversely, the fake class in ISOT shows frequent use of named entities such as “Trump”, “Obama”, “Hillary”, “GOP”, and “Americans”. These terms reflect a focus on politically charged figures and groups, which are often used in fake news articles to create sensational or emotionally engaging narratives. This contrast between the formal and institutional terms in the real class and the politically driven, attention-grabbing language in the fake class illustrates how the models differentiate the two based on token patterns.

The Kaggle Fake or Real dataset exhibits similar patterns, with the real class frequently incorporating institutional terms such as “president”, “state”, and “house”, aligning with trends observed in the ISOT dataset’s real class. Similarly, the fake class prominently features terms like “Hillary”, “Obama”, and “Russia”, reflecting a focus on named entities. This emphasis mirrors the ISOT dataset, where the fake

class highlights high-profile individuals and events, suggesting a bias toward politically charged content. While the real class focuses on broader institutional themes, the fake class leverages narratives centered on the recognition and emotional impact of well-known figures, potentially prioritising engagement or ideological messaging over balanced reporting.

The Kaggle Competition dataset presents a notable shift in term associations, with words like “2016”, “Hillary”, “Clinton”, “October” and “Obama” appearing prominently in the ‘real’ column, a reversal from their frequent presence in the fake columns of the ISOT and Kaggle Fake or Real datasets. This discrepancy highlights a fundamental issue in the construction of fake news detection datasets: the presence of topical biases, where certain terms or events may be inconsistently labelled based on their association with specific topics or individuals. Such inconsistencies can arise from differences in the dataset’s sources, curation methods, or labelling strategies, and they pose a significant challenge to model generalisability. Models trained on one dataset may incorrectly associate certain terms with fake or real news based on the biases or conventions of the training data, leading to poor performance when applied to other datasets. This inconsistency underscores the need for more standardised approaches to dataset creation, ensuring that labels reflect the intrinsic qualities of fake or real news rather than dataset-specific idiosyncrasies or biases linked to particular terms or events.

These patterns underscore the influence of topical biases on classification outcomes, which is further exacerbated by the inclusion of source-specific terms introduced during the data collection process. Terms such as “Reuters” in the ISOT dataset or “Breitbart” in the Kaggle Competition dataset exemplify how dataset-specific provenance influences the models’ decision-making. These terms, tied directly to the source of the articles, can lead to over-reliance on patterns unique to individual datasets rather than features that capture the intrinsic characteristics of fake or real news. While it could be argued that removing such terms during pre-processing might mitigate this bias, the systematic review conducted in Chapter 3 revealed that most studies utilizing these datasets do not typically exclude source-specific words during pre-processing. This omission reflects a common reliance on these terms as implicit indicators of classification, despite the risk they pose in reinforcing dataset-specific biases. By failing to address this issue, these studies may inadvertently hinder the models’ ability to generalize across datasets, limiting their effectiveness in detecting fake news in diverse and real-world contexts.

5.6 Discussion

This study was motivated by the systematic review conducted in Chapter 3, which highlighted that current fake news detection models struggle to generalise effectively. Despite advancements in the field, much of the existing research focuses on model performance within a single dataset or across different domains, often overlooking the more fundamental question of intra-domain generalisability, that is, how well these models perform on different datasets within the same domain. Addressing this issue is crucial for real-world applications, where models need to perform consistently across similar datasets before they can effectively progress to being trained in broader contexts. This study therefore aimed to explore this gap by focusing on the intra-domain generalisability of models trained on political news from the same time period.

In the first experiment of this study, a set of models was developed and tested using Stratified K-Fold Cross-Validation in order to provide a baseline for comparison in subsequent analyses. It was found that the models trained and tested with the same dataset produced high performances, which were comparable to, and replicated results reported in the literature. While these results are encouraging and frequently highlighted in prior studies, as discussed in Chapter 3, such experiments assume that the underlying data used in training and testing is representative and consistent. However, this assumption may not hold true in practice, as different datasets can vary significantly in terms of language, context, and the specific characteristics of the news articles.

To investigate this issue, the second experiment (presented in Section 5.5.2), tested these models with the remaining datasets not used in training to determine how well they performed when tested on different datasets of the same topic and time period. The core finding of this analysis provided additional evidence that models do not generalise well, even on a more fundamental test on data within the same topic and time period. This raises questions around the efficacy of current techniques in the real-world, given that models must be able to perform outside the datasets on which they are trained. Additionally, this adds to the body of evidence that suggests that current publicly available datasets are not suitable to train generalisable models. It is possible this is down to two factors. First, many existing datasets are simply too small (as evidenced in the systematic review in Chapter 3) to generate models that can generalise effectively across different contexts. The limited variety in language and context within these datasets can lead to overfitting, where models learn to identify patterns specific to the training data rather than developing the flexibility needed to adapt to new, unseen datasets. Second, coarsely labelling datasets can significantly limit a model's ability to generalise effectively. When datasets are

labelled according to broad protocols, such as labelling based solely on an article’s publisher, they fail to capture the nuanced distinctions that exist within the content itself. For example, articles from a reputable publisher might still contain misleading information, while those from lesser-known sources could provide accurate reporting. This simplistic approach to labelling overlooks the complexities of language, style, and context that are critical for accurate classification.

Evidence for this is provided in the LIME analysis conducted in Section 5.5.3, which revealed that certain words and phrases that were heavily weighted in determining whether an article was classified as fake or real were often tied to specific datasets rather than the inherent qualities of the articles themselves. This analysis showed that models tended to over-rely on specific terms associated with the training data, leading to biased predictions when faced with new articles from other datasets. For instance, in the Kaggle (Fake News) dataset, words that were associated with the ‘real news’ class were associated with the ‘fake’ news class in the Kaggle Fake or Real dataset. This overlap illustrates the dangers of using coarsely labelled datasets, as it can result in models that fail to discern the true nature of the articles, instead relying on potentially misleading keywords that do not accurately reflect the content’s reliability or credibility. Such dependencies on dataset-specific terminology hinder the models’ ability to generalise effectively across different datasets, emphasising the need for more sophisticated data collection strategies, labelling practices and more robust testing protocols.

This analysis also highlighted a weakness of token-representations (BoW, TFIDF, Word2Vec and BERT). These methods often rely on the frequency and context of specific words in the training data without fully considering their semantic significance in different contexts. While these methods are designed to capture linguistic patterns, they can become overly dependent on terms that may not be universally indicative of fake or real news. Consequently, when presented with new articles, models may misclassify content based on the presence or absence of certain keywords rather than evaluating the overall context and meaning of the text. As such, it could be argued that stylistic features may be more robust in terms of generalisability. Unlike token-based methods, which focus primarily on word usage, stylistic features consider the broader characteristics of the text. While such features also failed to generalise effectively in this study, these features demonstrated more consistent performance in comparison to token-representations suggesting such features are less sensitive to biases within datasets. This is supported by Castelo et al. (2019) which found positive results for generalisability over-time and across domains, using linguistic features. Further evidence supporting the argument that stylistic features can perform better in terms of generalisability can be found in Gautam and Jeripothula (2020); Janicka et al. (2019a).

Given that the use of coarsely labelled datasets is likely to continue to be necessary in order to create sufficiently large datasets—due to the substantial effort required to label such datasets manually—extracting value from these datasets using robust features becomes increasingly important. As only a relatively small number of stylistic features have been investigated in this study, further exploration into these features is essential for enhancing the generalisability of fake news detection models. For example, more selective, finer-grained experiments could pinpoint which of the 34 stylistic features chosen, or combinations of stylistic features, are the most essential and effective in this classification problem. Moreover, it could be argued that, given the right combination of stylistic features and additional novel features, good generalisability may be achievable. Examples of novel features include frequency of URL redirections (Chen and Freire, 2021), volume of advertising (as profit for advertising is often a motivation for producing fake news (Allcott and Gentzkow, 2017) and reverse image search to determine if images have been manipulated or used out of context (Saez-Trumper, 2014). As such, exploring such combinations of features with the view of improving generalisability should be the primary focus of future research.

5.7 Chapter Summary

This chapter focused on the investigation of intra-domain generalisability in fake news detection models, motivated by the findings from the systematic review in Chapter 3. It highlighted the limitations of current research, which predominantly emphasises model performance on single datasets or across different domains, often neglecting the more fundamental question of how well models can generalise within the same domain.

The first experiment involved developing and testing a set of models using Stratified K-Fold Cross-Validation. While the results showed high performance within the same dataset, the chapter emphasised that this success does not guarantee effective generalisation to unseen datasets. To address this concern, the second experiment tested the models against datasets not used in training, revealing that the models struggled to generalise even within the same topic and time period. This finding raises questions about the efficacy of current techniques in real-world applications, where models must operate across diverse datasets.

Two primary factors contributing to the lack of generalisability were identified: the limited size of existing datasets, which can lead to overfitting, and the use of coarsely labelled datasets that fail to capture the nuanced distinctions in content. The chapter also underscored the weaknesses of traditional token-representation methods, such as Bag-of-Words, TF-IDF, Word2Vec, and BERT, which tend to rely

heavily on specific word patterns and therefore exacerbate biases within datasets. In contrast, the chapter proposed that stylistic features might offer a more robust approach to generalisability, as they consider broader text characteristics beyond mere word usage. The exploration of such features could provide valuable insights, especially given the ongoing reliance on coarsely labelled datasets for training. Overall, the chapter concluded by identifying key areas for future research, including the need for more rigorous testing methodologies, the exploration of stylistic and novel features, and the development of larger, more representative datasets.

Chapter 6

Study 2: Exploring Features for Generalisable Fake News Detection

6.1 Introduction

Chapter 3’s literature review highlighted the ongoing challenge of generalisability in fake news detection, particularly concerning cross-domain generalisability. Chapter 5 narrowed this focus to the more fundamental task of intra-domain generalisability, examining whether models trained within a single domain can generalise effectively across related datasets. This investigation, centred on political news from the 2016 U.S. Presidential Election, revealed that token-based features, such as Bag-of-Words and TF-IDF, as well as advanced methods like Word2Vec and BERT, often struggle to generalise within the same domain due to dataset biases, including topical and narrative skews.

To address these limitations, stylistic features—capturing elements like sentence structure and punctuation patterns—emerged as potentially more robust to these biases, offering an alternative to token-based methods for enhancing model resilience. This second study, therefore, employs the Facebook URLs dataset, a manually labelled set curated by a third-party fact-checking organisation, for external validation. Testing models trained on coarsely labelled data against this high-quality dataset aims to evaluate whether models can generalise to real-world content. By investigating the potential of stylistic features, this study seeks to advance fake news detection models that are both adaptable and reliable across varied datasets, reducing the impact of inherent topical biases. Additionally, this chapter aims to introduce novel features, characterized as “social-monetisation” features, which capture the economic incentives driving the creation and spread of fake news, with the

goal of contributing to more generalisable models.

This chapter begins by outlining the motivation (Section 6.2), which discusses the limitations of current fake news detection models and the potential of stylistic and social-monetisation features to improve generalisability. Section 6.3 presents the specific thesis research questions addressed by the investigation, focusing on how alternative features might enhance model resilience. Section 6.4 details the experimental setup, including datasets, feature sets (including the proposed social-monetisation features), and machine learning algorithms, as well as the introduction of the Facebook URLs dataset for external validation. Section 6.5 provides an analysis of model performance across feature sets, highlighting their effectiveness on real-world data. Finally, Section 6.6 interprets the findings, considering their implications for model deployment and suggesting directions for future research in developing more robust fake news detection models.

6.2 Motivation

Study 1 in Chapter 5 identified two key challenges in fake news detection: the pronounced topical biases present in coarsely labelled datasets and the sensitivity of token-based features, such as Bag-of-Words and TF-IDF, to these biases. Coarsely labelled datasets often rely on simplified labelling methods, such as using publisher reputation as a proxy for accuracy, which introduces bias toward specific topics or narratives. This bias impacts model training, leading to overfitting to topic-specific patterns that do not generalise well to diverse content. While token-based features effectively capture dataset-specific word patterns, they are particularly vulnerable to such biases, resulting in models that perform well on training data but struggle in varied, real-world contexts. To evaluate the generalisability of models trained on biased datasets, the study introduces a manually labelled external dataset, the Facebook URLs dataset, as a validation tool. This curated, real-world dataset offers a clearer perspective on model robustness and adaptability beyond the confines of the training data.

With the introduction of this curated dataset, Chapter 6 adopts a dual approach. First, it re-evaluates token-based representations on the Facebook URLs dataset by repeating experiments with features like Bag-of-Words, TF-IDF, Word2Vec, and BERT. This aims to determine whether the limitations identified in Chapter 5 persist when applied to more granularly labelled, real-world data. Using this high-quality validation set, the study examines whether the challenges in model generalisability stem from the inherent limitations of token-based representations or the quality of the datasets used for training and testing.

The second part of the dual approach focuses on stylistic features, which Chapter

5 highlighted as potentially more robust indicators for fake news detection. Unlike token-based methods, stylistic features are less dependent on topic-specific vocabulary, making them more consistent across datasets and less influenced by topical biases. Chapter 6 explores stylistic features in greater depth, evaluating their capacity to improve model generalisability and adaptability to diverse content. By comparing several groups of stylistic and token-based features, the study investigates whether stylistic markers can serve as reliable indicators for fake news detection across varied contexts.

The systematic review in Chapter 3 emphasised the importance of incorporating features beyond text to address challenges in generalisability. Chapter 2 provided key background by highlighting the economic motivations behind fake news production, such as profit generation through advertising. Chapter 5 reinforced this perspective with LIME analysis, which revealed that terms like “twitter” and “share” significantly influenced fake news classification. Building on these insights, Chapter 6 introduces a novel set of social-monetisation features, including indicators like advertisements, social media share buttons, and affiliate links. These features capture the economic incentives often driving disinformation and extend feature engineering to include contextual signals beyond text patterns. By integrating these features, the study enhances model robustness, offering valuable cues for fake news detection, particularly in scenarios where text-based indicators may be insufficient.

6.3 Research Questions Addressed

Table 6.1: Study 2 - Thesis Research Questions Addressed

RQ	Description
RQ1	What are the current methods to detect fake news?
RQ2	How effective are current methods to detect fake news?
RQ3	To what extent do existing fake news detection methods generalise across datasets?
RQ4	What current features contribute to more generalisable models in the context of fake news detection?
RQ5	How can novel features that extend beyond the text—such as social dissemination behaviours and economic incentives—enhance the generalisability of fake news detection models?

This section highlights the thesis research questions addressed in this study, with particular emphasis on RQ3, RQ4, and RQ5. RQ3 examines the extent to which

existing fake news detection methods generalise across datasets, addressing a critical challenge in achieving broader model applicability. While this RQ was previously addressed in Chapter 5, owing to the introduction of the manually labelled Facebook URLs dataset, it was determined to repeat this experiment, as outlined in Section 6.2. RQ4 investigates which existing features most effectively contribute to the development of generalisable fake news detection models. Building on this, RQ5 explores the potential of novel features—such as social dissemination behaviours and economic incentives—that extend beyond textual content to enhance model robustness. Initially introduced in Section 1.4, these questions align with the study’s overarching aim of advancing generalisability in fake news detection by identifying effective feature sets and expanding beyond traditional token-based approaches.

6.4 Method

This section details the methodology used to evaluate the generalisability of different groups of stylistic features, along with outlining the proposed novel features. Similar to the previous chapter, these experiments build upon the text-classification process presented in Chapter 3. Two experiments were conducted (summarised in Figure 6.1 and detailed below) to address the relevant research questions outlined in Section 6.3 and assess the effectiveness of these features. Section 6.4.1 outlines the data collection process and resulting datasets that were used in the experiments. Section 6.4.2 to 6.4.3 describe the features that were extracted from these datasets in relation to Experiment 1 and Experiment 2 respectively. Sections 6.4.4 to 6.4.5 outlines the machine learning algorithms that were used in these two experiments and how they were trained and tested.

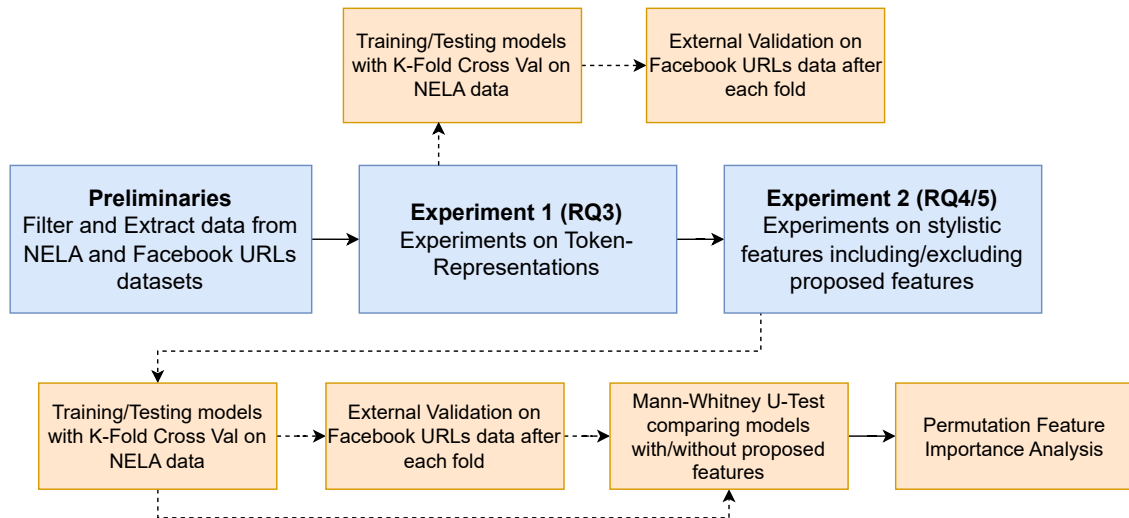


Figure 6.1: Study 2 - Overview

Experiment 1: Baseline Evaluation of Token-Based Features

The first experiment, similar to the experiment conducted in the previous chapter, focused on evaluating the generalisability of token-based features such as Bag of Words (BoW), TF-IDF, Word2Vec, and BERT, addressing RQ3. Models were initially trained on the NELA 2020-21 dataset, a comprehensive dataset used for fake news detection, and their performance was evaluated using K-fold cross-validation. Following this, the models were externally validated using the Facebook URLs dataset, which was manually labelled by a third-party fact-checking organisation, distinct from the more coarsely labelled datasets commonly found in the literature. This external validation established a baseline for comparison with the feature sets explored in Experiment 2.

Experiment 2: Evaluating Stylistic and Social-Monetisation Features

The second experiment examined the generalisability of five groups of stylistic features and the impact of the newly proposed social-monetisation features, addressing RQ4 and RQ5. As in the first experiment, models were trained on the NELA 2020-21 dataset and evaluated with K-fold cross-validation. Subsequently, the models were externally validated on the Facebook URLs dataset to test their generalisability. This experiment was performed twice: first using only the stylistic features and then with the inclusion of social-monetisation features, such as the frequency of ads, external links, and social media share buttons. The aim was to determine if these newly proposed features led to a statistically significant improvement in generalisability performance. The Mann-Whitney U-Test was used to assess the significance of the performance difference when the social-monetisation features were included. The following subsections elaborate on the methodology.

6.4.1 Datasets and Data Processing

This section outlines the datasets and data extraction methods used. Owing to the nature of the proposed social-monetisation features, the dataset required the source URL of the articles to facilitate the extraction of these features. The systematic review conducted as part of Chapter 3 identified several datasets used in content-based fake news detection however only three—FakeNewsNet, Buzzfeed, and Celebrity fake news—include the article’s source URL. These datasets are relatively small, which limits the likelihood of producing a generalisable model. To develop a more comprehensive and reliable model, a larger dataset is necessary. Therefore, the NELA series of datasets was chosen for its large size and inclusion of article URLs, providing a more extensive and diverse data source for training. Using a dataset of this size also ensures that a significant number of articles can be extracted to

compensate for pages that are no longer available. While not as frequently used in the literature, a number of studies make use of this dataset including Horne et al. (2019); Raj et al. (2023) and Raza and Ding (2022).

The latest iterations (at the time this experiment was conducted) of this dataset released in March 2023, NELA 2020 and 2021, were chosen for this study. Each dataset contains over 1 million articles from various sources and are coarsely labelled, with each article’s legitimacy derived from its source’s aggregated label from seven assessment sites: Media Bias Fact Check, Pew Research Center, Wikipedia, OpenSources, AllSides, BuzzFeed News, and Politifact. The labels are categorised as unreliable, mixed, and reliable. For this study, only ‘unreliable’ and ‘reliable’ labels were used, excluding the ‘mixed’ label to align with the binary labels in the external validation dataset.

The combined NELA 2020-21 dataset includes 3,635,636 records from 525 unique sources. After joining the labels file and excluding the ‘mixed’ category, the dataset consists of 1,013,808 ‘true’ and 551,051 ‘fake’ articles from 224 sources. To prevent any single source from dominating the training set (Figure 6.2), the number of URLs extracted from each source was reduced using the 1st quartile as a threshold (285 articles per source), resulting in a final set of 22,230 ‘true’ and 25,650 ‘fake’ articles from 168 sources (Figure 6.3).

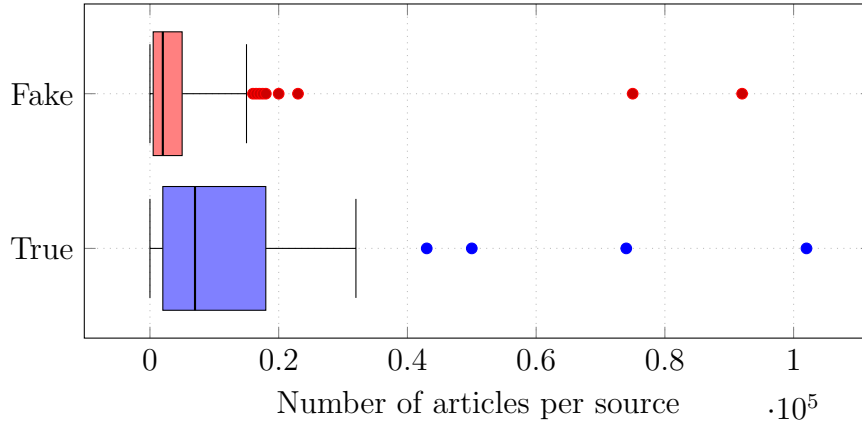


Figure 6.2: Articles per Source Prior to Extraction

The Facebook URLs Dataset was chosen as an external validation dataset owing to its unique position as a dataset collected in a ‘real-world’ context and granular labelling by a third-party fact-checking organisation. Its individual article labels provide a robust standard for assessing model accuracy and practical applicability in fake news detection. This stands in contrast to commonly used datasets in the field, which often employ coarse labels based on article publishers, potentially misrepresenting the true nature of fake news. This choice addresses an issue raised in the previous chapter, which highlighted the limitations of coarsely labelled datasets in misrepresenting the true nature of fake news, underscoring the importance of

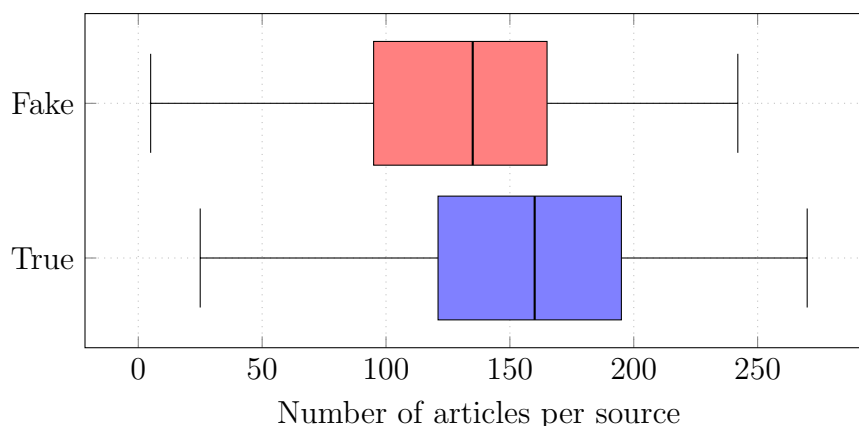


Figure 6.3: Articles per Source Post-Extraction

more precise labelling for developing generalisable detection models. By using a coarsely labelled dataset for training and a manually labelled dataset for testing, the aim is to demonstrate that despite the limitations of coarsely labelled datasets, meaningful features can still be extracted to develop robust models applicable in real-world scenarios.

The Facebook URLs dataset contains over 38 million URLs shared on Facebook since January 1, 2017, with 35,924 records identified as fake news. The dataset is protected with differential privacy, ensuring no information can be gathered regarding individuals (Messing et al., 2020). Given its restricted accessibility and limited usage in prior studies, this research represents one of the few to utilise the Facebook URLs Dataset for fake news classification, following a study by Barnabò et al. (2023). The dataset initially comprised 28,271 fake and 7,653 true records, with non-English articles filtered out based on 'US' and 'UK' values in the 'Public Shares Top Country' field, resulting in 14,354 fake and 1,468 true records. To enhance dataset quality, URLs referring to Tweets and videos were excluded. Class balancing was implemented during experimentation. Due to its size, the Facebook URLs Dataset served as a test set for external validation, complementing the larger training datasets to bolster the model's generalisability and validate its performance in diverse real-world scenarios.

In order to extract the raw textual data from the URLs in these datasets, the BeautifulSoup library was used. As many webpages in these datasets may no longer be available, particularly in relation to 'fake' news pages, initial extraction was attempted through the use of the Wayback Machine API (Internet Archive). This was done to increase the likelihood of extracting a webpage with a complete article and not a splash page indicating the article had since been deleted. In instances where webpages were not available in this archive, a final extraction attempt was made directly from the webpage using the URL provided in the dataset to account for cases where webpages may not yet have been added to the Internet Archive.

If through these methods a complete article was not extracted, the URL would be excluded from the resulting dataset.

In cases where full articles were available, rather than attempt to accurately extract only the text pertaining to the news articles from these URLs, all textual elements are extracted from the body of the webpage. While this may introduce additional noise to the feature-sets, it was a deliberate choice. Websites have different layouts, styles and coding structures, making it challenging to consistently and accurately extract only the article text. It is argued that models that extract all textual elements from the webpage body are more adaptable to the varying structures and formats of webpages and, as such, have the potential to be more robust and scalable across a wider range of online content. Following this data extraction phase, pages returning <3KB of data were excluded, as it was observed that pages with less than this amount of data had typically had their articles removed. The resulting datasets are summarised in Table 6.2:

Table 6.2: Dataset Summary

Category	NELA 2020-21 (Training Dataset)	Facebook URLs Dataset (External Validation)
Fake	10,529	5,355
True	10,487	798

6.4.2 Experiment 1 Features: Token-Representations

This section outlines the features used in the first experiment, which examines how well models utilising token-representations generalise between the NELA and Facebook datasets, as described in Section 6.4. An overview of the procedure for this experiment is provided in Figure 6.4. Each token-representation method and the corresponding libraries used to extract these features from the datasets are outlined below. Although this section mirrors the experiment conducted in the previous chapter, its inclusion is necessary for establishing a baseline to compare against the stylistic and proposed social-monetisation features in the subsequent analysis. The results of this experiment are presented in Section 6.5.1.

Like the previous experiment, this study uses the token-representations Bag of Words (BoW), TF-IDF, Word2Vec, and BERT to assess model generalisability. The same implementations of these techniques are used here to ensure consistency, with BoW and TF-IDF vectorizers implemented using scikit-learn, Word2Vec using a pre-trained model trained on a Google News corpus using the Gensim library (see

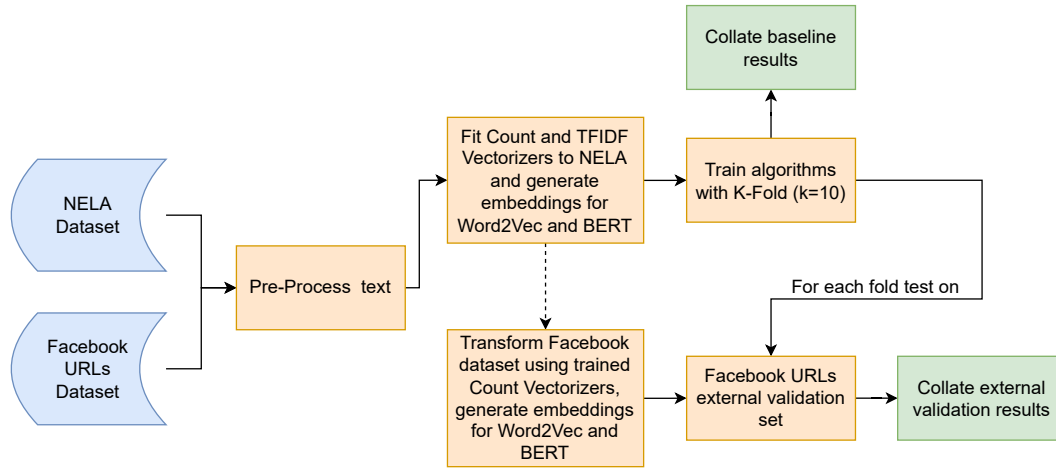


Figure 6.4: Experiment 1 - Flowchart

Appendix A.1), and BERT (base-uncased) with the Transformers library (see Appendix A.2). For BoW and TF-IDF, several preprocessing steps were employed to remove unwanted noise from the text: (i) converting the text to lowercase to ensure uniformity, (ii) lemmatising words to reduce them to their base forms, and (iii) removing punctuation, URLs, Twitter handles, extra whitespace, and stop words. These preprocessing steps were not applied to Word2Vec and BERT, as these methods require contextual information to generate accurate embeddings.

These representations establish a baseline for evaluating the effectiveness of the stylistic and social-monetisation features, as discussed in the next experiment.

6.4.3 Experiment 2 Features: Stylistic and Proposed Social-Monetisation Features

This section outlines the stylistic features used to evaluate how well they generalise compared to token-representations. Experiment 2 follows a structure similar to Experiment 1, assessing the generalisability of five groups of stylistic features identified in previous research and comparing the results with those from Experiment 1, which focused on token-representations. The experiment also examines whether the four social-monetisation features proposed in Section 6.2 improve generalisability. An overview of the procedure for Experiment 2 is shown in Figure 6.5, and the stylistic and social-monetisation features used are detailed in the following subsections. The results are presented in Section 6.5.2.

The study evaluates five groups of stylistic features proposed in the literature, each varying in complexity. The first group focuses on linguistic features, while the later groups incorporate additional dimensions, such as psycholinguistics and document complexity. To ensure consistent input for machine learning algorithms,

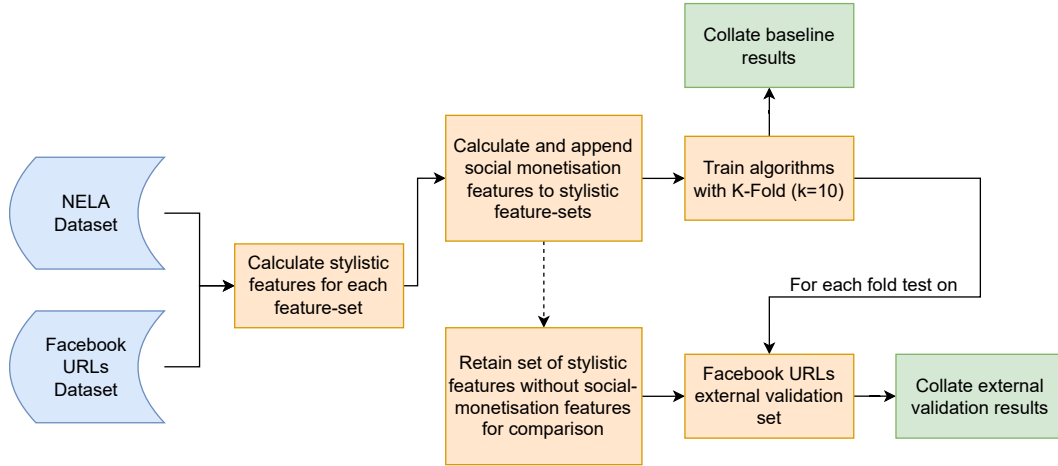


Figure 6.5: Experiment 2 - Flowchart

StandardScaler was applied to account for the differing scales of these features. The selection of these groups was driven by their inclusion in a small number of studies on generalisability in fake news detection models, with the exception of the NELA feature set, which was chosen because of its relevance to the NELA dataset used in this study. Unlike previous studies that used coarsely labelled datasets for external validation, this experiment evaluates the performance of these features using real-world data from the Facebook URLs dataset, highlighting their potential for practical fake news detection. A complete table of these features is provided in Appendix B.

Group 1: Fernandez and Devaraj Stylistic Features

Similar to token-representations, this study builds upon the previous chapter, which employed this set of stylistic features to explore the intra-domain generalisability of current approaches to fake news detection trained and tested on coarsely labelled datasets. With the introduction of the manually labelled Facebook URLs dataset in this study, it was determined that these features would be revisited to further examine their effectiveness.

These features include a collection of 34 linguistic attributes (referred to as ‘Linguistic Dimensions’ and ‘Punctuation Cues’) that demonstrated the highest efficacy in classifying fake news, as determined through a sequence of tests outlined by Fernandez and Devaraj (2019). The two groups of features can be summarised as follows:

- **Linguistic Dimensions:** Based on the Linguistic Dimensions of LIWC, this group aims to capture the complexity of news through inclusion of features such as word-per-sentence, average word size and type-token ratio (a measure

of lexical variety) as well as the different types of words used such as the ratio of adjectives, nouns, verbs and named-entities.

- **Punctuation Cues:** Focuses solely on the different types of punctuation used relative to all punctuation in a given article.

As linguistic features are a staple in NLP, the other groups of features described below also include similar groups of features. A comprehensive list of these features is provided in Table B.1

Group 2: Abonizio Features

This study leverages 21 features organised into three groups: complexity, stylometric and psychological. The inclusion of these features, similar to the previous group, is motivated by their use in another generalisability test on coarsely labelled datasets in Abonizio et al. (2020). Similar groups of features can also be found in Paschalides et al. (2019); Garg and Kumar Sharma (2022) and Reis et al. (2019) further justifying their inclusion in this study.

The ‘complexity’ and ‘stylometric’ features overlap with some of the features used in the previous ‘Fernandez’ feature-set however it should be noted that the Abonizio feature-set is not as granular. However, unlike the Fernandez feature-set, the Abonizio feature-set does extend to include a psychological group, capturing the sentiment analysis score of a given article. A list of these features is provided in Table B.2.

Group 3: Linguistic Inquiry Word Count (LIWC)

As previously mentioned in Section 2, LIWC is a dictionary-based approach comprising linguistic elements, punctuation characteristics, and psycholinguistic features organised into several groups. These groups can be summarised as follows:

- **Summary Variables:** These features aim to summarise the attributes from the groups listed below and attempt to capture document complexity as well as psychological characteristics.
- **Linguistic Dimensions:** These features capture different types of words, such as pronouns, verbs, and adjectives, as well as words indicating grammatical person (e.g., first-person and second-person) and numerical terms.
- **Psychological Processes:** This group includes words related to psychological aspects such as sentiment (e.g., “good” and “bad”), cognition (e.g., “know” and “think”), and social processes (e.g., “love” and “fight”).

- **Expanded Dictionary:** This group includes words related to a variety of topics such as culture, lifestyle, and temporal concepts (e.g., “when,” “now,” and “then”).

The total number of features in this feature-set amount to 118. Similar to the previous feature groups, LIWC is used in a generalisability study by Pérez-Rosas et al. (2017) which observed the performance of LIWC trained on the FakeNewsAMT dataset and tested Celebrity news datasets and vice versa. A number of other studies also leverage these features, thus further justifying their inclusion (Ahmad et al., 2020; Spezzano et al., 2021; Shu et al., 2019a). An exhaustive list of these features is provided in Table B.3.

Group 4: NELA Feature Extractor

The NELA feature extractor is a tool hosted on GitHub, designed by the authors of the NELA dataset, which has been used throughout this study. This motivated the inclusion of these features in this research. It includes a rich, hand-crafted feature set of 91 features, which can be summarised into the following groups:

- **Style:** Similar to features from the previous studies, this group focuses primarily on part-of-speech (POS) tags.
- **Complexity:** Similar to the *Linguistic Dimensions* and *Complexity* group of the Abonizio feature set, this group captures the complexity of an article by analysing lexical diversity, readability metrics, and the average length of words and sentences.
- **Bias:** Based on the work of (Recasans, 2013), this group identifies subjective text elements by counting the presence of hedges, factives, assertives, implicatives, and opinion words.
- **Affect:** Using VADER sentiment analysis, this group captures the emotional tone and sentiment present in the text.
- **Moral:** This group encompasses the ethical content of a text, leveraging the principles of Moral Foundation Theory (MFT) introduced by Graham et al. (2013). It employs the lexicon developed by Lin et al. (2018) to assess the morality of a text based on categories like care, fairness, loyalty, authority, and sanctity.
- **Event:** This group focuses on identifying words related to dates, times, and locations.

A full list of these features is provided in Table B.4.

Group 5: Modified NELA Features

Through the use of the NELA Feature Extractor, it was observed that several features were either duplicated or returned zero values, particularly when extracting punctuation. To address these issues, the NELA Feature Extractor was modified to resolve such inconsistencies and to include additional punctuation marks, such as #, @, £, \$, &, and %.

Additionally, the normalisation process was adjusted depending on the features. For example, instead of scaling punctuation features by the word count of an article, punctuation was normalised based on the total number of punctuation marks. This adjustment ensured a more accurate representation of punctuation use within the text.

Proposed ‘Social-Monetisation’ Features

As motivated in Section 6.2, a number of additional novel features were explored to improve the generalisability of fake news detection models. These features include the frequency of advertisements, affiliate links and social media sharing links. These features and their justifications for inclusion are outlined below:

- **Frequency of Ads:** One of the primary motivations behind the creation and dissemination of fake news is financial gain through advertising. According to Allcott and Gentzkow (2017), fake news websites often rely on sensationalist and misleading content to attract high volumes of traffic, which in turn increases their advertising revenue. These sites typically feature a large number of advertisements, as their business model is heavily reliant on generating ad impressions and clicks. Therefore, the number of adverts associated with a given article could be a significant indicator of fake news. Articles that contain an unusually high number of ads may be designed to maximise revenue rather than to provide factual information, making this a critical feature to include in fake news detection models.
- **External Links:** Similar to advertising, the prevalence of external links in an article can also be an indicator of fake news, especially when these links are intended for affiliate marketing purposes. Fake news articles often include numerous external links that direct readers to other sites, which can generate affiliate income for the publisher each time a link is clicked. This tactic is particularly common in disinformation related to healthcare and other high-interest topics, as noted by Rehman et al. (2022).
- **Social Media Share Links:** The role of social media in the spread of fake news is well-established, with platforms like Facebook and X/Twitter being

primary channels for disinformation dissemination. One of the mechanisms that facilitate this spread is the use of visual cues, such as share buttons, which prompt habitual behaviour in social media users (Ceylan et al., 2023). When users encounter these visual cues, they are more likely to share the content without critically evaluating its veracity. Including 'call to action' links that lead to social media platforms in the analysis is essential, as these links can significantly amplify the reach of fake news articles. By encouraging readers to share content on social media, these articles can quickly go viral, spreading disinformation at an unprecedented rate. Therefore, factoring in Facebook and X/Twitter links is expected to be important in identifying articles that are designed to exploit social media behaviour for rapid dissemination. It is important to note these social media features are distinct from others seen in the literature, which typically focus on user profiles and relationships between tweets and users.

The *number of ads* was extracted through the use of EasyList, an open-source project that compiles a list of the most popular adblocking filters. Using this list enables searching the webpage's LXML tree and counting the frequency of various ads. *External links* were identified through a combination of extracting 'hrefs' in the webpages and comparing their domains to the host domain using the 'tldextract' library. Links whose domains did not match the host domain were used to calculate the frequency of external links. Links that pointed to *Facebook* and *Twitter/X* were each counted separately.

By leveraging these features, the study aims to provide a more nuanced understanding of the mechanisms driving the dissemination of fake news and improve the models' ability to generalise across different datasets. Ultimately, this exploration seeks to contribute to the development of more robust and effective fake news detection systems that can better adapt to the complexities of real-world information environments.

6.4.4 Machine Learning Algorithms

As this study prioritises the exploration of stylistic features for generalisable fake news detection, less emphasis has been put on exploring the effect of different machine learning algorithms and their respective hyperparameters. However, for completeness and to offer an opportunity for comparison to the literature, a number of machine learning algorithms including Logistic Regression, SVM, Gradient Boosting, Decision Trees, Random Forest, a feed-forward neural network (FFNN) and LSTM are employed. Similar to the previous study in Chapter 5, each of these algorithms was implemented using default hyperparameters in SKLearn. The exception

to this was the neural network, where default hyperparameters are not available. As such, a shallow Sequential model was used with a single hidden layer of 10 neurons, a sigmoid activation function, binary cross-entropy loss, and the Adam optimizer in PyTorch. To protect against overfitting, the EarlyStopping technique was implemented manually by monitoring the validation loss and stopping training if the loss did not improve by 0.01.

Similar to the previous Chapter, the LSTM was also implemented using PyTorch and applied exclusively to Word2Vec and BERT embeddings to leverage their sequential and contextual information, while Bag-of-Words, TF-IDF and the groups of stylistic features were excluded due to their lack of sequential structure. The model consisted of an LSTM layer with 128 hidden units, a 40% dropout rate to prevent overfitting, and a fully connected layer for classifying fake and real news.

Unlike the previous chapter, Naïve Bayes was excluded from this analysis. This decision was made owing to its inconsistent performance on the stylistic feature set in Chapter 5. The algorithm's reliance on strong independence assumptions between features proved unsuitable for the nuanced and interdependent stylistic features explored in this study. As a result, the focus was placed on algorithms better suited to capturing complex relationships within the data, ensuring a more robust evaluation of the proposed feature sets.

6.4.5 Evaluation

The primary training and testing methodology in this study mirrors that of the previous chapter, integrating K-fold cross-validation and external validation to assess model performance and generalisability in fake news detection. Here, K was set to 10, enabling a comprehensive evaluation across multiple data folds. Unlike the previous study, which treated external validation as a separate experiment by training on the full training dataset and testing on distinct datasets, this study refines the approach by conducting external validation on a random sample of 500 articles per class from the Facebook URLs dataset following each training fold on the NELA dataset. This adjustment offers a more approach to evaluating model generalisability by integrating external validation directly into the K-fold cross-validation process. By using a random sample of 500 articles per class from the Facebook URLs dataset for external validation after each training fold on the NELA dataset, this methodology provides an iterative assessment of the models' performance on new data after every training phase, rather than only at the end.

Evaluation metrics such as Accuracy, Precision, Recall, Specificity and F1-Score were employed to assess the models' ability to distinguish between 'true news' and 'fake news.' Alongside these metrics, the Mann-Whitney U-test was applied to

evaluate the impact of social-monetisation features on model performance across the folds, testing the hypothesis that these features enhance accuracy in real-world conditions.

In addition to the Mann-Whitney U-test, Permutation Feature Importance (PFI) was used to identify the stylistic and novel social-monetisation features that contributed most to improving model generalisability. PFI works by randomly shuffling a feature to break its relationship with the target variable and then measuring the change in model performance. By applying this method to all features, PFI helps reveal the significance of each feature in the overall model, pinpointing which elements had the strongest positive impact on performance.

6.5 Results

This section outlines the results of the two experiments outlined in Section 6.4. The overarching objective of the experiments is to demonstrate whether models using different sets of stylistic and the proposed social-monetisation features are able to detect ‘real-world’ fake news (achieved by using the Facebook URLs dataset for evaluation) and comparing them to state-of-the-art approaches relying on token-representations.

6.5.1 Experiment 1: Generalisability of Token-Representations

Experiment 1 aimed to address this by examining how well token-representations (BoW, Word2Vec, BERT, and TF-IDF, as detailed in Section 6.4.2), combined with different machine learning models, generalise. The experiment evaluated the performance of these representations across various models to determine their effectiveness in generalising between a coarsely labelled datasets (NELA) and a manually labelled dataset (Facebook URLs dataset). This experiment establishes a baseline for comparison with the stylistic and social-monetisation features explored in subsequent experiments.

Table 6.3 presents the performance metrics of various models using token-based representations (BoW, TF-IDF, Word2Vec, and BERT) for fake news detection on the NELA dataset. Notably, the models tested with BoW and TF-IDF consistently outperform those using Word2Vec and BERT (with the exception of the BERT-trained LSTM), showing that simpler token-based approaches can be more effective in these specific conditions. BoW and TF-IDF models achieved high levels of accuracy, precision, recall, specificity, and F1 scores, with the best models reaching nearly perfect performance (0.98-0.99 accuracy, 0.97-100 recall, and similar scores

Table 6.3: Token-Representations Baseline Results

Features	Model	Acc.	Prec.	Rec.	Spec.	F1
BoW	Logistic Regression	0.98	0.97	0.99	0.97	0.98
	Decision Tree	0.96	0.95	0.96	0.95	0.96
	SVM	0.91	0.85	0.99	0.84	0.92
	Gradient Boosting	0.97	0.96	0.99	0.96	0.97
	Random Forest	0.99	0.97	1.00	0.97	0.99
	Neural Network	0.99	0.99	0.98	0.98	0.99
TF-IDF	Logistic Regression	0.98	0.96	1.00	0.96	0.98
	Decision Tree	0.95	0.95	0.96	0.95	0.95
	SVM	0.97	0.94	1.00	0.94	0.97
	Gradient Boosting	0.98	0.96	0.99	0.97	0.98
	Random Forest	0.99	0.97	1.00	0.97	0.99
	Neural Network	0.99	0.98	1.00	0.97	0.99
Word2Vec	Logistic Regression	0.89	0.86	0.93	0.86	0.90
	Decision Tree	0.87	0.85	0.88	0.88	0.86
	SVM	0.91	0.86	0.96	0.85	0.91
	Gradient Boosting	0.95	0.94	0.97	0.94	0.95
	Random Forest	0.95	0.92	0.98	0.92	0.95
	Neural Network	0.88	0.91	0.86	0.83	0.88
	LSTM	0.88	0.91	0.84	0.91	0.87
BERT	Logistic Regression	0.88	0.86	0.90	0.86	0.88
	Decision Tree	0.81	0.82	0.80	0.82	0.81
	SVM	0.90	0.89	0.92	0.89	0.90
	Gradient Boosting	0.84	0.83	0.85	0.83	0.84
	Random Forest	0.84	0.82	0.86	0.82	0.84
	Neural Network	0.85	0.89	0.81	0.83	0.84
	LSTM	0.99	0.99	0.99	0.99	0.99

for other metrics). This is particularly evident with models such as Random Forest, Gradient Boosting, and Neural Networks, where TF-IDF and Count-based features achieve accuracy scores up to 0.99, and F1 scores close to 0.99.

In contrast, Word2Vec and BERT models generally underperform relative to the BoW and TF-IDF representations. Word2Vec-based models see accuracy scores between 0.87 and 0.95, with a slightly more balanced performance in terms of recall and precision, though Neural Network performance was relatively lower (0.88 accuracy, 0.86 recall). BERT, as a more context-aware model, showed slightly higher precision for certain algorithms, such as Logistic Regression (0.86) and SVM (0.89), yet its performance remained more variable overall with accuracy scores generally ranging between 0.81 and 0.99. However, the highest-performing model, the LSTM trained on BERT embeddings, achieved a near-perfect accuracy and F1 score of 0.99, reflecting its capacity to leverage contextual information effectively when appropriately configured.

This discrepancy in performance could be attributed to the feature extraction process. Since BoW and TF-IDF representations capture the occurrence and relative importance of specific terms in a straightforward, frequency-based manner, they may be less susceptible to noise than embedding-based methods like Word2Vec and BERT. Given that the entirety of the text from each webpage was used to build

these models, BoW and TF-IDF’s exclusionary nature may have helped filter out non-informative terms, which are otherwise considered in the contextual representations of Word2Vec and BERT. Contextual embeddings are generally more suited for capturing nuanced meanings in sentences or paragraphs but may also introduce additional noise when applied to large, heterogeneous text, as with news webpages. The embeddings may thus capture unrelated context or topic shifts within articles, contributing to slightly poorer results. Furthermore, embeddings such as BERT may generally be better suited to more complex algorithms like LSTMs, which are specifically designed to leverage the sequential and contextual information captured by these embeddings. Therefore, performance using embeddings such as BERT may be hampered by the limitations of simpler algorithms, which may not fully exploit the rich contextual and sequential information embedded in these representations.

Table 6.4: Token-Representations Cross-Dataset Results

Features	Model	Acc.	Prec.	Rec.	Spec.	F1
BoW	Logistic Regression	0.66	0.68	0.60	0.72	0.64
	Decision Tree	0.64	0.66	0.57	0.71	0.61
	SVM	0.61	0.58	0.78	0.44	0.67
	Gradient Boosting	0.63	0.60	0.76	0.50	0.67
	Random Forest	0.61	0.57	0.88	0.34	0.69
	Neural Network	0.68	0.72	0.58	0.78	0.65
TF-IDF	Logistic Regression	0.68	0.68	0.71	0.66	0.69
	Decision Tree	0.64	0.67	0.58	0.71	0.62
	SVM	0.68	0.66	0.74	0.62	0.70
	Gradient Boosting	0.64	0.62	0.76	0.53	0.68
	Random Forest	0.64	0.60	0.86	0.42	0.70
	Neural Network	0.70	0.74	0.63	0.78	0.68
Word2Vec	Logistic Regression	0.67	0.70	0.57	0.76	0.63
	Decision Tree	0.60	0.62	0.56	0.66	0.59
	SVM	0.65	0.68	0.57	0.73	0.62
	Gradient Boosting	0.65	0.67	0.58	0.72	0.62
	Random Forest	0.65	0.66	0.62	0.67	0.64
	Neural Network	0.66	0.70	0.55	0.77	0.61
	LSTM	0.62	0.64	0.54	0.7	0.59
BERT	Logistic Regression	0.65	0.67	0.60	0.70	0.63
	Decision Tree	0.60	0.62	0.53	0.67	0.57
	SVM	0.66	0.69	0.59	0.74	0.64
	Gradient Boosting	0.63	0.64	0.61	0.66	0.62
	Random Forest	0.62	0.62	0.62	0.62	0.62
	Neural Network	0.66	0.69	0.56	0.75	0.62
	LSTM	0.68	0.75	0.55	0.81	0.63

Table 6.4 illustrates the impact of external validation on token-based models for fake news detection, revealing a notable drop in performance compared to initial K-fold results. On average, models experienced a 28% reduction in accuracy under these external validation conditions, highlighting the challenges of applying models trained on coarsely labelled datasets, such as the NELA dataset, to manually labelled datasets such as the Facebook URLs dataset. This outcome supports findings from the prior experiments in Chapter 5, emphasising the difficulty of achieving gen-

eralisability in fake news detection models when they are exposed to new datasets with different linguistic or stylistic properties.

While BoW and TF-IDF models demonstrated superior performance in K-fold testing, Word2Vec and BERT models showed a lesser decline in accuracy in external validation. This outcome suggests that while embedding-based representations like Word2Vec and BERT may capture more nuanced linguistic features, they still fall short of generalising well to new data. Nevertheless, the slightly improved resilience of these embeddings in external validation suggests they may be less dependent on dataset-specific patterns than BoW and TF-IDF.

Among the models tested, the highest external validation accuracy was achieved by the TF-IDF-based Neural Network, reaching 70% accuracy. However, this model showed poor recall, indicating it struggled to identify all instances of the positive (true news) class consistently. Similar patterns were observed in the other neural networks, which, despite strong performance in holdout testing, underperformed in external validation. This decline suggests possible overfitting to the NELA dataset, despite the use of early stopping, and raises questions about the practical applicability of neural networks trained solely on coarsely labelled data. The relatively high accuracy with low recall implies that these models may be biased toward correctly predicting the negative (real news) class, potentially overlooking fake news instances in new data.

In contrast, SVM and Logistic Regression models trained on TF-IDF features demonstrated a more balanced performance between recall and specificity, suggesting they may be more reliable in real-world applications where generalisability is critical. Although these models did not achieve the highest accuracy, their balanced metrics indicate that they could better distinguish between fake and real news across datasets. This balance between recall and specificity is crucial for fake news detection, as it suggests the models are less prone to source-specific biases and better equipped to generalise beyond the NELA dataset.

Notably, the LSTM model trained on BERT embeddings achieved 68% accuracy in external validation, with the highest precision (0.75) and specificity among all tested models (0.81). This performance highlights the model's ability to accurately identify true news instances while maintaining a strong capability to avoid false positives. However, the relatively low recall suggests that the model is potentially too conservative, favouring precision over a comprehensive identification of true news instances, which may result in many true news articles being overlooked. This may not be desirable, as accurately predicting true news may come at the expense of inadvertently censoring legitimate content, undermining the goal of ensuring that reliable information is widely accessible.

Overall, the results from this experiment underscore the limitations of token-

based representations when attempting to generalise across datasets. While the experiments in Chapter 5 highlighted the difficulties of generalising between coarsely labelled datasets, which often contain inherent biases, this experiment demonstrates the need for models that can effectively manage the linguistic diversity and complexity present in real-world, manually labelled fake news.

6.5.2 Experiment 2: Generalisability of Stylistic and Social-Monetisation Features

The second experiment aimed to determine whether the stylistic features suggested in the literature and the social-monetisation features introduced in this study are more generalisable than the token-level representations tested in Section 6.5.1. As detailed in Section 6.4.3, the following groups of stylistic features were evaluated: Fernandez; Abonizio; LIWC; NELA; and the modified NELA groups. Each of these groups was tested with and without the proposed social monetisation features identified in Section 6.4.3. A K-fold test was first performed with using the same splits used in the first experiment, using the NELA dataset to provide a baseline for comparison and the Facebook dataset to perform an external validation test for each model trained in each fold.

As shown in Table 6.5, the selected stylistic features performed comparably to token-based representations under K-fold cross-validation conditions (see Table 6.3). Across the various stylistic feature groups and machine learning algorithms, the mean accuracy reached 90%, with a range from 78% to 98%. This result highlights that stylistic features, traditionally considered less informative than token-based approaches, can achieve competitive performance in fake news detection. However, the performance varied significantly across models and feature sets. For instance, Logistic Regression models using the Fernandez and Abonizio feature sets (excluding the proposed social monetisation features) exhibited the lowest performance within this category, underscoring potential limitations when relying solely on these feature sets. In contrast, the Random Forest model trained on the Abonizio feature group demonstrated the highest accuracy, illustrating how model choice and feature set impact effectiveness even within the same class of features.

The introduction of the proposed social monetisation features resulted in marginal yet consistent improvements across all feature groups and machine learning models. The Mann-Whitney U-test confirmed the statistical significance of these improvements in most cases, with p-values indicating a meaningful increase in mean accuracy for models that included the proposed features compared to those that did not. This suggests that the proposed social monetisation features add a valuable layer of information, potentially capturing aspects of content dissemination and engage-

Table 6.5: Stylistic Features & S-M Features Baseline Results

Feature-Set	Model	Without proposed S-M Features					With proposed S-M Features					p-value
		Acc.	Prec.	Rec.	Spec.	F1	Acc.	Prec.	Rec.	Spec.	F1	
Fernandez	Logistic Regression	0.83	0.80	0.89	0.75	0.84	0.84	0.81	0.90	0.78	0.85	<0.001
	Decision Tree	0.88	0.88	0.88	0.87	0.88	0.93	0.93	0.93	0.93	0.93	<0.001
	SVM	0.90	0.86	0.95	0.84	0.91	0.92	0.89	0.96	0.87	0.93	<0.001
	Gradient Boosting	0.89	0.87	0.93	0.85	0.90	0.93	0.92	0.96	0.91	0.94	<0.001
	Random Forest	0.93	0.91	0.96	0.90	0.94	0.96	0.95	0.98	0.95	0.97	<0.001
	Neural Network	0.89	0.87	0.92	0.85	0.90	0.93	0.92	0.95	0.91	0.93	<0.001
Abonizio	Logistic Regression	0.78	0.77	0.82	0.73	0.79	0.78	0.77	0.82	0.74	0.79	0.5678
	Decision Tree	0.85	0.86	0.86	0.85	0.86	0.92	0.92	0.92	0.91	0.92	<0.001
	SVM	0.91	0.89	0.94	0.88	0.91	0.94	0.93	0.97	0.92	0.95	<0.001
	Gradient Boosting	0.86	0.85	0.90	0.82	0.87	0.93	0.91	0.96	0.90	0.93	<0.001
	Random Forest	0.93	0.92	0.95	0.91	0.94	0.98	0.97	0.99	0.97	0.98	<0.001
	Neural Network	0.90	0.90	0.92	0.88	0.91	0.94	0.94	0.96	0.93	0.95	<0.001
LIWC	Logistic Regression	0.91	0.90	0.92	0.89	0.91	0.91	0.91	0.92	0.90	0.92	0.2017
	Decision Tree	0.88	0.88	0.88	0.88	0.88	0.90	0.91	0.90	0.91	0.91	<0.001
	SVM	0.97	0.96	0.98	0.95	0.97	0.97	0.96	0.98	0.96	0.97	0.03
	Gradient Boosting	0.94	0.92	0.96	0.91	0.94	0.95	0.94	0.97	0.93	0.95	<0.001
	Random Forest	0.95	0.93	0.98	0.92	0.95	0.96	0.94	0.99	0.94	0.96	<0.001
	Neural Network	0.95	0.95	0.96	0.94	0.95	0.96	0.95	0.96	0.95	0.96	0.03
NELA Feature Extractor	Logistic Regression	0.85	0.85	0.88	0.83	0.86	0.86	0.85	0.88	0.84	0.87	0.023
	Decision Tree	0.85	0.86	0.85	0.85	0.86	0.89	0.89	0.89	0.89	0.89	<0.001
	SVM	0.94	0.92	0.96	0.91	0.94	0.95	0.93	0.97	0.92	0.95	<0.001
	Gradient Boosting	0.90	0.88	0.93	0.86	0.91	0.93	0.91	0.96	0.89	0.93	<0.001
	Random Forest	0.93	0.90	0.97	0.88	0.93	0.95	0.93	0.98	0.92	0.95	<0.001
	Neural Network	0.92	0.91	0.93	0.91	0.92	0.94	0.94	0.95	0.93	0.94	<0.001
Modified NELA Feature Extractor	Logistic Regression	0.87	0.87	0.89	0.86	0.88	0.88	0.88	0.89	0.86	0.88	0.023
	Decision Tree	0.88	0.89	0.88	0.88	0.88	0.90	0.91	0.91	0.90	0.91	<0.001
	SVM	0.96	0.95	0.98	0.94	0.96	0.97	0.96	0.98	0.95	0.97	<0.001
	Gradient Boosting	0.92	0.90	0.94	0.89	0.92	0.94	0.93	0.96	0.92	0.94	<0.001
	Random Forest	0.95	0.93	0.97	0.92	0.95	0.97	0.95	0.98	0.94	0.97	<0.001
	Neural Network	0.94	0.94	0.95	0.94	0.95	0.95	0.95	0.96	0.95	0.96	<0.001

ment that are not well-represented by purely stylistic features. The added predictive power provided by these features may enhance the model’s ability to identify patterns associated with fake news, such as monetisation strategies or social sharing behaviours, which are often subtle yet impactful indicators in real-world detection scenarios.

Moreover, the modified NELA feature set outperformed the original NELA set, both with and without the inclusion of the proposed social monetisation features. This consistent improvement suggests that the refinements made to the NELA set

increase its utility for fake news detection. The higher performance of the modified feature set across different models implies that strategic adjustments to established feature sets can lead to measurable gains in performance.

Table 6.6: Stylistic Features & S-M Features Cross-Dataset Results

Feature-Set	Model	Without proposed S-M Features					With proposed S-M Features					p-value
		Acc.	Prec.	Rec.	Spec.	F1	Acc.	Prec.	Rec.	Spec.	F1	
Fernandez	Logistic Regression	0.66	0.66	0.67	0.65	0.67	0.68	0.67	0.70	0.66	0.68	<0.001
	Decision Tree	0.66	0.65	0.67	0.65	0.66	0.68	0.69	0.64	0.71	0.66	<0.001
	SVM	0.68	0.66	0.71	0.64	0.69	0.69	0.69	0.71	0.68	0.70	<0.001
	Gradient Boosting	0.73	0.73	0.73	0.73	0.73	0.74	0.75	0.72	0.76	0.74	<0.001
	Random Forest	0.73	0.76	0.67	0.79	0.71	0.74	0.76	0.70	0.78	0.72	<0.001
	Neural Network	0.70	0.70	0.69	0.71	0.69	0.71	0.71	0.69	0.72	0.70	0.4446
Abonizio	Logistic Regression	0.59	0.59	0.62	0.56	0.60	0.62	0.61	0.66	0.57	0.63	<0.001
	Decision Tree	0.57	0.57	0.56	0.58	0.57	0.60	0.60	0.60	0.59	0.60	<0.001
	SVM	0.63	0.63	0.62	0.64	0.62	0.65	0.66	0.64	0.67	0.65	<0.001
	Gradient Boosting	0.61	0.60	0.64	0.58	0.62	0.65	0.65	0.67	0.63	0.66	<0.001
	Random Forest	0.64	0.65	0.61	0.67	0.63	0.67	0.68	0.67	0.68	0.67	<0.001
	Neural Network	0.60	0.60	0.59	0.61	0.60	0.64	0.64	0.61	0.66	0.63	<0.001
LIWC	Logistic Regression	0.69	0.71	0.65	0.73	0.68	0.67	0.69	0.62	0.72	0.65	N/A
	Decision Tree	0.62	0.62	0.63	0.61	0.62	0.62	0.62	0.63	0.61	0.62	0.3564
	SVM	0.67	0.68	0.65	0.70	0.66	0.68	0.68	0.67	0.68	0.67	0.02
	Gradient Boosting	0.69	0.70	0.65	0.72	0.68	0.74	0.74	0.72	0.75	0.73	<0.001
	Random Forest	0.67	0.67	0.69	0.66	0.68	0.69	0.69	0.68	0.69	0.69	<0.01
	Neural Network	0.68	0.69	0.64	0.71	0.66	0.69	0.71	0.63	0.74	0.67	0.0319
NELA Feature Extractor	Logistic Regression	0.65	0.65	0.63	0.67	0.64	0.68	0.69	0.65	0.71	0.67	<0.001
	Decision Tree	0.61	0.61	0.58	0.63	0.60	0.62	0.63	0.61	0.64	0.62	0.0258
	SVM	0.63	0.63	0.61	0.65	0.62	0.65	0.66	0.61	0.69	0.63	<0.001
	Gradient Boosting	0.69	0.72	0.64	0.75	0.68	0.70	0.71	0.67	0.73	0.69	0.1312
	Random Forest	0.66	0.68	0.62	0.70	0.65	0.68	0.69	0.64	0.72	0.66	<0.001
	Neural Network	0.63	0.64	0.60	0.67	0.62	0.65	0.67	0.61	0.70	0.64	<0.001
Modified NELA Feature Extractor	Logistic Regression	0.67	0.68	0.63	0.70	0.65	0.71	0.73	0.67	0.75	0.70	<0.001
	Decision Tree	0.64	0.64	0.63	0.65	0.64	0.66	0.66	0.65	0.66	0.66	<0.001
	SVM	0.66	0.66	0.65	0.67	0.66	0.69	0.70	0.64	0.73	0.67	<0.001
	Gradient Boosting	0.70	0.70	0.70	0.71	0.70	0.75	0.76	0.72	0.77	0.74	<0.001
	Random Forest	0.69	0.70	0.67	0.72	0.69	0.75	0.76	0.73	0.77	0.74	<0.001
	Neural Network	0.64	0.65	0.62	0.67	0.63	0.69	0.70	0.64	0.73	0.67	<0.001

In terms of generalisability (Table 6.6), models utilising stylistic features demonstrated slightly better performance in external validation compared to those relying on token representations (see Table 6.4), with an average accuracy drop of 24% from baseline to cross-dataset testing—about 4% better than the performance drop seen with token representations. Some stylistic feature-based models even surpassed the

cross-dataset performance of the best token-based model (Neural Network trained on TF-IDF features). Among models trained without the proposed social monetisation features, Gradient Boosting and Random Forest algorithms trained on the Fernandez feature set achieved higher mean accuracy than any token-representation-based models, indicating the potential effectiveness of stylistic features in handling the complexity of external validation scenarios.

With the inclusion of the proposed social monetisation features, the number of models surpassing token-based approaches in mean accuracy increased, underscoring the relevance of these features for developing more generalisable fake news detection models. Models that incorporated these features and outperformed token representations include Logistic Regression, Gradient Boosting, and Random Forest models trained on the modified NELA feature set, Gradient Boosting models trained on the LIWC feature set, and Gradient Boosting, Random Forest, and Neural Network models trained on the Fernandez feature set. This suggests that the social monetisation features capture unique aspects of fake news that contribute to enhanced model adaptability, particularly when applied to unseen datasets.

The statistical significance of these findings is further supported by the Mann-Whitney U-test, which indicates that the proposed social monetisation features have a statistically significant positive effect on model generalisability across various feature sets and algorithms. The improvements observed across multiple model types suggest that the proposed features capture relevant patterns related to fake news dissemination and monetisation that are not well-represented by token-based methods alone. This additional layer of information appears to provide stylistic models with the flexibility needed to maintain accuracy across different datasets, highlighting the potential for social monetisation features to address real-world variability and improve fake news detection models' resilience.

6.5.3 Analysis with Permutation Feature Importance

Experiment 2 presented evidence that the proposed social monetisation features contribute to producing more generalisable models. Permutation Feature Importance (PFI) analysis will further assess the impact of these features on the generalisability of the model. To prevent redundancy, the most successful model was selected for this analysis, which was Gradient Boosting trained on the modified NELA feature set, due to its higher mean accuracy (75%) in external validation conditions compared to other models. Although Random Forest trained on the same feature set demonstrates similar superior mean accuracy, Gradient Boosting was preferred due to its better performance across the other feature sets when compared to Random Forest. PFI was implemented by training the model on the NELA dataset and cal-

culating the feature importance on both an unseen portion of the NELA dataset and a random balanced sample of the Facebook URLs dataset. This allows us to observe the features that are relevant to both models, and therefore what features can be considered the most generalisable between the coarsely labelled NELA dataset and the manually-labelled Facebook URLs dataset.

Figures 6.6 and 6.7 display the feature importance plots for the Gradient Boosting models used in fake news detection, highlighting features that contribute meaningfully to model predictions. Due to the nature of the Gradient Boosting algorithm, certain features with ‘zero’ importance were excluded from the plots. This exclusion likely results from the algorithm’s tendency to select only one feature among highly correlated ones, thereby focusing on features with distinct positive or negative impacts on model performance.

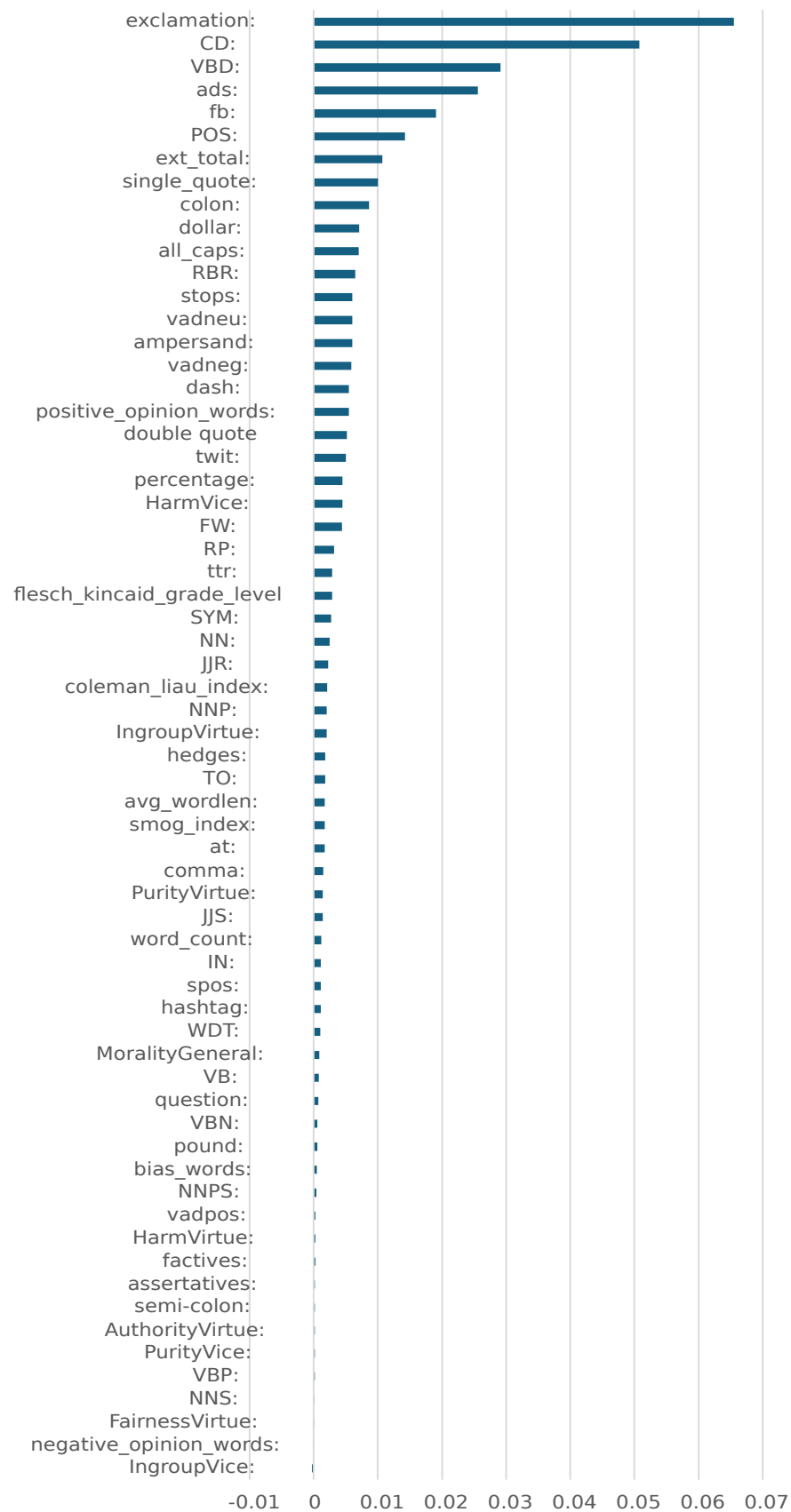


Figure 6.6: NELA Feature Importance

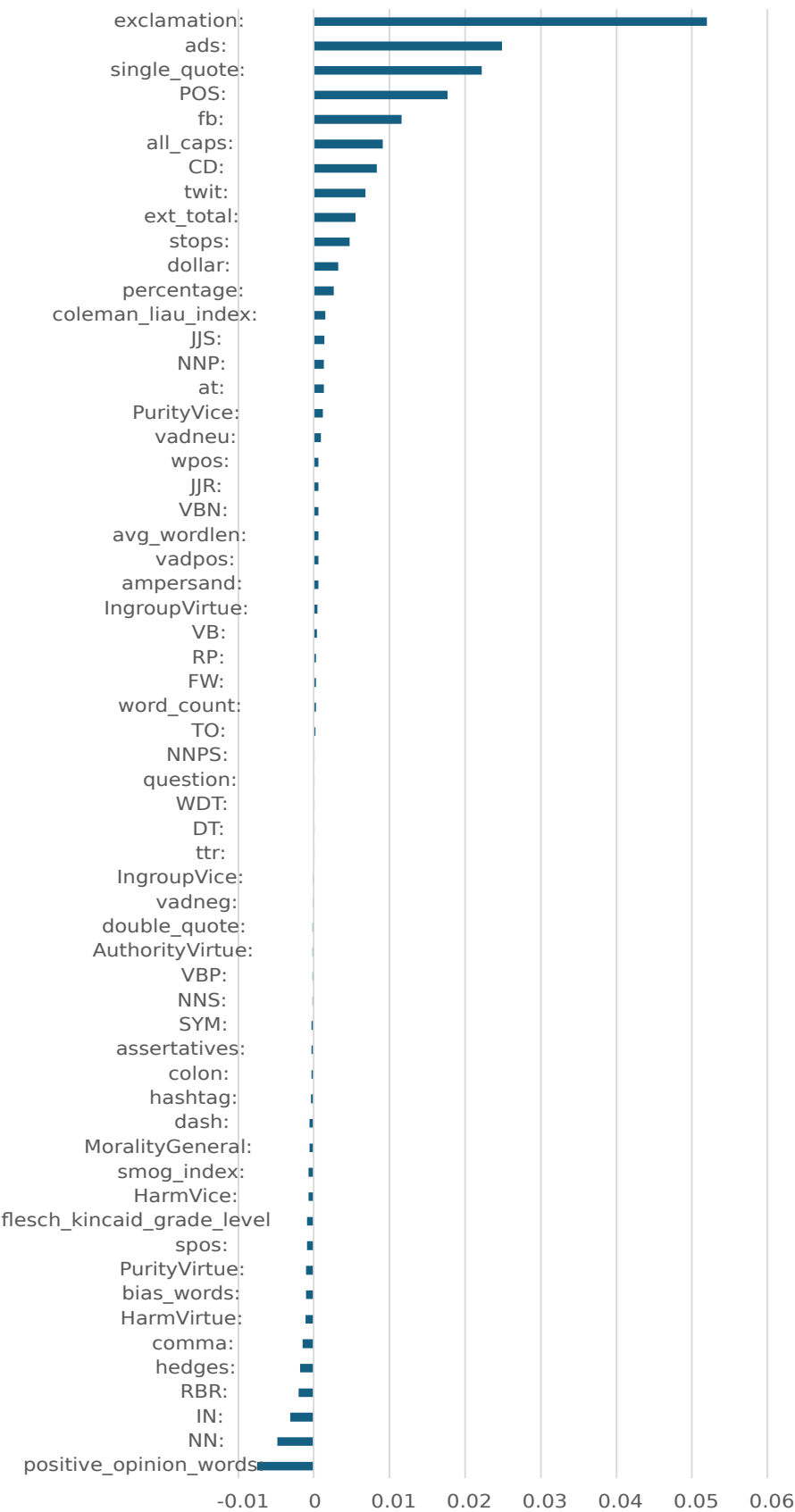


Figure 6.7: External Validation Feature Importance

Feature	Description
ads	Number of advertisements
all_caps	Words written entirely in uppercase
ampersand	Frequency of ampersand characters (&)
at	Frequency of the “at” symbol (@)
avg_wordlen	Average length of words
CD	Cardinal numbers
coleman_liaw_index	Readability metric indicating grade level
dollar	Dollar signs (\$)
exclamation	Exclamation marks (!)
ext_total	Total number of external links
fb	Presence of Facebook-related content
FW	Foreign words
IngroupVirtue	Words conveying positive group associations
JJR	Comparative adjectives (e.g., better)
JJS	Superlative adjectives (e.g., best)
NNP	Singular proper nouns
NNPS	Plural proper nouns
percentage	Percentage signs (%)
POS	Part-of-speech tags
PurityVice	Words indicating impurity or moral vice
question	Question marks (?)
RP	Particles
single_quote	Single quotation marks (')
stops	Stop words
TO	Infinitive marker “to”
ttr	Type-token ratio (lexical diversity)
twit	Presence of Twitter-related content
vadneu	Neutral valence in sentiment analysis
vadpos	Positive valence in sentiment analysis
VB	Base form verbs
VCN	Past participle verbs
WDT	Wh-determiners (e.g., which)
word_count	Total number of words

Table 6.7: Relevant features to both datasets

In examining these plots, we can identify 33 features (Table 6.7) that hold relevance across both datasets, including all four proposed social-monetisation features. This overlap provides additional evidence supporting the viability of social-monetisation features in enhancing the generalisability of fake news detection models. Notably, the ‘ads’ feature ranks highly in both datasets, reinforcing the idea that a key motivation for creating disinformation is often profit through advertising. This high ranking for ‘ads’ aligns with findings in fake news literature that connect monetisation tactics, such as heavy ad placement, with disinformation. Additionally, the Facebook feature ranks highly in both datasets, indicating the prominent role social media platforms play in the dissemination of fake news. The consistent relevance of these features suggests that economic incentives, captured through social-monetisation indicators like advertisements and Facebook links, are signifi-

cant drivers of disinformation. This aligns with prior research that highlights the exploitation of digital platforms for financial gain as a core characteristic of fake news.

From a stylistic perspective, exclamation marks consistently rank as the most important feature in both datasets, with ‘all caps’ words also ranking prominently. These features are frequently associated with fake news, particularly in sensation-alist headlines or emotionally charged content. This emphasis on exclamations and capitalised words aligns with prior research that links these stylistic cues to disinformation. Additionally, features like ‘CD’ (cardinal numbers) and ‘single quotes’ also show high importance in both datasets, which could reflect the tendency of fake news content to use specific numbers or quotations for added emphasis or perceived authority.

These findings underscore the value of both social-monetisation and stylistic features in identifying fake news. The strong presence of social-monetisation features, combined with stylistic cues like exclamations and all-caps text, suggests that these elements are integral to creating and detecting disinformation. Together, they enhance model accuracy and contribute to the broader objective of building more generalisable fake news detection models.

Table 6.8: Reduced Feature-Set Results

Test	Original Feature-Set					Reduced Feature-Set				
	Acc.	Prec.	Rec.	Spec.	F1	Acc.	Prec.	Rec.	Spec.	F1
K-Fold Test	0.94	0.93	0.96	0.92	0.94	0.91	0.89	0.94	0.89	0.91
Cross-Dataset Test	0.75	0.76	0.72	0.77	0.74	0.76	0.78	0.73	0.79	0.75

Further analysis, involving the repetition of K-Fold cross-validation and external validation using the 33 features that demonstrated positive feature importance across both datasets, revealed a slight decrease in K-Fold testing performance but slight improvements in external validation on the Facebook URLs dataset. Specifically, accuracy, recall, and F1 score increased by 0.01, while precision and specificity each improved by 0.02. These findings indicate that the reduced feature set, while slightly compromising K-Fold testing performance, enhances generalisability when applied to external datasets. This underscores the value of prioritising features with consistent positive importance across datasets.

Compared to word embeddings such as Word2Vec and BERT, the reduced set of stylistic features offers notable advantages in terms of computational efficiency. Word embeddings typically require significant resources for both feature extraction and model training, particularly when fine-tuning pre-trained models on large datasets. In contrast, the streamlined stylistic feature set demands less computa-

tional overhead, enabling faster training and evaluation while maintaining competitive performance.

These results highlight the practical and efficient nature of stylistic features for real-world applications, where resource constraints and model scalability are critical considerations. By balancing performance, generalisability, and efficiency, the reduced feature set provides a compelling alternative to computationally intensive word embedding approaches.

6.6 Discussion

The motivation for this study stems from the previous chapter which highlighted the poor generalisability of current approaches to fake news detection. These models often rely on token-based representations and are typically trained on coarsely labelled datasets, where the publisher of an article is used as a proxy for determining whether the article is ‘true’ or ‘fake.’ Recognising the challenges posed by such datasets, this study builds on the premise that stylistic features may be less affected by the biases introduced by coarsely labelled datasets. Acknowledging the necessity of using coarsely labelled datasets, due to the substantial effort required for manual labelling, this study aimed to develop a model trained on a coarsely labelled dataset (NELA) that could perform effectively on real-world data, specifically the Facebook URLs dataset. Additionally, the study proposed four novel social-monetisation features aimed at improving model generalisability and evaluated their effectiveness in this context.

One of the key findings of this study confirms that the challenge of generalisability using token-representations extends to real-world, manually labelled data. Unlike the previous chapter, which focused on generalisability between coarsely labelled datasets of the same topic and time period, this work shifts the focus to manually fact-checked, real-world data. Experiment 1 demonstrates that commonly used token-representations, such as Bag-of-Words (BoW), TF-IDF, Word2Vec, and BERT, exhibit a significant decline in accuracy when applied to the Facebook URLs dataset, echoing the patterns observed in the previous chapter and studies focused on dataset generalisability (e.g., Silva et al., 2020; Lakshmanarao et al., 2019). Moreover, there is considerable variability in recall and specificity across the models. For instance, models such as the Random Forest trained on BoW and TFIDF features, as well as the BoW SVM, displayed much higher recall compared to specificity, while the Word2Vec Logistic Regression model showed the opposite. Despite similar accuracy across these models, the disparity in these metrics underscores a fundamental challenge in fake news research: whether models should prioritise recall (accurately capturing all instances of true news) or specificity (accurately capturing all instances

fake news). Optimising for recall can prevent unintentional censorship of legitimate news, but it may increase the risk of spreading disinformation. On the other hand, optimising for specificity reduces the spread of fake news but may inadvertently suppress true news. Therefore, balancing false positives and false negatives becomes critical for the ethical and effective development of fake news detection systems.

The previous study posited that stylistic features may be a potential solution to the generalisability challenge. As such, a series of these stylistic features was tested as part of Experiment 2. While they did not outperform token-based models in terms of raw accuracy, they offered more balanced recall and specificity. This balance indicates that stylistic features can help models avoid the extremes of misclassifying true or fake news, making them more suitable for broader applicability. Additionally, the resilience of these features against dataset biases and concept drift adds to their utility in real-world applications (Przybyla, 2020). From a feature engineering and interpretability standpoint, stylistic features also provide more transparent insights into what drives model decisions, unlike the complexity of token-based methods (Qiao et al., 2020). This transparency is particularly valuable in models where explainability is crucial.

The study also introduced ‘social-monetisation’ features as a novel approach to improving model performance. As seen in Experiment 2, these features—designed to capture monetisation strategies such as the presence of advertisements, affiliate links, and social media calls-to-action—contributed to a statistically significant improvement in accuracy during external validation on the Facebook URLs dataset. Models like Random Forest and Gradient Boosting, which incorporated these features, achieved balanced recall and specificity, with a mean accuracy of 75%. This suggests that incorporating elements beyond the article’s text can lead to more robust and generalisable models. The feature importance analysis further supports this, showing that the frequency of advertisements, in particular, played a key role in enhancing model generalisability. Given that the previous study in Chapter 5 focused on generalisability across datasets with similar topics and time periods, and that the Facebook URLs dataset presents a more challenging standard with its broader scope and more detailed labelling, this result represents a notable improvement over earlier models in the literature.

The feature importance analysis also highlighted that a simplified feature set, consisting of 33 features, performed comparably to the full, more comprehensive feature sets used earlier in the study. The simplified model not only maintained the performance gains made with social-monetisation features but also contributed to more efficient model retraining and feature extraction. This efficiency is crucial for keeping up with the rapidly evolving news landscape, in contrast to the time and resource demands of fine-tuning large language models.

Overall, this study makes important strides in addressing the under explored issue of model generalisability in fake news detection. It add further evidence to the limitations of token-based models trained on coarsely labelled datasets, demonstrates the potential of stylistic features to provide balanced performance, and introduces novel social-monetisation features that produce a statistically significant improvement in model accuracy and generalisability. These findings underscore the value of multimodal approaches and offer a pathway for future research to further enhance the robustness and applicability of fake news detection models in real-world scenarios.

6.7 Chapter Summary

This chapter addresses the key challenges in generalising fake news detection models, particularly those trained on coarsely labelled datasets. It highlights the limitations of token-based models, which rely on representations such as Bag-of-Words, TF-IDF, Word2Vec, and BERT. These models show a significant drop in performance when tested on manually fact-checked, real-world datasets, specifically the Facebook URLs dataset. The study provides further evidence that while these token-based models can achieve high accuracy on their training datasets, their ability to generalise across different datasets is limited, with considerable variability in recall and specificity. This raises an ethical question about whether models should prioritise recall to avoid censoring legitimate news or specificity to minimise the spread of fake news.

Stylistic features emerge as a promising alternative. While they do not considerably outperform token-based methods in raw accuracy, they provide a more balanced trade-off between recall and specificity. This balance is crucial in preventing models from misclassifying either true or fake news. Additionally, stylistic features are shown to be more resilient to dataset biases and offer greater transparency in model decision-making, making them valuable in real-world applications where interpretability is key.

The study also introduced novel social-monetisation features—such as the presence of advertisements and social media links—which significantly improve model generalisability. These features, when incorporated into models like Random Forest and Gradient Boosting, lead to better performance in external validation conditions. The study’s feature importance analysis reveals that a simplified set of 33 features, including these novel features, performs comparably to a larger, more complex feature sets, highlighting the efficiency of these approaches in rapidly evolving news landscapes.

In summary, the chapter underscores the limitations of current token-based models and coarsely labelled datasets, while demonstrating the potential of stylistic and

social-monetisation features to improve model robustness and generalisability. It advocates for a multimodal approach in future fake news detection research to better address real-world challenges.

Chapter 7

Conclusions and Future Work

7.1 Introduction

The proliferation of fake news has emerged as a critical issue in today’s digital world, posing serious risks to public discourse, media credibility, and democratic processes. The rapid growth of social media and online platforms has not only accelerated the flow of information but has also expanded the reach of disinformation. This shift has transformed how people consume and share content, making it increasingly difficult to distinguish between accurate reporting and manipulated narratives. The pervasive nature of fake news has led to widespread consequences, contributing to social and political polarisation, eroding trust in credible news sources, and influencing election outcomes. Given the magnitude of the problem, manual fact-checking, while essential, is inadequate to keep up with the vast amount of information circulating online. The sophistication of disinformation strategies, such as clickbait and coordinated campaigns, adds further complexity to the challenge. These developments underscore the urgent need for automated and scalable solutions to detect and combat the spread of fake news. Although human-driven fact-checking plays a vital role in verification, its limitations in scope prevent it from meeting the demands of the rapidly evolving information landscape.

Driven by these concerns, this thesis focused on machine learning approaches to fake news detection, particularly addressing the shortcomings of existing methods. The research conducted throughout this thesis aimed to identify and fill key gaps in current literature, particularly around the generalisability of fake news detection models. A major challenge in the field has been the reliance on coarsely labelled datasets and token-based methods, which often lead to models that perform well on training data but struggle to generalise to unseen datasets. This lack of generalisability poses a significant obstacle to the practical deployment of machine learning models in real-world environments, where the diversity of news sources and topics

requires more adaptable detection capabilities.

This chapter synthesises the key findings from the empirical studies and theoretical explorations conducted in this thesis. It also reflects on the research limitations, identifying areas where the developed models and methods could be further refined. Although the issue of generalisability was addressed through the introduction of alternative feature sets, such as stylistic and social-monetisation features, generalisability remains an important focus for future research. Moreover, while the thesis primarily concentrated on datasets and feature engineering, less emphasis was placed on the exploration of algorithms, suggesting that future research could benefit from a deeper investigation into algorithmic approaches. In addition to these technical challenges, the chapter identifies opportunities for future work, such as exploring more advanced model architectures, integrating multimodal approaches that incorporate image and video data as well as the development of more refined datasets.

The chapter will also explore the ethical considerations associated with this research. As machine learning models become more advanced, concerns about their potential to inadvertently censor legitimate news or, conversely, allow harmful disinformation to slip through, continue to grow. This raises important ethical considerations in the design and deployment of fake news detection systems, particularly in balancing recall and specificity. Prioritising recall emphasises correctly identifying true news, but it risks misclassifying fake news. Conversely, focusing on specificity ensures that fake news is accurately flagged, though some legitimate news may be mistakenly marked as false. These ethical challenges are critical to the development of automated fake news detection technologies, and the chapter reflects on how future research should remain mindful of these concerns.

The chapter is organised as follows. Section 7.2 provides a summary of the earlier chapters, followed by Sections 7.3 and 7.4 which revisit and address the research objectives and questions of the thesis. Section 7.5 presents the main contributions of the thesis. Sections 7.6 and 7.7 address the limitations of the research, suggesting avenues for future work. Finally, the chapter concludes with a discussion on the ethical implications of the findings.

7.2 Summary of Thesis

Chapter 1

The introduction chapter provided the motivation and context for this thesis, emphasising the significant impact of fake news on public opinion, governance, and societal trust, as well as the urgent need for effective detection systems. It outlined key limitations of current fake news detection approaches, particularly their limited

generalisability across diverse datasets and real-world contexts.

To address this issue, the chapter framed the thesis’s primary aim of improving the adaptability and robustness of fake news detection models through novel feature selection and evaluation techniques. This included defining five research objectives: conducting a comprehensive literature review, testing diverse feature sets, developing a novel evaluation framework, and proposing features that capture the motivations behind fake news creation and dissemination. The chapter concluded with research questions designed to guide the investigation, setting a foundation for developing more generalisable fake news detection models.

Chapter 2

Elaborating on the motivation and current approaches outlined in Chapter 1, Chapter 2 provided an in-depth background on the issue of fake news. It traced the evolution of fake news from its early roots in traditional media to its current prominence in the digital era. The chapter explored how the rise of social media platforms has accelerated the spread of disinformation, creating new challenges for detecting and combating fake news. It highlighted the role of digital platforms in amplifying false information at an unprecedented speed and scale, often bypassing traditional editorial controls. The chapter also examined the technological advancements that have reshaped how news is shared and consumed, with a focus on the transition to digital platforms and the emergence of disinformation campaigns. This shift has not only changed the nature of news dissemination but has also introduced new complexities in identifying and addressing fake news.

In addition, the chapter reviewed different approaches to fake news detection, summarising both human-based methods—such as manual fact-checking and community reporting—and machine-based approaches that leverage artificial intelligence and machine learning. Human-driven methods were noted for their accuracy but limited in scalability, while machine learning approaches have emerged as essential for automating the detection process and handling the vast amount of content generated online. This discussion highlighted the growing need for automated solutions in an increasingly digital media landscape. This overview laid the contextual foundation for the thesis, offering key insights into the scale of the fake news problem and illustrating why automated detection methods have become essential.

Chapter 3

Extending from the broader context outlined in Chapter 2, Chapter 3 presented a more focused and detailed systematic review of the literature on machine learning approaches to fake news detection. This chapter critically examined the key

datasets, features, and algorithms used in the field, providing a thorough evaluation of their strengths and limitations.

The review highlighted the predominant use of small, coarsely labelled datasets, many of which were sourced from platforms like Kaggle. While these datasets are easily accessible and widely adopted, they often lack the granularity and quality needed for robust model training. This raised concerns about their ability to support the development of models that can generalise effectively in real-world scenarios. The review emphasised that this reliance on low-quality data could limit the applicability of the resulting models in practical settings, where content can vary significantly.

Regarding feature sets, the review categorised them into three main groups: content-based, social-context, and fused features. Content-based features focused on the intrinsic properties of the news articles, including token representations such as Bag-of-Words, TF-IDF, and Word2Vec, as well as stylistic and visual elements. Social-context features, on the other hand, examined the interactions surrounding news articles, such as user-based attributes (who shared the news), network-based features (how the news spread), and engagement metrics like comments and likes. Fused features integrated both content and social-context dimensions, offering a more comprehensive analysis. While token-based features were the most commonly explored, the review noted that stylistic, social-context, and fused features had been underutilised, despite their potential to provide deeper contextual insights and improve model performance.

The review also examined a variety of algorithms used in fake news detection, ranging from traditional machine learning methods like Support Vector Machines, Logistic Regression, and Naïve Bayes to more advanced models such as LSTMs. It found that most algorithms, with the exception of K-Nearest Neighbour (KNN), performed well under hold-out test conditions and cross-validation. However, the strong performance observed in these controlled testing environments raised concerns about the models' ability to generalise to unseen datasets.

The high performance of models under hold-out and cross-validation conditions prompted a deeper exploration of generalisability in fake news detection. Although these models performed well on their training datasets, the review revealed a limited number of studies addressing cross-domain generalisability, where models trained on one dataset (e.g., politics) are tested on a dataset from a different domain (e.g., health). These studies consistently found a significant drop in accuracy when models were exposed to new, unseen domains, underscoring the challenge of transferring knowledge across different contexts. This limitation raised fundamental concerns about the ability of fake news detection models to generalise effectively beyond the controlled environments in which they were initially tested.

Given these findings, the thesis identified key issues that needed further investiga-

tion, including the prevalence of low-quality datasets, the reliance on content-based features, and the need to better understand generalisability in the context of fake news detection. These challenges became central to the focus of the thesis, shaping the direction of the subsequent empirical studies.

Chapter 4

Chapter 4 detailed the methodological approach used to investigate intra-domain generalisability in fake news detection models. The research questions were defined at the outset, motivated by the findings from the systematic review in Chapter 3, which highlighted gaps in the literature related to the issue of generalisability in fake news detection models. This was followed by an overview of the text-classification process, rooted in the scientific method, that formed the core methodology of this thesis.

The first stage of this process was the data collection, where it was noted from the literature there are three predominant approaches. The first relies on using already established datasets, the second creating a dataset from scratch and the third a combination of these two methods. Given this thesis's focus on evaluating established fake news detection methods, it was decided to use existing datasets for the analysis because they allow for direct comparison with previous research and provide a reliable foundation for testing model performance. This approach enabled a thorough exploration of model generalisability without the resource-intensive process of creating a new dataset, aligning with the thesis's goal of evaluating current methodologies within real-world constraints.

The second stage of the text classification process outlined the pre-processing steps. These steps included text cleaning to remove noise such as punctuation, special characters, and URLs, followed by converting all text to lowercase and tokenisation to ensure consistency. It was noted that certain feature extraction methods, such as BERT and Word2Vec, would require less intensive pre-processing, as these models are designed to handle language in its more natural form and can capture contextual relationships without extensive modification of the text. This distinction in pre-processing requirements allowed the thesis to tailor the approach according to each feature extraction method, ensuring optimal model performance and preserving meaningful linguistic information where needed.

Following pre-processing, the different methods of feature-extraction were defined, motivated by the findings of the systematic review in Chapter 3. The review noted that token-representation methods, such as Bag-of-Words, TF-IDF, Word2Vec, and BERT, were among the most popular approaches in the literature. Consequently, these methods were selected for this study to ensure consistency with established research and to facilitate a meaningful comparison with existing models.

Additionally, stylistic features were included to address gaps identified in the literature, specifically in capturing broader textual characteristics beyond simple word patterns.

The third stage of the methodology involved selecting machine learning algorithms, a choice also informed by the systematic review in Chapter 3. The review highlighted the popularity of certain algorithms in the literature, reflecting their effectiveness in fake news detection. Consequently, this thesis included a mix of traditional models—such as Naïve Bayes, Logistic Regression and Decision Trees, as well as more complex methods like Gradient Boosting and LSTMs. This selection enabled a balanced evaluation of both simpler models, valued for their interpretability and efficiency, and complex models, which could leverage the contextual and stylistic nuances captured in the chosen feature sets. By using these well-established algorithms, the study aimed to ensure comparability with existing research and facilitate a comprehensive assessment of model generalisability within intra-domain contexts.

The chapter concluded by outlining the training and evaluation methods, alongside model interpretability techniques. Training and evaluation involved the use of K-fold cross-validation to assess model performance within the primary dataset, providing a robust measure of consistency across different data splits. Additionally, external validation was employed to evaluate model generalisability on datasets not included in the training phase, offering insights into how well each model performed on unseen data. Model interpretability techniques, such as Local Interpretable Model-Agnostic Explanations (LIME) and Permutation Feature Importance (PFI), were also outlined with the goal to clarify how specific features influenced predictions.

Overall, the chapter provided a framework for investigating intra-domain generalisability in fake news detection models. By systematically outlining data collection, pre-processing, feature extraction, algorithm selection, and model evaluation, the chapter established a structured approach to assess model effectiveness. The inclusion of interpretability techniques further enriched this framework, enabling insights into feature importance and model decision-making. This comprehensive methodology served as the foundation for testing and understanding how well fake news detection models generalise within similar data contexts, addressing the gaps identified in the literature and guiding the thesis’s empirical investigation.

Chapter 5

Motivated by the systematic review in Chapter 3 and the research questions formalised in Chapter 1, the first empirical chapter investigated intra-domain generalisability in fake news detection models. This chapter addressed the issue of generalisability within a single domain. The experiments in this chapter were designed to assess how well fake news detection models could generalise within a single domain,

using datasets related to the 2016 US Presidential Election.

The first experiment, employing Stratified K-Fold Cross-Validation, revealed high model performance within individual datasets, aligning with the findings of the literature. However, while these results were promising, they did not guarantee the models' ability to generalise beyond the training data. This limitation became evident in the second experiment, where models were tested against datasets not used during training. The results demonstrated that even within the same domain and time period, the models struggled to generalise effectively, raising concerns about the practical application of these models in real-world scenarios where diverse datasets are encountered.

Two primary factors were identified as contributing to the lack of generalisability: the limited size of existing datasets, which led to overfitting, and the use of coarsely labelled datasets, which failed to capture nuanced distinctions in content. Furthermore, the reliance on traditional token-representation methods—such as Bag-of-Words, TF-IDF, Word2Vec, and BERT—exacerbated these limitations. These methods, by focusing on specific word patterns, tended to reinforce biases within the datasets.

Reflecting on these findings, the chapter proposed that stylistic features, which capture broader text characteristics beyond word patterns, might offer a more promising solution for improving generalisability. By shifting the focus from what is said to how it is said, stylistic features could mitigate the biases that token-representation methods reinforce. The chapter concluded by identifying key areas for future research. These included the need for larger, more representative datasets to reduce overfitting, the exploration of novel feature sets—especially stylistic features—to enhance generalisability, and the development of more rigorous testing methodologies to ensure models perform effectively across diverse datasets, even within the same domain.

Chapter 6

Building on the findings from the previous chapter, which emphasised the limitations of token-based models and coarsely labelled datasets in fake news detection, Chapter 6 aimed to explore stylistic features more comprehensively. Recognising the ongoing reliance on coarsely labelled datasets, the NELA dataset was used for training, while the manually-labelled Facebook URLs dataset was crucially employed for testing. This method provided a more rigorous assessment of model generalisability across different dataset types.

While stylistic features emerged as a promising alternative to token-based models, they did not exceed them in raw accuracy. However, they offered a more balanced trade-off between recall and specificity, helping to reduce the misclassification

of both true and fake news. Notably, these features were more resistant to dataset biases and enhanced model transparency, making them particularly beneficial for real-world applications.

The chapter also introduced novel social-monetisation features, such as the presence of advertisements and social media links, which significantly boosted the models' ability to generalise. When incorporated into models like Random Forest and Gradient Boosting, these features led to superior performance in cross-dataset evaluations. A feature importance analysis showed that a streamlined set of 33 features, including the novel features, achieved performance comparable to more complex feature sets, highlighting their efficiency in dynamic news environments.

Chapter 6 reinforced the limitations of token-based models and coarsely labelled datasets, while showcasing the potential of stylistic and social-monetisation features to enhance model robustness and generalisability. It advocated for a multimodal approach in future fake news detection research to better address the practical challenges of real-world detection systems.

7.3 Research Objectives Revisited

This thesis set out to address the challenge of improving the generalisability of fake news detection models across diverse datasets and real-world contexts. The primary aim was to enhance the adaptability and effectiveness of these models by exploring new approaches to feature selection and model evaluation. To achieve this aim, five key objectives were established, each contributing a specific focus to the research. This section explains how each of these objectives was addressed.

- **Objective 1:** *Conduct a comprehensive literature review on fake news detection using machine learning, identifying current approaches, evaluating their effectiveness, and highlighting specific challenges and gaps related to model generalisability.*

A systematic review was conducted in Chapter 3, which provided a comprehensive analysis of current approaches and techniques. This review identified key limitations in existing models, particularly regarding generalisability, and highlighted gaps in feature selection and evaluation methods that informed the direction of this research.

- **Objective 2:** *Systematically test and compare the impact of different feature sets and ML algorithms on generalisability.*

Motivated by the systematic review conducted in Chapter 3, the empirical chapters in Chapter 5 and Chapter 6 systematically explored a range of popular token-based representations and stylistic features. These chapters tested

and compared the impact of these feature sets and various algorithms on generalisability, offering insights into which features most effectively enhance model robustness across different datasets. This systematic evaluation helped identify key feature sets that contribute to more adaptable and reliable fake news detection models.

- **Objective 3:** *Create a novel evaluation framework that combines training on widely available datasets with testing on manually labelled data, simulating real-world scenarios and enabling more accurate assessments of model performance.*

Recognising the need to work with coarsely labelled datasets, Chapter 6 introduced a more robust evaluation framework that combined K-fold cross-validation with external validation. This method involved training on a coarsely labelled dataset and testing on both an unseen portion of the training dataset for each fold and a random sample from an external, manually labelled dataset. By incorporating both familiar and independent data, this dual-layered approach provided a more accurate assessment of model performance, better simulating real-world scenarios and enhancing the evaluation of model generalisability.

- **Objective 4:** *Propose and test novel feature sets specifically designed to improve generalisability, with a focus on features beyond text that capture the motivations for fake news creation and dissemination.*

To address Objective 4, novel social-monetisation features were proposed and tested, focusing on elements beyond traditional text-based features to capture the motivations behind fake news creation and dissemination. These features included the frequency of advertising, affiliate links, and social media share links. Testing in Chapter 6 revealed that these features contributed a statistically significant improvement in model generalisability, demonstrating the value of incorporating economic incentives as a means to enhance the generalisability of models tested on real-world data.

- **Objective 5:** *Provide practical guidelines and recommendations for developing generalisable fake news detection models.*

To fulfil Objective 5, practical guidelines and recommendations were discussed in the closing sections of Chapter 3, Chapter 5, and Chapter 6. These guidelines underscore the importance of incorporating manually labelled datasets for external validation, prioritising feature selection over algorithm choice to enhance generalisability, and adopting robust testing frameworks. Together,

these insights offer a foundation for developing more adaptable and reliable fake news detection models capable of handling diverse real-world scenarios.

7.4 Research Questions Revisited

This thesis was guided by a set of research questions aimed at systematically addressing the limitations in fake news detection models, with a particular focus on enhancing model generalisability. Here, each research question is revisited in light of the findings and contributions of the thesis:

- **RQ1:** *What are the current methods to detect fake news?*

The systematic review in Chapter 3 explored the range of methods currently used in fake news detection, focusing on the types of datasets, features, and machine learning algorithms that underpin these approaches. The review demonstrated that established, coarsely labelled datasets are the most frequently used in the field, often paired with token-based approaches such as Bag-of-Words and TF-IDF, as well as embeddings like Word2Vec and BERT. Other features include stylistic indicators, such as sentiment, linguistic complexity, and readability, which capture patterns in language style that may distinguish fake news from real news, and social-context features, which analyse user engagement metrics like shares, comments, and likes to provide insight into how fake news propagates and resonates with audiences.

While coarsely labelled datasets and token-based features are clearly established as the most widely used approach for fake news detection, a variety of algorithms have been commonly applied. Traditional algorithms such as Naïve Bayes, Support Vector Machines, and Logistic Regression are widely favoured for their interpretability and computational efficiency. In contrast, more advanced algorithms, including Gradient Boosting and Long Short-Term Memory (LSTM) networks, are also utilised, as they can capture complex patterns and sequential dependencies within the data. This range of algorithms reflects the diverse approaches adopted to address the complexities of fake news detection, with each method offering unique strengths depending on the specific features and data characteristics used.

- **RQ2:** *How effective are current methods to detect fake news?*

The findings of the systematic literature review indicated that a broad range of features and algorithms perform well on a variety of common datasets in the literature in holdout testing or K-fold cross-validation conditions. Across

studies, both traditional algorithms, such as Naïve Bayes and Support Vector Machines, and advanced approaches like LSTMs, consistently achieved strong performance in terms of accuracy. Commonly used features, including token-based approaches (e.g., Bag-of-Words, TF-IDF) and embeddings (e.g., Word2Vec, BERT), also demonstrated effectiveness across multiple datasets. Stylistic features demonstrated greater variability, however, likely due to the broad range of different stylistic features available, which may not be as consistent compared to other token-based approaches.

- **RQ3:** *To what extent do existing fake news detection methods generalise across datasets?*

The analysis in Chapters 5 and 6 empirically demonstrated that generalisability remains a significant challenge in fake news detection. Models trained on coarsely labelled datasets frequently performed well within their original contexts but struggled with external datasets, such as manually labelled real-world data. Token representations were found to be particularly susceptible to capturing dataset-specific vocabulary and biases that limited their effectiveness when applied to new data sources. This outcome reinforces the need for rigorous external validation and the development of models capable of adapting to diverse data sources, languages, and topics.

- **RQ4:** *What current features contribute to more generalisable models in the context of fake news detection?*

The empirical results in Chapters 5 and 6 underscored the value of stylistic features in enhancing model generalisability. Unlike token-based approaches, which are often influenced by the specific vocabulary and topics in the training data, stylistic features—such as linguistic complexity, sentiment, and readability—provided a more stable foundation across datasets. These features showed greater resilience against dataset biases and concept drift, enabling models to maintain balanced recall and specificity in detecting fake news. Additionally, stylistic features contributed to model transparency, offering interpretability advantages over token-based models, where decision-making is often opaque. Together, these findings suggest that stylistic features support more adaptable and interpretable models.

- **RQ5:** *How can novel features that extend beyond the text—such as social dissemination behaviours and economic incentives—enhance the generalisability of fake news detection models?*

The novel social-monetisation features examined in Chapter 6—such as the presence of advertisements, social media share links, and affiliate content—were

shown to significantly improve model generalisability. By capturing economic motivations and dissemination patterns associated with fake news, these features added a valuable dimension to model analysis beyond text content alone. Importantly, social-monetisation features were less prone to topical biases, allowing models to focus on underlying monetisation patterns rather than dataset-specific vocabulary. This approach enabled the models to detect fake news with greater consistency across datasets, highlighting the potential of such features to contribute to more robust, generalisable fake news detection models.

7.5 Contributions

The research conducted in this thesis makes several contributions to the field of fake news detection and the issues of generalisability in this context. These contributions are as follows:

Table 7.1: Key Contributions of the Thesis

ID	Contribution	Type
C-1	Demonstrated the advantages of stylistic features in the context of fake news detection	Empirical
C-2	Introduced a novel category of features relating to social dissemination behaviours and economic incentives for fake news detection	Empirical
C-3	Produced a reduced and simplified set of features for more generalisable and efficient fake news detection models	Empirical
C-4	Developed a novel evaluation approach of models in the context of fake news detection	Methodological

7.5.1 C-1 – Stylistic Features

The first contribution of this thesis was the demonstration of the advantages of stylistic features in fake news detection. While traditional token-based methods such as Bag-of-Words and TF-IDF focus on word patterns and frequencies, this research highlighted how stylistic features offer a more nuanced and effective approach to improving generalisability and performance in fake news detection models. Stylistic features—capturing elements like sentence complexity, tone, readability, and writing style—emerged as a valuable alternative to token-based approaches, particularly in addressing the limitations of generalisability across datasets.

One of the key advantages of stylistic features identified in this thesis was their ability to provide a more balanced trade-off between recall and specificity, as evidenced in Chapter 6. While token-based models often prioritise word patterns, stylistic features focus on how content is presented, offering a broader perspective that captures subtle cues indicative of fake news. This balance is critical in preventing the misclassification of both true and fake news, as stylistic features are less reliant on the specific vocabulary of a dataset and more attuned to the underlying presentation of the information. As fake news often relies on exaggerated or sensationalised writing styles to manipulate readers, stylistic features help models identify such patterns, making them more adaptable to new, unseen datasets.

The thesis also showed that stylistic features demonstrated greater resilience to dataset biases compared to token-based models. Token-based models are heavily influenced by the word patterns within the training data, which can lead to overfitting and poor generalisability when applied to different datasets, particularly when datasets are biased. In contrast, stylistic features are less likely to overfit to a particular dataset because they focus on the structure and presentation of text rather than the specific words used. This makes stylistic features more robust when applied to a variety of datasets, as they are better able to generalise across different topics and writing styles. The empirical findings revealed that models incorporating stylistic features performed more consistently when tested across diverse datasets, including the manually labelled Facebook URLs dataset, highlighting their potential for real-world application.

Beyond demonstrating that stylistic features offer advantages over token-based methods, the thesis also identified specific stylistic features that contribute to the generalisability of models tested on ‘real-world’ data through a permutation feature importance analysis. This analysis allowed the thesis to pinpoint the most influential stylistic attributes—such as exclamations and words in all-caps—that enhanced the model’s ability to detect fake news across diverse datasets. By understanding which features had the most significant impact on performance, the research provided practical insights into how models can be fine-tuned for better performance.

These findings underline the importance of incorporating stylistic features into fake news detection models to address the limitations of token-based approaches. The ability of stylistic features to capture broader contextual cues, rather than relying solely on word patterns, makes them a valuable addition to fake news detection systems aimed at operating in dynamic and evolving news environments.

7.5.2 C-2 – Social-Monetisation Features

The second contribution of this thesis was the development of novel social-monetisation features for fake news detection. These features were introduced to address a critical gap in the existing research: the lack of attention to the economic drivers behind the creation and spread of fake news. While most models have focused on token-based or stylistic features, social-monetisation features capture the financial motivations that often underpin disinformation, offering a new perspective on how fake news can be detected more effectively.

Empirical testing in Chapter 6 showed that the inclusion of social-monetisation features, such as the presence of advertisements, social media share buttons and affiliate links, produced a statistically significant increase in model accuracy, in both K-fold cross validation and external validation conditions. By focusing on the financial mechanisms that encourage the spread of disinformation, the thesis moves beyond purely linguistic or stylistic analysis, grounding its approach in the real-world economic context of fake news production.

This advancement not only produces a statistically significant improvement in detection accuracy but also brings a more comprehensive understanding of the factors driving the fake news phenomenon, which is particularly important for addressing the root causes of disinformation in the digital age. This contribution has practical implications for future research and model development, highlighting the importance of considering economic incentives in the fight against disinformation. It also lays the groundwork for further exploration of monetisation-related features, encouraging the integration of social and economic indicators in future fake news detection models. Furthermore, the integration of social-monetisation features represents a shift towards a more holistic understanding of fake news detection. Unlike traditional token-based approaches, which are often limited to surface-level text patterns, this contribution delves deeper into the structural and economic aspects of how disinformation propagates online. It demonstrates that fake news is not just a linguistic phenomenon but also a financially motivated one, where the incentives to maximise user engagement can drive the creation and dissemination of false information. By capturing these incentives through features like advertisements, affiliate links, and social share buttons, the thesis provides a richer feature set that not only improves classification accuracy but also enhances the model’s ability to detect fake news across diverse contexts.

7.5.3 C-3 – Simplified Feature Set

The third contribution of this research centres on the Permutation Feature Importance analysis conducted as part of Study 2 in Chapter 6, which uncovered that a

refined, streamlined feature set of 33 key indicators could deliver similar performance to the original, comprehensive feature set. This finding underscores the potential to reduce the model’s complexity without sacrificing accuracy, paving the way for a more efficient, practical approach to fake news detection.

This streamlined feature set, despite its reduced size, preserves critical information needed for accurate classification on real-world, manually labelled data. By leveraging only the most influential features, the model maintains robust performance, whilst also improving computational efficiency. This improvement is especially valuable when applying machine learning algorithms like Gradient Boosting, which benefit from the computational efficiency and reduced memory requirements associated with a leaner set of features. The model’s performance remains strong, showing that carefully selected features can retain predictive power even without the extensive scope of the full feature set.

In practical terms, this optimised feature set enhances the model’s ability to keep pace with the rapid evolution of news content. The simplified model can be retrained more swiftly, which is crucial in a domain where new information surfaces constantly and disinformation spreads quickly. This efficiency in retraining allows for faster model updates, facilitating its application to new or evolving data with minimal delay. Moreover, a smaller feature set streamlines feature extraction, making it possible to extract features faster, a critical advantage in high-stakes environments like social media monitoring and news verification.

In contrast to the computationally intensive process of fine-tuning large language models (LLMs), which can be both time-consuming and costly, this streamlined approach offers a balanced solution, retaining high accuracy while reducing resource demands. This makes it ideal for applications that require both scalability and adaptability, addressing the need for quick updates and reliable performance in the dynamic and often unpredictable landscape of news and social media.

7.5.4 C-4 – Evaluation Approach for Fake News Detection Models

The fourth contribution of this thesis was the improvement in the process of training and evaluation of fake news detection models. While techniques such as external validation have already been established, this thesis represents the first to combine training on coarsely labelled data with testing on manually labelled data in the fake news detection literature. This approach addresses a key challenge in fake news detection: the limitations of coarsely labelled datasets, which often misclassify content based on broad publisher-level assumptions rather than the actual veracity of the information, as well as inflate the perceived performance of models for the fake news

detection task. By introducing manually labelled data into the evaluation process, the thesis offers a more granular and accurate method for assessing model performance. This contribution provides a significant enhancement to the field, as models trained solely on coarsely labelled data may fail to capture the nuanced distinctions between fake and real news, leading to overfitting and reduced generalisability.

Furthermore, combining K-fold cross-validation with external validation (i.e., testing each model trained for each fold of K-fold cross-validation against an external dataset) strengthens the reliability of the findings. This dual approach allows for a more rigorous evaluation of the model's performance by ensuring that it is tested on both held-out data from the training set and an entirely independent dataset. As a result, this method provides a more comprehensive understanding of the model's generalisability and its robustness in real-world applications.

This novel approach not only enhances the credibility of the evaluation process by providing a more accurate assessment of model performance, but also reveals the limitations of training solely on coarsely labelled data. Testing on manually labelled data allows for a clearer understanding of a model's ability to generalise to real-world disinformation, offering insights that go beyond the inflated performance metrics often associated with coarsely labelled datasets. This approach sets a more realistic benchmark for future studies in fake news detection.

7.6 Limitations and Future Research

While this thesis has made significant contributions to the field of fake news detection, particularly in terms of exploring and developing a broad range of feature sets, several limitations remain, providing opportunities for future research.

7.6.1 Features over Model Architectures

One key limitation of the thesis is it has concentrated focus on features rather than the models themselves. The research has primarily aimed at identifying, implementing, and evaluating various feature sets—such as token-based, stylistic, and novel social-monetisation features—to improve the generalisability and robustness of fake news detection systems. By concentrating on these feature sets, the thesis has demonstrated how different types of information within a text contribute to the effectiveness of detection models. However, this feature-focused approach has meant that less attention has been given to the selection and optimisation of the machine learning models used to process these features.

As part of this limitation, model hyperparameter tuning was not a central component of the research. The models used throughout the thesis generally relied on

default or standard hyperparameters, with limited exploration into optimising these settings. Hyperparameter tuning plays a crucial role in achieving the best possible performance for machine learning models, and future studies could investigate the impact of systematic hyperparameter optimisation on the performance of fake news detection models when combined with the various feature sets explored in this thesis. Techniques such as grid search, random search, or more advanced methods like Bayesian optimisation could be employed to fine-tune models for improved results.

Given these limitations, future research may look to incorporate further exploration into novel features as well as novel model architectures. Previous research, such as the studies by Kozik et al. (2021); Raghavendra and Niranjnamurthy (2024) and Wanda and Diqi (2024), have proposed innovative architectures for fake news detection. By combining advanced feature sets with novel architectures and testing under external validation conditions, future studies could yield insights into how these models perform in practical applications, ultimately leading to more adaptable and robust fake news detection systems.

7.6.2 Expanding Beyond Textual Features

While this thesis has focused on a range of textual features—including token-based, stylistic, and social-monetisation attributes—there remains a significant opportunity to explore other non-textual features that could further enhance fake news detection. Extending this research to include external validation on real-world, manually labelled datasets could offer a more rigorous examination of how well non-textual features, such as visual and social-context features, perform under conditions that reflect the complexity and variability of actual disinformation.

In particular, visual features could provide valuable insights for fake news detection, as disinformation often uses images and videos to amplify its reach and influence. Integrating features such as image manipulation detection, visual sentiment analysis, or consistency checks between textual and visual content could strengthen models by capturing the multimodal nature of fake news. For instance, images accompanying fake news articles are often chosen or altered to elicit strong emotional responses, which can increase the persuasive impact of the content. Incorporating visual cues could make detection models more robust and reflective of the ways disinformation operates on multimedia-driven platforms like social media.

Social-context features, while becoming increasingly difficult to gather owing to the increased restrictions by social media platforms, could also significantly enhance fake news detection models by providing context beyond the content itself. Fake news often spreads widely due to social dynamics, with engagement metrics like likes, shares, and comments amplifying its visibility and reach. Additionally, source

credibility and network characteristics—such as the influencers or groups promoting the content—can reveal patterns associated with disinformation. Features that capture these social elements could allow models to better assess the potential reach and impact of fake news. Testing these features on externally validated, manually labelled datasets would reveal how social dynamics contribute to model performance, helping to identify which social-contextual attributes are most effective across various datasets and platforms.

Incorporating visual and social-context features into fake news detection systems, as well as additional novel features, represents a promising direction for future research. By moving beyond textual features and testing non-textual attributes under real-world conditions, researchers can build more holistic and adaptive detection models. These models would be better suited to handle the evolving tactics of disinformation, capturing both the multimodal and socially driven aspects of fake news that are increasingly prevalent in today’s digital media landscape.

7.6.3 Reliance on Existing Datasets

A key limitation of this thesis is its reliance on existing datasets for model training and evaluation, which, while practical and relevant to the focus on current approaches, may impact the adaptability and robustness of the models when applied to emerging and evolving forms of fake news. Established datasets offer a stable foundation and facilitate comparison with previous research, allowing for consistency in evaluating model performance. However, these datasets often contain inherent limitations: they may lack diversity in topic range, geographic scope, and sources, and they may not fully represent the complex tactics of modern disinformation campaigns. This reliance may inadvertently narrow the scope of model effectiveness, as existing datasets may not include the latest forms of disinformation, such as deep-fakes, coordinated social media campaigns, or mixed-media disinformation where visuals and text work together to mislead audiences.

Moreover, this thesis has only leveraged one manually-labelled dataset for evaluation. While the Facebook URLs dataset is currently the only available dataset labelled in this manner, relying on a single manually-labelled source limits the depth and diversity of the evaluation process. Manually-labelled datasets offer a higher level of accuracy and detail compared to coarsely-labelled datasets, providing insights into nuanced distinctions within fake news content. However, using just one dataset may restrict the thesis’s ability to generalise findings across varied forms of disinformation and across different platforms.

As such, future research would benefit from the development or inclusion of additional manually-labelled datasets spanning different domains, topics, and disinform-

mation strategies. Expanding the evaluation to include multiple manually-labelled datasets could provide a more comprehensive understanding of model effectiveness, allowing for a richer assessment of generalisability and robustness in real-world contexts. This approach would enhance confidence in the model’s ability to adapt to diverse forms of fake news beyond those represented in a single dataset.

7.7 Ethical Considerations

While this thesis has produced encouraging results in the field of fake news detection, it is important to acknowledge the ethical considerations inherent in developing and deploying such models. As such, this section will explore the ethical considerations associated with fake news detection, examining issues related to model accuracy and bias, transparency and accountability, freedom of speech and censorship, and the risk of misuse. By examining these considerations, this thesis not only seeks to enhance the technical capabilities of fake news detection but also aims to promote responsible practices that align with democratic values, safeguard public trust in information systems, and mitigate potential societal risks.

7.7.1 Model Accuracy and Bias

This thesis has identified critical challenges related to model accuracy and bias in fake news detection, particularly stemming from the use of coarsely labelled datasets and traditional evaluation approaches. Coarsely labelled datasets often assign labels such as “fake” or “real” based solely on the source of an article rather than its content (Torabi Asr and Taboada, 2019). This approach assumes that all content from reputable publishers is entirely reliable, while all content from less reputable or biased sources is entirely false. Such assumptions fail to account for variations in reporting quality and accuracy within individual sources, oversimplifying the inherently complex nature of disinformation (Horne et al., 2023).

Relying on source-based labelling results in models that perform well in controlled evaluation settings but struggle to generalise effectively to real-world scenarios. As noted in Chapter 3, much of the existing literature highlights strong results when models are tested on such datasets. However, these results often fail to reflect the true robustness of the models, as their predictions are largely driven by source-specific patterns rather than meaningful features that differentiate fake from real news. Additionally, datasets labelled in this manner frequently exhibit topical biases, where certain topics, entities, or events are overrepresented. This imbalance can lead to unreliable model performance when applied to new datasets, even within the same domain, as the models struggle to adapt to varied distributions of content.

Chapter 5 further illustrated these limitations, revealing that while models often achieve high accuracy under holdout or cross-validation conditions, they fail to generalise effectively to other coarsely labelled datasets addressing similar topics. This inconsistency exposes the limitations of traditional evaluation methods like K-fold cross-validation, which assume that the training and testing splits adequately represent real-world data. In practice, such methods often enable models to exploit dataset-specific biases, resulting in predictions that lack robustness when exposed to diverse content or new sources (Steyerberg and Harrell Jr, 2015). Consequently, model accuracy is frequently overestimated, creating a misleading perception of reliability and generalisability. This overestimation carries ethical risks, particularly when these models are deployed in environments where fairness and impartiality are crucial.

7.7.2 Transparency and Accountability

Given the prevalent use of supervised machine learning algorithms, transparency and accountability emerge as critical ethical challenges in fake news detection. While these algorithms are highly effective, they rely on large, labeled datasets that can inadvertently embed biases or unintended patterns into the decision-making process. When these systems operate as “black boxes,” offering little to no insight into how classifications are determined, users and stakeholders are left without a clear understanding of the reasoning behind their decisions (Rudin, 2019). This lack of transparency not only undermines trust but also limits the ability to identify, correct, or refine errors.

The opaque nature of many machine learning models, particularly those employing complex architectures like deep learning networks, compounds the challenge of transparency. These models process data in intricate, nonlinear ways that are difficult to interpret. Consequently, when legitimate content is misclassified as fake or sophisticated disinformation goes undetected, users and other affected parties are left without a way to trace or resolve the root causes. This lack of interpretability fosters mistrust, particularly if the model’s outputs appear inconsistent or biased.

Accountability in fake news detection is inherently linked to these transparency issues. The absence of clear interpretability makes it challenging to determine who is responsible for the outcomes of such systems. This is particularly problematic when errors—such as misclassifying genuine content—have real-world implications, including influencing public discourse or damaging reputations. Addressing these concerns requires a focus on both technical solutions for interpretability and ethical frameworks to ensure responsible deployment.

7.7.3 Freedom of Speech and Censorship

Fake news detection models face a unique challenge in navigating the nuanced landscape of online content, where factual reporting often coexists with speculation, opinion, satire, and commentary. A significant ethical concern is that, without careful consideration of context, these models may inadvertently classify legitimate expressions—such as opinion pieces, speculative articles, or satirical content—as disinformation. This potential for mis-categorisation not only risks unjustly censoring certain viewpoints but also raises concerns about the preservation of free speech in a digital environment where automated systems increasingly mediate information access (Hasimi and Poniszewska-Maranda, 2024).

Speculation and opinion are natural parts of news and commentary, providing audiences with diverse perspectives and interpretations. However, fake news detection models, particularly those relying on linguistic or token-based features, may struggle to differentiate between factually misleading information and subjective content that intentionally offers interpretation or hypothesis. As a result, opinionated or speculative articles may be flagged as disinformation if they contain language or stylistic patterns similar to those in fabricated content. This overreach risks limiting public discourse by imposing rigid boundaries around what constitutes “truthful” content, effectively stifling discussions that may be controversial, provocative, or counter-narrative but still legitimate.

Moreover, satire, parody, and humour often rely on exaggeration and irony to critique current events and social issues. These forms of expression, though not intended to deceive, could be misconstrued by detection models as false information due to their use of hyperbole and unconventional framing. Mislabelling satire as disinformation can lead to the suppression of a valuable form of social commentary, undermining the role of satire in fostering critical thinking and societal reflection.

7.7.4 Risk of Misuse

The risk of misuse is a critical ethical concern in the deployment of fake news detection models. While these models are developed to address disinformation, they also present opportunities for manipulation by powerful entities, such as governments, corporations, or interest groups, to control narratives, suppress dissenting voices, or discredit opposing viewpoints. For instance, an authoritarian government could exploit these models to label critical journalism or opposition voices as fake news, silencing legitimate discourse and undermining democratic freedoms. Similarly, corporations may use these tools to downplay or remove critical content that could harm their reputation, prioritising corporate interests over public transparency (Blauth et al., 2022).

To mitigate such risks, establishing ethical guidelines and clear usage limitations is essential. These guidelines should prevent models from being deployed as instruments of censorship, particularly in politically sensitive or controversial contexts. Transparency in model usage is equally important, allowing public oversight and accountability to help prevent abuses and ensure these tools serve their intended purpose without infringing on freedom of expression.

7.7.5 Recommendations

To address the ethical complexities surrounding fake news detection, this thesis offers a set of recommendations focused on enhancing model fairness, transparency, and responsible application. These recommendations aim to guide the ethical deployment of fake news detection models, balancing the need to address disinformation with the imperative to protect democratic values.

1. *Use Models to Flag, Not to Decide*

Given the limitations of current models in accurately discerning between disinformation and legitimate content, it is not recommended that these systems be relied upon as the sole method of limiting the spread of disinformation. Instead, fake news detection models should serve as tools for flagging potential disinformation, with final decisions left to human moderators who can consider context, intent, and nuance. This human-in-the-loop approach minimises the risk of misclassification and supports fairer, more accurate decision-making.

2. *Enhance Dataset Diversity and Labelling Practices*

To improve model generalisability and reduce bias, diverse and nuanced datasets should be prioritised in model training. Moving beyond binary “fake” and “real” labels, more datasets should include categories that reflect the range of real-world content, such as opinion, speculation, and satire. Ethically robust labelling practices help ensure that models differentiate between harmful disinformation and legitimate forms of expression, supporting balanced and responsible outcomes.

3. *Prioritise Transparency in Model Design*

Transparency is crucial to fostering accountability and public trust. Clearly documenting model processes, including data sources, feature selection, and decision-making criteria, allows stakeholders to understand model limitations and make informed assessments. Such transparency also enables users to see models as support tools rather than unquestioned authorities, reducing the risk of overreliance and promoting critical engagement with flagged content.

4. *Provide Ongoing Ethical Oversight*

The potential for misuse of fake news detection models necessitates ethical oversight and clear usage guidelines. Regular audits and reviews should be implemented to ensure that models are not employed for censorship or to suppress legitimate content. Usage guidelines should emphasise fair application and restrict contexts where models might infringe upon freedom of speech, particularly in politically sensitive areas.

By adopting these recommendations, fake news detection models can better support efforts to manage disinformation without infringing on freedom of expression. These measures promote a responsible, human-centered approach that respects ethical standards and reinforces public trust in AI-driven systems.

Appendix A

Embedding Algorithms

A.1 Word2Vec Embedding Algorithm

Algorithm 1 `get_word2vec_embeddings`

Require: List of texts (`texts`), Pre-trained Word2Vec model (`word2vec_model`),
Maximum token length (`max_length`, default = 300)

Ensure: Array of embeddings for each text

```
1: Initialize embeddings as an empty list
2: for text in texts do
3:   Split text into tokens (limit to first max_length tokens)
4:   Initialize text_embeddings as an empty list
5:   for token in tokens do
6:     if token exists in word2vec_model then
7:       Retrieve token embedding from word2vec_model
8:     else
9:       Use a zero vector of size word2vec_model.vector_size
10:    end if
11:    Append the embedding to text_embeddings
12:  end for
13:  if len(text_embeddings) < max_length then
14:    Compute padding_length = max_length - len(text_embeddings)
15:    Create padding of padding_length zero vectors
16:    Extend text_embeddings with padding
17:  end if
18:  Append text_embeddings to embeddings
19: end for
20: Convert embeddings to an array
21: return embeddings
```

A.2 BERT Embedding Algorithm

Algorithm 2 `get_bert_embeddings`

Require: List of texts (`texts`), Tokenizer (`tokenizer`), Pre-trained BERT model (`bert_model`), Device (`device`), Maximum token length (`max_length`)

Ensure: Array of BERT embeddings for each text

- 1: Set `bert_model` to evaluation mode
- 2: Initialize `embeddings` as an empty list
- 3: Disable gradient computations using `torch.no_grad()`
- 4: **for** `text` in `texts` **do**
- 5: Tokenize `text` with `tokenizer`, specifying:
 - 6: Maximum token length = `max_length`
 - 7: Padding = "`max_length`"
 - 8: Truncation = `True`
 - 9: Return tensors = "`pt`"
- 10: Move tokenized data to `device`
- 11: Pass tokenized input through `bert_model` to obtain output
- 12: Extract the token embedding from `output.last_hidden_state[:,0,:]`
- 13: Convert the embedding from GPU to CPU and append to `embeddings`
- 14: **end for**
- 15: Convert `embeddings` to a NumPy array
- 16: **return** `embeddings`

Appendix B

Stylistic Feature-Sets

B.1 Fernandez Feature-Set

Table B.1: Fernandez Feature-Set

Feature	Description
Word Count	Total number of words
Syllables Count	Total number of syllables
Sentence Count	Total number of sentences
Word/Sent	Total words divided by total sentences
Long Words Count	Number of words with more than 6 characters
All Caps Count	Number of words in all caps
Unique Words Count	Number of unique words
Personal Pronouns %	Percentage of words such as ‘I, we, she, him’
First Person Singular %	Percentage of words such as ‘I, me’
First Person Plural %	Percentage of words such as ‘we, us’
Second Person %	Percentage of words such as ‘you, your’
Third Person Singular %	Percentage of words such as ‘she, he, her, him’
Third Person Plural %	Percentage of words such as ‘they, them’
Impersonal Pronouns %	Percentage of words such as ‘it, that, anything’
Articles %	Percentage of words such as ‘a, an, the’
Prepositions %	Percentage of words such as ‘below, all, much’
Auxiliary Verbs %	Percentage of words such as ‘have, did, are’
Common Adverbs %	Percentage of words such as ‘just, usually, even’
Conjunctions %	Percentage of words such as ‘until, so, and, but’
Negations %	Percentage of words such as ‘no, never, not’
Common Verbs %	Percentage of words such as ‘run, walk, swim’
Common Adjectives %	Percentage of words such as ‘big, small, silly’
Comparisons %	Percentage of words such as ‘better, greater, larger’
Concrete Figures %	Percentage of words that represent real numbers
Punctuation Count	Total number of punctuation marks per document
Full Stop Count	Total number of full stops
Commas Count	Total number of commas
Colons Count	Total number of colons
Semi-Colons Count	Total number of semi-colons
Question Marks Count	Total number of question marks
Exclamation Marks Count	Total number of exclamation marks
Dashes Count	Total number of dashes
Apostrophe Count	Total number of apostrophes
Brackets Count	Total number of brackets ‘()’

B.2 Abonizio Feature-Set

Table B.2: Abonizio Feature-Set

Group	Feature	Description
Complexity	Word_per_sents	Average number of words per sentence
	Avg_word_size	Average length of the words in the text
	Sentences	Number of sentences
	TTR	Type-Token Ratio – a metric of lexical variety
Stylometric	POS_diversity_ratio	Ratio of words with POS tags to length of text
	Entities_ratio	Ratio of named entities to length of text
	Upper_case	Number of upper-case letters
	Oov_ratio	Words that are OOV in Spacy’s language model
	Quotes_count	Number of quotation marks
	Quotes_ratio	Ratio of quotation marks to length of text
	Ratio_ADJ	Ratio of adjectives to text size
	Ratio_ADP	Ratio of adpositions to text size
	Ratio_ADV	Ratio of adverbs to text size
	Ratio_DET	Ratio of determiners to text size
	Ratio_NOUN	Ratio of nouns to text size
	Ratio_PRON	Ratio of pronouns to text size
	Ratio_PROPN	Ratio of proper nouns to text size
	Ratio_PUNCT	Ratio of punctuation to text size
	Ratio_SYM	Ratio of symbols to text size
	Ratio_VERB	Ratio of verbs to text size
Psychological	Polarity	Sentiment analysis score

B.3 LIWC

Table B.3: LIWC

Group	Feature	Description
Summary Variables	WC, Analytic, Clout, Authentic, Tone, WPS, Big-Words, Dic	Word Count, Metric of logical/formal thinking, language of leadership/status, degree of +ve/-ve tone, average words per sentence, percentage of words >7 letters, percentage of words captured by LIWC dictionary
Punctuation Marks	Period, comma, qmark, exclam, apostro, otherp	Full stops, commas, question marks, exclamations, apostrophes, other punctuation
Linguistic Dimensions	Function, pronoun, ppron, I, we, you, shehe, they, ipron, det, article, number, prep, auxverb, adverb, conj, negate, verb, adj, quantity	Total function words, total pronouns, personal pronouns, personal pronouns (1st person singular), personal pronouns (1st person plural, singular), personal pronouns (2nd person), personal pronouns (3rd person singular), personal pronouns (3rd person plural), impersonal pronouns, determiners, articles, numbers, prepositions, auxiliary verbs, adverbs, conjunctions, negations, common verbs, common adjectives, quantities
Psychological Processes	Drives, affiliation, achieve, power, Cognition, allnone, cogproc, insight, cause, discrep, tentat, certitude, differ, memory, Affect, tone_pos, tone_neg, emotion, emo_pos, emo_neg, emo_anx, emo_anger, emo_sad, swear, Social, secbehav, prosocial, polite, conflict, moral, comm, socrefs, family, friend, female, male	Drives, Affiliation, Achievement, Power, Cognition, All-or-none, Cognitive processes, Insight Causation, Discrepancy, Tentative, Certitude, Differentiation, Memory, Affect, Positive tone, Negative tone, Emotion, Positive emotion, Negative emotion, Anxiety, Anger Sadness, Swear words, Social processes, Social behaviour, Prosocial behaviour, Politeness, Interpersonal conflict, Moralization, Communication, Social referents, Family, Friends, Female references, Male references
Expanded Dictionary	Culture, politic, ethnicity, tech, lifestyle, leisure, home, work, money, relig, physical, health, illness, wellness, mental, substances, sexual, food, death, need, want, acquire, lack, fulfil, fatigue, reward, risk, curiosity, allure, perception, attention, motion, space, visual, auditory, feeling, time, focuspast, focuspresent, conversation, netspeak, assent, nonflu, filler	Words pertaining to the following categories: Culture, Politics, Ethnicity Technology, Lifestyle, Leisure, Home, Work, Money Religion, Physical, Health, Illness, Wellness, Mental health, Substances, Sexual, Food, Death, States, Need, Want, Acquire, Lack, Fulfilled, Fatigue, Motives, Reward, Risk, Curiosity, Allure, Perception, Attention, Motion, Space, Visual, Auditory, Feeling, Time orientation, Time, Past focus, Present focus, Future focus, Conversational, Netspeak, Assent, Nonfluencies, Fillers

B.4 NELA Feature-Set

Table B.4: NELA Feature-Set

Group	Feature	Description
Style - Largely similar to those from the previous two studies, focusing on POS tags	'quotes', 'exclaim', 'allpunc', 'allcaps', 'stops', CC, CD, DT, EX, FW, IN, JJ, JJR, JJS, LS, MD, NN, NNS, NNP, NNPS, PDT, POS, PRP, PRP\$, RB, RBR, RBS, RP, SYM, TO, UH, VB, VBD, VBG, VBN, VBP, VBZ, WDT, WP, WP\$, WRB, ('\$'), ('"',), ('('), (')'), (';'), ('-',), ('.',), (':'), ('"',)	Quotes, Exclamations, Punctuation Count, All Caps Count, Coordinating conjunction, Cardinal number, Determiner, Existential 'there', Foreign word, Preposition or subordinating conjunction, Adjective, Adjective (comparative), Adjective (superlative), List item marker, Modal, Noun (singular or mass), Noun (plural), Proper noun (singular), Proper noun (plural), Predeterminer, Possessive ending, Personal pronoun, Possessive pronoun, Adverb, Adverb (comparative), Adverb (superlative), Particle, Symbol, 'to', Interjection, Verb (base form), Verb (past tense), Verb (gerund or present participle), Verb (past participle), Verb (non-3rd person singular present), Verb (3rd person singular present), Wh-determiner, Wh-pronoun, Possessive wh-pronoun, Wh-adverb, Dollar signs, Double Quotations Marks, Open Parentheses, Closing Parentheses, Commas, Dashes, Sentence Terminators, Colons, Single Quotation Marks
Complexity - Assesses an article's complexity by analyzing lexical diversity, reading-difficulty metrics, and the average length of words and sentences.	'ttr', 'avg_wordlen', 'word_count', 'flesch_kincaid_grade_level', 'smog_index', 'coleman_liau_index', 'lix'	Type-token ratio (variation of vocabulary), Average Word-Length, Word Count, Flesch Kincaid Grade (readability metric), SMOG Index ('Simple Measure of Gobbledygook'), Coleman-Liau Index (readability metric), LIX (readability metric)
Bias - based on (Recasens et al., 2013), capture text subjectivity by identifying hedges, factives, assertives, implicatives, and opinion words.	'bias_words', 'assertatives', 'factives', 'hedges', 'implicatives', 'report_verbs', 'positive_opinion_words', 'negative_opinion_words'	Bias words (word that introduce prejudice), assertatives (words stating facts with confidence), factives (words that imply truth), hedges (that determine the strength of a statement), implicatives (words that imply), report verbs (e.g., 'report' or 'declare'), positive opinion words, negative opinion words
Affect - Relying on VADER sentiment analysis, this group aims to capture the emotion and sentiment of the text	'vadneg', 'vadneu', 'vadpos', 'wneg', 'wpos', 'wneu', 'sneg', 'spos', 'sneu'	VADER Negative sentiment, VADER Neutral sentiment, VADER Positive Sentiment. The remaining tags refer to different types of words (positive, negative and neutral) that appear in a dictionary based on Recasens et al.'s work
Moral - Evaluates the ethical content of text using a lexicon developed from Moral Foundation Theory by Graham et al., further elaborated by Lin et al.	'HarmVirtue', 'HarmVice', 'FairnessVirtue', 'FairnessVice', 'IngroupVirtue', 'IngroupVice', 'AuthorityVirtue', 'AuthorityVice', 'PurityVirtue', 'PurityVice', 'MoralityGeneral'	Words pertaining to the following categories: Caring for others, causing harm, fairness, unfairness, loyalty, disloyalty, authority, subversion, purity, degradation and general words in relation to morality
Event - Aims to capture words relating to dates, times and locations.	Num.locations, num_dates	Number of geographical locations, number of dates

Bibliography

- Athira A, Abhishek Tiwari, S D Madhu Kumar, and Anu Mary Chacko. Multimodal data fusion framework for fake news detection. In *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, nov 2022.
- A D L Abeynayake, A A Sunethra, and K A D Deshani. A stylometric approach for reliable news detection using machine learning methods. In *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, nov 2022.
- Hugo Queiroz Abonizio, Janaina Ignacio de Moraes, Gabriel Marques Tavares, and Sylvio Barbon Junior. Language-Independent Fake News Detection: English, Portuguese, and Spanish Mutual Features. *Future Internet*, 12(5):87–105, may 2020. ISSN 1999-5903. doi: 10.3390/fi12050087.
- Maja Adena, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. Radio and the Rise of The Nazis in Prewar Germany. *The Quarterly Journal of Economics*, 130(4):1885–1939, nov 2015. ISSN 0033-5533. doi: 10.1093/qje/qjv030.
- Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake News Detection Using Machine Learning Ensemble Methods. *Complexity*, 2020:1–11, October 2020. ISSN 1099-0526. doi: 10.1155/2020/8885861.
- Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10618 LNCS:127–138, 2017. ISSN 16113349. doi: 10.1007/978-3-319-69155-8_9/TABLES/6.
- Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9, jan 2018. ISSN 2475-6725. doi: 10.1002/SPY2.9.

- Yahya Albalawi, Jim Buckley, and Nikola S Nikolov. Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting arabic health information on social media. *Journal of big Data*, 8(1):95, 2021.
- Stuart Allan. *News Culture*. Open University Press, 2nd edition, 2004. ISBN 9780335210732.
- Douglas Allchin. Alternative facts & fake news. *The American Biology Teacher*, 80(8):631–633, 2018.
- Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, may 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.211.
- Manisha Aluri, Divya Panchumarthi, Bhargav Boddupalli, Murali Krishna Enduri, Sumana G Sree, and Satish Anamalamudi. An empirical study on fake news prediction with machine learning methods. In *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, dec 2022.
- A B Athira, Abhishek Tiwari, S D Madhu Kumar, and Anu Mary Chacko. Multimodal Data Fusion Framework For Fake News Detection. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–4. IEEE, November 2022. ISBN 978-1-6654-7350-7. doi: 10.1109/INDICON56171.2022.10039737.
- Muhammad Babar, Awais Ahmad, Muhammad Usman Tariq, and Sarah Kaleem. Real-time fake news detection using big data analytics and deep neural network. *IEEE Trans. Comput. Soc. Syst.*, 11(4):5189–5198, aug 2024.
- Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. Deep active learning for misinformation detection using geometric deep learning. *Online Soc. Netw. Media*, 33(100244):100244, jan 2023.
- José Manuel Benítez, Juan Luis Castro, and Ignacio Requena. Are artificial neural networks black boxes? *IEEE Transactions on neural networks*, 8(5):1156–1164, 1997.
- Himani Bhavsar and Mahesh H Panchal. A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1(10):185–189, 2012.
- Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

- Przemyslaw Biecek and Tomasz Burzykowski. Local interpretable model-agnostic explanations (lime). *Explanatory Model Analysis Explore, Explain and Examine Predictive Models*, 1:107–124, 2021.
- Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.
- Ciara Blackledge and Amir Atapour-Abarghouei. Transforming Fake News: Robust Generalisable News Classification Using Transformers. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3960–3968. IEEE, dec 2021. ISBN 978-1-6654-3902-2. doi: 10.1109/BigData52589.2021.9671970.
- Taís Fernanda Blauth, Oskar Josef Gstrein, and Andrej Zwitter. Artificial intelligence crime: An overview of malicious use and abuse of ai. *Ieee Access*, 10: 77110–77122, 2022.
- Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55, September 2019. ISSN 00200255. doi: 10.1016/j.ins.2019.05.035.
- Xavier Bouthillier and Gaël Varoquaux. Survey of machine-learning experimental methods at neurips2019 and iclr2020. Technical report, Inria Saclay Ile de France, 2020.
- Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7, 2019.
- Max Bramer. Avoiding overfitting of decision trees. *Principles of data mining*, pages 119–134, 2007.
- Sofia Bratu. Fake news, health literacy, and misinformed patients: the fate of scientific facts in the era of digital medicine. *Analysis and Metaphysics*, 17:122–127, 2018.
- Paul R. Brewer and Emily Marquardt. Mock News and Democracy: Analyzing The Daily Show. *Atlantic Journal of Communication*, 15(4):249–267, nov 2007. ISSN 1545-6870. doi: 10.1080/15456870701465315.
- Asa Briggs and Peter Burke. *A Social History of the Media: From Gutenberg to the Internet*. Polity Press, 2009. ISBN 978074564494.
- Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. Misunderstanding the harms of online misinformation. *Nature*

2024 630:8015, 630(8015):45–53, jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07417-w.

C. Bisailon. Fake and Real News Dataset. <https://www.kaggle.com/datasets/clmentbisailon/fake-and-real-news-dataset>, n.d. Accessed: 2024-10-04.

Federico Cabitza, Andrea Campagner, Felipe Soares, Luis García de Guadiana-Romualdo, Feyissa Challa, Adela Sulejmani, Michela Seghezzi, and Anna Carobene. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, 208:106288, 2021.

Dustin P. Calvillo, Justin D. Harris, and Whitney C. Hawkins. Partisan bias in false memories for misinformation about the 2021 U.S. Capitol riot. *Memory*, 31(1): 137–146, jan 2023. ISSN 14640686. doi: 10.1080/09658211.2022.2127771.

W. Joseph Campbell. Yellow Journalism. *The International Encyclopedia of Journalism Studies*, pages 1–5, apr 2019. doi: 10.1002/9781118841570.IEJS0159.

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. Exploring the Role of Visual Content in Fake News Detection. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161. Springer, Cham, mar 2020. ISBN 978-3-030-42699-6. doi: 10.1007/978-3-030-42699-6_8.

Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Francesco David Nota. Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review. *Neurocomputing*, 530:91–103, apr 2023. ISSN 09252312. doi: 10.1016/j.neucom.2023.02.005.

Jeffrey C. Carver, Edgar Hassler, Elis Hernandez, and Nicholas A. Kraft. Identifying barriers to the systematic literature review process. In *International Symposium on Empirical Software Engineering and Measurement*, pages 203–213, 2013. doi: 10.1109/ESEM.2013.28.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of The 2019 World Wide Web Conference, WWW '19*, pages 975–980, New York, NY, USA, may 2019. ACM. ISBN 9781450366755. doi: 10.1145/3308560.3316739.

Jaydine M Castillo, Kyla Dann F Fadera, Alexana Alian A Ladao, Jeline G Go, Melecia B Tamayo, and Manolito V Octaviano. Fake news detection on English news article’s title. In *2021 1st International Conference in Information and Computing Research (iCORE)*. IEEE, dec 2021.

- Gizem Ceylan, Ian A. Anderson, and Wendy Wood. Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences of the United States of America*, 120(4):e2216614120, jan 2023. ISSN 10916490. doi: 10.1073/PNAS.2216614120/SUPPL\FILE/PNAS.2216614120.SAPP.PDF.
- Christine P Chai. Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3):509–553, 2023.
- Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as "false news". In *WMDD 2015 - Proceedings of the ACM Workshop on Multimodal Deception Detection, co-located with ICMI 2015*, pages 15–19, 2015. ISBN 9781450339872. doi: 10.1145/2823465.2823467.
- Zhouhan Chen and Juliana Freire. Discovering and Measuring Malicious URL Redirection Campaigns from Fake News Domains. *Proceedings - 2021 IEEE Symposium on Security and Privacy Workshops, SPW 2021*, pages 1–6, may 2021. doi: 10.1109/SPW53761.2021.00008.
- Rajdipa Chowdhury, Sriram Srinivasan, and Lise Getoor. Joint Estimation of User And Publisher Credibility for Fake News Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20*, pages 1993–1996, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3412066.
- Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.
- Jonathan Clarke, Hailiang Chen, Ding Du, and Yu Jeffrey Hu. Fake News, Investor Attention, and Market Reaction. *SSRN Electronic Journal*, sep 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3213024.
- L Cui, S Wang, and D Lee. SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 41–48, 2019. ISBN 2473-991X VO -. doi: 10.1145/3341161.3342894.
- Mansour Davoudi, Mohammad R Moosavi, and Mohammad Hadi Sadreddini. A novel method for fake news detection based on propagation tree. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*. IEEE, oct 2021.

- Mansour Davoudi, Mohammad R Moosavi, and Mohammad Hadi Sadreddini. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Syst. Appl.*, 198(116635):116635, jul 2022.
- Barry De Ville. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455, 2013.
- Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.
- Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. Drink Bleach or Do What Now? COVID-HeRA: A Study of Risk-Informed Health Decision Making in the Presence of COVID-19 Misinformation. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1218–1227, may 2022. ISSN 2334-0770. doi: 10.1609/ICWSM.V16I1.19372.
- Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544, 2022.
- Ross Eaman. *Historical dictionary of journalism*. Rowman and Littlefield, 2nd edition, 2021. ISBN 9781538125038. doi: 10.5860/choice.47-0014.
- Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. Fake News Detection on Social Media: A Systematic Survey. In *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 1–8. IEEE, August 2019. ISBN 978-1-7281-2794-1. doi: 10.1109/PACRIM47961.2019.8985062.
- Aaron Carl T Fernandez and Madhavi Devaraj. Computing the Linguistic-Based Cues of Fake News in the Philippines Towards its Detection. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, WIMS2019, pages 1–9, New York, NY, USA, jun 2019. ACM. ISBN 9781450361903. doi: 10.1145/3326467.3326490.
- Victor C Ferreira, Sandip Kundu, and Felipe M G Franca. Analysis of fake news classification for insight into the roles of different data types. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*. IEEE, jan 2022a.

- Victor C. Ferreira, Sandip Kundu, and Felipe M. G. França. Analysis of fake news classification for insight into the roles of different data types. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 75–82, 2022b. doi: 10.1109/ICSC52841.2022.00018.
- Rochelle Forrester. History of Printing - From Gutenberg to the Laser Printer. *SSRN Electronic Journal*, jan 2020. doi: 10.2139/ssrn.3512249.
- Carla M Forster and Nicole Wong. Rightfully-Placed Blame: How social media algorithms facilitate post-truth politics. *Bristol Institute for Learning and Teaching (BILT) Student Research Journal*, 5, jul 2024.
- Damien François, Vincent Wertz, and Michel Verleysen. The permutation test for feature selection by mutual information. In *ESANN*, pages 239–244, 2006.
- Paulo Márcio Souza Freire and Ronaldo Ribeiro Goldschmidt. Fake news detection on social media via implicit crowd signals. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*, New York, NY, USA, oct 2019. ACM.
- Sonal Garg and Dilip Kumar Sharma. Linguistic features based framework for automatic fake news detection. *Computers and Industrial Engineering*, 172:108432, oct 2022. ISSN 03608352. doi: 10.1016/j.cie.2022.108432.
- Akansha Gautam and Koteswar Rao Jerripothula. SGG: Spinbot, Grammarly and GloVe based Fake News Detection. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 174–182. IEEE, sep 2020. ISBN 978-1-7281-9325-0. doi: 10.1109/BigMM50055.2020.00033.
- Joma George, Shintu Mariam Skariah, and T Aleena Xavier. Role of Contextual Features in Fake News Detection: A Review. In *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE, February 2020. ISBN 978-1-7281-4210-4. doi: 10.1109/ICITIIT49094.2020.9071524.
- A Giachanou, G Zhang, and P Rosso. Multimodal Multi-image Fake News Detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654, 2020. ISBN VO -. doi: 10.1109/DSAA49011.2020.00091.
- Marcos Gôlo, Mariana Caravanti, Rafael Rossi, Solange Rezende, Bruno Nogueira, and Ricardo Marcacini. Learning Textual Representations from Multiple Modalities to Detect Fake News Through One-Class Learning. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '21, pages 197–204, New York, NY, USA, November 2021. ACM. ISBN 9781450386098. doi: 10.1145/3470482.3479634.

- Juan M Gorriz, Fermín Segovia, Javier Ramirez, Andrés Ortiz, and John Suckling. Is k-fold cross validation the best model selection method for machine learning? *arXiv preprint arXiv:2401.16407*, 2024.
- Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47:55–130, 2013. ISSN 0065-2601. doi: <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>.
- Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213, August 2019. ISSN 09574174. doi: [10.1016/j.eswa.2019.03.036](https://doi.org/10.1016/j.eswa.2019.03.036).
- Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The Future of False Information Detection on Social Media. *ACM Computing Surveys*, 53(4):1–36, July 2021. ISSN 0360-0300. doi: [10.1145/3393880](https://doi.org/10.1145/3393880).
- Quanjiang Guo, Zhao Kang, Ling Tian, and Zhouguo Chen. TieFake: Title-text similarity and emotion-aware fake news detection. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, jun 2023.
- Maissae Haddouchi and Abdelaziz Berrado. A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pages 1–6. IEEE, 2019.
- Luke Harper, Katherine W Herbst, Dàrius Bagli, Martin Kaefer, Goedeke MA Beckers, Magdalena Fossum, Nicolas Kalfa, et al. The battle between fake news and science. *Journal of pediatric urology*, 16(1):114–115, 2020.
- Lumbardha Hasimi and Aneta Poniszewska-Maranda. Detection of disinformation and content filtering using machine learning: implications to human rights and freedom of speech. In *ROMCIR@ ECIR*, pages 68–77, 2024.
- Ebtihal A Hassan and Farid Meziane. A Survey on Automatic Fake News Identification Techniques for Online and Socially Produced Data. In *2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pages 1–6. IEEE, September 2019. ISBN 978-1-7281-1006-6. doi: [10.1109/ICCCEEE46830.2019.9070857](https://doi.org/10.1109/ICCCEEE46830.2019.9070857).
- Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146, 2022.

- Vaishali Vaibhav Hirlekar and Arun Kumar. Natural Language Processing based Online Fake News Detection Challenges – A Detailed Review. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 748–754. IEEE, June 2020. ISBN 978-1-7281-5371-1. doi: 10.1109/ICCES48766.2020.9137915.
- May Me Me Hlaing and Nang Saing Moon Kham. Defining news authenticity on social media using machine learning approach. In *2020 IEEE Conference on Computer Applications (ICCA)*, pages 1–6, 2020. doi: 10.1109/ICCA49400.2020.9022837.
- Joe Hoft. EXCLUSIVE: Analysis of Votes in Illinois Makes No Sense - Indicates Election Fraud Occurred Across the Entire Country — The Gateway Pundit — by Joe Hoft, 2020.
- Isabel Holmes, Timothy Cribbin, and Nelli Ferenczi. Style over substance: A psychologically informed approach to feature selection and generalisability for author classification. *Computers in Human Behavior Reports*, 9:100267, 2023.
- Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, 2017.
- Benjamin D Horne, Jeppe Norregaard, and Sibel Adali. Robust Fake News Detection Over Time and Attack. *ACM Trans. Intell. Syst. Technol.*, 11(1), dec 2019. ISSN 2157-6904. doi: 10.1145/3363818.
- Benjamin D Horne, Jeppe Nørregaard, and Sibel Adali. Robust Fake News Detection Over Time and Attack. *ACM Transactions on Intelligent Systems and Technology*, 11(1):1–23, February 2020. ISSN 2157-6904. doi: 10.1145/3363818.
- Benjamin D Horne, Dorit Nevo, and Susan L Smith. Ethical and safety considerations in automated fake news detection. *Behaviour & Information Technology*, pages 1–22, 2023.
- Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 873–882, 2012.
- Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and

- infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, 2021.
- Maria Janicka, Maria Pszona, and Aleksander Wawer. Cross-Domain Failures of Fake News Detection. *Computación y Sistemas*, 23(3):1089–1097, October 2019a. ISSN 2007-9737. doi: 10.13053/cys-23-3-3281.
- Maria Janicka, Maria Pszona, and Aleksander Wawer. Cross-Domain Failures of Fake News Detection. *Computación y Sistemas*, 23(3):1089–1097, oct 2019b. ISSN 2007-9737. doi: 10.13053/cys-23-3-3281.
- Nael Jebril, Matthew Loveless, and Vaclav Stetka. Media and democratisation: Challenges for an emerging sub-field. *Medijske Studije*, 6(11):84–98, 2015. ISSN 18485030.
- Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. Nothing stands alone: Relational fake news detection with hypergraph neural networks. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, dec 2022.
- Yuqian Jiang, Huaizhong Lin, Xuesong Wang, and Dongming Lu. A technique for improving the performance of naive bayes text classification. In *Web Information Systems and Mining: International Conference, WISM 2011, Taiyuan, China, September 24-25, 2011, Proceedings, Part II*, pages 196–203. Springer, 2011.
- Jillani. Fake or Real News. <https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news/data>, n.d. Accessed: 2024-10-05.
- Rohit Kumar Kaliyar and Navya Singh. Misinformation Detection on Online Social Media-A Survey. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, July 2019. ISBN 978-1-5386-5906-9. doi: 10.1109/ICCCNT45670.2019.8944587.
- Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. EchoFakeD: improving fake news detection in social media with an efficient deep neural network. *Neural Comput. Appl.*, 33(14):8597–8613, jan 2021.
- Aravind Kumar Kalusivalingam, Amit Sharma, Neha Patel, and Vikram Singh. Leveraging random forests and gradient boosting for enhanced predictive analytics in operational efficiency. *International Journal of AI and ML*, 3(9), 2022.
- Jackson Kamiri and Geoffrey Mariga. Research methods in machine learning: A content analysis. *International Journal of Computer and Information Technology (2279-0764)*, 10(2):78–91, 2021.

- Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. Scrutinizer. *Proceedings of the VLDB Endowment*, 13(12):2965–2968, aug 2020. ISSN 2150-8097. doi: 10.14778/3415478.3415520.
- Harveen Kaur. Using network analysis to detect fake news in social media. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, jul 2023.
- Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.
- Barbara Kitchenham, Rialette Pretorius, David Budgen, O Pearl Brereton, Mark Turner, Mahmood Niazi, and Stephen Linkman. Systematic literature reviews in software engineering – A tertiary study. *Information and Software Technology*, 52(8):792–805, August 2010. ISSN 09505849. doi: 10.1016/j.infsof.2010.03.006.
- Vitaly Klyuev. Fake News Filtering: Semantic Approaches. In *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 9–15. IEEE, August 2018. ISBN 978-1-5386-4692-2. doi: 10.1109/ICRITO.2018.8748506.
- Shimon Kogan, Tobias J. Moskowitz, and Marina Niessner. Social Media and Financial News Manipulation. *SSRN Electronic Journal*, mar 2022. doi: 10.2139/SSRN.3237763.
- Rafał Kozik, Michał Choraś, Sebastian Kula, and Marek Pawlicki. Distributed Architecture for Fake News Detection. *Advances in Intelligent Systems and Computing*, 1267 AISC:208–217, 2021. ISSN 2194-5365. doi: 10.1007/978-3-030-57805-3_20. URL https://link.springer.com/chapter/10.1007/978-3-030-57805-3_20.
- Naresh Kumar, Meetu Malhotra, Bharti Aggarwal, Dinesh Rai, and Gaurav Aggarwal. Leveraging natural language processing and machine learning for efficient fake news detection. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*. IEEE, nov 2023.
- Mayank Kumar Jain, Dinesh Gopalani, Yogesh Kumar Meena, and Rajesh Kumar. Machine Learning based Fake News Detection using linguistic features and word vector features. In *2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, nov 2020.
- Ohjoon Kwon, Dohyun Kim, Soo-Ryeon Lee, Junyoung Choi, and SangKeun Lee. Handling out-of-vocabulary problem in hangeul word embeddings. In Paola

- Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3213–3221, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.280. URL <https://aclanthology.org/2021.eacl-main.280>.
- Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. A survey on stylistic text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE, 2019.
- Yasmine Lahlou, Sanaa El Fkihi, and Rdouan Faizi. Automatic detection of fake news on online platforms: A survey. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pages 1–4. IEEE, October 2019. ISBN 978-1-7281-4368-2. doi: 10.1109/ICSSD47982.2019.9002823.
- Tom Leighton and Paul Sagan. The internet & the future of news. *Daedalus*, 139 (2):119–125, apr 2010. ISSN 00115266.
- Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13 (2):1–41, 2022.
- Zhiping Liang. Fake news detection based on multimodal inputs. *Comput. Mater. Contin.*, 75(2):4519–4534, 2023.
- W. Lifferth. Kaggle (Fake News) Dataset. <https://kaggle.com/competitions/fake-news>, 2018. Accessed: 2024-10-04.
- Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. Acquiring background knowledge to improve moral value prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559, 2018. doi: 10.1109/ASONAM.2018.8508244.
- Alexandra L’heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5:7776–7797, 2017.
- M. Risdal. Getting Real about Fake News Dataset. <https://www.kaggle.com/datasets/mrisdal/fake-news>, n.d. Accessed: 2024-10-04.

- S Madhusudhan, S Mahurkar, and S K Nagarajan. Attributional analysis of Multi-Modal Fake News Detection Models (Grand Challenge). In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 451–455, 2020. ISBN VO -. doi: 10.1109/BigMM50055.2020.00074.
- Zaitul Iradah Mahid, Selvakumar Manickam, and Shankar Karuppayah. Fake News on Social Media: Brief Review on Detection Techniques. In *Proceedings - 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018*, 2018. ISBN 9781538671672. doi: 10.1109/ICACCAF.2018.8776689.
- Fahim Belal Mahmud, Mahi Md Sadek Rayhan, Mahdi Hasan Shuvo, Islam Sadia, and Md Kishor Morol. A comparative analysis of Graph Neural Networks and commonly used machine learning algorithms on fake news detection. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*. IEEE, mar 2022.
- Deepak Mangal and Dilip Kumar Sharma. Fake News Detection with Integration of Embedded Text Cues and Image Features. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 68–72. IEEE, June 2020. ISBN 978-1-7281-7016-9. doi: 10.1109/ICRITO48877.2020.9197817.
- Syed Ishfaq Manzoor, Jimmy Singla, and Nikita. Fake News Detection Using Machine Learning approaches: A systematic Review. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 230–234. IEEE, April 2019. ISBN 978-1-5386-9439-8. doi: 10.1109/ICOEI.2019.8862770.
- Elio Masciari, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. Detecting fake news by image analysis. In *Proceedings of the 24th Symposium on International Database Engineering & Applications, IDEAS '20*, pages 1–5, New York, NY, USA, aug 2020. ACM. ISBN 9781450375030. doi: 10.1145/3410566.3410599.
- Katharine Eisaman Maus. Fake News. *New Literary History*, 51(1):249–252, dec 2020. ISSN 1080-661X. doi: 10.1353/NLH.2020.0014.
- George McIntire. How to Build a "Fake News" Classification Model - Open Data Science - Your News Source for AI, Machine Learning & more. <https://opendatascience.com/how-to-build-a-fake-news-classification-model/>, 2017.

- Priyanka Meel and Dinesh Kumar Vishwakarma. HAN, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences*, 567:23–41, August 2021. ISSN 00200255. doi: 10.1016/j.ins.2021.03.037.
- Alessio Miaschi and Felice Dell’Orletta. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, and Hannaneh Hajishirzi, editors, *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.15. URL <https://aclanthology.org/2020.repl4nlp-1.15>.
- Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.
- Amy Moreno. BREAKING : Pepsi STOCK Plummets After CEO Tells Trump Supporters to “Take Their Business Elsewhere” – TruthFeed, 2016.
- Susan Morgan. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, 3(1):39–43, jan 2018. ISSN 2373-8871. doi: 10.1080/23738871.2018.1462395.
- Muhammad Imran Nadeem, Kanwal Ahmed, Dun Li, Zhiyun Zheng, Hend Khalid Alkahtani, Samih M. Mostafa, Orken Mamyrbayev, and Hala Abdel Hameed. Efnd: A semantic, visual, and socially augmented deep framework for extreme fake news detection. *Sustainability*, 15(1), 2023a. ISSN 2071-1050. doi: 10.3390/su15010133.
- Muhammad Imran Nadeem, Kanwal Ahmed, Zhiyun Zheng, Dun Li, Muhammad Assam, Yazeed Yasin Ghadi, Fatemah H Alghamedy, and Elsayed Tag Eldin. SSM: Stylometric and semantic similarity oriented multimodal fake news detection. *J. King Saud Univ. - Comput. Inf. Sci.*, 35(5):101559, may 2023b.
- Manu Nandan, Pramod P Khargonekar, and Sachin S Talathi. Fast svm training using approximate extreme points. *The Journal of Machine Learning Research*, 15(1):59–98, 2014.
- Lisa Napoli. *Up All Night: Ted Turner, CNN, And the Birth of 24-Hour News*. Abrams Press, dec 2020. ISBN 9781417943061.
- Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

- Keshav Nath, Priyansh Soni, Anjum, Aman Ahuja, and Rahul Katarya. Study of Fake News Detection using Machine Learning and Deep Learning Classification Methods. In *2021 6th International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2021*, pages 434–438. Institute of Electrical and Electronics Engineers Inc., August 2021. ISBN 9781665435598. doi: 10.1109/RTEICT52294.2021.9573583.
- John C. Nerone. The mythology of the penny press. *Critical Studies in Mass Communication*, 4(4):376–404, dec 1987. ISSN 0739-3180. doi: 10.1080/15295038709360146.
- Todd G Nick and Kathleen M Campbell. Logistic regression. *Topics in biostatistics*, pages 273–301, 2007.
- Halyna Padalko, Vasyl Chomko, and Dmytro Chumachenko. A novel approach to fake news classification using LSTM-based deep learning models. *Frontiers in Big Data*, 6:1320800, jan 2023. ISSN 2624909X. doi: 10.3389/FDATA.2023.1320800/BIBTEX.
- Shivam B Parikh and Pradeep K Atrey. Media-Rich Fake News Detection: A Survey. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 436–441. IEEE, April 2018. ISBN 978-1-5386-1857-8. doi: 10.1109/MIPR.2018.00093.
- D Paschalides, C Christodoulou, R Andreou, G Pallis, M D Dikaiakos, A Kornilakis, and E Markatos. Check-It: A plugin for Detecting and Reducing the Spread of Fake News and Misinformation on the Web. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 298–302, 2019. ISBN VO -.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic Detection of Fake News. In *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, pages 3391–3401, aug 2017. ISBN 9781948087506.
- Francesco Pierri and Stefano Ceri. False News On Social Media: A Data-Driven Survey. *ACM SIGMOD Record*, 48(2):18–27, December 2019. ISSN 0163-5808. doi: 10.1145/3377330.3377334.
- AMIT PURUSHOTTAM Pimpalkar and R Jeberson Retna Raj. Influence of pre-processing strategies on the performance of ml classifiers exploiting tf-idf and bow features. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 9(2):49, 2020.

- Julliano Trindade Pintas, Leandro A.F. Fernandes, and Ana Cristina Bicharra Garcia. Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review* 2021 54:8, 54(8):6149–6200, feb 2021. ISSN 1573-7462. doi: 10.1007/S10462-021-09970-6.
- G. Pontes. Fake News Sample Dataset. <https://www.kaggle.com/datasets/pontes/fake-news-sample>, n.d. Accessed: 2024-10-04.
- Wisam A Qader, Musa M Ameen, and Bilal I Ahmed. An overview of bag of words; importance, implementation, applications, and challenges. In *2019 international engineering conference (IEC)*, pages 200–204. IEEE, 2019.
- Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. Deception detection on social media: A source-based perspective. *Knowl. Based Syst.*, 256(109649):109649, nov 2022.
- R. Jain. Kaggle ‘Fake News Detection’ Dataset. <https://www.kaggle.com/datasets/jruvika/fake-news-detection>, n.d. Accessed: 2024-10-04.
- R. Raghavendra and M. Niranjanamurthy. An Effective Hybrid Model for Fake News Detection in Social Media Using Deep Learning Approach. *SN Computer Science*, 5(4):1–15, apr 2024. ISSN 26618907. doi: 10.1007/S42979-024-02698-4/FIGURES/15.
- Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*, 3:98–105, June 2022. ISSN 26663074. doi: 10.1016/j.ijcce.2022.03.003.
- Lee Rainie, Janna Anderson, and Jonathan Albright. The Future of Free Speech, Trolls, Anonymity and Fake News Online — Pew Research Center. *PEW Internet*, 2017.
- Chahat Raj and Priyanka Meel. ConvNet frameworks for multi-modal fake news detection. *Appl. Intell.*, 51(11):8132–8148, nov 2021.
- Chahat Raj, Anjishnu Mukherjee, and Ziwei Zhu. True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 962–963, New York, NY, USA, aug 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604760.
- Dipti P Rana, Isha Agarwal, and Anjali More. A Review of Techniques to Combat The Peril of Fake News. In *2018 4th International Conference on Computing*

- Communication and Automation (ICCCA)*, pages 1–7. IEEE, December 2018. ISBN 978-1-5386-6947-1. doi: 10.1109/CCAA.2018.8777676.
- Shaina Raza and Chen Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362, may 2022. ISSN 2364-415X. doi: 10.1007/s41060-021-00302-z.
- H Reddy, N Raj, M Gala, and A Basava. Text-mining-based Fake News Detection Using Ensemble Methods. *International Journal of Automation and Computing*, 17(2):210–221, 2020. doi: 10.1007/s11633-019-1216-5.
- Tanveer Rehman, Gayathri Surendran, and Yuvaraj Krishnamoorthy. Developing Counter Strategy for Information Warfare in Health Sector–Sifting ‘Real’ from ‘Fake’ News. *International Journal of Medicine and Public Health*, 12(2):46–49, apr 2022. ISSN 22308598. doi: 10.5530/ijmedph.2022.2.10.
- Julio C S Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benvenuto. Explainable Machine Learning for Fake News Detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci ’19, pages 17–26. ACM, jun 2019. ISBN 9781450362023. doi: 10.1145/3292522.3326027.
- Philip Resnik and Jimmy Lin. Evaluation of nlp systems. *The handbook of computational linguistics and natural language processing*, pages 271–295, 2010.
- Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health (Germany)*, 31(7):1007–1016, jul 2023. ISSN 16132238. doi: 10.1007/S10389-021-01658-Z/TABLES/4.
- Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2019.
- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
- Diego Saez-Trumper. Fake tweet buster. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 316–317, New York, NY, USA, sep 2014. ACM. ISBN 9781450329545. doi: 10.1145/2631775.2631786.

- Pallabi Saikia, Kshitij Gundale, Ankit Jain, Dev Jadeja, Harvi Patel, and Mohendra Roy. Modelling social context for fake news detection: A graph neural network based approach. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, jul 2022.
- Dietram A Scheufele and Nicole M Krause. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669, 2019.
- Karl-Michael Schneider. Techniques for improving the performance of naive bayes for text classification. In *International conference on intelligent text processing and computational linguistics*, pages 682–693. Springer, 2005.
- Bernhard Schölkopf. The kernel trick for distances. *Advances in neural information processing systems*, 13, 2000.
- Noureddine Seddari, Abdelouahid Derhab, Mohamed Belaoued, Waleed Halboob, Jalal Al-Muhtadi, and Abdelghani Bouras. A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media. *IEEE Access*, 10:62097–62109, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3181184.
- S Selva Birunda and R Kanniga Devi. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, pages 267–281, 2021.
- Sanchari Sen and Anand Raghunathan. Approximate computing for long short term memory (lstm) neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(11):2266–2276, 2018.
- Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating Fake News: A Survey on Identification and Mitigation Techniques. *arXiv*, 10(3), January 2019a. ISSN 23318422.
- Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3), apr 2019b. ISSN 2157-6904. doi: 10.1145/3305260.
- Sunidhi Sharma and Dilip Kumar Sharma. Fake News Detection: A long way to go. In *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, pages 816–821. IEEE, November 2019. ISBN 978-1-7281-3651-6. doi: 10.1109/ISCON47742.2019.9036221.

- Saeid Sheikhi. An effective fake news detection method using WOA-xgbTree algorithm and content-based features. *Appl. Soft Comput.*, 109(107559):107559, sep 2021.
- Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, sep 2017. ISSN 1931-0145. doi: 10.1145/3137597.3137600.
- Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The Role of Social Context for Fake News Detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, New York, NY, USA, jan 2019a. ACM.
- Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439, New York, NY, USA, August 2019b. ACM. ISBN 9781450368681. doi: 10.1145/3341161.3342927.
- Abubukar Siddique, KN Harikishan, Mohammed Kaif, Nachiketha N Gowda, VM Aparanji, and S Karthik. Crop prediction using npk sensor: Novel approach using machine learning algorithms. In *2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, pages 1–6. IEEE, 2024.
- Shivangi Singhal, Mudit Dhawan, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Inter-modality discordance for multimodal fake news detection. In *ACM Multimedia Asia*, New York, NY, USA, dec 2021. ACM.
- Dylan Slack, Sorelle A Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501*, 2019.
- Chris Snijders, Rianne Conijn, Evie de Fouw, and Kilian van Berlo. Humans and Algorithms Detecting Fake News: Effects of Individual and Contextual Confidence on Trust in Algorithmic Advice. *International Journal of Human–Computer Interaction*, 39(7):1483–1494, 2023. ISSN 15327590. doi: 10.1080/10447318.2022.2097601.

- Kayato Soga, Soh Yoshida, and Mitsuji Muneyasu. Confirmation bias-aware fake news detection with graph transformer networks. In *2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*. IEEE, oct 2023.
- Francesca Spezzano, Anu Shrestha, Jerry Alan Fails, and Brian W. Stone. That’s Fake News! Reliability of News When Provided Title, Image, Source Bias & Full Article. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–19, April 2021. ISSN 2573-0142. doi: 10.1145/3449183.
- Nitin Srinivasan, Kishore Kumar Perumalsamy, Praveen Kumar Sridhar, Gowthamaraj Rajendran, and Adithyan Arun Kumar. Comprehensive study on bias in large language models. *International Refereed Journal of Engineering and Science*, 13(2):77–82, 2024.
- Ewout W Steyerberg and Frank E Harrell Jr. Prediction models need appropriate internal, internal-external, and external validation. *Journal of clinical epidemiology*, 69:245, 2015.
- Xing Su, Jian Yang, Jia Wu, and Yuchen Zhang. Mining User-aware Multi-relations for Fake News Detection in Large Scale Online Social Networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM ’23*, pages 51–59, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570478.
- Mihai Surdeanu and Marco Antonio Valenzuela-Escárcega. *Feed-Forward Neural Networks*, page 73–86. Cambridge University Press, 2024.
- Simen Sverdrup-Thygeson and Pauline C Haddow. Feature selection for fake news classification. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, dec 2021.
- Ayisha Tabassum and Rajendra R Patil. A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, 7(06):4864–4867, 2020.
- Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.
- Ludovic Terren Ludovic Terren and Rosa Borge-Bravo Rosa Borge-Bravo. Echo chambers on social media: A systematic review of the literature. *Review of Communication Research*, 9, 2021.

- Kai Ming Ting and Zijian Zheng. A study of adaboost with naive bayesian classifiers: Weakness and improvement. *Computational Intelligence*, 19(2):186–200, 2003.
- Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big data & society*, 6(1):2053951719843310, 2019.
- Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, 13(9):e0203958, 2018.
- Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake News Detection in Social Networks via Crowd Signals. In *Companion Proceedings of the The Web Conference 2018*, WWW ’18, pages 517–524, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3188722.
- Sunny Emmanuel Udeze and Emmanuel Uzuegbunam. Sensationalism in the media: the right to sell or the right to tell? *Journal of Communication and Media Research*, 5(1):69–78, 2013.
- A Uppal, V Sachdeva, and S Sharma. Fake news detection using discourse segment structure analysis. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 751–756, 2020. ISBN VO -. doi: 10.1109/Confluence47617.2020.9058106.
- Sander van der Linden, Jon Roozenbeek, and Josh Compton. Inoculating Against Fake News About COVID-19. *Frontiers in Psychology*, 11, oct 2020. ISSN 16641078. doi: 10.3389/fpsyg.2020.566790.
- Pankaj Kumar Varshney and Ganesh Kumar Wadhwani. Systematic approach for fake news detection using machine learning. *Multimedia Tools and Applications*, pages 1–10, dec 2023. ISSN 15737721. doi: 10.1007/S11042-023-17913-2/FIGURES/6.
- Gaurav Verma and Balaji Vasan Srinivasan. A lexical, syntactic, and semantic perspective for understanding style in text. *arXiv preprint arXiv:1909.08349*, 2019.
- Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893, 2021. doi: 10.1109/TCSS.2021.3068519.

- István Kornél Vida. The "Great Moon Hoax" of 1835. *Hungarian Journal of English and American Studies (HJEAS)*, 18(1/2):431–441, 2012. ISSN 12187364.
- S Vijayarani, Ms J Ilamathi, Ms Nithya, et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.
- Dinesh Kumar Vishwakarma and Chhavi Jain. Recent State-of-the-art of Fake News Detection: A Review. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, June 2020. ISBN 978-1-7281-6221-8. doi: 10.1109/INCET49848.2020.9153985.
- Nguyen Vo and Kyumin Lee. Learning from fact-checkers: Analysis and generation of fact-checking language. *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344, jul 2019. doi: 10.1145/3331184.3331248/SUPPL_FILE/CITE3-17H40-D1.MP4.
- Putra Wanda and Mohammad Diqi. DeepNews: enhancing fake news detection using generative round network (GRN). *International Journal of Information Technology (Singapore)*, 16(7):4289–4298, oct 2024. ISSN 25112112. doi: 10.1007/S41870-024-02017-3/TABLES/7.
- Carol A. Watson. Information Literacy in a Fake/False News World: An Overview of the Characteristics of Fake News and its Historical Development. *International Journal of Legal Information*, 46(2):93–96, jul 2018. ISSN 0731-1265. doi: 10.1017/JLI.2018.25.
- Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714, 2010.
- Johannes Weber. Strassburg, 1605: The Origins of the Newspaper in Europe. *German History*, 24(3):387–412, jul 2006. ISSN 0266-3554. doi: 10.1191/0266355406GH380OA.
- Thomas Welchowski, Kelly O Maloney, Richard Mitchell, and Matthias Schmid. Techniques to improve ecological interpretability of black-box machine learning models: Case study on biological health of streams in the united states with gradient boosted trees. *Journal of Agricultural, Biological and Environmental Statistics*, 27(1):175–197, 2022.
- Kathryn S. Wenner. Peeling the Onion. *American Journalism Review*, 24(7):49–55, sep 2002. ISSN 10678654.

- Joel H. Wiener. The Beginnings of Sensationalism. *Palgrave Studies in the History of the Media*, pages 28–53, 2011. ISSN 2634-6583.
- Ian H Witten, Eibe Frank, Mark A Hall, Christopher J Pal, and Mining Data. Practical machine learning tools and techniques. In *Data mining*, volume 2, pages 403–413. Elsevier Amsterdam, The Netherlands, 2005.
- Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. Prompt-and-Align: Prompt-Based Social Alignment for Few-Shot Fake News Detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23*, pages 2726–2736, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615015.
- Lei Wu, Steven CH Hoi, and Nenghai Yu. Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing*, 19(7):1908–1920, 2010.
- Shangyuan Wu. What Motivates Audiences to Report Fake News?: Uncovering a Framework of Factors That Drive the Community Reporting of Fake News on Social Media. *Digital Journalism*, 2023. ISSN 2167082X. doi: 10.1080/21670811.2023.2243489.
- Xiaojun Wu and Jimin Wang. SAFS: Social-article features-stacking model for fake news detection. In *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, may 2021.
- Yi Xie, Xixuan Huang, Xiaoxuan Xie, and Shengyi Jiang. A Fake News Detection Framework Using Social User Graph. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering, BDE 2020*, pages 55–61, New York, NY, USA, May 2020a. ACM. ISBN 9781450377225. doi: 10.1145/3404512.3404515.
- Yi Xie, Xixuan Huang, Xiaoxuan Xie, and Shengyi Jiang. A Fake News Detection Framework Using Social User Graph. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering, BDE 2020*, pages 55–61, New York, NY, USA, may 2020b. ACM. ISBN 9781450377225. doi: 10.1145/3404512.3404515.
- Shufeng Xiong, Gupei Zhang, Vishwash Batra, Lei Xi, Lei Shi, and Liangliang Liu. TRIMOON: Two-Round Inconsistency-based Multi-modal fusion Network for fake news detection. *Inf. Fusion*, 93:150–158, may 2023.
- Lanxin Yang, He Zhang, Haifeng Shen, Xin Huang, Xin Zhou, Guoping Rong, and Dong Shao. Quality Assessment in Systematic Literature Reviews: A Software Engineering Perspective. *Information and Software Technology*, 130:106397, February 2021. ISSN 09505849. doi: 10.1016/j.infsof.2020.106397.

- Seyhmus Yilmaz and Sinan Toklu. A deep learning analysis on question classification task using word2vec representations. *Neural Computing and Applications*, 32(7): 2909–2928, 2020.
- Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *Journal of Data and Information Quality*, 11(3):1–37, September 2019. ISSN 1936-1955. doi: 10.1145/3309699.
- Guobiao Zhang, Anastasia Giachanou, and Paolo Rosso. SceneFND: Multimodal fake news detection by modelling scene context information. *J. Inf. Sci.*, page 016555152210876, apr 2022.
- Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.
- Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2): 102025, March 2020. ISSN 03064573. doi: 10.1016/j.ipm.2019.03.004.
- Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52, 2010.
- Guifen Zhao, Yanjun Liu, Wei Zhang, and Yiyou Wang. Tfidf based feature words extraction and topic modeling for short text. In *Proceedings of the 2018 2nd international conference on management engineering, software engineering and service sciences*, pages 188–191, 2018.
- Yunze Zhao and Ping Sun. *The communication mechanism in ancient China*. Routledge, sep 2018. doi: 10.4324/9781315720555-2/COMMUNICATION-MECHANISM-ANCIENT-CHINA-YUNZE-ZHAO-PING-SUN.
- Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.
- Xinyi Zhou and Reza Zafarani. Network-based fake news detection. *SIGKDD Explor.*, 21(2):48–60, nov 2019.
- Xinyi Zhou and Reza Zafarani. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.*, 53(5), sep 2020. ISSN 0360-0300. doi: 10.1145/3395046.

Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake News: Fundamental Theories, Detection Strategies and Challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 836–837, New York, NY, USA, January 2019. ACM. ISBN 9781450359405. doi: 10.1145/3289600.3291382.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. Multimodal fake news detection via CLIP-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, jul 2023.