MULTILINGUAL SENTIMENT ANALYSIS OF ARABIC, BAHRAINI DIALECTS AND ENGLISH

A Thesis Submitted for the Degree of Doctor of Philosophy

By

Thuraya Mohamed Maki Omran

Department of Computer Science, Brunel University London

October 2022

Declaration

I want to acknowledge that some parts of this thesis have been published and others have been accepted for publishing.

The first published part was in (Omran, T.M., Sharif, B.T., Grosan, C. and Li, Y., 2022. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, p.102106). This part includes the following sections:

Section 1.1"Introduction", section 2.3"Modern standard Arabic (MSA) and dialects-definition and distinction", section 2.4"Available Arabic dataset for sentiment analysis", section 2.6 "Multilingual sentiment analysis", section 2.12.2 "Related works" of transfer learning, section 4.2 "Proposed LSTM model", section 4.2.1"LSTM model design, implementation and configuration", section 4.2.2"LSTM model compiling and training", section 4.3 "Experiments and results", section 6.2"The pre-trained model and transfer learning", section 6.3"Dataset design and preparation", section 6.4"Dataset preprocessing", section 6.5"Pre-trained LSTM model creation", section 6.6"Experiments and results", chapter 7, and appendices.

The other published part includes section 2.2 "Challenges of Arabic sentiment analysis", section 2.4 "Available Arabic datasets for sentiment analysis", and section 2.8 "Deep learning in sentiment analysis of Arabic and other languages". These sections were published as a part of a chapter in a book in (Omran, T., Sharef, B.T. and Grosan, C., 2021. Sentiment analysis of Arabic sequential data using traditional and deep Learning: A Review. *The Fourth Industrial Revolution: Implementation of Artificial Intelligence for Growing Business Success*, pp.439-459).

The accepted parts that will be published in IEEE Xplore are as follows:

- Section 2.10 "Data Augmentation" and its experimental part conducted in the preprocessing steps of section 3.3. "Dataset preprocessing". This part was

submitted as a conference paper entitled “The impact of data augmentation on sentiment analysis of translated textual data”.

- Section 2.5 “Impact of translation on sentiment analysis”, section 2.11 “Ensemble learning and sentiment analysis, section 5.2 “Proposed LSTM model and ensemble learning”, section 5.3 “Experiments and results “,and part of chapter 7 “Conclusion and future work”. This part was submitted as a conference paper entitled “Ensemble learning for sentiment analysis of translation-based textual data”.

Acknowledgment

First, all thanks to God for providing me health, wellness, patience, success and facilitation and helping me at all times, including working on this research.

Secondly, my great thanks, gratitude, and appreciation to all of my supervisors, starting from Dr.Baraa Sharif, for his enthusiasm, exceptional support, and interest in details which have been an inspiration and kept my work on track, from proposal preparation to the final draft of this research, for his time during holidays, continuous encouragement and cooperation, especially in completing the approval procedures of collecting research data at Ahlia University.

I am also grateful to Dr.Crina Grosan for her moral support and continuous suggestions that opened up a more comprehensive research horizon. All the gratitude to Dr.Yongmin Li, who took over my supervision in my third year, for offering valuable and insightful comments, and constructive criticism on the research papers and documents.

Thirdly, I greatly appreciate all participants for their time and efforts in enriching this research by rewriting the standard Arabic dataset into their Bahraini dialects,

Finally, I thank my family members for their continued interest and support. Thanks to everyone who has cared about my success in my PhD journey.

Contents

List of Acronyms

Abbreviation	Description
AD	Arabic Dialect
BDs	Bahraini Dialects
DA	Data Augmentation
GRU	Gated Recurrent Unit
LSTM	Long Short- Term Memory
MSA	Modern Standard Arabic
RCNN	Recurrent Convolutional Neural Network
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SVM	Support Vector Machine
TL	Transfer Learning

List of Tables

List of Figures

Abstract

Sentiment analysis is a crucial natural language processing (NLP) task to analyze the user's emotions and opinions towards entities such as events, services, or products. Arabic NLP faces numerous challenges, some of which include: (1) the scarcity of resources, especially in modern standard Arabic and Arabic dialects, particularly the Bahraini one; (2) the lack of multilingual deep learning models; and (3) insufficient transfer learning studies on Arabic dialects in general and Bahraini dialects specifically. This research aims to create a balanced dataset of Bahraini dialects that covers product reviews by translating English Amazon product reviews to modern standard Arabic, which were then converted to Bahraini dialects. Another aim of this research is to provide a multilingual deep learning long short-term memory (LSTM) model to analyze the parallel dataset of English, modern standard Arabic, and Bahraini dialects, which differ in linguistic properties. Many experiments were conducted using train-validate-test split and k-fold cross-validation to evaluate the model performance using accuracy, F1 score, and AUC metrics. The average accuracy of the model on all datasets ranged from 96.72% to 97.04% and 97.91% to 97.93% in the F1 score, while in AUC was 98.46% to 98.7% when utilizing an augmentation technique. The LSTM model was incorporated in a stacking ensemble learning process that includes other LSTM architectures as base learners and a decision tree (DT) as a meta-learner. Interestingly, promising results were obtained, such as 99.52%, 99.25%, and 98.52% of mean accuracy for English, MSA, and BDs datasets. Moreover, the LSTM model was utilized as a pre-trained model in the transfer learning process to exploit the knowledge gained from analyzing the product reviews in Bahraini dialects to perform another sentiment analysis task on a small dataset of movie comments in the same dialects. The pre-trained model performance was 96.97% accuracy, 96.65% F1 score, and 97.94% AUC.

Chapter 1 – Introduction

1.1 Introduction

Part of the work included in this chapter was published in (Omran, T.M., Sharif, B.T., Grosan, C. and Li, Y., 2022. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, p.102106).

With the web revolution, expressing opinions is not restricted to questionnaires or surveys but extended to a broader range with the emergence of social media, blogs, forums and review sites. These opinions led to automated opinion mining or sentiment analysis (SA) applications (Al Sallab *et al*., 2015).

SA is a computational study of peoples' emotions, attitudes, and opinions toward entities such as organizations, individuals, topics, events, services, and products (Zhang, Wang & Liu, 2018) (Munezero *et al*., 2013), whether these emotions were positive or not (Pawar, Jawale & Kyatanavar, 2016).

In the field of NLP, there is a need for more clarity between SA and emotion detection. SA analyzes the text to certain polarities such as positive, negative, and neutral, whereas emotion detection analyzes the text to several emotional states like anger, fear, joy, disgust, or surprise (Kaur, Saini, 2014).

In recent years, SA has received much attention as an area of natural language processing (NLP) (Fouadi *et al*., 2020). It gained significant importance in academics, industry, social media, business, and politics (Badaro *et al*., 2019).

SA represents a powerful tool for governments, researchers, and businesses to investigate the public mood and opinions of people and analyze them, which leads to better business insights (Hajiali, 2020) (Birjali, Kasri & Beni-Hssane, 2021), and make better decisions (Birjali, Kasri & Beni-Hssane, 2021).

SA can be carried out at different levels like document level, sentence level and aspect level to classify sentiment polarity as positive, negative or neutral. This research interested in the document level, whereas the concerned polarities are positive and negative. The Positive and negative polarities are the most considered ones by the

stakeholders. The positive polarity of the text reflects the accepted features that need to be kept, the negative polarity reflects the undesirable aspects that need to be improved, and the neutral ones were not considered here due to the absence and ambiguity of the sentiment.

There are many techniques, methods, and tools for conducting SA, such as lexicons, machine learning (Fouadi *et al*., 2020) and hybrid techniques (Ahmad *et al*., 2017). Lexicons are techniques where each word is associated with an emotion or a feeling, and they can be created from corpora of existing dictionaries (Fouadi *et al*., 2020, Alessia *et al*., 2015). Machine learning techniques are categorized as supervised, unsupervised, and semi-supervised (Ahmad *et al*., 2017, Boudad *et al*., 2018).

The supervised method requires several thousands of labelled text as training data to build a model capable of predicting the label of unseen data (Fouadi *et al*., 2020). Unlike the supervised method, the unsupervised does not need a labelled dataset but requires a lexicon that covers most emotion's words. Semi-supervised is a method where annotated data and much-unlabelled data are used to build a better learning model (Fouadi *et al*., 2020). With the abundance of opinion lexical resources and natural language processing solutions, research on English SA has achieved considerable improvement and success. In contrast, in other languages like Arabic, much more effort is needed to achieve the same level of performance (Al-Sallab *et al*., 2017).

Arabic SA researches improve slowly due to many factors such as lack of Arabic datasets and pre-processing tools (Alayba *et al*., 2018), ambiguity, morphological richness system, and a large number of dialects (Al Sallab *et al*., 2015) that differ from modern standard Arabic (MSA) (Mohammed, Kora, 2019, Badaro *et al*., 2019). Hence, the Arabic language and its dialects resources have become limited compared to the English language despite the spread of the Arabic language (Mohammed, Kora, 2019).

Arabic is ranked as the fifth most used language in the world (Mohammed, Kora, 2019) (Elnagar *et al*., 2021). It is categorized into three varieties: 1- Classic Arabic (CA) used in Quran and literary texts (Sharaf, Atwell, 2012). 2- Modern Standard Arabic (MSA) is a regulated and standardized variety used in writing and formal communication and

taught in schools (Zaidan, Callison-Burch, 2014), and 3- Arabic Dialect (AD) is the spoken form of Arabic used in daily communication and unofficial exchange (Guellil, Azouaou & Mendoza, 2019). It varies from country to country and from region to region (Elnagar *et al*., 2021).

The literature review revealed a gap in SA and opinion mining in one of the Arabic dialects, such as Bahraini, where studies and datasets are lacking. There is a notable lack of empirical research focusing mainly on Bahraini dialects (BDs) SA. In addition, datasets in such dialects are scarce. Creating a dataset is difficult, especially in resource rarity, the difficulty of collecting contents, and the cost of labelling. It is expensive and requires excellent collaboration (Darwish *et al*., 2021).

There are some approaches to creating a dataset for languages that suffer from resource scarcity. One of these methods is translation.

This research presents a dataset by translating an existing dataset of rich-sources language such as English into modern standard Arabic (MSA), then converting the Arabic dataset to BDs, to tackle the scarcity of resources in the Arabic language in general and Bahraini dialects precisely.

The final SA process might be affected by the cultural differences between the target and source language, even if the translation was done perfectly (Becker *et al*., 2017). To capture the data of multilingual cultures, a multilingual SA model was created to sentimentally analyze the resulting parallel dataset of English, MSA, and BDs to provide a SA tool for more than one language. In other words, this tool uses one model for different languages instead of using a different model for each language.

Obtaining a robust model for sentiment analysis still represents a challenge (Kazmaier, van Vuuren, 2022). One of the solutions to address this challenge is ensemble learning. Ensemble learning is a technique used in machine learning that combines several base learners of a specific learning task to generate a model with a more generalized prediction than the individual learners (Ortiz *et al*., 2020). Ensemble learning methods were categorized as dependent and independent based on the interaction between the base learners. The dependent method entails a base learner that affects the construction of the

other, such as boosting. In contrast, in the independent method, the base-learner is built separately using different subsets of the dataset, followed by combining the results using some fusion methods (Mohammed, Kora, 2021). Stacking is an example of the independent method. In this research, the stacking ensemble was applied, where three LSTM base learners were trained on the English, MSA, and BDs datasets, and the results were combined using a decision tree (DT) classifier as a meta-learner.

To make an efficient ensemble, the base learners have to be independent or with a negative correlation (Mohammed, Kora, 2021, Chandra, Yao, 2006). Different biases provided by different models represent the underlying idea behind the ensemble technique. In case of these biases' un-correlation, the ensemble constituent members will try to compensate for the error of the other, which leads to decreasing overall error when the results are combined, which reduces the variance (Rokach, 2010, Kazmaier, van Vuuren, 2022). One of the obstacles that Arabic natural language processing (NLP) researchers face is the morphology richness of the Arabic language (Guellil, Azouaou & Mendoza, 2019). Added to that, it is polysemic. For example, the word "أكبر", which is pronounced as “Akbar” in Arabic that means "larger" in English, may be used as a person's noun or as a comparative adjective. These characteristics of the Arabic language and its dialects as Bahraini one necessitate and impose a robust model that can be achieved by employing an ensemble learning method to enhance the SA process.

The knowledge gained from the SA task by the multilingual model can be exploited in another related task characterized by fewer data by applying transfer learning (TL), where a pre-trained model was created. The presence of the pre-trained model that achieves well in a source task is one of the main requirements of transfer learning (Sarkar, Bali & Ghosh, 2018).

Transfer learning is a technique where knowledge gained from a rich training data source can be invested in another related but different domain (target) (Omara, Mosa & Ismail, 2019). Transfer learning has many applications using deep learning, and it could be used with audio or speech, computer vision, and textual data (Sarkar, Bali & Ghosh, 2018). TL of textual data is the case in this research, which applies a pre-trained LSTM model to analyze the sentiments of a small generated dataset of movie comments in

Bahraini dialects, which was utilized as a target domain dataset. In contrast, the Bahraini dataset of Amazon products that resulted from the translation process was utilized as a source domain dataset.

LSTM is one of the RNN feedforward networks which can predict the next word in a sequence (Abdullah, Hadzikadicy & Shaikhz, 2018) due to its architecture composed of an input gate, output gate, forget gate, and memory cell. The memory cell task stores a state or value for a long or short time using an activation function (Abdullah, Hadzikadicy & Shaikhz, 2018).

LSTM is an example of a deep learning model. The deep learning model represents inductive learning, where the model learns a mapping of the input features to the class labels through a set of assumptions related to the distribution of the training data (Sarkar, Bali & Ghosh, 2018). This learning of features occurs through a layered architecture (Sarkar, Bali & Ghosh, 2018) network to produce state-of-the-art prediction results (Zhang, Wang & Liu, 2018).

Deep learning methods are data-hungry (Sarkar, Bali & Ghosh, 2018). Subsequently, they require extensive training data to overcome data sparsity and overfitting (Sun, He, 2020). Due to the fundamental role the dataset size plays in determining the accuracy of polarity in SA dialects of Arabic (Algburi *et al*., 2019), and due to the scarcity of resources for this research, the data augmentation technique was used to enhance the performance of the deep learning model and data handling (Luque, 2019). Four powerful data augmentation consisted of the easy augmentation technique (EDA) of (Wei, Zou, 2019), such as random swap, random insertion, random deletion, and synonym replacement. The random swap augmentation technique at the word level was applied in this research to take advantage of the data augmentation (DA).

This chapter is subdivided into four sections. This first section is an introduction; section 2 presents the problem statement. Section 3 presents the aim and objectives, while section 4 describes the thesis organization.

1.2 Problem Statement

The available resources of the Arabic NLP community suffer from the scarcity of dialectal resources in general and Bahraini dialects in specific, which represents a big challenge. In addition, the NLP research community resources are dataset and language-dependent, which means that for every language model for SA, some tasks should be followed, like extracting features, model training, and tuning model parameters, which means repeating the efforts for each language model. The dataset and language-dependent model need to provide a model for different languages, which can capture the nature of data, especially with translated datasets. Therefore, this research here worked on tackling the above issues. Addressing such a challenge will enrich the Arabic NLP community with a new dataset and provide it with a model that can be utilized in more NLP applications.

1.3 Aim and Objectives

Nowadays, social media are used by many people from different countries with Multilanguage and dialects to express their opinions about various topics. Over the years, many sentiment analysis studies have been done about the English language, with insufficient research in other languages like Arabic, which seems to lack such substantial research despite the rapid growth of its use on social media outlets. That means modern standard Arabic should be studied besides specific Arabic dialects. This research bridges the gap due to its focus on one of the Arabic dialects, Bahraini. Additionally, the morphology and polysemic of the Arabic language and its dialects necessitate or impose a robust model that can be achieved by applying an ensemble learning technique to enhance the SA process. On the other hand, the lack of Arabic NLP studies that consider Multilanguage sentiment analysis motivated the researcher to conduct this research to develop a deep learning approach for sentiment analysis of a dataset of English and its corresponding ones in modern standard Arabic and Bahraini dialects.

The aim of this research is to address the challenges of NLP in the Arabic language and its dialects, specifically the Bahraini ones. To achieve this, we shall address the following objectives:

1 To determine the critical aspects of Arabic NLP and its dialects.

2 To enrich the NLP community with a parallel English, modern standard Arabic, and Bahraini dialects dataset.

3 To develop a deep learning approach using a recurrent neural network including LSTM and enhance its achievement using ensemble learning for sentiment analysis of English, standard Arabic and Bahraini dialects datasets.

4 To provide a pre-trained model in Bahraini dialects that might be used in the transfer learning process.

5 To evaluate the proposed LSTM models using the appropriate metrics.

This research significance was represented by providing a model for more than one language; in other words, avoid building a model for each language, which saves the effort of developing a separate model for standard Arabic and another for Bahraini Dialects. This research mainly proposes a deep learning approach for analyzing and predicting the included sentiment in a dataset composed of English Amazon product reviews and their corresponding ones in MSA, and BDs, in addition to boosting the SA process through one of the ensemble learning techniques. The knowledge gained from the analysis of BDs Products' reviews will be further used in SA of a small dataset of Bahraini dialect that falls in the movies domain.

1.4 Organization of Thesis

This research is organized into seven chapters; chapter 1 is an introduction chapter that gives an overview of this research, problem statement, research aim, objectives, motivation, and significance. A survey of literature reviews, including Arabic NLP issues and related works and the discussion of deep learning studies, including multilingual, ensemble learning, and transfer learning, are all presented in chapter 2. Chapter 3 describes datasets design and processing, chapter 4 covers the methods and experiments of LSTM multilingual SA, chapter 5 introduce the methods and experiments of ensemble learning, while the transfer learning methods and experiments were covered in chapter 6. Finally, chapter 7 presents the conclusion of this research and the suggested future works.

Chapter 2- Background and Related Work

2.1 Introduction

Part of the work included in this chapter was previously published in (Omran, T., Sharef, B.T. and Grosan, C., 2021. Sentiment Analysis of Arabic Sequential Data Using Traditional and Deep Learning: A Review. *The Fourth Industrial Revolution: Implementation of Artificial Intelligence for Growing Business Success*, pp.439-459).

Another part was published in (Omran, T.M., Sharif, B.T., Grosan, C. and Li, Y., 2022. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, p.102106).

This chapter covers some issues of the Arabic SA, one of which is the challenges mentioned in section 2.2. The definition of both MSA and dialects and their distinction is in section 2.3. At the same time, section 2.4 shed light on the available Arabic datasets in MSA and dialects. Section 2.5 highlights the impact of translation on SA. Section 2.6 reviewed some studies of multilingual SA. While the deep learning and neural networks section, deep learning in SA of Arabic and other languages are covered in section 2.7 and 2.8, respectively, followed by section 2.9 and its subsection that explain the theoretical background of RNN LSTM, training of LSTM, the architecture of LSTM, and the topologies of LSTM. Section 2.10 was devoted to data augmentation. Section 2.11 reviewed ensemble learning, while section 2.12 and its subsections were devoted to transfer learning. And finally, section 2.13 for the chapter summary.

2.2 Challenges of Arabic Sentiment Analysis

SA of unstructured textual data poses many challenges, especially for a morphologically rich, ambiguous orthographic language like Arabic (Badaro *et al.*, 2019), which suffers from a lack of annotated data compared to English (Baly *et al.*, 2017a). The datasets of Arabic are limited in number and size, covering limited domains such as news reviews and movie reviews, with a minimal exploration of the diverse dialects of Arabic (Algburi *et al.*, 2019).

The efforts to create a dataset, especially on tweets, are limited to particular dialects such as Jordan and the Gulf (Baly *et al.*, 2017a). Creating annotated corpora requires a more intensive effort in the Arab world (Darwish *et al.*, 2021), and incorporating more

dialects is necessary to achieve better results (Badaro *et al.,* 2019). Bahraini and Mauritanian dialects, in particular, suffer from a dearth of resources and studies for SA (Elnagar *et al.,* 2021).

Moreover, the dialects used in social media, especially Twitter, are unstandardized, besides using arabizi, special characters like mentions, hashtags, URLs, and miss-spelling of words that abide by the restricted length of tweets. These factors may implicitly impact the sentiment, mainly when modelling text semantics (Soufan, 2019) and (Baly *et al.,* 2017a).

More details are mentioned by (Alahmary, Al-Dossari & Emam, 2019), (Alsayat, Elmitwally, 2020), and (Soufan, 2019) regarding MSA and Arabic dialects' challenges, such that one word has multi and various meanings which differ from one dialect to another.

Added to that, the morphology property, grammar rules, Arabic diacritics, affixing, suffixing, prefixing, and position of the word in the sentence, resulting in different meanings for the word, adding the nature of a sentence as being verbal or nominal. All these difficulties contribute to changing the opinion polarity.

It was mentioned by (Elnagar *et al.,* 2017) that names and idioms also represent a challenge in Arabic SA. For example, the name 'سعيد' that means 'happy' in English may lead to a triggering of false-positive sentiment.

More open issues and challenges are listed by (Alsayat, Elmitwally, 2020), including 1-Translating mechanism of a figurative language while keeping its actual essence, 2-Detecting sentiment in irony, sarcasm and figurative expression, 3-Identifying euphemism, hyperbole, and metonymy, 4-Setting a mechanism for understanding the rhetorical questions, and 5- Polysemy and synonymy.

In a survey by (Guellil, Azouaou & Mendoza, 2019), more challenges were inferred. Some of these challenges are 1-The small size of manual Arabic resources and datasets, which gives accurate and best results in SA, is not adequate for deep machine learning

because of its small size; 2- Most of the works in reviewed approaches were in document level and sentence level, where aspect level seems as unexplored.

In addition to 1- Arabic words' agglutination, 2- Polysemic words in sentiment analysis may cause ambiguity in corpus-based approaches, 3- Switching in writing from Arabic to Arabizi, and 4- Stemming in lexicon-based approaches is a complex task where different stems can be estimated or thought proper for the same word.

2.3 Modern Standard Arabic (MSA) and Dialects - Definition and Distinction

Arabic is a language that includes 28 letters, consisting of 25 consonant letters and three long vowels. These letters are written from right to left, and their shape changes according to their position in the word. The meaning of words changed depending on the diacritics used (Alnawas, Arici, 2018).

Arabic represents the official language of 22 countries. It is ranked as the 1- fourth most used language on the internet (Boudad, Faizi *et al*. 2018) 2- fifth most used language around the world (Mohammed, Kora, 2019), 3- the second language in 58 countries (Alshuaibi, Mohd Shamsudin *et al*. 2015).

Arabic is categorized into three varieties: 1- Classic Arabic (CA) used in Quran and literary texts. (Sharaf, Atwell 2012). 2- Modern Standard Arabic (MSA) is a regulated and standardized variety used in writing and formal communication and taught in schools (Zaidan, Callison-Burch, 2014). 3- Arabic Dialect (AD) is a spoken form of Arabic used in daily communication and unofficial exchange (Guellil, Azouaou & Mendoza, 2019).

MSA differs significantly from the Arabic-spoken dialects, which in turn differ from one to another. There are no explicit grammatical rules for regional dialects, but some kinds of grammatical concepts exist. Despite multi varieties of spoken dialects, there are spelling rules used in creating texts of AD, which are phonetics in most cases. Exposure to Arab literature and culture, in addition to the person's dialect, are the main factors in having the ability to understand other dialects (Zaidan, Callison-Burch, 2014).

Nowadays, AD emerged as the primary language of communication in social media, both in verbal and written forms (Obeid *et al.*, 2019). The classification of AD is often done in terms of geographical regions (Obeid *et al.*, 2019), such as 1- Maghrebi (MAGH), 2- Egyptian (EGY), 3- Levantine (LEV) 4- Gulf (GLF) 5- Iraqi (IRQ) and 6- Other of the remaining countries (Guellil, Azouaou & Mendoza, 2019).

Maghrebi: The most spoken ambiguous dialects in the Middle East, affected by Barber and French languages (Tilmatine, 1999).

Egyptian: The most understood dialect, because of the movies industry and Egyptian television (Haeri, 2003)

Levantine: A group of dialects similar in written form and differ in pronunciation (Abdul-Latif, 2016).

Gulf: This is the closest dialect to MSA due to the evolution of the current form of MSA that is originated in the gulf region (Versteegh, 2014).

Iraqi: A dialect with features of pronunciation, prepositions, and verb conjugation. However, it is sometimes considered one of the gulf dialects (Mitchell, 1990).

MSA is used for communication when a native speaker cannot understand the speaker of the other dialect (Alnawas, Arici, 2018).

It is worth mentioning that regional dialects and MSA differ in some linguistic characteristics such that: MSA has singular, dual, and plural forms where the dialects lack the dual property. In addition, MSA has a gender distinction (feminine and masculine) in plural form, while the dialects have no such distinction. MSA and dialects have differences in lexical choice that exceed standardization in orthography and verb conjugation, even in the case of preserving the trilateral root (Zaidan, Callison-Burch, 2014).

The Bahraini dialect is one of the dialects of the Arabic language used by most of the inhabitants of Bahrain (اللهجات الشعبية في البحرين وجذورها التاريخية, 2016). According to Holz as cited in (Ameen, 2020), Arabic is pronounced by the residents of Bahrain in two different dialects: The "Arab" and the "Baharna" dialect. The "Arab" dialect is closely

related to diving and the practices related to this economic activity which flourished in Bahrain before the discovery of oil. At the same time, (Matar as cited in Ameen, 2020) categorized the Bahraini dialects into the Muharraq and Sitra dialects. The Muharraq dialect is spread in the city of Muharraq, the eastern and western Rifa'in, the Hidd and the villages of Zallaq, Jasra, Jaww, Askar, and parts of the city of Manama. Meanwhile, the Sitra dialect is spoken on the island of Sitra and in the villages of Tubli, Kawarah, Ma'amir, Jidhafs, Sanabis, Sanad, Nabih Saleh Island and parts of the city of Manama.

This section has attempted to summarise the definition and some distinctions of MSA and dialects.

2.4. Available Arabic Datasets for Sentiment Analysis

Nowadays, sentiment analysis of Arabic represents an active area of research; however, there is still a scarcity of resources for SA tasks (Al-Twairesh *et al*., 2017).

In a paper entitled "A corpus for Arabic sentiment analysis of Saudi tweets", the details of collecting and constructing a corpus of Arabic tweets were presented by (Al-Twairesh *et al.,* 2017), where the annotation process, Cleaning techniques, and preprocessing process of collected data were explained.

2.2 million tweets were collected by (Al-Twairesh *et al.,* 2017) over three months, from which the Saudi tweets were extracted to create a corpus called "AraSenti_tweets". Three annotators conducted a process of annotation to classify the Saudi tweets into five classes (positive, negative, mixed, neutral and intermediate). The annotators were trained for a one-hour session about the annotation guidelines.

The Inter Annotation Agreement (IAA) was used and measured by Fleiss's Kappa to ensure the reliability of the annotation process.

The cleaning and preprocessing of the dataset were represented at 1- Excluding all tweets containing media, mentions, URLs or retweets. 2- Normalization and tokenization of tweets via MADAMIRA.

Several experiments were conducted by (Al- Twairesh *et al.,* 2017) to establish a benchmarking baseline for AraSenti_tweets for multi-way sentiment classification using

Support Vector Machines (SVM) and TF_IDF. These experiments were 1-Two way classifications where positive and negative tweets were used. 2- Three-way classification of positive, negative and neutral tweets. 3- Four-way classification of positive, negative, neutral and mixed.

F1-Score was the evaluation metric used in the experiment. It was observed that the classifier performance was highly affected by the number of classes, i.e., the performance of the classifier degrades when the classification classes are more, concluding that a more sophisticated model is needed for three and four-way classification.

Another dataset, called 'Saudi Dialect Corpus from Twitter' (SDCT), was created by (Alahmary, Al-Dossari & Emam, 2019) to enhance the SA at the sentence level of Saudi Arabia dialect using an approach of deep learning.

Sixty thousand tweets from different scopes were collected using Twitter API, and only 32063 tweets were extracted to create the dataset. In order to classify the tweets manually, an annotation process took place, which resulted in 17707 positive tweets and 14356 negative ones.

The annotation process was followed by some preprocessing steps such as 1-removing all special symbols like (#, &,%,$), diacritics, punctuations, single Arabic letters, and non-Arabic letters, 2-Normalizing some characters by replacing multi-variant of (أ,إ,آ) letter to be one letter (ا), (ة)by (هـ), and (ي ,ئ) by(ى), and 3- eliminating repeated characters, for example, شكررررررا is replaced by (شكرا).

Some techniques and models were employed by (Alahmary, Al-Dossari & Emam, 2019) to perform SA of Saudi dialects. These techniques are 1- Continuous bag of words (CBOW), which is a Word2Vec model to learn the words vector representation in an unsupervised way, 2- LSTM and Bidirectional LSTM (Bi-LSTM), which are deep learning models in a supervised manner, and 3- SVM to evaluate and compare its performance with LSTM and Bi-LSTM.

Many experiments were conducted by (Alahmary, Al-Dossari & Emam, 2019) to explore how the SA of the Saudi dialects could be enhanced using deep learning. These experiments were conducted by 1-Training the CBOW in an unsupervised way using a gensim library of python and feeding it to the Bi-LSTM via TensorFlow. 2-Selecting 70% of the dataset as training while 30% as testing. 3- Utilizing SVM after applying vectorization using CountVectorizer and TF-IDF.

The obtained results showed that Bi-LSTM outperforms the other classifiers in terms of accuracy, which was 94%, whereas the accuracy of LSTM was 92% and 86.4% for SVM.

In 2017 a corpus called SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments) was created through a study conducted by (Rahab, Zitouni & Djoudi, 2017). The corpus was created by collecting comments on different domains (sport, news, politics, and culture) from the Echorouk newspaper website. The comments included in SIAAC were used to be classified into negative or positive classes through a proposed approach by (Rahab, Zitouni & Djoudi, 2017) using NB and SVM classifiers and through a software platform called RapidMiner.

One of the challenges faced by (Rahab, Zitouni & Djoudi, 2017) was the imbalance between positive and negative comments. The negative comments predominate the positive ones. There are 92 negative ones, whereas positive comments are 32. Besides that, the rating system gives a negative or positive point instead of giving points in a specific range leading to difficulty in the annotation task.

The methodology that was followed by (Rahab, Zitouni & Djoudi, 2017) was represented as 1- corpus creation, 2- Corpus preprocessing, and 3- Feature selection. Through the preprocessing stage, the French words and Algerian dialect words were translated to MSA because the stemmer would not work properly with dialectal words.

After the preprocessing stage, the corpus passed through the processing steps such as tokenization, removing of stop words, stemming, and filtering of short tokens.

Four parameters were used to create word vectors, Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF), Term Occurrences (TO), and Binary Term Occurrences (BTO). In addition to the vector creation, several parameters of feature selection have been used besides unigram and bigram word representation.

Cross-validation of 10 folds was used to evaluate the NB and SVM classifiers. The performance of the classifiers was measured using Precision, Recall, and F1 metrics.

The results obtained showed promising results in terms of recall and precision, whereas more improvement is needed in terms of F1.

The obtained results by (Rahab, Zitouni & Djoudi, 2017) on their created SIAAC corpus were compared to that one on opinion corpus Arabic (OCA) using SVM. The results showed that the (Rahab, Zitouni & Djoudi, 2017) approach outperformed that of OCA in terms of precision metrics.

Another corpus that is publically available from the Arabic Maghreb coming from Tunisia is called the Tunisian Sentiment Analysis Corpus (TSAC), which was created by (Mdhaffar *et al.,* 2017) by collecting 17,000 Facebook comments. The comments were written by the users of Facebook pages of Tunisian TV channels and radio channels in a period spanning from January 2015 to June 2016. The TSAC was annotated manually to negative and positive polarity. The TSAC covers multi-domain comments (political, social, and educational).

During the development of TSAC, many challenges faced (Mdhaffar *et al.,* 2017) due to: 1- Scarcity of free Tunisian dialect resources. 2- Limited number of studies regarding Tunisian dialect. 3- No standard orthographies and tools for this type of dialect.

It is worth mentioning that TSAC, which covers multi-domain comments (political, social, and educational), passed through a manual cleaning stage before the annotation process. The cleaning stage includes 1-Removing of all non-Arabic comments. 2-Deleting usernames, URLs, and hash characters.

In addition to their TSAC corpus, two additional corpora and three classifiers, (Mdhaffar *et al.,* 2017) were used in their study. The two other corpora are 1-Opinion

Corpus for Arabic (OCA). 2- Large Scale Arabic Book Review (LABR). The three classifiers were SVM, Naïve Bayes (NB), and Multi-Layer Perceptron (MLP).

Knowing that all of the experiments were implemented using Python scikit learn and gensim libraries. The Scikit learn library for classification and the gensim to represent the learning vectors.

All the classifiers were evaluated using Precision and Recall metrics. The results obtained showed the best classification was achieved when using the Tunisian dialect TSAC as a training set with an error rate of 0.22 by MLP, 0.42 by NB and 0.23 by SV.

In addition to the Algerian and Tunisian datasets from Arabian Maghreb countries, a large dataset of about 25,000 Arabic tweets like Levantine, Maghrebi, Egyptian, Iraqi, and Gulf was presented by (Alsarsour *et al*., 2018). The dataset was called Dialectal ARabic Tweets (DART). The mechanism of collecting the DART dataset was as follows: 1- Collecting 1000 words per target dialect (Levantine, Maghrebi, Egyptian, and Gulf) by referring to specific previous researchers. 2- Extending the collected list of dialectal words by randomly selecting phrases representing the target dialect from the mo3jam.com website, and 3- Filtering and cleaning the resultant list by a- Manually removing inappropriate phrases and b-Removing Arabizi phrases, retweets, and short tweets. It ended up with an average of 246 for Gulf dialects phrases, 121 for Iraqi, 244 for Levantine, 278 for Egyptian, and 273 for Maghrebi.

Unlike the previous corpora and datasets, the DART was manually annotated using a crowd sourcing platform called CrowdFlower and evaluated using two methods: a-Final label accuracy and b-Inter annotator agreement. DART has no implementation for any classifier.

Unlike the previous studies presenting Arabic corpora, a significantly large dataset of size greater than 0.25 billion tweets was presented by (Abdul-Mageed, Alhuzali & Elaraby, 2018). The presented dataset is tagged at a city level and covers 29 dialects representing 10 Arabic countries (Egypt, Iraq, Jordan, Palestine, Yemen, Kuwait, Oman, Qatar, KSA, and UAE) for which little to no datasets are available. Some cues were relied on to label the dataset with geographical labels. One of these cues is the

information of the user-provided location on the Twitter platform. The developed dataset was collected by using Twitter API for a period of five years spanning from 2013 to 2018.

The study of (Abdul-Mageed, Alhuzali & Elaraby, 2018) differs from the other studies in labelling the tweets based on location by acquiring location labels geopy, which is a geocoding library in Python that helps in locating cities, countries, and landmarks in coordinates form, based on another data source as a third party like "OpenStreetMap Nominatim".

All tweets which are non-Arabic were excluded via the character count method, followed by a pilot annotation task performed by native Arabic speakers and college-educated. Each annotator was provided with dialect data representing a single country and instructions regarding the labelling process.

To see the inter-annotator agreement of the annotation task, a Cohen's Kappa was used with an obtained value of 67%.

To tackle the problem of the issues associated with the MSA model when applied to the dialect dataset, a skip-gram SG word vector model was built by (Abdul-Mageed, Alhuzali & Elaraby, 2018). The skip-gram model had parameters like window size =5 words, wordcount=100, which was implemented using the gensim tool.

The preprocessing steps which took place are a- removing all non-Unicode and non-Arabic characters, b- normalizing all Alif maksura to ya, c- reducing all Alif with hamza to plain Alif, and d- cleaning the noise that is in non-standard typography form.

The results showed that the (Abdul-Mageed, Alhuzali & Elaraby, 2018) model outperforms the available models in the distributed representation.

Another Multi Dialect Arabic Sentiment Twitter Dataset called (MD_ArSenTD) was created by (Baly *et al.*, 2017a). (MD_ArSenTD) is composed of tweets from 4 regions (Gulf, Levant, North Africa, and Egypt). Both Gulf and Levant tweets covered four countries. Gulf tweets covered (Kuwait, KSA, Qatar, and UAE), whereas Levant tweets covered (Jordan, Lebanon, Palestine, and Syria). The covered countries of North Africa

were (Algeria, Morocco, Tunisia), and finally, Egypt, with no specification regarding the included countries, resulting in a total of 12 covered countries.

Tweets in (MD_ArSenTD) were assigned sentiment labels using a scale of 5 points to provide intensity information besides the polarity.

The focus of (Baly *et al.*, 2017a) was on 1- describing the characteristics and specificities such as discrete features and structure of Egyptian and UAE tweets and 2- highlighting the discussed topics in tweets of both countries.

Some steps were followed to create the (MD_ArSenTD). These steps were as follows 1- collecting 470,000 tweets from 12 countries using Twitter4J API, from the first of March 2017 till the end of April 2017, using specific-gio locations to enforce the retrieving of the tweets from the four regions' countries. 2-Selecting the tweets according to a pre-defined target size of a- 14400 for (MD_ArSenTD), b- 1200 tweets per country, c- removing all duplicated tweets and those with less than 30 characters, d- applying a pre-trained model that won SemEval-2017 task4 for the remaining of tweets, and e- selecting the top 1200 tweets that were predicted as negative, positive, and neutral with high confidence to decrease irrelevant tweets. 3-Annotating the selected tweets for both sentiments and dialect using CrowdFlower.

Regarding the sentiment notation, the annotators were asked to use the polarity of 5 points scale (very negative, negative, neutral, positive, and very positive). Regarding the dialect notation, the annotators were guided to identify the country and region for each country, otherwise choosing a foreign language or MSA.

The performance of the annotators was monitored using a test of a gold set of 100 tweets per country and calculating Kohen's Koppa, which was 0.65 in sentimental annotation, and 0.8 in the annotation of region-level dialects.

Two types of models were the focus of (Baly *et al*., 2017a): 1-Feature engineering by evaluating an equivalent model to the one that won the SemEval-2017 task4, trained SVM with a group of handcrafted features that covered semantic, syntactic and surface information. The evaluated model extracts features such as emoticons, URL, mentions

of a user, ngram of the lemma, counting of POS, both questions and exclamation marks, negated context, positive and negative emotions, and 2- Deep learning, throughout evaluating LSTM, using Skip-gram (SG) model of word2vec to generate the input feature. Two types of embedding were used by (Baly *et al*., 2017a), specific and generic, to assess how the dialect influences the SA.

As (Al- Twairesh *et al.*, 2017), MADAMIRA software was used in the study of (Baly *et al*., 2017a). MADAMIRA was used by (Al- Twairesh *et al.*, 2017) to tokenize and normalize the tweets, whereas it was used by (Baly *et al*., 2017a) to train the SG.

Some preprocessing steps took place to get an improved quality of input text. The preprocessing steps included normalizing repeated characters such as word elongation, replacing the emoticons with global equivalent sad or happy tokens, and replacing parenthesis with squared brackets.

Other software was used for implementation, such as LibSVM to train and evaluate SVM, whereas Keras with TensorFlow and accompanied libraries were used to train and evaluate LSTM.

Results of (Baly *et al.*, 2017a) models indicated the outperforming of LSTM over SVM in terms of feature engineering, and the embedding of dialect-specific gave better results than generic embedding, and better performance by SVM and LSTM was achieved on Egyptian tweets.

More Arabic SA resources were provided by (Itani, Roast & Al-Khayatt, 2017) by developing corpora and lexicon in Arabic dialect by collecting comments from the Facebook platform. Two corpora were developed in the domain of news and arts, and each one includes 1000 comments. The corpus of arts was created by collecting comments from the Facebook page of The Voice, whereas the news corpus was created by collecting comments from the AlArabiyya news Facebook page. At the same time, lexicons of words and phrases were developed for the two corpora. The collected comments were preprocessed by removing repeated contents and irrelevant data such as timestamps, resulting in 12053 words in 1000 posts of arts and 8423 words in the news.

After the preprocessing steps, an annotation process took place where four Arabic native expert speakers of (Lebanese, Syrian, Palestinian, and Egyptian) labelled the collected comments with positive, negative, dual, neutral, and spam labels. Rules of annotating were provided to the annotators, and only the comments classified similarly by all annotators were used. 97% is the similarity value between the annotators using Inter Annotator Agreement (IAA).

Regarding the lexicon development, they were created by the annotators, who extract the sentences and words from each comment that has an influent effect on sentiment classification as negative, positive, or spam. The extracted words and sentences were added to the lexicon as lexemes. After all, the lexemes were modified in two steps: Factoring and Repetition extracting. In (Itani, Roast & Al-Khayatt, 2017) study, no information was provided about the classifier implementation, except that its performance was compared with the annotators' results.

A corpus of Sudanese Dialect Arabic (SDA) was designed by (Abo *et al.*, 2019). The SDA is designed and presented as a lexical resource. It comprises 5456 tweets, collected through Twitter API, and covers a political scope.

The dataset passed through a preprocessing and tokenization process. The Preprocessing was represented by removing the repeated character, numerals, punctuation, URL and correcting the misspelling character using a Packet of R-Studio. In contrast, tokenization occurs using an Operator Script by breaking the text into chunks. The operator Script was also used in filtering the text from the stopwords list. After the preprocessing, an AYLIEN API connection, part of R-Studio tools, was used for classification. Two classification steps were adopted in the (Abo *et al.,* 2019) study, one for SA and one for subjectivity classification. The sentiment analysis is a two-way classification, i.e. binary polarity (positive and negative), while the subjectivity classification is OBJ and SUBJ.

Two annotation methods were applied by (Abo *et al.,* 2019), one by recruiting educated native speakers of Arabic who were assigned the task of annotating 5456 tweets, while the other method was online using the tool of RapidMiner. A Kohen's Kappa was applied to measure the reliability and agreement between the annotators, giving 92.5%.

The classification experiment included two machine learning classifiers: Decision Tree (DT) and Naïve Bayes (NB). The classifier's performance was evaluated using the accuracy metric, and the accuracy value was 60.8% and 59.4% for DT and NB, respectively, in sentiment analysis. In contrast, the obtained accuracy regarding subjectivity classification was 82.1% for NB and 83.5% for DT, indicating that the classifier's performance is better in terms of subjectivity sentiment analysis (SSA) than SA when applied using SDA.

Another corpus of 40,000 tweets was constructed by (Mohammed and Kora, 2019). The 40,000 tweets are a mix of Egyptian dialects and MSA. These 40,000 tweets were collected using Twitter API from 11 April to 12 December 2015, covering multiple topics such as proverbs, poetry, sarcastic jokes, social topics, politics, health, sports, and product opinions. The collected tweets were filtered, processed, annotated manually, and validated by two experts. The manual filtering and processing of the corpus include removing: spam tweets, duplicated tweets, gulf Arabian countries' tweets, words' diacritics, and elongation, in addition to replacing different forms of letters with one form, adding space to combined words, and finally, correction process for words that are wrongly written or have missing letters. The corpus' tweets were classified into positive and negative by (Mohammed, Kora, 2019), who proposed three deep learning models, namely convolutional neural network (CNN), long short term memory (LSTM), and recurrent convolutional neural network (RCNN). RCNN is a neural network that contains LSTM as a layer with the layers of CNN. CNN is used for strongly extracting features, while the LSTM layer memorises and applies the architecture of recurrent neural networks on extracted features.

In the proposed approach of (Mohammed and Kora, 2019), Aravec was used. Aravec is a pre-trained CBOW model for generating the matrix of word embedding. Three different data splits were applied in (Mohammed and Kora, 2019) : (60%, 40%), (70%, 30%), and (80%, 20%) on each model for training and testing purposes. 10% was considered as validation data for tuning the hyper-model parameters. The evaluation metrics were f1-score, precision, recall, and accuracy. Data augmentation (DA) with shuffling was used to randomly change words’ order in a small window of text

sequence. The results showed that LSTM outperformed the other two models by achieving an accuracy of 81.3% compared to 78.46% and 75.72% for accuracy measures achieved by RCNN and CNN, respectively. LSTM also achieved the highest accuracy of 88.05% when applying the DA technique.

Given all that has been reviewed so far, one notices that despite the conducted studies and created corpora that consider the Arabic languages and dialects, there is an absence and complete neglect of such resources in Mauritanian and Bahraini dialects, as mentioned by (Elnagar *et al*., 2021) in their systematic literature review of identification and detection of colloquial Arabic. Additionally, most of the created datasets were generated by collecting tweets using the Twitter API, while in this research, different tools and procedures were created and followed. This research applied the machine and manual translation approaches using 500 custom forms created with https://getfoureyes.com.

2.5 Impact of Translation on Sentiment Analysis

A study for improving SA in scarce resources languages was carried out by (Mohammad, Salameh & Kiritchenko 2016) to study the effect of translation on sentiment, where two approaches were systematically examined. The two approaches are -1 Translating a dataset of target or insufficient resources languages to a language of rich resources such as English and applying English SA to the translated text.2- Translating sentiment lexicons or labelled corpus from a language with ample resources , such as English and using them as extra resources in the target language's SA system, i.e. (Arabic).

Two experimental techniques corresponding to approaches one and two were proposed to achieve the study purpose. The first technique was represented by translating Arabic text to English in two ways: automatically and manually, then annotating the English text automatically and manually, and then comparing the label obtained by the system for translated text with that of manually annotated Arabic text. The second technique was represented by automatically translating the English annotated resources such as lexicons and corpora and utilizing them as supplementary resources in the supervised classification of SA, then comparing the labels predicted by the system with the manual

annotation of the Arabic text. In the first technique, 1200 sentences fell in Levantine dialects called BBN posts were utilized. The BBN posts were randomly chosen from an already existing translated social media corpus, namely BBN Arabic Dialect-English parallel text, which includes approximately a 3.5million tokens from sentences of Arabic dialects and the corresponding one of the English translation.

In addition to BBN posts, a Syrian dataset was created by collecting 2000 Levantine tweets through Twitter API in May 2014, knowing that this dataset has no manual translation. The automatic translation was implemented by using Statistical Machine Translation (SMT).

In the first technique experiments, the Syrian and BBN posts datasets were split into training and testing folds as a portion of cross-validation. The training dataset was preprocessed using Penn Arabic Tree Bank (PATB) for tokenization. Some characters are normalized, such as the character Alif which has multi forms (إ,ا,أ,آ) is normalized to(ا) and Ya (ى,ي) was normalized to dotless Ya (ى). The manual annotation of the BBN posts and Syrian dataset was achieved through the CrowdFlower crowdsourcing platform, where more than ten annotators annotated each post.

Support Vector Machine (SVM) with linear kernel was trained on training data.

In the second technique, a dataset of English tweets of SemEval-2013 and all Lexicons' words were translated into Arabic using Google translate. Three lexicons were created by (Mohammad, Salameh & Kiritchenko, 2016), namely (Arabic Emotion Lexicon, Arabic Hashtag Lexicon, and Arabic Hashtag Lexicon (Dialectal). The translated text from English into Arabic was preprocessed with the CMU Twitter API tool to tokenize the tweet components such as emoticons, user names, and URLs. The Lemmas were generated using MADA and followed the same normalizing procedure of the first technique of Alif and Ya.

The datasets with labelled sentiment include RR and MD, which refer to the researcher (Refaee and Raiser) and (Murad and Darwish).

The Arabic text sentiment analysis system was on RR, MD, NBB and Syrian datasets.

In the second technique, different experiments were conducted with various combinations that were resulting convergent accuracy values, while in the experiments of technique one, the highest obtained accuracy value was 79.35 and 65.31 for Syrian and BBN posts, respectively, using the features of the Dialectal Arabic hashtag lexicon with ten cross-validations.

(Mohammad, Salameh & Kiritchenko, 2016) experiments showed similar results to state-of-the-art systems of Arabic SA when Arabic text is translated into English. The results also showed that automatic SA of machine-translated Arabic text into English outperforms the manually annotated text.

Regarding the results of technique two, it was found that translating English lexicons and using them as extra resources improves the accuracy while adding the English-translated sentiment tweets to Arabic training data leads to a fall in the accuracy. Also, it has been shown that when Arabic text is translated into English, the SA produces promising and competitive results. In addition, an automatic and manual translation could affect the sentiment in some situations. One of these situations is translating positive and negative reviews into neutral ones. It was also shown that automatic SA is not affected by some automatic translation properties, which mislead the human regarding the true sentiment of the source text.

Similarly, (Gangula and Mamidi, 2018) conducted a set of experiments to study the impact of translation on the sentiment of a text when it is translated automatically and manually from English into low-resource south Asian languages such as Telugu. The manual and automatic annotated English text at a 5-value scale as a benchmark was utilized to determine the loss in sentiment and its predictability in the translated text.

(Gangula and Mamidi, 2018) decided to translate the English reviews of books and products into Telugu using automatic and manual translation and annotated them manually using the 5-value scale (highly negative, negative, neutral, positive, and highly positive) through Telugu native speakers.

To translate the reviews automatically, the Google translate API was used by (Gangula and Mamidi, 2018), while the manual translation was conducted by some translators

who were provided with instructions such as keeping the exact meaning and sentiment of the translated text and considering the grammar and syntax of the Telugu language.

The obtained Telugu language text was sentimentally analyzed using a lexicon of 6000 words which was created using Telugu SentiWordNet with certain specifications in assigning scores for every word in a sentence. In contrast, the sentiment of the English reviews was analyzed using SVM with a linear kernel after normalizing and generating the TF-IDF vector for each sentiment in the data.

The classifier performance was measured using the accuracy metric, giving 67.5%.

The results showed the following: 1- The SA system had difficulty assigning positive and negative marks, which was justified by the lack of negative and positive participation in the training data. 2- The system predicts a high positive score in a dominant manner, which is also justified by the amount of training data.

Finally, it was concluded by (Gangula and Mamidi, 2018) that the impact of translation is very high at a fine-grain level such that significant loss in sentiment occurs when using automatic translation. It works much better when done manually.

However,(Can, Ezen-Can & Can, 2018) claimed that automatic translation retains most of the necessary and required information for sentiment analysis, especially when reusing a model for multi-languages, since it causes reducing in the requirements of data.

(Guo, Xiao, 2012) mentioned that many problems may arise as a result of translation in SA, such as the discrepancy or contradiction in the distribution of data between the target language and the source one, which is generated as a result of what is called word drifting or word deviation, i.e., the word rarely appears in the target language. In contrast, it frequently appears in the source language.

As stated by (Becker *et al.*, 2017) and (Wehrmann, Becker & Barros, 2018), the final classification process may be affected due to the distance of cultures between the target and source language, even if the translation was done perfectly.

The multilingual SA model is one of the solutions to address the translation process's impact on sentiment analysis.

2.6 Multilingual Sentiment Analysis

The multilingual SA model is a solution used to tackle issues like capturing the data of multi-cultures and multilanguage and utilizing the same SA model for different languages (Agüero-Torales, Salas & López-Herrera, 2021).

A multi-language sentiment classification approach was proposed by (Becker *et al.*, 2017) to classify tweets in 4 languages without relying on machine translation. The proposed approach was a deep neural model based on a convolutional neural network's optimized convolutions and character embedding. The proposed approach is Conv-Char-R, which optimizes and reduces the original architecture of Conv-Char proposed by other prior researchers.

The approach of (Becker *et al.*, 2017) was obtained by reducing 6x of the original architecture parameters by improving the model linearity while reducing tensor dimensionality. About 1.6 million positive and negative tweets were used to evaluate the proposed architecture. These tweets fall in 4 languages (German, Portuguese, Spanish, and English)—knowing that the dataset provides the tweet's URL instead of the tweet itself.

Several experiments were conducted by (Becker *et al.*, 2017) to compare their Conv-Char-R model with 1-LSTM-EMB: according to (Becker *et al.*, 2017), it is complex recurrent neural network architecture with input and forget gates, which can learn long-term dependency and forget the useless information within a sentence, 2-Conv-EMB: is a faster architecture than LSTM that employs convolutional layer jointed with an operation called max-pooling over time, 3-Conv-Char: an architecture that composed of many layers of convolutions learning step that avoid large vocabulary by acquiring small alphabet in memory, and 4-SVM.

The performance measures used in the proposed approach were the Accuracy and F1 score. The results showed that 1- LSTM-EMB achieved the best F1 score, followed by the accuracy metric, with 0.753 and 0.713, respectively. 2-The convolutional models

achieved better when using a word embedding instead of character embedding, and they learn better using back propagation instead of embedding freezing. 3-The proposed Conv-Char-R model gave the same results as its larger version, i.e., Conv-Char, with a trade-off parameters amount and performance among the other neural network architectures. It is ~ 3% less than the best model LSTM-EMB in terms of F1 score and 2% in terms of Accuracy.

A helpful approach for SA was proposed by (Can, Ezen-Can & Can, 2018) by building a single reusable model to avoid replicating feature engineering efforts and determine the correct set of features for each SA model. The model was trained on a large dataset in English and reused with limited data in other languages to evaluate how well a generic model can detect unseen opinions of data in different languages.

After translating them to English, the generic model was evaluated on different datasets that cover different languages. The datasets used by (Can, Ezen-Can & Can, 2018) covered Amazon reviews, Yelp restaurant reviews and competition reviews. At the same time, the used languages were Spanish, Turkish, Dutch and Russian, where the model was RNN, including Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).

The results obtained by (Can, Ezen-Can & Can, 2018) model outperform the baseline in different languages.

Another study aimed at creating a multiclass multilingual sentiment analysis model was conducted by (Attia *et al*., 2018), where they used CNN with simple architecture composed of five layers, namely (input, convolutional neural network, max pooling, dense layer, and output layer).

Three datasets were utilized: Arabic, English, and German, each with specific classification labels to evaluate the CNN model. The Arabic dataset was called the Arabic Senti Tweets Dataset (ASTD), which consists of 10,006 tweets that were classified into four classes (objective, positive, negative, and mixed).

The English dataset was called the Standers Twitter Sentiment Corpus, composed of 5,513 tweets that cover the product domain of famous companies like Twitter, Microsoft, Google, and Apple. These tweets were manually classified into four classes (neutral, positive, negative, and irrelevant). In contrast, the third German dataset was called Deutch Bahn, which includes 21,284 customer reviews from web sources and social media about train operators of German trains. This dataset was divided into training, development, test1 and test2 sets and classified into three classes which are (positive, negative, and neutral).

Some preprocessing steps were performed, like manual tokenization by putting a space between the punctuation marks and the preceding word. Other preprocessing steps were represented by removing the punctuations, prepositions, determiners, and URL addresses and ignoring the words with low frequency.

The results were compared with the best results and scores for each dataset in the literature. The system of (Attia *et al*., 2018) outperforms the other systems when using the English, and German datasets, where the accuracy achieved was 78.3% and 75.45%, respectively, while the accuracy obtained with ASTD was 67.93% which is below the value achieved by the other systems. The justification by (Attia *et al*., 2018) for getting these results was the lack of test data that were randomly selected.

It was noticed by (Attia *et al*., 2018) that the prediction result was biased to the major classes at the expense of the minor classes, so they applied a macro average of F1 score and manual oversampling to resolve this issue.

When comparing the studies of (Attia *et al*., 2018) and (Can, Ezen-Can & Can, 2018), it is found that (Can, Ezen-Can & Can, 2018) had trained their model using English training dataset that was obtained by translation approach, while (Attia *et al*., 2018) had trained their model using different datasets each in a different language without any translation. In this research, the translation approach was used.

According to (Can, Ezen-Can & Can, 2018), using a multilingual sentiment analysis model eliminates word embedding, lexicon usage, and model training per language. This information was confirmed by (Medrouk and Pappa, 2017), who stated that the

multilingual neural network model could learn and predict in a multilingual environment as if it worked separately for each language.

In this research, the multilingual aspect has been achieved by developing a single LSTM model that was trained and tested individually on each of our datasets: English, MSA, and BDs, by applying the same values of tuned parameters such as the number of layers, the number of nodes per layer, the optimizer, learning rate, and the loss function.

2.7 Deep Learning and Neural Networks

Deep learning is an application of neural networks (Zhang, Wang & Liu, 2018). The neural network is a linear classifier (Chaudhari *et al.*, 2016) composed of harmonic processing units called neurons. The neuron is the fundamental element of the computational process called the activation function in the neural network. These neurons are organized in layers (Zhang, Wang & Liu, 2018) and fed by inputs representing the word frequency and a neuron weight. These neurons perform the non-linear function and pass the results (output) to the next layer (Chaudhari *et al.*, 2016). There are some topologies for neural networks architectures as listed below:

1. Feedforward network with fully connected layers like the multilayer perceptron (Goldberg, 2016). An example of the feedforward network is shown in figure 2.1. Layer 1(L1) is the input layer that includes the vectors of x1, x2, and x3 and a bias value of +1. Layer 2 (L2) is a hidden layer that takes the vectors of x1, x2, and x3 and the bias as input from layer 1. It does some calculations using the *f* function and output $f(W^t x) = f(\sum_{i=1}^{3} W_i x_i + b)$, where W_i denotes the weight of the connection and b the bias. f function varies from hyperbolic(tanh), sigmoid, and rectified linear (ReLU) (Zhang, Wang & Liu, 2018). Figure 2.2 shows an example of a sigmoid function unit. The *f* function yields Xi documents' class label (*yi*) in binary classification (Medhat *et al.*, 2014). Layer 3 (L3) represents the output vector (S_1). The circle in Layers 2 and 3 represents neurons, while a circle in Layer 1 represents an element. Weighted connection lines represent the data flow between neurons. (Zhang, Wang & Liu, 2018).

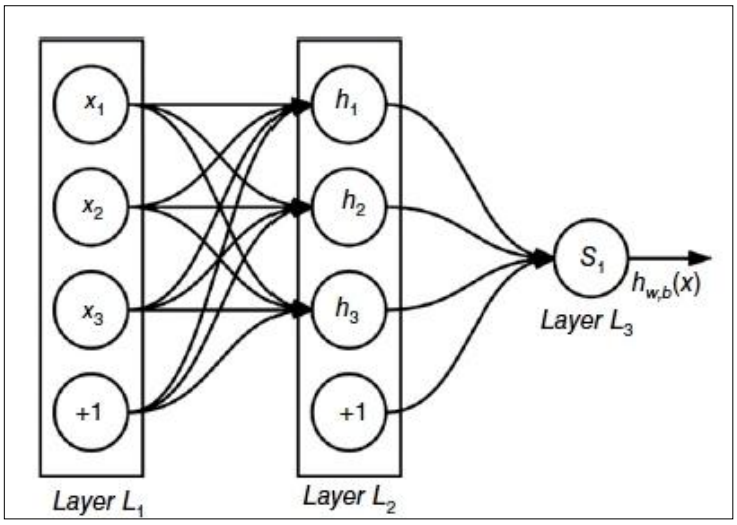


Figure 2.1 Feedforward neural network (Zhang, Wang & Liu, 2018)

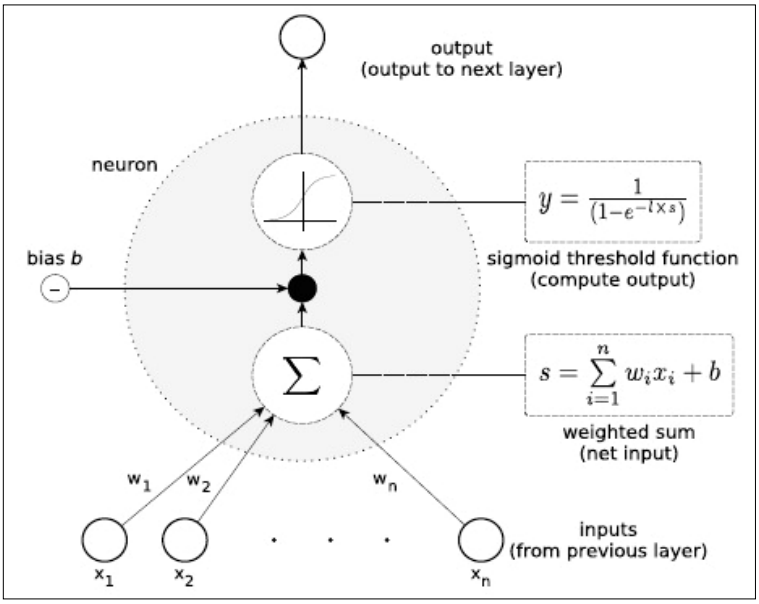


Figure 2.2 A neuron with Sigmoid function (Staudemeyer, Morris, 2019).

2. Recursive Neural Networks (RvNNs) are machine learning models that can learn and process deep structured information and parse trees in natural language processing (Chinea, 2009). It recursively produces parent representation at the bottom-up style by joining the tokens of a sentence's components until completing the whole sentence (Zhang, Wang & Liu, 2018).

3. Recurrent Neural Networks (RNN) has a particular architecture with internal memory; this internal memory can remember the previous information. Recurrent Neural Network (RNN) has many sophisticated types like Bidirectional RNN, deep Bidirectional RNN, Long Sort Term Memory (LSTM) and Gated Recurrent Unit (GRU), which is like LSTM with slight variation (Zhang, Wang & Liu, 2018).

There are examples of neural networks such as Convolutional Neural Network (CNN), Deep Belief Network (DBN) (Ain *et al.*, 2017), and Long Short Term Memory (Zhang, Wang & Liu, 2018).

a) Convolutional Neural Network (CNN)

CNN is a common type of feedforward neural network (Zhang, Wang & Liu, 2018), consisting of multiple Conv layers. Each layer abstracts the input data and consecutively generates a feature map (fmap) pertaining to the unique, essential information. CNN's best performance can be achieved using layers in a deep hierarchy (Krizhevsky, Sutskever & Hinton, 2012).

b) Deep Belief Network (DBN)

It is a neural network with a stack of hidden layers composed of a Restricted Boltzmann Machine (RBM). In DBN, the unlabelled data fulfils the deficiency issue of labelled data. DBN is efficient in the representation of features (Ruangkanokmas, Achalakul & Akkarajitsakul, 2016).

4. Transformers are deep learning models widely used in fields such as speech processing, computer vision, and NLP. Three different architectures for the transformers: encoder-decoder, encoder-only, and decoder-only (Lin *et al.*, 2022). The encoder comprises a set of multi-head attention layers and normalization and feed-forward layers. The multi-head attention mechanism role is represented by enabling the model to coordinate the information from the different subspaces of representation at different positions in the sentence. (Vaswani *et al.*, 2017). Depending on the transformer architecture, the encoder outputs are fed to another encoder or a decoder. Some examples of transformers are: BERT, RoBERTa,

ALBERT, XLnet, DistilBERT (Braşoveanu, Andonie, 2020), and various versions of GPT (Zhao *et al.*, 2023).

2.8 Deep Learning in Sentiment Analysis of Arabic and other languages

Deep learning is a powerful technique of machine learning by which the learning process occurs through multiple layers of data features representation to produce a state of the art prediction results (Zhang, Wang & Liu, 2018). This section highlights various studies on SA of Arabic and English texts using deep learning networks and techniques.

An ensemble model combined both the LSTM model and CNN used by (Heikal, Torki & El-Makky, 2018) to predict the sentiment polarity of Arabic tweets as (positive, negative or neutral) at the sentence level. (Heikal, Torki & El-Makky, 2018) explored a model of (Cliche, 2017) that is ranked first in the competition of SemEval-2017 by adjusting and tuning multiple hyperparameters to improve the model accuracy.

To evaluate their model, (Heikal, Torki & El-Makky, 2018) used the Arabic Sentiment Tweets Dataset (ASTD) composed of 10,000 tweets categorized into four classes, namely objective, subjective negative, subjective positive, and subjective mixed.

The system architecture of (Heikal, Torki & El-Makky, 2018) is represented by 1- Preprocessing phase, including removing objective tweets. 2-Training phase for two deep learning models, LSTM and CNN, by varying the hyperparameters for each model. 3- Selecting the best model based on calculated F1- highest score and accuracy.4- Build the ensemble model using soft voting.

The deep learning model of (Heikal, Torki & El-Makky, 2018) relied on a vector representation of pre-trained words without using any features' engineering. The results showed that: the 1- LSTM model achieved the best results (acc=64.75%, F1 score=62.08%) using a drop rate of 0.2 while keeping the default configuration of other parameters. 2- The CNN model achieved the best results (acc=64.30%, F1 score=64.09%) using a fully connected layer of size 100 while keeping the other parameters at default configuration. The ensemble model of (Heikal, Torki & El-Makky, 2018) outperformed the model that ranked first by 6.55% accuracy and 10.86% F1 score.

Similar to (Heikal, Torki & El-Makky, 2018), (Alayba *et al.*, 2018) conducted a study on a model by integrating LSTM with CNN.(Heikal, Torki & El-Makky, 2018) explained their goals in brief, where (Alayba *et al.*, 2018) further clarified their study objectives at 1- Investigating the benefits of combining the mentioned models of LSTM and CNN and reporting the obtained results using different Arabic datasets. 2-Considering the morphological variance of certain words utilizing different levels of sentiment classification.

Four datasets were used by (Alayba *et al.*, 2018), where one dataset is a subset of another. These datasets were 1- The Arabic Health Service Dataset (AHS), an imbalanced dataset containing 2026 tweets (628 positives and 1398 negatives). 2- A sub dataset (AHS) containing 502 positive tweets and 1230 negatives 3- Twitter Dataset (Ar-Twitter) containing 1000 positive and 975 negative tweets covering different scopes and topics in Arabic such as arts, communities and politics. 4- Arabic Sentiment Tweets Dataset (ASTD) containing 2479 tweets, 795 positive tweets and 1684 negative ones.

There are three different SA levels: character level to increase the feature number per tweet, character N-gram level (ch five gram-level) and word level for each dataset.

The proposed model of (Alayba *et al.*, 2018) comprises multiple layers: input layer, convolutional, Max-pooling, LSTM, a fully connected layer and finally, the output layer. A fixed dimension matrix represents the input layer with different vector embedding based on the SA level.

Some filters slide over the matrix in the convolutional layer to produce a feature map. Different features could be obtained using various sizes filters in the convolutional layer. The max-pooling computes the maximum value and assigns it as a feature to a specific filter. The Maximum pooling layer will feed the LSTM network, giving the final output as negative or positive.

The results obtained by (Alayba *et al.*, 2018) model showed that the performance accuracy has improved by achieving 94.24% regarding (AHS) dataset and 95.68% regarding the sub-AHS dataset compared with other previous models that achieved

92.00% for (AHS) and 95.00% for the sub dataset. It is noticed that (Alayba *et al.*, 2018) got better results than (Heikal, Torki & El-Makky, 2018) regarding the accuracy metrics.

Besides the studies carried out on ensemble learning of the mentioned deep learning models CNN and LSTM, (Al-Azani, El-Alfy, 2018) in their study of sentiment classification based on emojis of Arabic microblogs aimed to assess nonverbal features extracted from the 2091 microblogs dataset, in addition to comparing the performance to a baseline deep neural network and traditional learning methods by evaluating two state-of-the-art-models of deep RNN considering uni and bidirectional model of LSTM and its simplified variant Gated Recurrent Unit (GRU).

Five datasets were considered by (Al-Azani, El-Alfy, 2018): ArTwitter, ASTD, Syria, QCRI and SemEval-2017 task4 subtask#A, besides collecting 843 Arabic microblogs containing emojis from YouTube and Twitter and annotating them using a manual way. For feature extraction, (Al-Azani, El-Alfy, 2018) used a lexicon called Emoji Sentiment Ranking (ESR) and a Principal Component Analysis (PCA) to reduce features dimensionality.

The proposed approach by (Al-Azani, and El-Alfy, 2018) was composed of input, hidden and output layers. The input layer contains some neurons equal to the number of features which is 100. In contrast, the hidden layer comprises 100 GRU or LSTM, which are considered for applying different modes like concatenation, multiplication, summation and average. The output layer is fully connected to the hidden layer by applying the sigmoid activation function. The Nadam optimizer and logarithmic loss function minimize the prediction error.

The evaluation metrics of the (Al-Azani, El-Alfy, 2018) proposed model were the accuracy, precision, recall, F1-score, Geometric Mean and MCC correlation. The results showed that bidirectional GRU obtained the highest results, followed by the bidirectional LSTM when concatenation mode is used.

One thing that makes the (Al-Azani, El-Alfy, 2018) study distinct is the mentioning of the implementation environment used, which is python (Keras and Theano libraries).

(Altaher, 2017) conducted a study for Arabic tweets SA by proposing a hybrid approach using: 1- Feature weighting algorithms like (chi-square and information gain) to assign a high score (weight) to the essential features as a step of preprocessing stage. 2- The deep learning model is called H2O, which has feed-forward multi-layers trained using back-propagation gradient descent.

Deep learning with H2O is scalable and fast, and it is used with innovative applications like PayPal to get accurate predictions faster without data sampling (Arora *et al.*, 2015). The datasets used about 500 tweets related to the education field.

The obtained results by the model of (Altaher, 2017) were compared with the ones obtained by the Support Vector Machine (SVM) and Decision Tree (DT). The results showed that the deep learning model outperformed the SVM and DT in precision and accuracy by 93.7% and 90%, respectively.

Another deep learning framework for SA of Arabic text was proposed by (Al Sallab *et al.*, 2015), where four different architectures were explored, three of them based on Deep Auto Encoder (DAE), Deep Neural Network (DNN), and Deep Believe Networks (DBN), while the fourth one based on Recursive Auto Encoder (RAE), which was used to compensate the loss of text handling in the three models.

Each model performs specific tasks, DBN for applying a pre-training before the feeding step, DAE for generating a reduced dimensionality of representations, and RAE parsed words in the best order that reduces the error of reconstruction of the same words order, i.e. giving the best parsing tree. RAE and DAE are unsupervised learning techniques that provide a compact presentation of input sentences. DAE parses the words of whole sentences at once, while RAE considers the order and context of sentence parsing.

The sentence-level of sentiment classification was the focus of (Al Sallab *et al.*, 2015), where the Linguistic Data Consortium Arabic Tree Bank (LDC-ATB) dataset was used to evaluate the proposed models knowing that ArSenL lexicon sentiment scores were used for features vectors.

The results showed that a better representation of the input sparsing vector was obtained by DAE, while RAE obtained the best F1-score with an improvement of 9% compared with the literature models.

A recursive deep learning model was presented by (Al Sallab *et al.* 2107). It was called "AROMA", i.e. (A Recursive deep learning model for Opinion Mining in Arabic). The AROMA was proposed to tackle the limitations and challenges of the RAE model in performing SA in Arabic, such as parsing of language and morphological complexity. The AROMA proposed model addressed these challenges: 1- Implementing morphological tokenization. 2- Modeling semantic composition at the morpheme level.

For modeling the sentiment and semantic composition, (Al Sallab *et al.*, 2017) used a parser of phrase structure to generate a tree of syntactic parsing.

Different types of datasets were used to evaluate the proposed model, such as (tweets, online comments extracted from Qatar Arabic Language Bank (QALB) and newswire extracted from Arabic Tree Bank (ATB) with different styles of Arabic writing dialects and standard.

The experiments' results showed that the proposed model outperformed the 1- RAE model in tweets, QALB and ATB regarding accuracy improvement by 7.2%, 8.4%, and 12.2%, respectively. 2- The literature models on the same dataset (tweets, online comments extracted from Qatar Arabic Language Bank (QALB) and newswire extracted from Arabic Tree Bank (ATB) by 7.6%, 1.7% and 7.3%, respectively.

Another study in which the recursive neural network was utilized was carried out by (Baly *et al.*, 2017b) to characterize Arabic tweets to understand in better way the writing style, linguistic phenomena, topics of discussion, and the extent of variation from one region to another. The target tweets were collected from Arabian regions like the Levant, Egypt and Arabian Gulf. (Baly *et al.*, 2017b) explored two approaches to opinion mining. The traditional approach is based on feature engineering to train the classifier, and the approach of the deep learning model is based on compositionality.

Two models were used in (Baly *et al.*, 2017b) experiments; the SVM model that won SemEval-2016 task4 on SA on Twitter was used for opinion mining in English. The second model was a recursive neural tensor network (RNTN) based on compositionality models.

The RNTN was trained on Arabic Sentiment Treebank (ArSenTB), which was developed by (Baly *et al.*, 2017b) from Qatar Arabic Language Bank (QALB). The sentiment tree is a collection of trees dedicated to parsing all constituency annotations that the SVM was trained on by lemmatizing and POS tagging. The performance of the two models was compared, showing that the deep learning model performed better than that based on feature engineering.

Overall, these studies consistently indicate the positive aspect of the deep learning approach. The recurrent neural network is one of these approaches that is useful in many natural language processing tasks, particularly the LSTM network.

2.9 Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM)

RNN is one of the most famous methods for classifying and analyzing sequential data such as voice, videos, and sequence of text (Thomas, 2019). RNN, in its simple form, is densely connected hidden layers, where the output of the hidden layer is fed back into itself to preserve the data sequence while passing the time, as shown in figure 2.3. In other words, to model the sequence-dependent data in a way that distinguishes it from feed forward networks (Thomas, 2019).

For more clarification, the unrolled network is shown in figure 2.4, where the embedding vector of each input word is fed into the RNN network. The same *F* network at time *t=1* is fed by the previous output h0 in addition to the embedding vector of the next word, producing h1, the subsequent output, and so on (Thomas, 2019).

Due to the exploding or vanishing of the feedback signals, RNN is restricted to look back to approximately 5-10 time steps. LSTM addresses this issue. The LSTM can learn 1000 time steps and more by using a solution called Constant Error Carousels (CEC) that is based on enforcing a flowing of constant error inside specific cells, which can be

accessed through granting access process which is learned via multiplicative gate units (Staudemeyer, Morris, 2019).

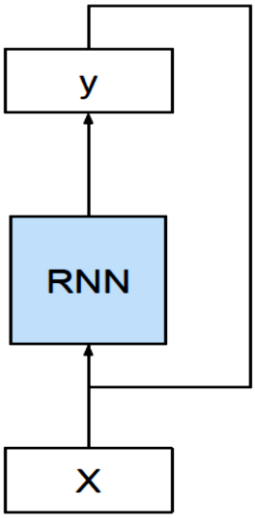


Figure 2.3 Example of simple RNN (Brownlee, 2017a)

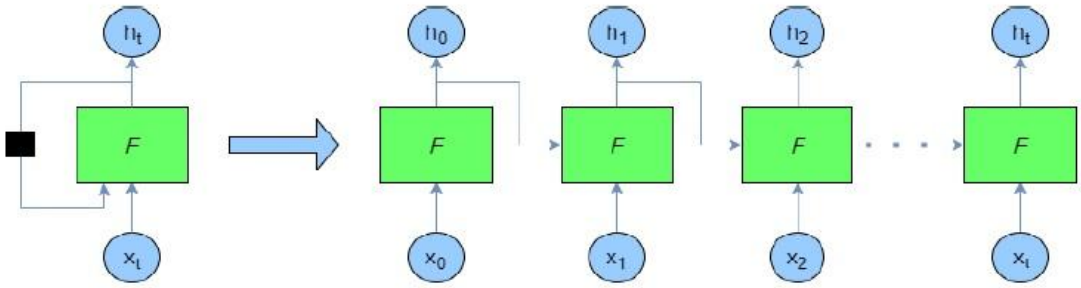


Figure 2.4 Example of unrolled RNN (Thomas, 2019)

2.9.1 LSTM Structure

Some authors like (Buduma, Locascio, 2017; Brownlee, 2017a; Thomas, 2019) have presented the LSTM architecture in various ways, but (Thomas, 2019) presented the LSTM components accompanied with mathematical equations. In chapter ten of (Thomas, 2019) book, the components of the LSTM were shown as in figure 2.5, where the output h_{t-1} of the previous cell and the current input X_t, are concatenated and squashed to be in the range of -1 and 1 through applying the following function :

$$g=tanh(x_t U^g + h_{t-1} V^g + b^g) \quad (2.1)$$

The weight of the current input and the weight of the output of the previous cell were expressed by U^g and V^g, respectively, and the bias is expressed by b^g, knowing that the exponent g is not a power. It merely indicates the weights and bias of inputs of each gate (input, forget, and output gate).

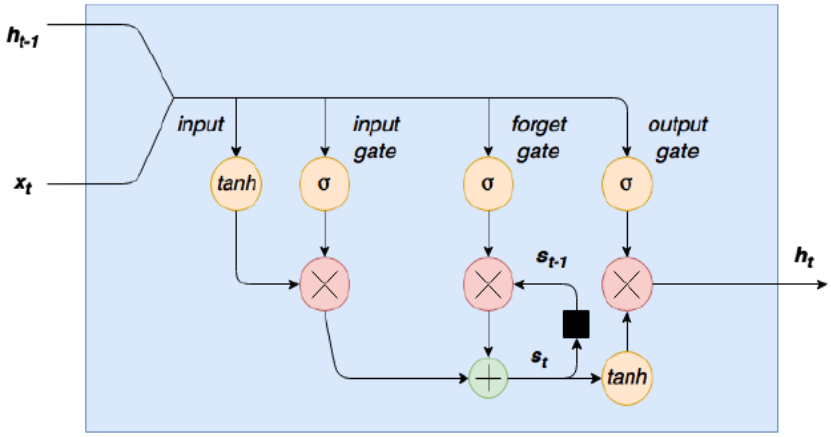


Figure 2.5 LSTM cell (Thomas, 2019)

(Brownlee, 2017a) and (Thomas, 2019) categorized the components of the LSTM into three gates: input gate, forget gate, and output gate. Contrary to (Brownlee, 2017a), (Thomas, 2019) gave more details about each gate as in the following subsections.

2.9.1.1 Input Gate

The input gate is a hidden layer represented by nodes activated by the sigmoid function. The inputs to the sigmoid function are the weighted x_t and $h_{t-1,}$ and the function's output is a value ranges from 0 to 1. The output of the input gate is multiplied element-wise with squashed input values to control the switching of input values between on and off (Thomas, 2019). The input gate expression is as follows:

$$i= \sigma\ (x_t U^i + h_{t-1}\, V^i + b^i) \tag{2.2}$$

And the output of the input gate is as follow:

$$g \circ i \tag{2.3}$$

Knowing that i represents the weight of g

2.9.1.2 Forget Gate and Internal State

Like the input gate, the forget gate is a hidden layer of a set of nodes activated by the sigmoid function. The output of the sigmoid function passes through element-wise multiplication with the delayed time step state −1, in a stage playing a role of remembering or forgetting the previous state depending on the output value of the sigmoid function, 1 for remembering and 0 for forgetting (Thomas, 2019). The forget gate can be expressed as follows:

$$f= \sigma\ (x_t U^f + h_{t-1} V^f + b^f) \quad (2.4)$$

And the final output of this stage can be expressed as:

$$s_t=(s_{t-1}\circ f)+g\circ i \quad (2.5)$$

It can be said that the forget gate controls the amount of the previous memory that should be kept, and the s_t related gradient can stay high very long time by using the gating mechanism (Goldberg, 2017)

2.9.1.3 Output Gate

The output gate is composed of two parts, the sigmoid function and tanh function, the sigmoid function for determining the state values that should be output from the cell (Thomas, 2019). The output gate expression is as follows :

$$o= \sigma\ (x_t U^o + h_{t-1} V^o + b^o) \quad (2.6)$$

And the output of the LSTM cell is expressed as follow:

$$h_t=tanh(s_t)\circ o \quad (2.7)$$

Unlike (Brownlee, 2017a) and (Thomas, 2019), (Buduma, Locascio, 2017) called the gates of the LSTM as the keep gate, write gate, and output gate.

2.9.2 Training of RNN-LSTM

In chapter two of his book, (Brownlee, 2017a) gave a brief introduction to the most used training algorithm of the LSTM that is called Truncated Back Propagation Through Time (TBPTT), where the processing of the sequence occurs one step at a time and an update for a fixed number of time steps occurs periodically. TBPTT is an adjusted form of Back Propagation Through Time (BPTT):

1. Presenting pairs of input and output of k1 time steps sequence to the network.
2. Unfolding of the network, calculating, and accumulating the errors of k2 time steps.
3. Rolling up of the network, and updating the weights to take a place.
4. Repeat the previous steps.

K1 denotes the number of time steps in the forward pass between updates, while k2 denotes the number of time steps that BBTT should be applied to them.

More descriptions about the LSTM training process with mathematical representation are fully detailed in (Staudemeyer, Morris, 2019)' paper.

2.9.3 LSTM Topologies

There are different topologies and varieties of LSTM-RNN, which can be obtained from the primary method. The most common topologies are vanilla LSTM, Gated Recurrent Unit (RGU), bidirectional LSTM, multidimensional LSTM, sequence_to_sequence, attention_based_learning (Staudemeyer, Morris, 2019), LSTM (Soliman, Eissa & El-Beltagy, 2017), and stacked LSTM (Soliman, Eissa & El-Beltagy, 2017; Staudemeyer, Morris, 2019).

2.9.3.1 Stacked LSTM

Stacked LSTM is a stack of LSTM layers one above the other to increase the network capacity (Staudemeyer, Morris, 2019) and improve the performance in some tasks than the shallower one (Goldberg, 2017). Such architecture that allows storing more information (Smagulova, James, 2019) is called deep RNN. If we have k RNNs, such as RNN_1,...., RNN_k, then the input of $_j$th RNN where j>=2 is the output of the preceding layer j-1 (Goldberg, 2017). Figure 2.6 shows an example of stacked LSTM.

LSTM is a data-hungry deep learning neural network requiring a data augmentation technique, especially with insufficient resources.

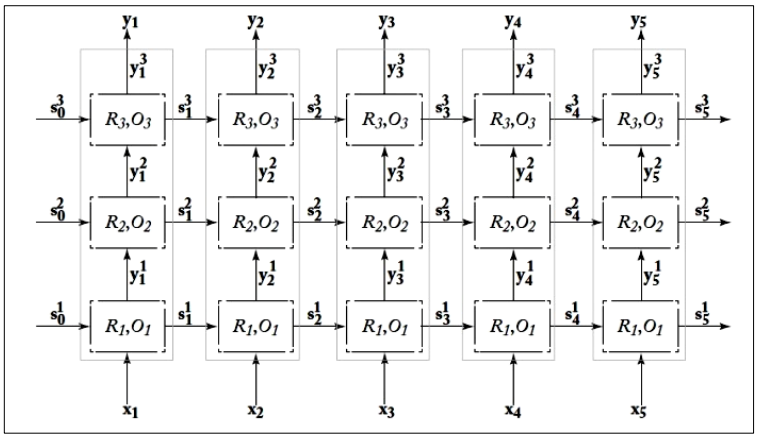


Figure 2.6 stacked LSTM (Goldberg, 2017)

2.10 Data Augmentation

Data augmentation (DA) is a technique that copes with the lack of data, and it contributes to boosting the data and model performance. Two data augmentation techniques were utilized by (Luque, 2019) in a shared task of his participation in TASS 2019 to detect the sentiment polarity of Spanish tweets. The two techniques were: 1- machine back translation from Spanish to other four pivot languages (English, Arabic, Portuguese, and French) and other 200 languages, and back to Spanish. The translation process is done using Google's cloud translation API services. 2- Instance crossover: a novel technique tested by (Luque, 2019) inspired by a genetic algorithm cross over operation, where a new instance is created by combining two halves, one half from the original instance while the other represents a half of another instance. The instance crossover technique augmentation showed feasible results, despite using the logistic regression algorithm, while the back translation of the four pivot languages (English, Arabic, Portuguese, and French) does not improve model performance.

In the same vein, another technique called contextual augmentation was proposed by (Kobayashi, 2018). The proposed technique supposed that a sentence stays invariant, even if words in it are replaced with words with a paradigmatic relation. The technique of (Kobayashi, 2018) provides a wide range of words to replace the original text words with predicted words from its context utilizing bi-directional CNN or RNN in multi-tasks of classification.

The learning model was altered to avoid word replacement with another one incompatible with the original label of the text.

Combinations of three augmentations were tested in the experiments of (Kobayashi, 2018) using six bench datasets 1 and 2, namely SST5 and SST2, which are movie reviews of Stanford sentiment tree bank that are binary labelled, 3- TREC, which is six question types classification dataset. 4- RT movie review sentiment dataset. 5- MPQA short phrase polarity prediction, and 6- Subjectivity dataset (whether the sentence is objective or subjective), knowing that 10% of the training part was used for validation.

The results showed that the proposed method boosted the model performance more than the augmentation based on synonyms, achieving an average accuracy of 78.20% across all datasets when using CNN while achieving a 77.83% value of the same metric when using RNN.

Similarly, (Sharami, Sarabestani & Mirroshandel, 2020) suggested three DA methods for extracting sentiments at sentence level from low resources Persian dataset, synonym DA, extra data augmentation to increase the small number of instances, and translation DA.

Contrary to (Luque, 2019)**,** (Sharami, Sarabestani & Mirroshandel, 2020) obtained results that showed that DA by translation contributes to achieving the highest performance of Bi-LSTM and CNN models used in binary multiclass classification when compared to their baseline.

A similar idea has been supported broadly by (Sun, He, 2020), who proposed multi-granularity DA techniques, including sentence level, phrase level, and word level, that enhanced their LSCNN hybrid neural network-based model, which is evaluated on a corpus of Chinese news headlines and a dataset of Chinese online comments.

At word-level DA, four parts were considered by (Sun, He, 2020): 1- Substituting words with synonyms using the thesaurus, where Stanford parser parsed sentences and substituted certain words with their synonyms. 2- Substituting word2vec where similar words were searched for in the semantic space to enrich the substitution process. 3-

Translate the sentence by adding some words that have no meaning to either side of the sentence, left or right, resulting in a sentence that retains its semantic. 4- Insert meaningless words into the sentence, where the number and positions of inserted words are randomly selected. At phrase-level data augmentation, (Sun, He, 2020) focused on two attributes of contemporary Chinese, adverbial phrases and attribute head. More details regarding these attributes in (Sun and He, 2020), while at sentence-level DA, the focus was on deleting the sentence that does not affect the sentiment of the text, using two methods one of them is filtering the words utilizing sentiment dictionary, where the other one is using an automatic classifier such as SVM. In the end, the final result of the voting is obtained.

(Chen, Ji, 2019) made a similar point in their study of addressing the problem of the explainability lack of the sentiment classifier. (Chen, Ji, 2019) proposed two methods of DA by creating more training instances. One of these methods was based on using a list of predefined words of sentiment as external knowledge, while the other method was based on adversarial instances. The methods of (Chen, Ji, 2019) were 1- tested using LSTM and CNN classifiers and three benchmark datasets (MR, IMDB, and SST). 2- Evaluated by both human evaluators and automatic measurements. The results showed that data augmentation by providing more examples improves the model prediction's explainability.

It is worth mentioning that (Mohammed, Kora, 2019) adopted data augmentation by shuffling technique in their proposed deep learning CNN, RCNN, and LSTM models, to analyze the sentiment of MSA and Egyptian dialects tweets.

Overall, these studies highlight the need for integrating DA techniques with deep learning models due to their impact on models' performance and achievement. In this research, the swapping augmentation technique was applied to this research dataset. Ensemble learning is another technique for enhancing the SA of textual data.

2.11 Ensemble Learning and Sentiment Analysis

Ensemble learning is a way of building a predictive model that combines the output of multiple learners by treating them as a decision-maker committee. The ensemble's idea

based on the committee's overall accuracy will be better than single-member (Zaman, Hirose, 2011). In a recent study of ensemble learning by (Haje *et al*., 2022), a hybrid model of ensemble learning and convolutional neural network(CNN) was used to improve the accuracy of emotion detection.

Three datasets were used in (Haje *et al*., 2022) study, namely IMDB, SemEval, and PL04. The datasets comprise 50,000, 3441, and 2000 instances, respectively. The base learner of the ensemble model is composed of a decision tree (DT), k nearest neighbour (KNN), Bayesian network (BN), and support vector machine (SVM). The proposed model was assessed using F1, precision, recall, and accuracy metrics. The results showed that the proposed model achieved the highest accuracy, precision, recall, and F1 score at 95%, 92%, 98%, and 95%, respectively, on the PL04 dataset, while the lowest accuracy, precision, recall, and F1 score on SemEval with 88%, 88%, 98%, and 89% respectively. No details about hybridizing the CNN with the ensemble model.

(Al-Hashedi *et al*., 2022) conducted a study to analyze the Arab emotions towards the conspiracy theory of COVID-19. To achieve their study objectives, (Al-Hashedi *et al*., 2022) collected 1026 tweets by tracking related keywords through the Twitter API. The tweets were annotated to positive and negative and passed through a preprocessing step, including noise removal, normalization, tokenization, streaming, and stop words removal. The preprocessing step has been followed by a word embedding and Synthetic minority oversampling technique for the nominal and continuous (SMOTENC) process. The dataset was divided into training and testing parts with 90% and 10%, respectively, with cross-validation of 10 folds. The evaluation metrics were accuracy, precision, and F1 score.

Two kinds of classification methods were implemented by (Al-Hashedi *et al*., 2022). The first one is based on single classifiers such as Bernouli Naïve Bays (BNB), Stochastic gradient descent(SGD), Logistic Regression(LR), and Linear support vector machine(LSVM). In contrast, the second method is based on ensemble classifiers that include multi combinations of classifiers that consider the majority voting. According to the results of the applied re-sampling technique, the Random Forest with voting was selected.

The features were represented using two pre-trained continuous bag of words (CBOW) models. One is called Arabian.news, while the other is called cc.ar, 300.

Many experiments were conducted by (Al-Hashedi *et al*., 2022) for both classification methods, single-based and ensemble-based, using CBOW pre-trained model, with and without SMOTENC.

The obtained results showed that the best performance was achieved when applying SMOTENC for both single-based and ensemble-based.

In a study by (Kazmaier, van Vuuren, 2022), a heterogeneous ensemble was constructed by applying many techniques, which were compared and evaluated across four datasets that fall into different domains. In Addition, an approach for selecting ensemble members was proposed.

The experiments that were conducted by (Kazmaier, van Vuuren, 2022) started with single classification models, which were evaluated on four benchmark datasets, namely PL04 (2000 binary movie reviews), Yelp (3801 binary business reviews), Twitter (4200 ternary social media tweets), and SMS(2500 text messages as comments on the services provided by South African retail bank). The single classification algorithms are LSTM, CNN, naïve Bayes, logistic regression, support vector machine, and a feed-forward neural network.

Nine different combinations of the bag of words features representation were implemented with the latter four learning algorithms, whereas word embedding was trained for LSTM and CNN models. In addition to four pre-trained models based on Hue and Liu's lexicons, Sentiwords, Vader, Pattern, and Opinion, which resulted in a pool of base-learners equal to ((4*9) +2+4), which were trained on 80% and tested on 20% of each dataset.

The results showed that the performance of combined pre-trained ensemble models with specifically trained models gave a better average than the members.

In a study for enhancing sentiment classification, a customized ensemble model was proposed by (Alsayat, 2022) to classify tweets on the COVID-19 pandemic. The

customized ensemble model is based on FastText for word embedding, and LSTM detects the relationship between the words and understands the unseen words by recognizing the prefixes and suffixes from the training data.

Three different datasets were used by (Alsayat, 2022). The first one is composed of 18,000 COVID-19 tweets. 70% of the tweets were used for training, while 30% were used for validation and testing. The dataset is balanced with 50% positive and 50% negative tweets. The second dataset includes Amazon reviews and Yelp reviews. The third dataset is web2.0 and contains six sub datasets from social media platforms such as Digg comments, Runners, BBC, YouTube, My Space, and Twitter.

The datasets were divided into training, validation, and testing by applying ten folds cross-validation. The customized ensemble model was created by changing the number of hidden layers and number of neurons per layer which were tuned using the GridSearch technique. The weighted voting method was applied to aggregate the prediction results.

In the study of (Alsayat, 2022), it was concluded that the best classification was achieved when using 200 neurons and two layers for the validation set.

Additional methods such as natural language processing APIs of Google, Microsoft Azure, and IBM are used as members of the ensemble learning models. The obtained accuracy of the ensemble model composed of IBM+ Microsoft+ Goggle+, a customized deep learning model, showed an improvement of 2-5% than others.

A study was conducted by (Luo *et al*., 2021) to apply stacking ensemble by adopting heterogeneous models, including deep learning and traditional learners. The deep learners entail a convolutional neural network (CNN), long short-term memory (LSTM), and a combined model of CNN and BiLSTM, while the traditional learner was the support vector machine (SVM). All these learners represent the base or first-level learners, while the second-level learner was a simple neural network. There is no mention of the name of the simple neural network.

(Luo *et al*., 2021) proposed a doc2Vec method for effective feature extraction using paragraph vector (PV-DM) algorithm, where a concatenation of word vectors with paragraph vectors was applied.

The ensemble model of (Luo, *et al*, 2021) was compared with a gated recurrent unit (GRU), attention mechanism in CNN (Att-CNN), CNN, LSTM, CNN-BiLSTM, and machine learning stacking that includes Bayes classifiers, KNN, and SVM.

All single and ensemble classifiers were evaluated on IMDB and Weibo datasets. The experimental results showed a significant improvement in terms of accuracy metrics of sentiment classification when integrating the SVM with neural networks.

In 2022, an ensemble learning model for textual sentiment analysis based on bidirectional encoder representation from the transformer (BERT) was proposed by (Lin, Kung & Leu) to identify the harmful news. To achieve their objective, (Lin, Kung & Leu, 2022) created a framework composed of two phases. Phase 1 includes constructing a harmful news dataset using publicly available political fake news datasets. This dataset was edited by adding new classification labels representing the sentiment and harmfulness for each instance through crowd sourcing strategy.

A model was established by (Lin, Kung & Leu, 2022) to analyze the correlation between text sentiment and harmful news. The correlation analysis took place between three attribute labels of the dataset with different combinations: fake news, harmful news, and sentiment polarity. The conducted correlation experiments were: 1- the correlation between harmful news and sentiment, 2- the correlation between fake news and sentiment, 3- and the correlation between harmful news and fake news.

Phase 2 of the framework included creating a model for classifying the harmful news with sentiment features and combining this model with ensemble learning. This phase entails three stages: 1- Input stage, where a pre-classification of the harmful news into positive, negative, and neutral took place. After that, the dataset was divided into training, validation, and testing sets with a ratio of 60%, 20%, and 20%, respectively. 2- Process stage, which inputs the sentences into the BERT model for feature representation, and the obtained sentiment features were passed into the ensemble

learning model. 3- Output stage, where the calculation of the harmful probabilities occurred and sigmoid activation function was used to the results of news identification.

The ensemble method used by (Lin, Kung & Leu, 2022) was a combination of bagging and stacking. The results showed that the pre-classification of sentiment with the ensemble was very useful in classifying the harmful news.

The effect of the ensemble on the performance in the case of domain shift was investigated by (Özbey, Dilekoğlu & Açiksöz, 2021). Many experiments were conducted using the Amazon dataset of product reviews to analyze the impact of changing bagging technique parameters, such as the sample size of training data and the number of weak classifiers. The probability prediction measure based on distance was investigated to find its impact on the number of classifiers in case cross-validation does not represent an option for target data.

The logistic regression (LR) classifier was trained in the experiments of (Özbey, Dilekoğlu & Açiksöz, 2021). Unigram and bigram were adopted for features representation, while the term frequency-inverse document frequency (TF-IDF) was used for weighing and vectorizing the data.

The results showed the robustness of the proposed bagging-based model for the ensemble.

In summary, most studies used different datasets that covered different domains; some are publicly available, and others are collected using Twitter API and other tools.

In this study, heterogeneous ensemble learning techniques represented at stacking were applied to analyze the sentiment of a parallel dataset of three languages, English, MSA, and BDs, that cover Amazon product reviews, applying word embedding for feature representation.

2.12 Transfer Learning

Transfer learning (TL) is a technique where knowledge gained from a rich training data source can be exploited in another related but different domain (Omara, Mosa & Ismail, 2019).

2.12.1 Theoretical Background

A set of terms must be used when talking about TL: source domain, source task, target domain, and target task.

A previous survey of (Weiss, Khoshgoftaar & Wang, 2016) identified the domain as $\mathcal{D}=\{\mathcal{X}, p(X)\}$ that is composed of two parts. A space of features $\mathcal{X}$ and a marginal probability P(X), where $X=\{x_i,\ldots,x_n\}\in\mathcal{X}$.

x_i is the ith sample or feature vector, n is the number of samples or instances, X is a specified learning instance. $\mathcal{X}$ is the feature space of all possible vectors.

The task $\mathcal{T}$ is also defined as $\mathcal{T}=\{\mathcal{Y}, f(.)\}$That is composed of two parts. A space of labels ($\mathcal{Y}$) and a prediction function f(.). The function f(.) is learned from feature vector and pairs of labels $\{x_i, y_i\}$where $x_i \in X$ and $y_i \in \mathcal{Y}$. y_i is a label in label space $\mathcal{Y}$. f(x) is the function used to predict the label value of (x)

The domain of source is defined as $\mathcal{D}_s = \{(x_{si},y_{si}),\ldots., (x_{sn},y_{sn})\}$, where x_{si} is the i-th instance $\in \mathcal{X}_s$ and y_{si} is the corresponding label of x_{si}.

The target domain is defined as $\mathcal{D}_T = \{(x_{T1}, y_{T1}), \ldots\ldots., (x_{Tn},y_{Tn})\}$, where x_{Ti} is the i-th instance $\in \mathcal{X}_T$ and y_{Tn} is the corresponding label of x_{Tn}.

According to previous definitions (Weiss, Khoshgoftaar & Wang, 2016) transfer learning is the process of using knowledge from $\mathcal{D}_s$ and $\mathcal{T}_s$ in $\mathcal{D}_T$ and $\mathcal{T}_T$ for the purpose of improving the f(.) when $\mathcal{D}_s$ not equal to $\mathcal{D}_T$ or $\mathcal{T}_s$ not equal to $\mathcal{T}_T$.

To date, several studies have shown that transfer learning reduces the training time of the model in the task of the target domain and improves its generalization ability.

2.12.2 Related Works

(Omara, Mosa & Ismail, 2019) applied their proposed CNN character-based model in TL between two domains, SA and emotion detection in Arabic textual data. The source data of SA are written in modern standard Arabic and dialects, covering reviews in

multi-domains such as hotels, movies, books, products, restaurants, and tweets beside a raw set composed of 492 entries.

The target dataset of emotion detection is constructed from two available Arabic datasets. The first is the Emotion Tone dataset, and the second is used in the SemEval-2018 task.

The synonymy replacement augmentation technique was used by (Omara, Mosa & Ismail, 2019) on the part of the second dataset by utilizing a lexicon of 250 sentiment words.

The proposed CNN model by (Omara, Mosa & Ismail, 2019) achieved an enhanced performance with 1.3% in the SA while achieving a 95.24% accuracy metric in emotion detection using transfer learning compared with 64.14% in traditional models.

In their study of transfer learning, a similar point was made by(Osama, El-Beltagy, 2019). They shed light on TL's importance in inducing information from pre-trained models of SA to the model of predicting emotion intensity.

In their study, (Osama, El-Beltagy, 2019) aimed to predict the intensity of a specific emotion included in a tweet using a score ranging from 0 to 1, where 1 indicates the maximum value of emotion intensity and 0 indicates no emotion in the text.

The dataset used by (Osama, El-Beltagy, 2019) is the one provided by the first task of SemEval-2018: Affect in tweets, which was divided into four categories: sadness, anger, fear, and joy. Every tweet was labelled with a score ranging from 0 to 1. The results obtained by the attention layer of DeepMoji achieved the best results across the other models of TL.

A broadly similar point has been made by (Dong, De Melo, 2018), whose CNN model achieved effective TL results on a different dataset that covers various languages.

To achieve their objectives, TL was utilized by (Dong, De Melo, 2018) to benefit from supervised pre-trained models on various tasks from various domains. TL was applied

using SVM on multi domain datasets that cover 25 categories of product reviews of Amazon customers.

In their survey of TL-based SA, different SA TL methods were presented by (Liu *et al*., 2019). These methods are 1- Parameter transfer, where the model's parameters trained on many datasets can be transferred from the source to the target. An example of this method is the word2vec technique, through which the TL process occurs in the first layer of the target model, 2- Instance-transfer method, where a reweighing process filters sharing of data between the target and the source. In this method, the labelled samples of the source can be used to augment the sample of the target domain, and 3- feature-representation transfer, where the source and target domain should have shared features. An essential step in this method is transforming the source and target data into the same features' space via feature transformation.

In conclusion, these studies show that TL plays a significant role in achieving better generalization due to using the pre-trained model on a large and multi-domain corpus, which in turn leads to shedding light on the investigation of TL by providing more pre-trained models for MSA and Arabic dialects which will contribute to addressing some of the Arabic NLP challenges.

2.13 Summary

Arabic NLP faces many challenges: the morphological nature of Arabic language sentences, lack of datasets, and limited exploration of dialects.

There are differences between the dialects and modern standard Arabic (MSA). MSA is used for formal communication, while dialects are used in daily informal communication.

Some of the created datasets of Arabic NLP covered different dialects and domains, some of which covered Maghreb countries like Algeria, Tunis, and Morocco, and others covered the Levantine, Egypt, and Gulf countries. These datasets are composed of tweets or comments covering different domains such as news, movies, books, etc. Twitter API was the most common tool to collect the dataset contents.

After creating the datasets, they pass through preprocessing steps and feature representation, followed by the SA process, which traditional or deep learning models perform.

Sometimes, there are no resources for particular dialects; in this case, the translation approach is used to create a dataset. The translation can be done manually or by using machine translation. Some translation issues may arise due to the target and source language data distribution. These problems can be addressed by using the multilingual SA.

Various deep learning approaches can be utilized to analyze the multilingual SA, some of which are convolutional neural networks (CNN) and recurrent neural networks (RNN). Recurrent neural networks (RNNs) are more useful for analyzing sequential data. One of these RNN networks is long short term memory (LSTM) due to its architecture characterized by an input gate, forget and memory state, and output gate.

Because of the greedy nature of the deep learning models for data, augmentation techniques could be used. Some techniques include back translation, deletion, insertion, synonyms replacement, and swap.

The data augmentation techniques and an excellent pre-trained learner model are good motivators for transfer learning, especially with scarce source dialects. Transfer learning methods are used in SA, including parameter transfer, instance transfer, or feature-representation transfer.

Other techniques can enhance the SA process, one of which is ensemble learning, which, based on the committee's overall accuracy, will be better than a single member.

Chapter 3- Dataset Design and Preprocessing

3.1 Introduction

Parts of the work included in this chapter were previously published in (Omran, T.M., Sharif, B.T., Grosan, C. and Li, Y., 2022. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, p.102106).

This chapter represents a part of phase 1 of this research framework. This research framework comprises two phases, as depicted in figure 3.1.

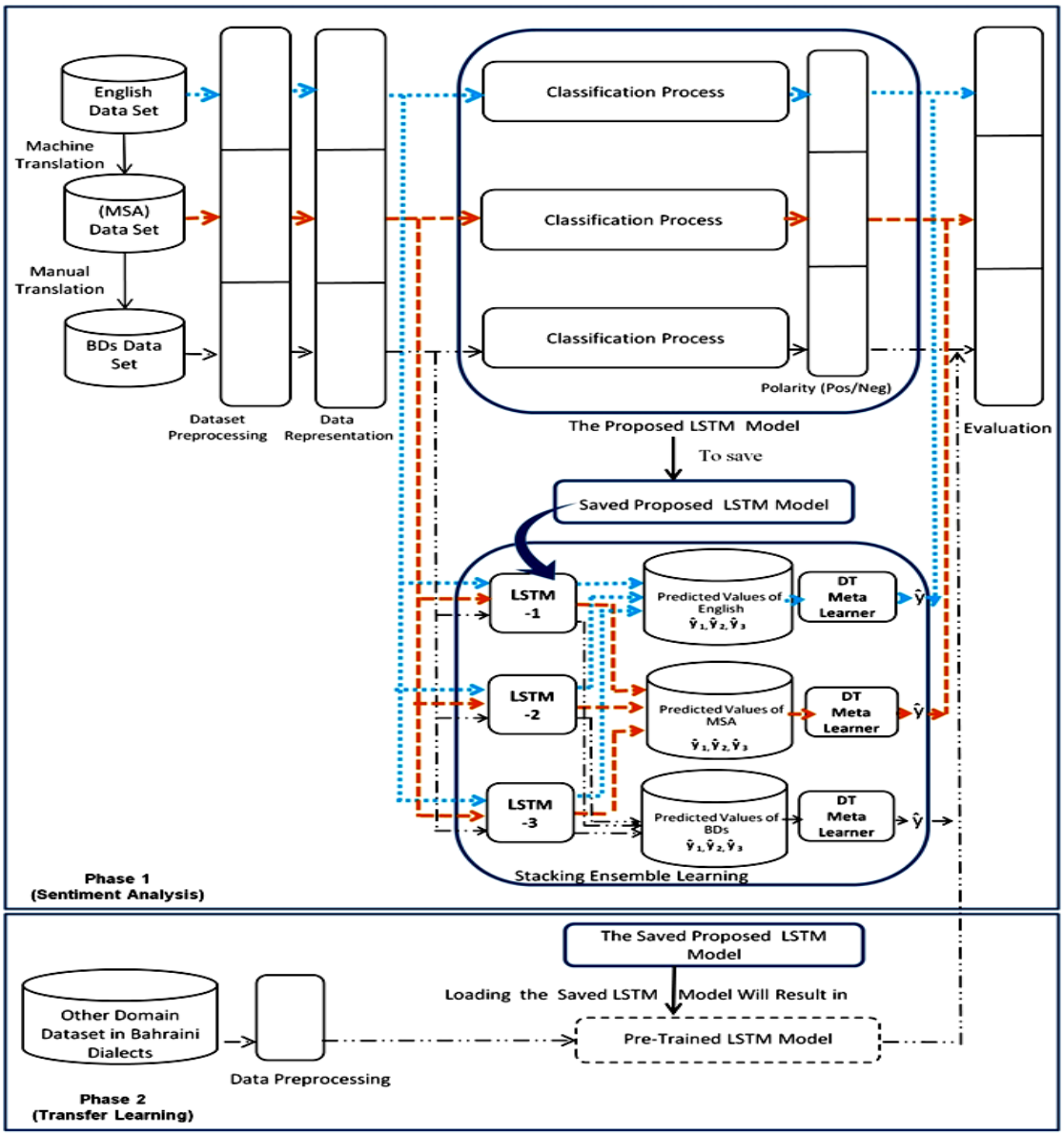


Figure 3.1 Research's framework

Phase 1 includes five sub-stages: 1- Creating a dataset of Bahraini dialects that suffers from a lack of resources by taking advantage of NLP-rich source language such as English through translating Amazon's product reviews dataset to MSA, which was then converted to its corresponding Bahraini dialects; 2- Preprocessing of the English, MSA, and BDs reviews. 3- Creating the word embedding of the preprocessed reviews through data representation. 4- Analyzing our parallel dataset of product reviews sentimentally using our proposed LSTM model and evaluating its performance, detailed in chapter 4. 5- Enhancing our proposed LSTM model by incorporating it as a base learner in an ensemble learning process. Chapter 5 describes this stage of this research framework. In contrast, Phase 2 represents the transfer learning process, where the knowledge gained in Phase 1 was exploited in another SA task of another domain. Detailed methods and experiments of phase 2 were described in chapter 6.

3.2 Dataset Design and Preparation

During the literature review, it was noticed that Arabic NLP faces many challenges: the scarcity of Arabic resources, either in MSA or Arabic dialects. Some resources in Arabic dialects include Egyptian, Levantine, Tunisian, and Saudi. Still, there is a big gap in BDs resources that needs bridging. In this research, a dataset of Bahraini dialects is created and presented to tackle this issue. The dataset's creation was done by taking advantage of the availability of English resources through the translation approach, where the English dataset of Amazon product reviews was translated by machine to MSA and validated manually. The resulting MSA reviews were converted to BDs by specific participants. The creation of the BDs dataset passed through a series of steps that spanned from December 2019 to February 2021.

1. An English dataset of Amazon reviews covering electronics, books, movies, and others, was downloaded from the following Kaggle link: https://www.kaggle.com/bittlingmayer/amazonreviews (Accessed on December 1, 2019). Kaggle is an online community that belongs to Google LLC. It comprises practitioners of machine learning and data scientists. It provides the users with many opportunities, one of which is finding and publishing datasets (Kaggle, 2019). The downloaded English Dataset was divided into two commas separated values (CSV)

training and testing files. The training file comprises more than 3milion negative and positive reviews labelled as label_1 and label_2, respectively. Similarly, the testing file in the number of reviews. Among these reviews, some covered the same features for the same product. For example, it might be found that many reviews include only the corrosion of the grill coating.

2. 2500 positive and 2500 negative reviews were selected manually from the training and testing files of the English dataset of Amazon reviews, resulting in a balanced dataset of 5000 reviews. These 5000 reviews have been selected with great care to cover most products that include multiple features and aspects of the product to ensure that the selected reviews are distinct and unique. For example, the selected grill reviews include ease of use, grilling time, temperature, browning of the meat, and other properties.
3. The selected 5000 reviews were copied to 25 separated Ms-word files. Each one includes a template of a table composed of six columns (number, label, review in English, corresponding review in MSA, corresponding review in BDs, and a city) as shown in figure 3.2. The Ms-word files have been saved as P1, P2, and P3…P25. Each of P*i* contains 200 balanced reviews. This numbering method helps in identifying each review in the stage of distributing the MSA reviews and collecting their corresponding ones in BDs, as will be explained later.

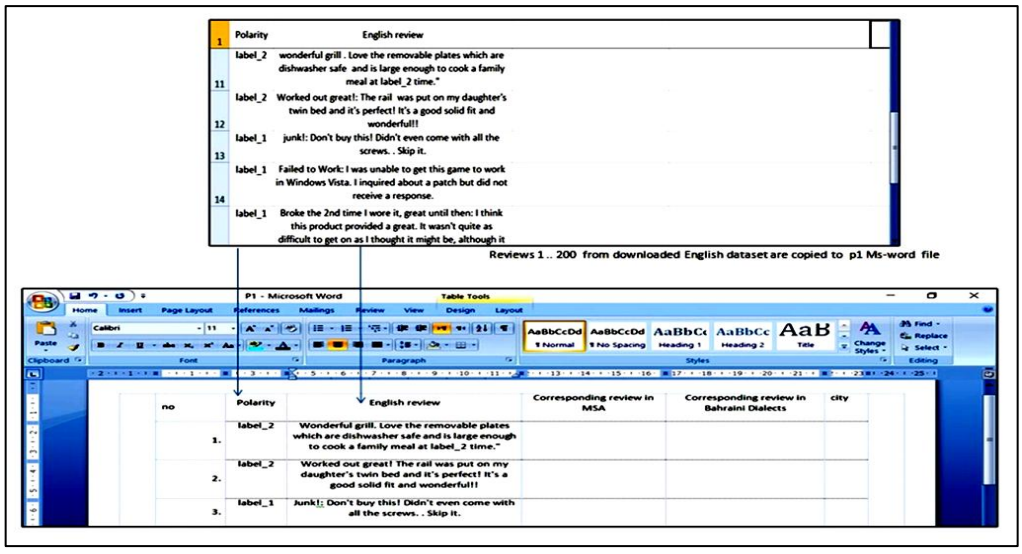


Figure 3.2 Example of a table template in Ms-word file that contains the copied reviews

4. The 200 reviews of P1 were translated to MSA one by one using https://translate.google.com/. The MSA translated review is copied to its dedicated cell in the table in the MS-word file, as shown in figure 3.3.

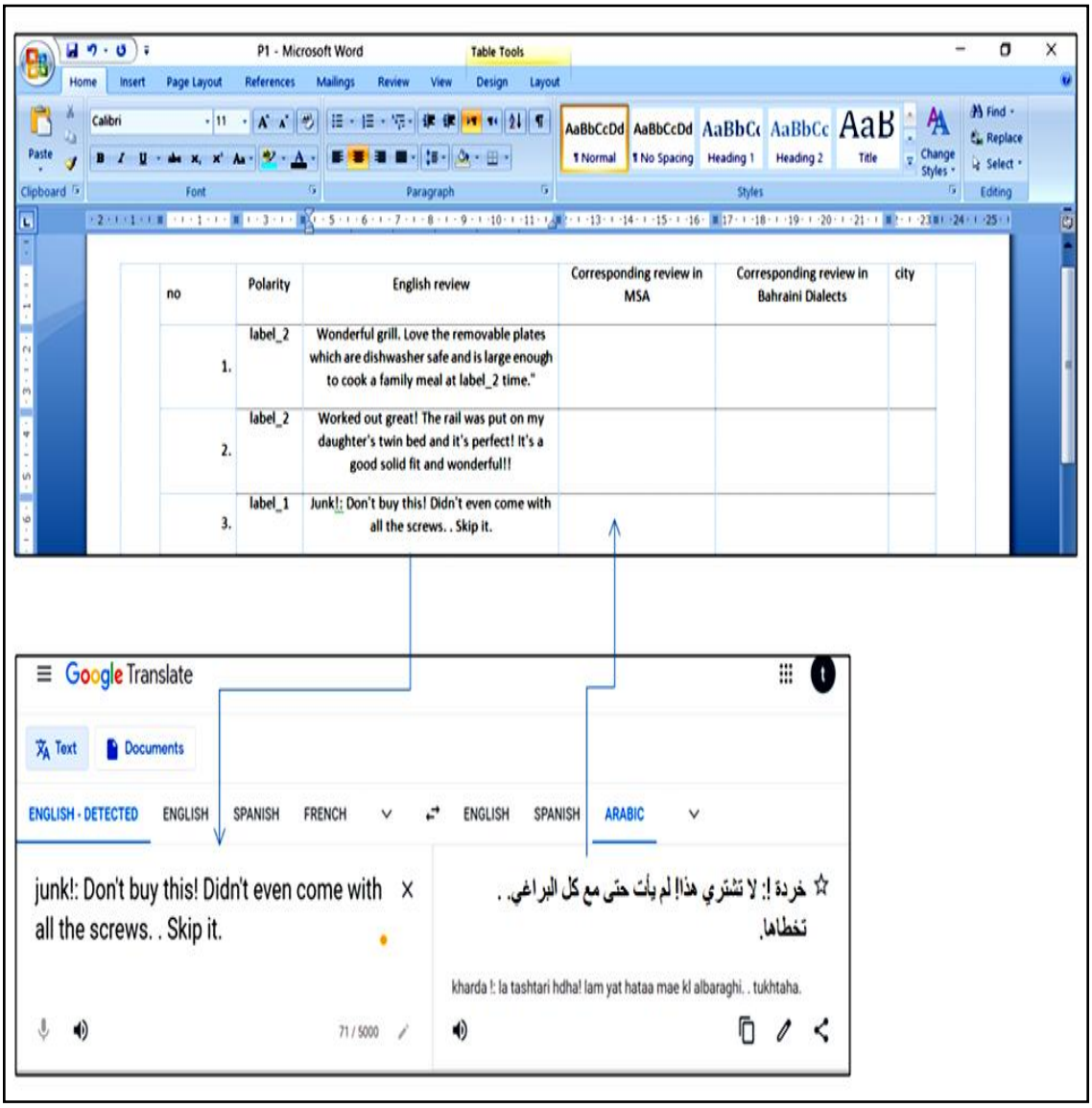


Figure 3.3 Example of translating an English review to MSA

5. The 5000 MSA reviews resulting from the translation process were distributed to 100 speakers of Bahraini dialects via 500 customized forms constructed through https://getfoureyes.com/ (Accessed on 5 December 2019). Each respondent or speaker has filled in 5 forms. Appendix1 shows an example of the full picture of one

of these forms. Each form includes 10 different and unique MSA reviews. Below each review, a text box was provided, where the respondent can rewrite the review in his spoken dialect, as depicted in figure 3.4.

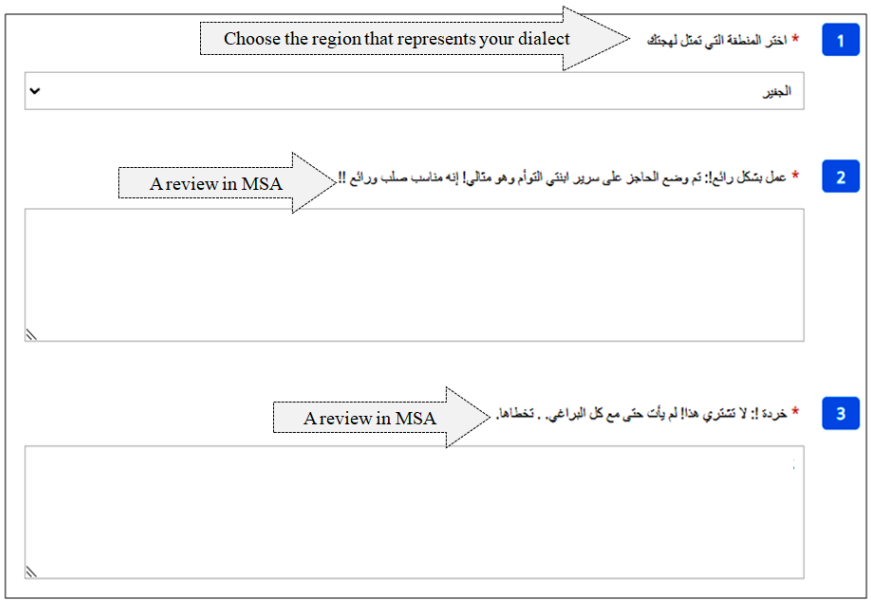


Figure 3.4 Example of MSA reviews and the provided text box

In addition, there is a combo box containing a list of Bahraini cities and villages from which the respondent select one of them that represents his dialect, as shown in figure 3.5.

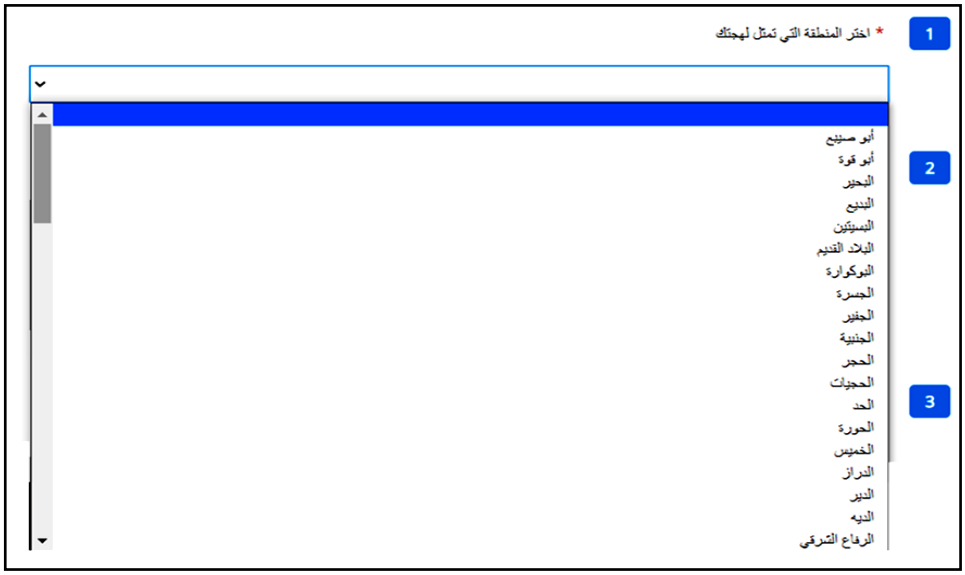
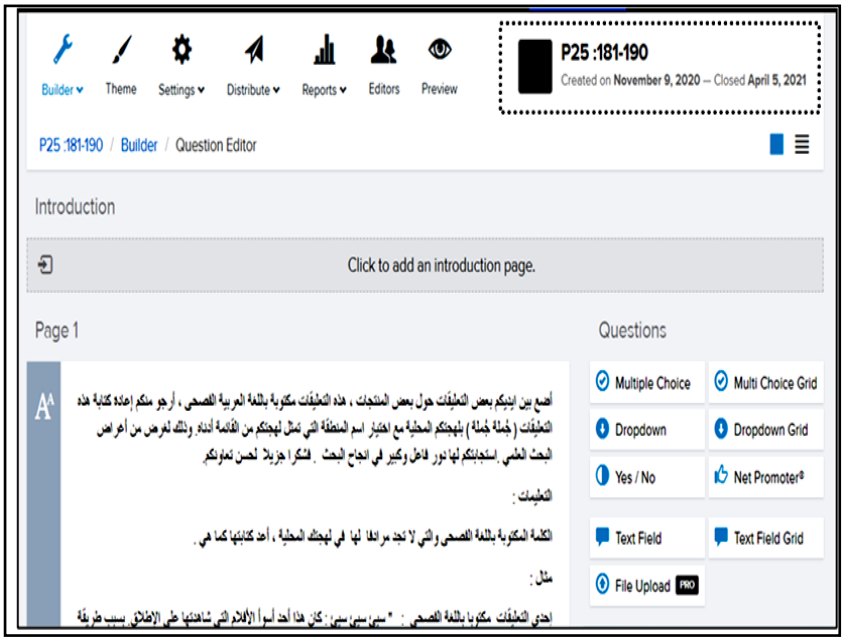


Figure 3.5 A Combo box shows some of Bahraini cities and villages

At the beginning of the form, the respondents were instructed to rewrite the provided MSA reviews in their dialects, sentence by sentence, and any word which is in MSA that they do not find any synonym for it in BDs; they have to rewrite it as it is.

Each form was identified according to the Ms-word file it affiliates to, followed by the range of reviews it contains, as shown with a dotted rectangle in figure 3.6. For example, the form identified by (P25:181-190) represents a form that belongs to the Ms-word file named P25 and contains reviews from 181-190, Keeping in mind that each file contains 200 reviews, which means creating 20 forms per Ms-word file. So, when the respondent submits the form, it is easy for the data collector to locate it and collect the responses.



Arabic text translation

I put in your hands some comments about some products. These comments are written in Standard Arabic. Please rewrite these comments (sentence by sentence) in your local dialect, and choose the region's name that represents your dialect from the list below. This is for scientific research.

Your response has an active and significant role in the success of the research, so thank you very much for your kind cooperation.

Instructions:
The word that was written in Standard Arabic, which has no synonym in your local dialect, rewrite it as it is.

Figure 3.6 The identification of the distributed form

Each form was automatically assigned a link during its setup by the https://getfoureyes.com/. All forms' links were listed in a created web page: https://lahajat866082393.wordpress.com/ (Accessed on 20 April 2021). The URL of the created web page was distributed to the respondents via text message and WhatsApp application. When the respondent visit the URL page, he/she will find the list of links as shown in figure 3.7, where he/she was asked to choose a link. If a message like "*This survey has ended and no longer accepting new responses*" is

popped out, he/she should choose another form link. Such a procedure was followed to guarantee that all forms have been responded to and no duplication in responses.

All respondents are holders of academic qualifications such as diploma, bachelor, master, and doctorate, in various disciplines.

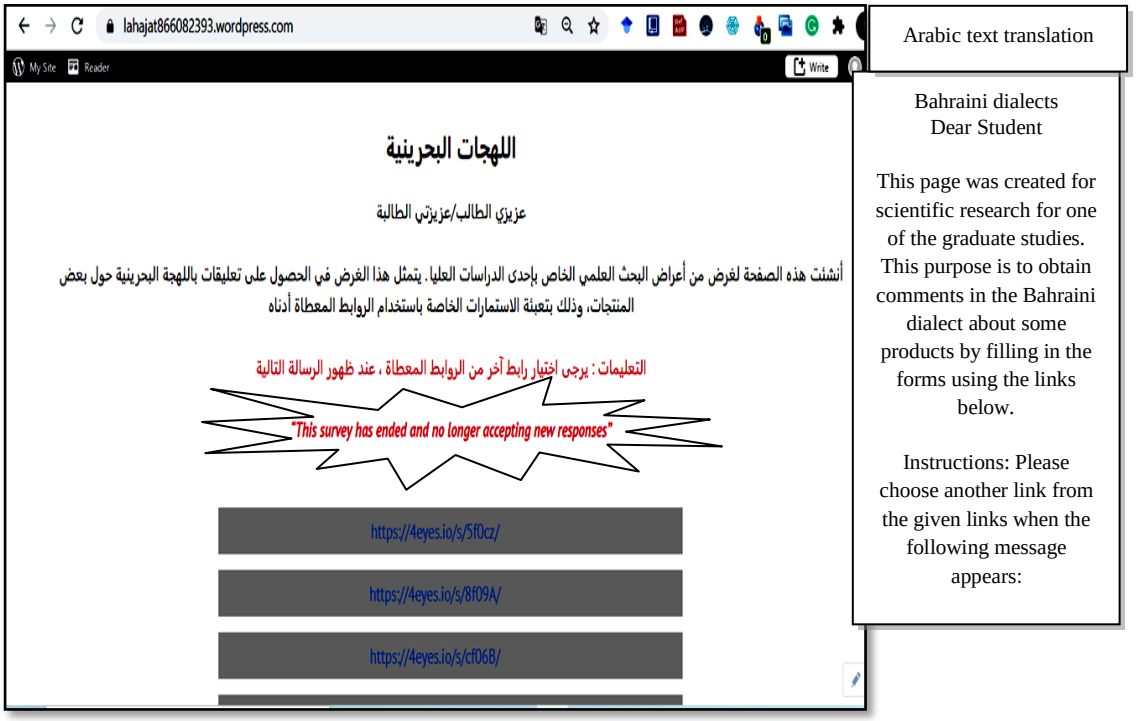


Figure 3.7 a snippet for the web page

Several criteria were considered when selecting the participants, such as filling out the forms on time, being aware of the research importance, and adequately responding as instructions were received. The overall responding was very positive. The respondents were thanked for their time and efforts. It is worth to mention that before the form distribution; all procedures for obtaining the ethical approvals were followed. Appendix 2 shows the corresponding messages of ethical approval.

6. The submitted forms that have been responded to were opened one by one by the data collector. Each submitted form contains 10 MSA reviews and their corresponding ones in BDs, as shown in figure 3.8. The reviews written in BDs were revised and

compared with the accompanying one in MSA by the data collector, a Bahraini native speaker and specialist in Arabic, to ensure that both the BDs review and the one in MSA are similar in meaning and the spelling correct. If the dissimilarity is more than 80%, the original form is re-sent to another participant; otherwise, a manual validation occurs, i.e. the spelling and meaning checking. The 80% is an approximated value; for example, if the review in MSA is composed of 60 words, and approximately 40-50 words using in the corresponding BDs reviews are not similar in meaning, the original form is re-sent to another participant, otherwise if the review in BDs is identical to its corresponding the one in MSA, the BDs review passes through a manual validation by checking its spelling and meaning. The submitted forms were re-sent to other participants in four to five cases.

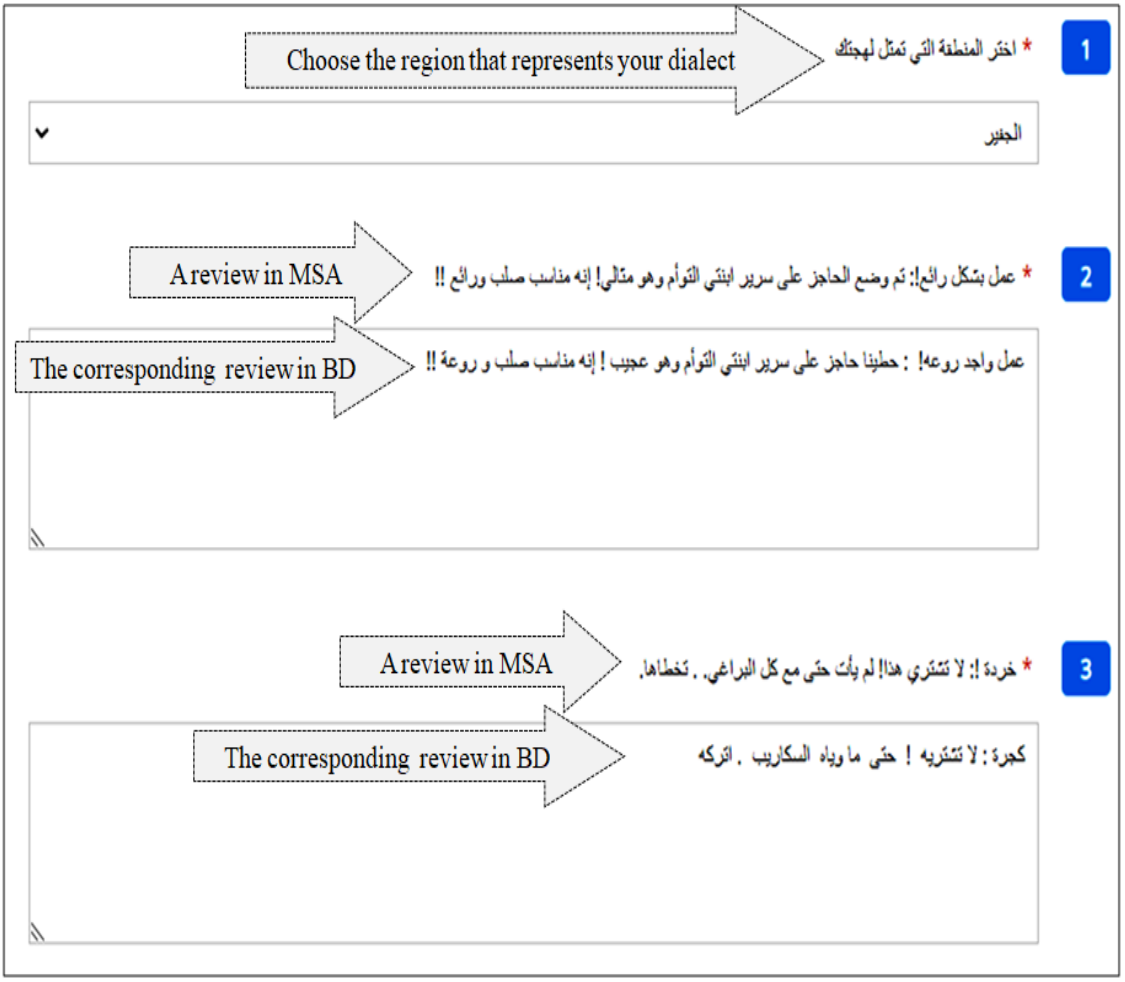


Figure 3.8 An example of MSA reviews and the corresponding ones in BDs

7. After the checking, revising, and validation process, the responses were collected by copying them to the dedicated cell in the table of the corresponding file, as shown in figure 3.9.

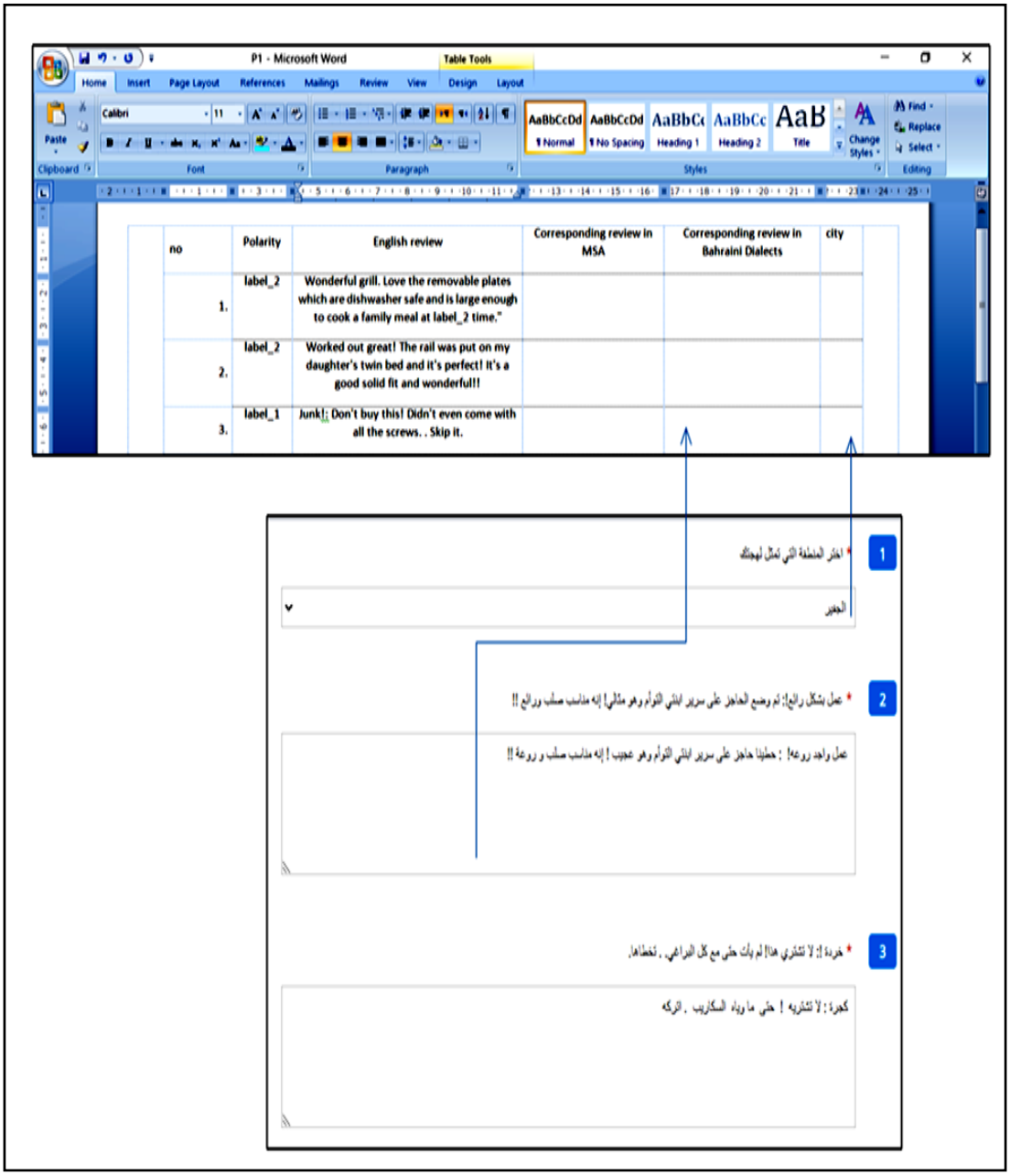


Figure 3.9 Copying the responses to the corresponding Ms-word file

The collected responses cover most Bahraini cities and villages such as Manama, Isa town, Al Nuwaidart, Sitra, East Riffa, and many others. A list of covered cities and

villages are attached in appendix 3. The complete steps of preparing the datasets are depicted in figure 3.10.

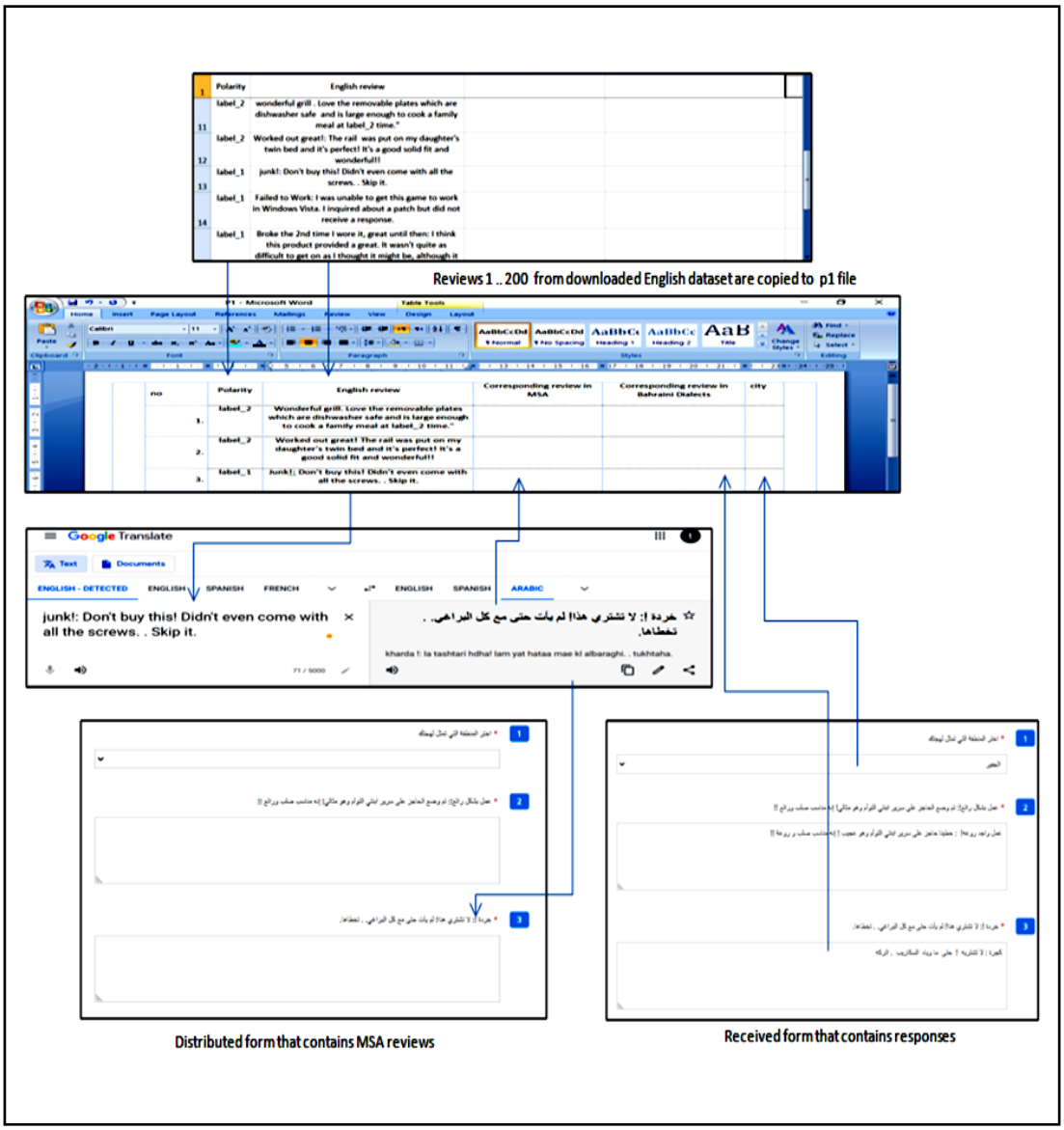
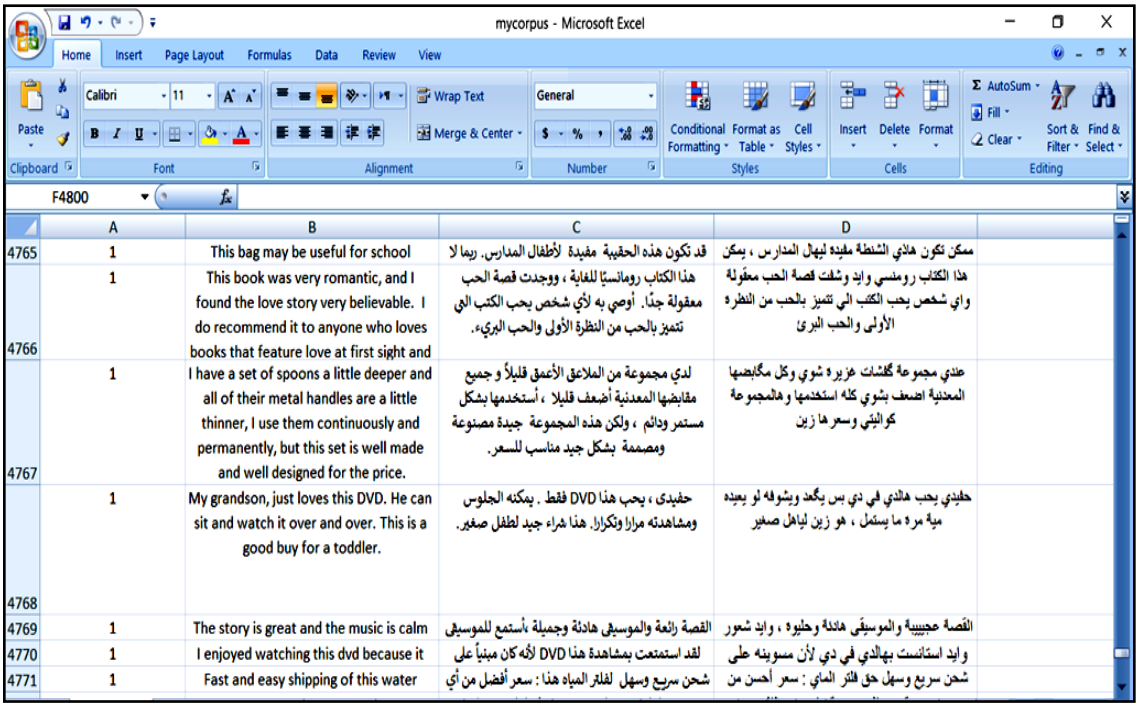


Figure 3.10 Steps of preparing this research's datasets

8. The resulting parallel dataset of English, MSA, and BDs of 5000 reviews was copied to the Ms-Excel file and converted to a text file with UTF-8 encoding, a suitable format for processing Arabic text in python. The polarity labels (label_1, Label_2)

were replaced by (0, 1), respectively. Part of the parallel dataset is shown in figure 3.11.

9. Each dataset was separated individually in a step to prepare them for SA process.



	A	B	C	D
4765	1	This bag may be useful for school	قد تكون هذه الحقيبة مفيدة لأطفال المدارس. ربما لا	ممكن تكون هالي الشنطة مفيده لهيال المدارس ، يمكن
4766	1	This book was very romantic, and I found the love story very believable. I do recommend it to anyone who loves books that feature love at first sight and	هذا الكتاب رومانسيًا للغاية ، ووجدت قصة الحب معقولة جدًا. أوصي به لأي شخص يحب الكتب التي تتميز بالحب من النظرة الأولى والحب البريء.	هذا الكتاب رومنسي وايد وشفت قصة الحب معقولة واي شخص يحب الكتب الي تتميز بالحب من النظره الأولى والحب البرئ
4767	1	I have a set of spoons a little deeper and all of their metal handles are a little thinner, I use them continuously and permanently, but this set is well made and well designed for the price.	لدي مجموعة من الملاعق الأعمق قليلاً و جميع مقابضها المعدنية أضعف قليلا ، أستخدمها بشكل مستمر ودائم ، ولكن هذه المجموعة جيدة مصنوعة ومصممة بشكل جيد مناسب للسعر.	عندي مجموعة گلشات عزيزه شوي وكل مگابضها المعدنية اضعف بشوي كلة استخدمها و هالمجموعة كوالبتي وسعرها زين
4768	1	My grandson, just loves this DVD. He can sit and watch it over and over. This is a good buy for a toddler.	حفيدى ، يحب هذا DVD فقط . يمكنه الجلوس ومشاهدته مرارا وتكرارا. هذا شراء جيد لطفل صغير.	حفيدي يحب هالدي في دي بس يگعد ويشوفه لو يعيده مية مره ما يسئمل ، هو زين لياهل صغير
4769	1	The story is great and the music is calm	القصة رائعة والموسيقى هادئة وجميلة ،أستمع للموسيقى	القصة عجيبية والموسيقى هادئة وحليوه ، وايد شعور
4770	1	I enjoyed watching this dvd because it	لقد استمتعت بمشاهدة هذا DVD لأنه كان مبنياً على	وايد استانست بهالدي في دي لأن مسوينه على
4771	1	Fast and easy shipping of this water	شحن سريع وسهل لفلتر المياه هذا : سعر أفضل من أي	شحن سريع وسهل حق فلتر الماي : سعر أحسن من

Figure 3.11 Parallel datasets (English, MSA, BDs)

3.3 Dataset Preprocessing

Two preprocessing steps took place to reduce the noise from all dataset reviews: manual and automatic preprocessing. The manual preprocessing was conducted during the translation stage, which is represented by spelling error correction and shortening of the lengthy reviews that contain sentences with a sum of words that exceed 200 by deleting sentences that do not affect the sentiment in the text. This deletion results in a review of a reasonable length by which the respondents or participants can easily capture the meaning of the sentences when converting the reviews to the Bahraini dialects. The automatic preprocessing steps were carried out using python regular expression libraries, as detailed in the following subsections.

An augmentation technique was used with dataset preprocessing. Data augmentation (DA) is one of the techniques that deal with the lack of data. It contributes to boosting

data (Luque, 2019) and model performance, especially with deep learning-based models, that need a large training data set to overcome some problems like data sparsity and overfitting (Sun, He, 2020).

There are four powerful methods of DA consisting of easy data augmentation (EDA) techniques of (Wei, Zou, 2019) such as random deletion, random insertion, synonym replacement, and random swap. The random swap augmentation method at the word level was applied in this research to take advantage of DA techniques. Random swap chooses any pair of words in the sentence in a random manner and interchange their positions. This process is done *n* times (Wei, Zou, 2019). Python NLP augmentation library provides various textual augmenters at character, word, and sentence levels (makcedward, 2021). The Random swap method was explicitly selected due to the restrictions of the BDs dataset.

The word random augmenter in python provides arguments like *min* and *max*. *Min* is the minimum number of words to be augmented, while the *max* is the maximum number. In this research, a random swap was applied two times. The first time applying augmentation using the random swap method, the minimum number of words was 1, and the maximum number of the words was 10. While in the second time, the minimum number of words was 1, and the maximum number of words was 5. Keeping in mind, that each augmented review was labeled with the same label as its corresponding original review. The maximum values of 5 and 10 were chosen because they were the best fitting for our dataset.

Figure 3.12 shows the steps of obtaining 10,000 reviews from the original 5000 reviews by applying the random swap method twice.

It is worth noting that applying the DA method consumes more execution time. It is observed that the step of saving a concatenated list of the original and augmented reviews, as showed in figure 3.12, to a file that is read, processed, and fed to the model, contributes to reducing the execution time from twenty minutes to 3 minutes. The augmentation experiments were run on Intel(R) Core (TM) i7-6500U CPU @2.50GH, with a 64-bit operating system- x64-based processor.

3.3. 1 English Dataset Preprocessing

The preprocessing steps for the English dataset were represented at:

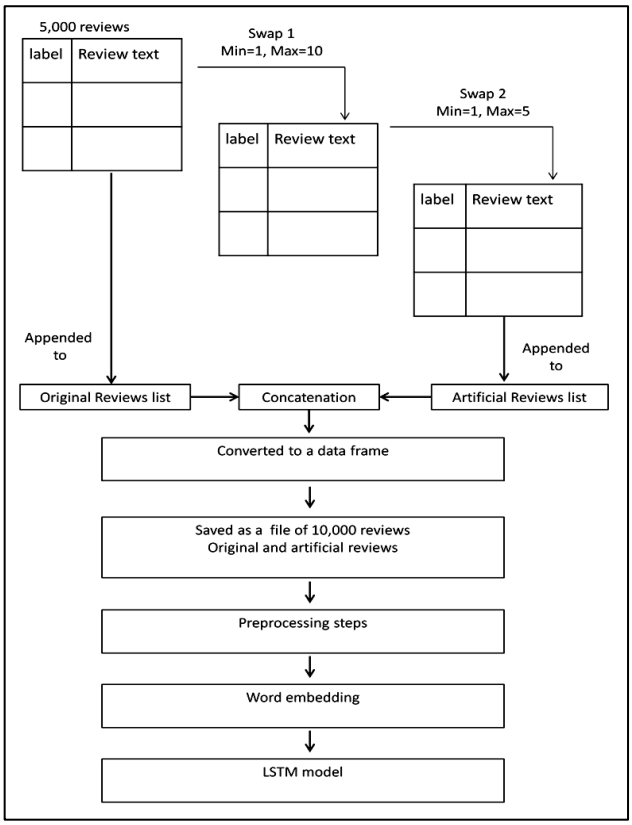


Figure 3.12 Augmentation process

1. Applying an augmentation technique called “swap” on each review in the dataset.
2. Applying another ”swap” augmentation technique on the augmented reviews that resulted from the previous step
3. Normalizing text (converting uppercase characters to lower case)
4. Replacing the emojis manually with its synonym words, for example “😊”or “☹”with “Happy”and “Sad”.
5. Removing the white spaces, non-alphanumeric characters such as “@”, tags, and digits.
6. Tokenizing the text (breaking it into words).
7. Removing the stopwords.

It is worth mentioning that the built-in list of English stopwords had been extended by adding some pronouns and words such as (I, this, it, the). Table 3.1 shows the preprocessing steps of an English review and its corresponding ones in MSA and BDs.

3.3.2 Modern Standard Arabic and Bahraini Dialect Datasets Preprocessing

The preprocessing steps of MSA dataset were as follows:

1. Applying an augmentation technique called “swap” on each review of the dataset.
2. Applying another “swap” augmentation technique on the augmented reviews that resulted from the previous step to change the word order further, which leads to considering it as a different review, thus avoiding duplicating the same review and increasing the size of the dataset, which boosts the model performance through reducing the overfitting and increasing the accuracy. The random swap randomly chooses any pair of words in the sentence and exchanges their positions. The augmentation technique was limited to the random swap method due to the scarcity of other BDs resources that enable us to use other augmentation techniques such as random insertion, random deletion, and synonym replacement. More details about the impact of data augmentation on sentiment analysis of translated textual data can be found in our published paper (Omran *et al*., 2023).
3. Normalizing some characters by replacing them by other ones, for example the character “ؤ” is replaced by “ء”, “ئ” is replaced by “ء”, “ة” is replaced by “ه”, and “گ” is replaced by “ك”.
4. Removing the punctuations, diacritics, repeated characters, English words, and digits.
5. Replacing the emojis manually with its synonym words, for example “😂”or “☹”with “سعيد”and “حزين”.
6. Tokenizing the text.
7. Removing the stopwords.

Similar to the English stopwords list, the MSA stopwords list has been extended by adding the following words:

احد, احدى ,اكون ,الأخرى ,الأخرى ,الى ,إلى ,ان ,أنا,انها, اني ,عندما, كان, لقد, مثل, و, وهى, وهي

Table 3.1 shows an example of the preprocessing steps for one of the MSA reviews.

Unlike the English and MSA stopwords list, the BDs stopwords list has been created in this research. The BDs stopwords are listed in appendix 4. An example of BDs review before and after the implementation of the preprocessing steps is shown in table 3.1.

Table 3.1 Preprocessing steps of an English review and its corresponding ones in MSA and BDs.

Stage	Preprocessing step	English Review	Corresponding Review in	
			MSA	Bahraini Dialect
Before the preprocessing (original review)	-	Boring : I bought 3 books because I read the review and they look interesting and funny. But after the first pages, I knew that this was a mistake. Is very repetitive and some of the content really ridiculous for this age and time.	مملة : اشتريت 3 كتبٍ لأنني قرأت المراجعةَ وهي تبدو ممتعة ومضحكة. لكن بعد الصفحات الأولى ، علمت أن هذا كان خطأ. مكرر جدا وبعض المحتويات سخيفة حقا لهذا العصر والزمان.	تملل : اشتريت 3 كتب لأني قريت تعليقات الناس عنها و كانت اتبين انها زينة و شيقة و تضحك بس عقب ما قرات الصفحات الاولية منها اكتشفت ان رايي غلط . الكتب فيها تكرار واااجد و محتواها تاااااااافه بالنسبة لهالزمن
After the preprocessing	First augmentation	Boring: I bought 3 books because I the read and review they look interesting and funny. But after the first pages, knew I that a this mistake was. very Is repetitive and the some of really content ridiculous for this age and time.	: مملة اشتريت 3 كتب لأنني قرأت المراجعة وهي ممتعة تبدو ومضحكة. لكن بعد ، الصفحات علمت الأولى أن هذا خطأ كان. مكرر وبعض جدا المحتويات حقا سخيفة لهذا العصر والزمان.	تملل: اشتريت كتب لأني 3 تعليقات قريت الناس و عنها كانت انها اتبين زينة و شيقة تضحك و بس عقب ما قرات الصفحات الاولية منها اكتشفت ان رايي غلط. الكتب تكرار فيها واااجد محتواها تاااااااافه بالنسبة و لهالزمن
	Second augmentation	Boring: I bought 3 books because I the read and they review look interesting and funny. after But the first pages, knew I that a this mistake was. very Is repetitive and some the of content ridiculous really for this age and time.	: مملة اشتريت 3 كتب لأنني قرأت المراجعة وهي ممتعة تبدو ومضحكة. لكن بعد ، الصفحات الأولى علمت أن هذا خطأ كان. مكرر وبعض المحتويات حقا جدا لهذا العصر سخيفة والزمان.	تملل اشتريت: لأني كتب 3 تعليقات قريت الناس و عنها كانت اتبين انها زينة و شيقة تضحك بس و عقب ما قرات الصفحات الاولية اكتشفت منها ان رايي غلط. الكتب تكرار فيها واااجد محتواها تاااااااافه بالنسبة و لهالزمن
	Normalizing letters	boring: i bought 3 books because i the read and they review look interesting and funny. after but the first pages, knew i that a this mistake was. very is repetitive and some the of content ridiculous really for this age and time.	مله 3 اشتريت قرات لاني كتب المراجعه وهي متعه تبدو ومضحكه لكن بعد الصفحات الاولى ان علمت كان هذا خطا مكر جدا وبعض المحتويات حقا سخيفه العصر لهذا والزمان	تمل اشتريت لاني كتب 3 تعليقات قريت الناس و عنها كانت اتبين انها زينه و شيقه تضحك بس و عقب ما قرات الصفحات الاوليه اكتشفت منها ان راي غلط الكتب تكرار فيها واجد محتواها تافه بالنسبه و لهالزمن
	Removing digits	boring i bought books because i the read and they review look interesting and funny after but the first pages knew i that a this mistake was very is repetitive and some the of content ridiculous really for this age and time	مله اشتريت قرات لاني كتب المراجعه وهي متعه تبدو ومضحكه لكن بعد الصفحات الاولى ان علمت كان هذا خطا مكر جدا وبعض المحتويات حقا سخيفه العصر لهذا والزمان	تمل اشتريت لاني كتب تعليقات قريت الناس و عنها كانت اتبين انها زينه و شيقه تضحك بس و عقب ما قرات الصفحات الاوليه اكتشفت منها ان راي غلط الكتب تكرار فيها واجد محتواها تافه بالنسبه و لهالزمن
	Tokenization	['boring', 'i', 'bought', 'books', 'because', 'i', 'the', 'read', 'and', 'they', 'review', 'look', 'interesting', 'and', 'funny', 'after', 'but', 'the', 'first', 'pages', 'knew', 'i', 'that', 'a', 'this', 'mistake', 'was', 'very', 'is', 'repetitive', 'and', 'some', 'the', 'of', 'content', 'ridiculous', 'really', 'for', 'this', 'age', 'and', 'time']	['مله', 'اشتريت', 'قرات', 'لاني', 'كتب', 'المراجعه', 'وهي', 'متعه', 'تبدو', 'ومضحكه', 'لكن', 'بعد', 'الصفحات', 'الاولى', 'ان', 'علمت', 'كان', 'هذا', 'خطا', 'مكر', 'جدا', 'وبعض', 'المحتويات', 'حقا', 'سخيفه', 'العصر', 'لهذا', 'والزمان']	['تمل', 'اشتريت', 'لاني', 'كتب', 'تعليقات', 'قريت', 'الناس', 'و', 'عنها', 'كانت', 'اتبين', 'انها', 'زينه', 'و', 'شيقه', 'تضحك', 'بس', 'و', 'عقب', 'ما', 'قرات', 'الصفحات', 'الاوليه', 'اكتشفت', 'منها', 'ان', 'راي', 'غلط', 'الكتب', 'تكرار', 'فيها', 'واجد', 'محتواها', 'تافه', 'بالنسبه', 'و', 'لهالزمن']
	Removing stopwords	['boring', 'bought', 'books', 'read', 'review', 'look', 'interesting', 'funny', 'pages', 'knew', 'mistake', 'repetitive', 'content', 'ridiculous', 'age', 'time']	['مله', 'اشتريت', 'كتب', 'لاني', 'قرات', 'المراجعه', 'متعه', 'تبدو', 'ومضحكه', 'الصفحات', 'الاولى', 'علمت', 'خطا', 'مكر', 'وبعض', 'المحتويات', 'حقا', 'جدا', 'لهذا', 'العصر', 'سخيفه', 'والزمان']	['تمل', 'اشتريت', 'لاني', 'كتب', 'تعليقات', 'قريت', 'الناس', 'اتبين', 'زينه', 'شيقه', 'تضحك', 'عقب', 'قرات', 'الصفحات', 'الاوليه', 'اكتشفت', 'راي', 'غلط', 'الكتب', 'تكرار', 'واجد', 'محتواها', 'تافه', 'بالنسبه', 'لهالزمن']

3.4 Word Embedding

A key concept of NLP is converting the words into numeric vectors that fed into the prediction model. Word2Vec is the technique used to do this kind of conversion. Word2Vec, or word embedding, is characterized by keeping the words' context information, consequently keeping the semantic (Thomas, 2019).

Word embedding is one of the most ways to represent text in a dense low dimensional vector learned from large quantities of plain text in an unsupervised manner (Luque, 2019). Word embedding could be obtained using two methods, Embedding Layer and pre-trained word embedding (Chollet, 2018). In this research, the word embedding layer was applied. The embedding layer is a kind of layer provided by Keras. It is the first layer of the network (Brownlee, 2017b). It is initialized with a random word vector followed by the learning process of embedding all other words in the text. This learning process is similar to learning network weights (Brownlee, 2017b) and (Chollet, 2018). Three variant uses of embedding layer were listed in (Brownlee, 2017b). These uses are 1- Learning word embedding, which can be used later with another model, 2- Loading a pre-trained word embedding model, and 3- Learning word embedding with a deep neural network, where the word embedding layer represents a part of the network. The third usage was applied in Phase 1 of our proposed model. The embedding layer requires its input data to be in integer encoded format that Keras Tokenizer API can achieve.

The working mechanism of the embedding layer is like a dictionary, where mapping of integer indices to dense vector takes place. When integers are fed to the Keras embedding layer, it searches for these integers in an internal dictionary and returns the corresponding vectors. The embedding layer takes a 2D tensor of integers as input (Number of samples, Length of each sample) and returns a 3D floating-point tensor (Number of samples, Length of each sample, embedding size) as output. The samples' Length in an input tensor should be the same, leading to truncating or padding the sample with zeros. The 3D tensor is then fed to the network to be processed (Chollet, 2018).

Defining the embedding layer imposes a specification of 3 arguments (*input_dim*, *output_dim*, and *input_length*). The *input_dim* is the vocabulary size, *output_dim* is the vector size, and the *input_length* is the input sequence length (Brownlee, 2017b). In this

research, the three arguments of the embedding layer are (*input_dim*, 50, 30). The *input_dim* represents the number of unique vocabularies. It *was* calculated automatically by the program, which gave 8913, 19143, and 22632 values for English, MSA, and BDs datasets, respectively. The variation in the obtained value of the *input_dim* from one dataset to another might be explained by the richness of polysemic per language.

In conclusion, the presented dataset will enrich the Arabic NLP community with a parallel dataset of product reviews in MSA and its corresponding ones in BDs. This will fill a gap in Arabic resources in general and Bahraini dialects in particular, which promotes research studies in such a field.

The accuracy check was implemented to our datasets by data validity via correcting grammatical and spelling errors of the review text in all datasets, in addition to missing values checking plus the cleaning process. The BDs dataset is reliable because it was obtained from qualified respondents.

The created dataset was deposited in a Mendeley repository with DOI: https://doi.org/10.17632/5rhw2srzjj.1 and dataset license CC-BY-NC. A detailed dataset descriptor has been published at https://www.mdpi.com/2306-5729/8/4/68 , while the entire experiments on the created dataset have been published at https://www.sciencedirect.com/science/article/abs/pii/S0169023X22000970. The latter publication has attracted NLP researchers, where the citations reach five till the resubmission day of this thesis.

Chapter 4 - LSTM Multilingual Sentiment Analysis

4.1 Introduction

Parts of the work included in this chapter were previously published in (Omran, T.M., Sharif, B.T., Grosan, C. and Li, Y., 2022. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, p.102106).

One of the objectives of this research is to develop a deep learning approach using LSTM for analyzing the sentiment of English, MSA, and BDs datasets. This chapter explains the stages of creating, configuring, compiling, and training our proposed deep learning LSTM model, as detailed in the following subsections.

4.2 Proposed LSTM Model

Our proposed model is composed of 2 stacked LSTM layers as shown in figure 4.1, where the entire sequence of the previous LSTM layer is fed to the second layer through setting the argument *return_sequences= true.* The LSTM layers are followed by a dense layer with a sigmoid function.

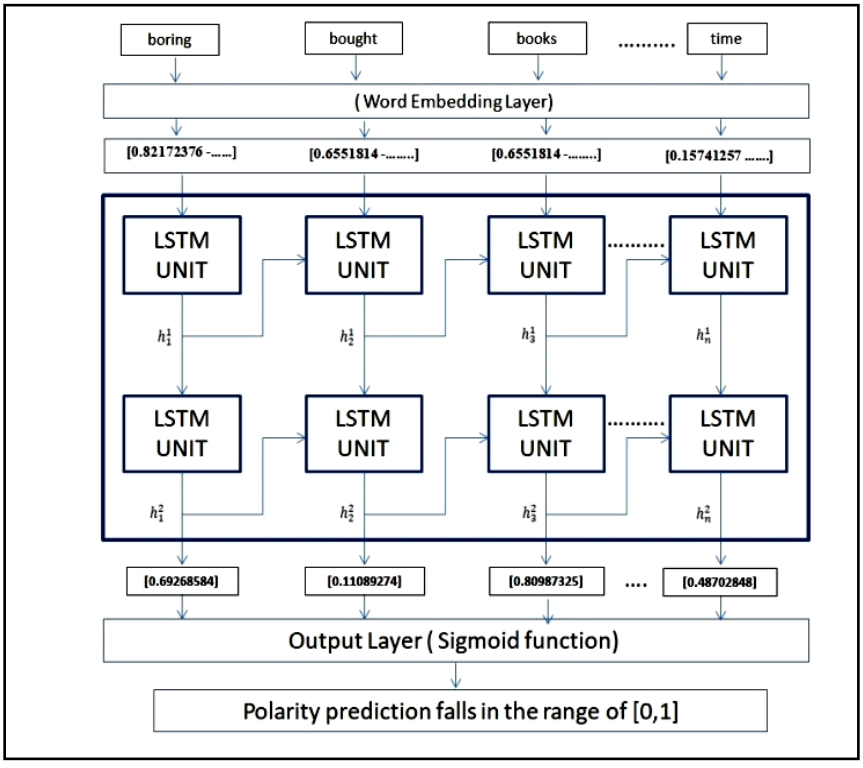


Figure 4.1 LSTM proposed model

4.2.1 LSTM Model Design, Implementation and Configuration

After applying the word embedding via the embedding layer, the word embedding vectors are fed into the nodes of the first LSTM layer, where the LSTM is trained on these embedded words and makes a prediction. The predicted words are connected to a dense layer where a sigmoid function is included (Han, Moraga 1995).

LSTM is a particular architecture of RNN (Goldberg, 2017), composed of multiplicative gates and internal memory (Smagulova, James, 2019) such as input gate, forget gate, and output gate, as seen in figure 4.2. The memory task is to remember data over time, while the multiplicative gates task is to control the entry and exit of information to the cell (Mohammed, Kora, 2019). This kind of architecture solved the problem of exploding and vanishing gradient (Mohammed, Kora, 2019), which in turn contributes to achieving a state of the art results of modelling sequential data (Goldberg, 2017).

Two factors control the capacity of the network, namely network depth and network width. The network depth is represented by the number of layers, while the layer's width is represented by the number of nodes per layer (Brownlee, 2019 a).

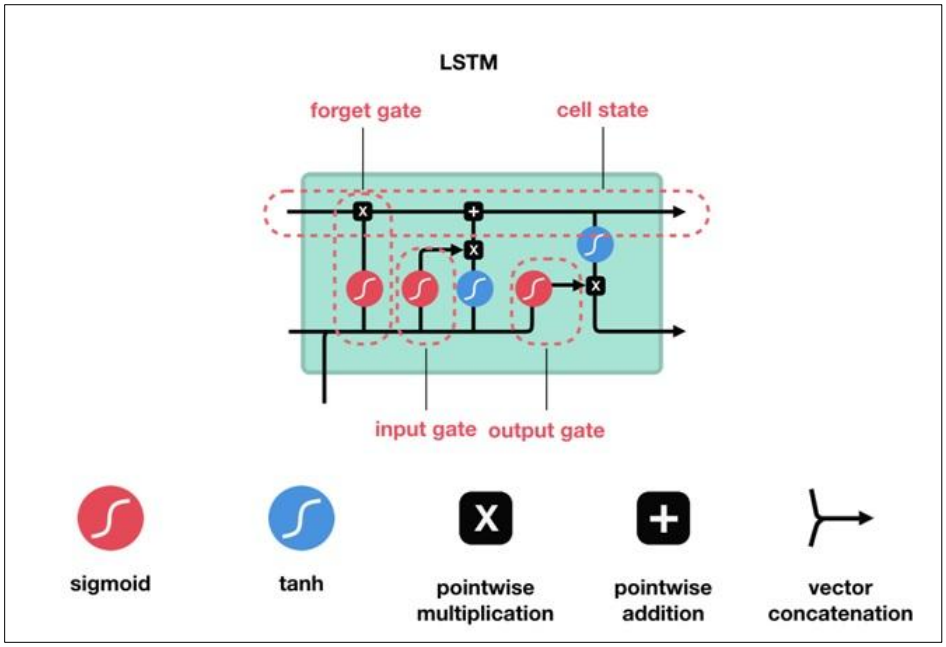


Figure 4.2 LSTM cell (Phi M, 2018)

Keras can add layers and nodes. Adding a layer to the model occurs through calling the *add ()* function, while the number of nodes is set using the first argument of the layer, as shown in figure 4.3.

```
model = Sequential()
e = Embedding(vocab3_size, embedding3_dim, input_length=max3_length, trainable=True)
model.add(e)

model.add(LSTM(10, dropout=0.1, recurrent_dropout=0.5, return_sequences=True,activation='relu'))
model.add(LSTM(10, dropout=0.1, recurrent_dropout=0.5,activation='relu'))
model.add(Dense(1, activation='sigmoid'))

opt=tf.keras.optimizers.Adamax(lr=0.01)
model.compile(optimizer=opt, loss='binary_crossentropy', metrics=['accuracy'])

model.fit(X_train,y_train, epochs=20, batch_size=80,validation_data=(X_val,y_val),verbose=0)


_, test_accuracy = model.evaluate(X3_test, y_test, verbose=0)
```

Figure 4.3 Code of creating this research LSTM model

Our model's number of hidden layers and nodes was determined by evaluating the model on a range of layers and nodes via a created function for optimization. The best accuracy returned by the function was achieved using two layers and ten nodes, where the minimum loss value has been achieved, as shown in figure 4.4 and figure 4.5, respectively.

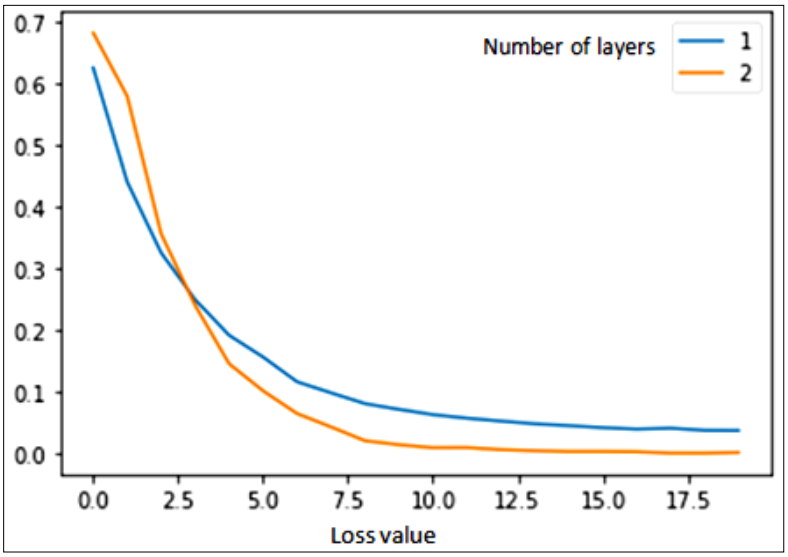


Figure 4.4 Number of layers and model loss history

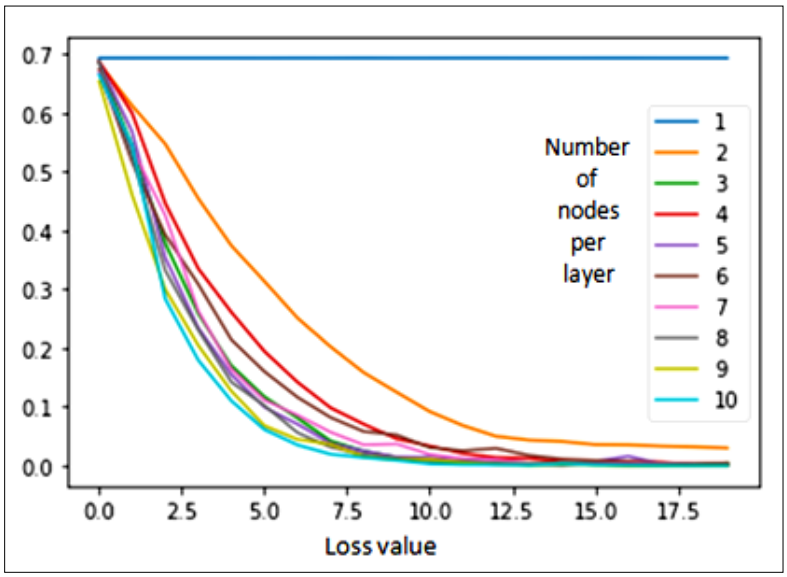


Figure 4.5 Number of nodes and model loss history

The hidden layers were followed by a dense output layer with sigmoid function. The sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ has an S shape depicted in figure 4.6 that transforms each value of x into the range of [0,1] (Goldberg, 2017).

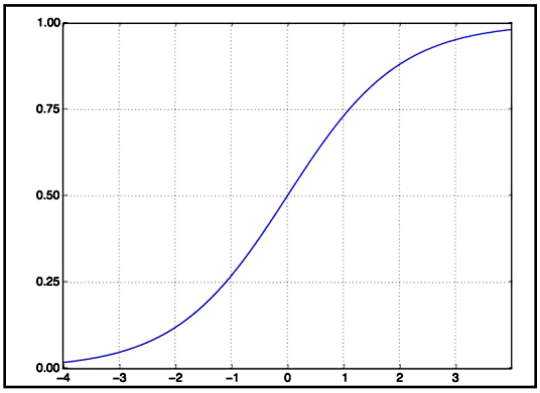


Figure 4.6 Sigmoid function (Chollet, 2018)

Other arguments of the LSTM layer were specified, such as *dropout*, *recurrent_dropout*, *return_sequences, and activation*. These arguments were tuned manually. The *dropout* is one of the common techniques used for regularizing neural networks and preventing them from overfitting (Goldberg, 2017). It drops out randomly some features of the layer output during the training process by setting those features to zeros. For example**,** if a

layer will return a vector of [0.5,1.1,1.3,0.2] for a specific sample of input, after applying the dropout, the vector becomes like [0,1.1,1.3,0] (Chollet, 2018). The dropout is the percentage of the zeroed features. In this research, the dropout value is 0.1. The *Recurrent_dropout* specifies the rate of the dropout of recurrent units. A value of 0.5 was used for the recurrent dropout in our model. The *return_sequences=True* is used to stack the recurrent layers by returning the full sequences for all prior layers output in 3D shape tensor instead of returning the output of the last step (Chollet, 2018),(Thomas, 2019), and (Brownlee, 2017c). The stacked LSTM, as its name suggests, is composed of a stack of LSTM layers on top of each other to increase the network capacity. In stacked LSTM, the i-th network is defined in equation 4.1 by (Staudemeyer, Morris, 2019):

$$X_i = \begin{bmatrix} \vec{y}_{h_i-1} \\ \vec{y}_{h_i} \end{bmatrix} \tag{4.1}$$

$\vec{y}_{h_i-1}$ which represents the hidden signal from the previous LSTM, replaces the input signal $\vec{x}$ from the previous LSTM. The *activation* is responsible for transforming the summed weighted input of the node into the output. One of the activation functions is rectified linear unit (ReLU). ReLU is a simple activation that performs well despite its simplicity. It maps 0 value for each x <0 as shown in equation 4.2 and figure 4.7. If the input to the node is positive, the ReLU function output it directly; otherwise, the output will be 0. (Brownlee, 2019b)

$$\text{ReLU}(x) = \max(0, x) = \begin{cases} 0 & x < 0 \\ x & \text{otherwise.} \end{cases} \tag{4.2}$$

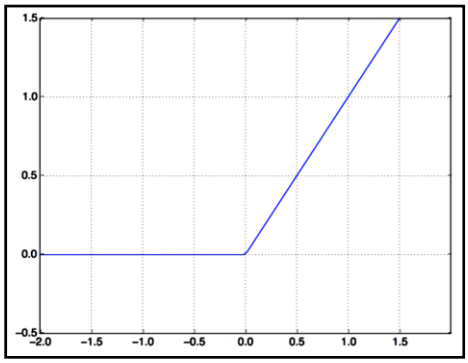


Figure 4.7 ReLU function (Chollet, 2018)

4.2.2 LSTM Model Compiling and Training

The configuration of the learning process of the LSTM model is done through the compiling step. The compiling step occurs using the *model.compile* command, where parameters like an optimizer, loss function, and evaluation metric are specified to be used by Keras to train the model. The adamax optimizer with learning rate (lr) =0.01, binary_crossentropy loss function, and accuracy, F1, and AUC metrics were applied in this research. The adamax and 0.01 learning rate were specified using the grid search optimizing technique. Grid search is provided in GridSearchCV class in sickit-learn. This class requires a dictionary of hyperparameters to be evaluated in the *param_grid* argument. This dictionary helps assign an array of values to try as model parameters (Brownlee, 2016). Grid search was defined by (Buduma, Locascio, 2017) as a technique for optimizing the hyperparameter. Each parameter value is picked from a list of options. The model is trained on all combinations or permutations of all choices. The best combinations that give the best accuracy on the validation set are chosen, and the accuracy of the test set is reported.

In this research, a list of optimizers and learning rates were provided to GridSearchCV class to specify the ones that achieve the best accuracy. The list of optimizers was [SGD, RMSprop, Adagrad, Adadlta, Adam, Adamax], while the list of learning rates was [0.001, 0.01, 0.2, 0.3].

The Adamax is a variant of adam optimizer. It is better than Adam, especially when using models with embedding (Team, 2021)**.** *Learning rate* is an updated amount of the network's weights during the training process of the neural network. It has a value ranging from 0.0 to 1.0, and it has a configurable parameter of the optimizer (Brownlee, 2019c). The size of the learning rate plays an essential role in network convergence. Minimal learning rate makes the network converges for a very long time. In contrast, a high learning rate will not allow the network to converge efficiently (Goldberg, 2017). *Loss* is the absolute value of the distance or difference between the predicted value and the actual value. It should be minimized during the training process using the loss function. *Binary cross entropy* is the most common function used to calculate the loss of

the problems that consider binary classification (Ruby, Yendapalli, 2020). The loss is calculated by (Setia, 2020) as:

$$\log loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \,.\log\big(p(y_i)\big) + \ (1-y_i).\log\,(1-p(y_i)) \tag{4.3}$$

y_i denotes the actual class 1, $\log\big(p(y_i)\big)$ denotes the probability of the actual class (y_i) , N is the number of instances, $p(y_i)$ is the probability of the positive class 1, and $(1-p(y_i)$ is the probability of the negative class 0.

After the compiling step, the learning process passes through an iteration process over the training data using a *fit ( )* method, where the *input_data*, *target_data*, *batch_size*, and *epochs* parameters are specified. The *input_data* and the *output_data were* represented by X_train and y_train, respectively, in our model. The X_train and y_train were determined using two methods, train-validate-test split and cross-validation methods. The *batch_size* and the number of *epochs* were specified using the Grid search technique. A list of batch sizes = [80, 81, 100]. The 81 value has been mainly listed to be more decisive in what value to choose, 80 or 81, because they give an equal accuracy value when implemented manually by try and error. The list of epochs = [20,50,100] to the GridSearchCV class. The *batch_size* and the number of *epochs* that achieved the best accuracy were obtained in 80 and 20, respectively. The batch size is the number of instances the model must work on before updating its other internal parameters. The batch varies in size, which accordingly changes the name of the learning algorithm. For example, if the batch size is one, the learning algorithm name will be stochastic gradient descent. When the batch size equals the size of the dataset, it is called batch gradient descent, while the learning algorithm is called mini-batch gradient descent when the batch size is greater than one and less than the size of the dataset. The mini-batch gradient descent was applied in this research. At the end of the batch size, the predicted values are compared with the actual values, and the prediction error is calculated (Brownlee, 2018a). The epoch represents the number of iterations over which the learning algorithm operates to be displayed on entire training samples. A single epoch consists of one or more batches (Brownlee, 2018).

The final configuration of our proposed LSTM model after tuning its parameters is shown in table 4.1. This LSTM model with the tuned parameters: the number of layers, nodes per layer, the optimizer, and the loss function, has been individually trained and tested on each dataset: English, MSA, and BDs.

Table 4.1 Final configuration of LSTM model

Number of Layers	nodes/ Layer	Dropout	Recurrent dropout	Activation function	Optimizer	Learning rate	Epochs	Batch size
2	10	0.1	0.5	ReLU	Adamax	0.01	20	80

It is worth mentioning that an LSTM model with un-tuned parameters was created to be used as a benchmark model to compare our proposed model performance with it, especially when keeping in mind that there is no benchmark model in the literature, specifically for sentiment analysis of MSA and BDs Amazon product reviews. The configuration of the un-tuned (benchmark) model is shown in Table 4.2.

Table 4.2 Final configuration of benchmark LSTM model

Number of Layers	Layer Number	nodes/Layer	Dropout	Recurrent dropout	Activation function	Optimizer	Learning rate	Epochs	Batch size
2	1	8	-	-	ReLU	Adamax	0.01	20	80
	2	4							

4.3 Experiments and Results

A literature review search revealed a lack of resources and studies that considered the SA of Bahraini dialects. One objective of this research was to develop a deep learning approach using a recurrent neural network, including LSTM, to analyze the SA of the Bahraini dialects and their corresponding ones in MSA and English. This section presents and discusses the results of Phase 1 experiments that were conducted to evaluate our proposed model on the parallel dataset of English, MSA, and BDs composed of 10,000 reviews obtained by augmenting the initial 5000 review datasets. Each review comprises a group of sentences with a maximum word total of 200. Three evaluation metrics were used to evaluate our proposed model performance. These metrics are Accuracy, F1 score, and AUC. The accuracy measures the ratio of correctly predicted instances to the total

prediction of all instances. The accuracy is calculated using equation 4.4 (Chicco, Jurman, 2020).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{4.4}$$

TP and TN represent the number of correctly predicted instances, while the FP and FN represent the number of incorrectly predicted instances. F1 score is the harmonic mean of recall and precision. It can be expressed as follows (Chicco, Jurman, 2020):

$$\text{F-Measure} = \frac{\text{2*Precision*Recall}}{\text{Precision+Recall}} \tag{4.5}$$

While the precision and recall are represented as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4.6}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.7}$$

The range of F1 score falls in [0, 1]. When the F1 score closes to 0, it means that TP=0; in other words all of the positive instances are misclassified, while the perfect classification is achieved when F1 reaches 1, that is, FP=FN=0 (Chicco, Jurman, 2020) (Fourure *et al*., 2021). AUC measures the model quality in binary classification (Team, 2022). It can be calculated using the *roc_auc_score()* function provided by the Keras library. This function takes two values: the true outcomes from the test set and the predicted probabilities of the positive class and returns the score of AUC between 0 and 1. It is also known as the area under the roc curve (AUROC). It can be expressed as in equation 4.8 (Fourure *et al*., 2021).

$$AUC = \int_{t=-\infty}^{\infty} \frac{tp(t)}{tp(t)+fn(t)} \frac{d}{dt}\left(\frac{fp}{fp+tn}\right)|_t \, dt. \tag{4.8}$$

ROC curve is Receiver Operating Characteristic Curve. It is an efficient tool for predicting the probability in binary classification. It is a plot of two axes, x and y. The x-axis represents the rate of false positive, while the true positive rate is represented on the y-axis of different threshold values that fall in the range of [0, 1] (Brownlee, 2018b).

All Phase 1 experiments were applied using: (1) train-validate-test split with a ratio of 75% for training, while the 25% was split into two parts, a quarter of it 6.26% as validation and the remaining part, which is 18.74% as testing; (2) different k-fold cross-validation, particularly k=3, k=5, and k=10; and (3) learning rate =0.01. (4) multiple runs of the proposed model on both types of data split train-validate-test and k-fold cross-validation. The train-validate-test split was used by calling the function *train_test_split ()* provided by scikit-learn. The purpose of the validation set is to reflect the model performance during the tuning process of the model parameters, which in turn improves the learning process. K-fold cross-validation is a technique where the training set is divided into *k* smaller partitions and a model is trained on *k-1* partitions and tested on the remaining parts. The performance metric is calculated by averaging the k accuracies that result from k-fold cross-validation (Wong, 2015). Figure 4.8 shows a visualization of k-fold cross-validation.

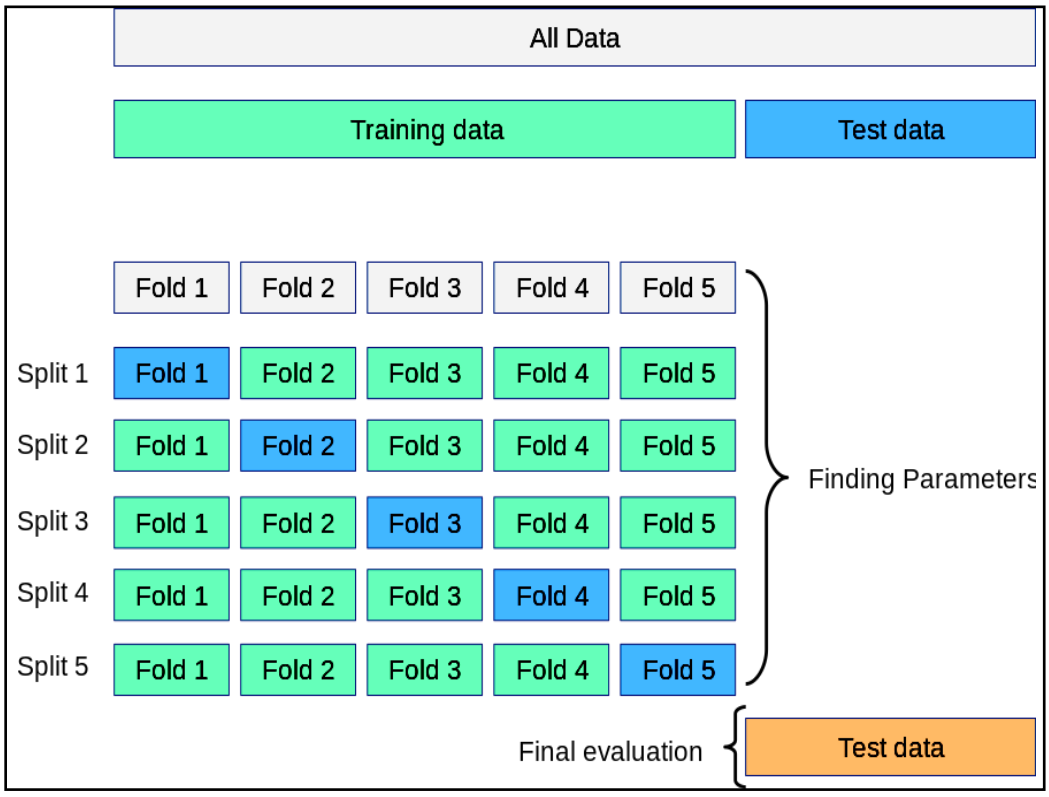


Figure 4.8 Cross-validation visualization (scikit-learn developers, 2020)

Thirty-six experiments were held out to evaluate the model performance. Nine used train-validate-test split and twenty-seven used K-fold cross-validation, all using a 0.01 learning rate. The nine experiments were represented by three experiments for evaluating the model performance using the accuracy, F1score, and AUC metrics on each dataset: English, MSA, and BDs. The twenty-seven used K-fold cross-validation experiments were represented by nine experiments on the English dataset, nine on MSA, and nine on the BDs. Each nine of these experiments were conducted to evaluate the model performance using accuracy, F1 score, and AUC metrics on 3, 5, and 10 folds, as shown in figure 4.10. Despite using the same dataset, whether English, MSA, or BDs, the deep learning model gives various results every time it is trained or run due to its flexibility. The multiple runs were applied to reduce the variation in model performance by calculating the average value of the evaluation metric, whether the accuracy, F1 score, or AUC. Figure 4.9 shows an example of the LSTM performance variation in test accuracy on the test set of 10,000 reviews in Bahraini dialects (BDs) using ten runs utilizing the train-validate-test split.

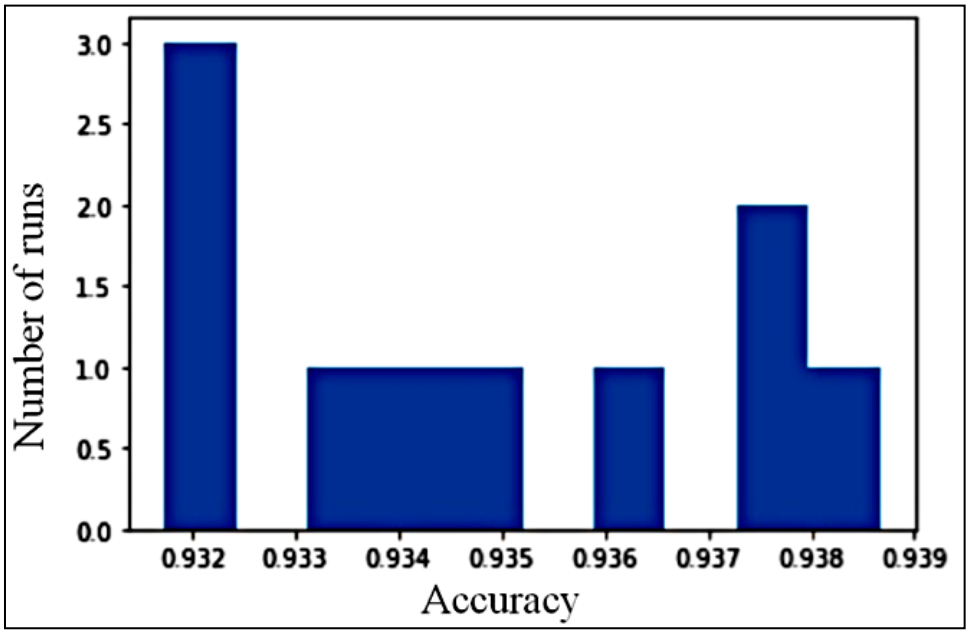
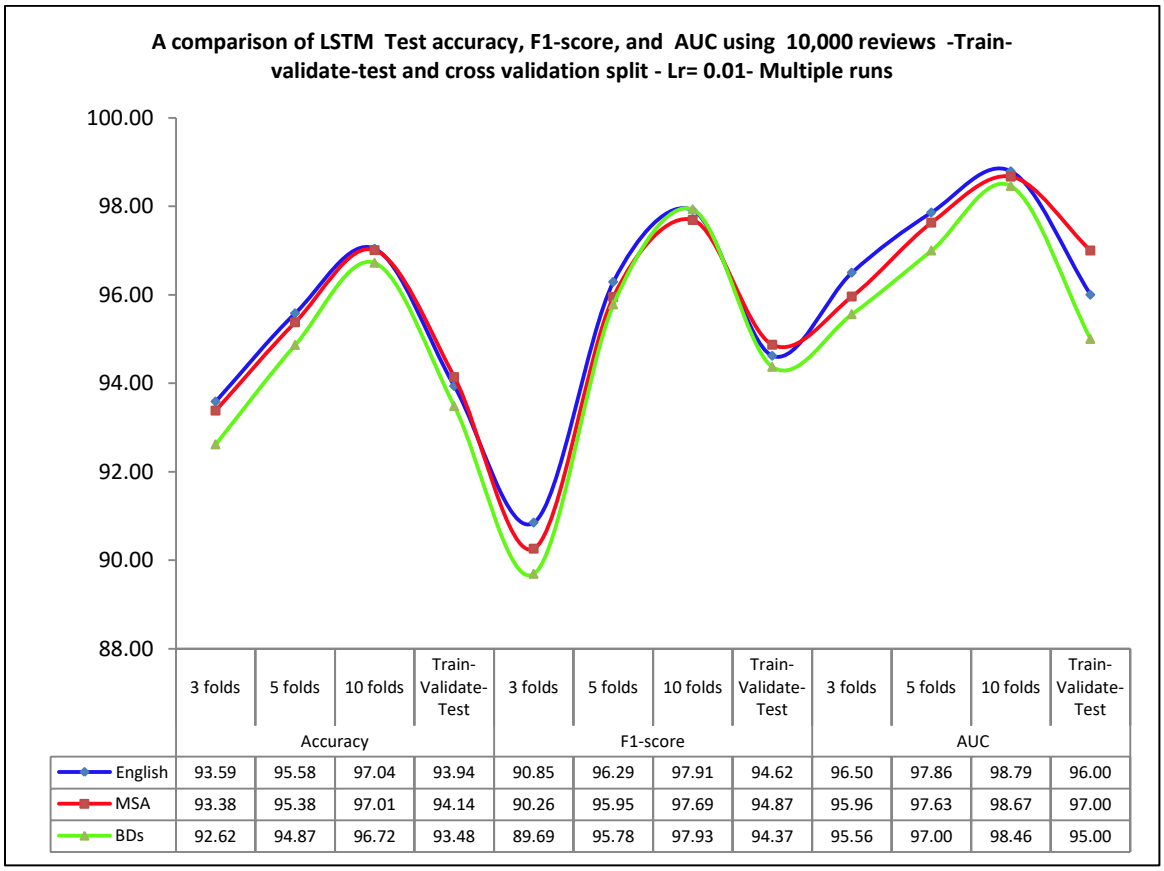


Figure 4.9 Proposed LSTM model test accuracy variation over 10 runs

Figure 4.10 shows a comparison of the average accuracy, F1 score, and AUC of the multiple runs of the LSTM model on the English, MSA, and BDs datasets.



	Accuracy				F1-score				AUC			
	3 folds	5 folds	10 folds	Train-Validate-Test	3 folds	5 folds	10 folds	Train-Validate-Test	3 folds	5 folds	10 folds	Train-Validate-Test
English	93.59	95.58	97.04	93.94	90.85	96.29	97.91	94.62	96.50	97.86	98.79	96.00
MSA	93.38	95.38	97.01	94.14	90.26	95.95	97.69	94.87	95.96	97.63	98.67	97.00
BDs	92.62	94.87	96.72	93.48	89.69	95.78	97.93	94.37	95.56	97.00	98.46	95.00

Figure 4.10 A comparison between multiple runs of LSTM using train-validate-test and cross-validation split, 10,000 reviews, and 0.01 learning rate

As can be seen from figure 4.10, the LSTM performance behaves in the same rhythm for the three evaluations metrics: Accuracy, F1 score, and AUC, as its performance rises gradually from the case of applying three folds to the case of 10 folds and then decreases in the case of the train-validate-test split on all datasets. In other words, the proposed model's best performance was achieved when applying ten folds of cross-validation. Hence our discussions of the results will focus on the case of 10 folds.

What is striking about the figures in figure 4.10 is the highest achievement performed by the proposed LSTM model when using the AUC metric, followed by the F1 score and finally the accuracy, where it was 98.79%, 97.91%, and 97.04%, respectively on English dataset, 98.67%, 97.69%, and 97.01%, respectively on MSA, while it was on BDs 98.46%, 97.93%, and 96.72%, respectively.

The most surprising aspect of the data, specifically in the F1 score figures in figure 4.10, is the outperformance of the LSTM model in classifying the BDs dataset over both English and MSA by 0.02% and 0.24, respectively, where the F1 score on English was 97.91%, and 97.69% on MSA. From this obtained value of 97.93% F1 score on BDs, it can be concluded that the misclassified instances, either the FP or FN, are low, which means that the model prediction is perfect. In other words, a higher value of the F1 score means a higher value of precision and recall. Also, the difference in the F1 score between BD and the English and MSA datasets, which was 0.02% and 0.24%, was very slight, and this can be explained by the fact that the terms and words that were used in constructing the Bahraini dialects dataset, were carefully and accurately chosen which reflects its validity and authenticity.

The most striking result to emerge from the data in figure 4.10 was the highest model performance in the AUC metric, which was 98.79% in English, 98.67% on MSA, and 98.46% in the BDs dataset. A possible explanation of this result may be the high ability of the model to distinguish between the positive and negative instances, which indirectly reflect the correctness and appropriateness of the words in the translated reviews per class, precisely the ones of BDs.

A closer look at the data table of figure 4.10 indicates that the highest accuracy, F1 score, and AUC achieved by the LSTM when using the train-validate-test split was on the MSA dataset.

A closer inspection of data in figure 4.10 shows the outperformance of the LSTM when using the accuracy metric, where LSTM achievement on the English dataset outperforms it on MSA by 0.03%, where the accuracy on the English dataset was 97.04% and 97.01% on MSA, while the amount of the outperformance on MSA over BDs was 0.29%, where the accuracy was 96.72 on BDs and 97.01% on MSA.

The accuracy difference of 0.29% between the MSA and BDs is greater than it is between the MSA dataset and English, which was 0.03%, despite the expectation that the difference between the MSA and BDs should be the least. The possible justification for these results can be as follows:

1. Automatic translation from English to MSA may be limited and governed by a range of certain words, even if they are many, while this does not govern the conversion from the MSA to BDs. The words in dialects are multiple and change, albeit at the end has one meaning. For example, the word "جميل" in MSA which is pronounced as "Jameel" in English was converted to"حلو","حليو","فلة". The word "آسف" which is pronounced as "Aasef" in English was converted to "متأسف", "متحسف". The MSA word "كثيرا" which is pronounced as "Katherann" in English, has also been converted in BDs to "وايد" and"واجد".

2. Some of the respondents convert some of the sentiment words of MSA reviews to their English synonyms using Arabic characters, for example, some reviews in MSA contains sentiments words like الأفضل,لطيف, مرن, جيد. Their conversion to BDs by the respondent was: التوب, كيوت, فليكسيبل, اوكي. This might make the sentiment words or the number of vocabulary in the BDs dataset less than that in MSA.
3. The highest percentage of the participants was from a village called Al Nuwaidrat. This percentage may use the exact words that express the positive and negative emotions, affecting the LSTM learning model extracted features. Appendix 2 shows the number of respondents from each covered city and village of Bahrain. In addition, the shared vocabulary between the villages and cities of Bahrain despite their difference in dialects of them. An example of this vocabulary is "عدل","زين","مريح", and "فله".This similarity in vocabulary might influence the number of vocabulary embedded by the LSTM embedding layer.

A comparison of this research's findings on the English dataset with those of other studies is shown in Table 4.3, where Amazon product reviews were one of the datasets that were utilized by (Can, Ezen-Can, and Can, 2018) to train an LSTM model on 9,478,095 instances with no more details about the percentage of the training and testing set. At the same time, Table 4.4 shows our model's results on MSA and BDs, which were compared to the un-tuned parameters model created for comparison because of the rarity of benchmark studies on Amazon product reviews in Modern Standard Arabic and Bahraini Dialects.

Table 4.3 A comparison between our proposed model accuracy and the ones of the literature

Article	Dataset	Accuracy	F1 score	AUC
(Can, Ezen-Can, and Can, 2018)	English	85.61%	-	-
Our Proposed Model		97.04%	97.91%	98.79%

Table 4.4 A comparison between our proposed model accuracy,F1 score, and AUC with the created benchmark model.

Model	Dataset	Accuracy	F1 score	AUC
Un-tuned Parameters Model(Benchmark)	MSA	83.49%	83.79%	85.54%
	BDs	88.89%	88.95%	92.53%
Our Proposed Model	MSA	97.01%	97.69%	98.67%
	BDs	96.72%	97.93%	98.46%

In summary, the difference in the obtained values of our proposed LSTM model on the English, MSA, and BDs dataset was very slight. This slight difference indicates the following:

1. Meaning similarity in the parallel reviews of English, MSA, and BDs, which indicate the maintaining of sentiment features despite using the translation approach.
2. Validity of a dataset that resulted from the translation approach despite its morphological richness.
3. The proposed model successfully performed the SA process on a rare source dialect by taking advantage of the rich source language and building a language-independent model.

The best-obtained results: 96.72% accuracy, 97.93% F1 score, and 98.46% AUC on BDs, revealed something about the nature of the BDs dataset, such as good preprocessing and good validation of the translation process. Accordingly, it can be inferred that the BDs dataset is qualified and reliable for future use by NLP researchers. In addition, these findings have important implications for developing more studies in Arabic NLP that consider the SA of sequential textual data, specifically in Bahraini dialects, not only for the researchers but for all stakeholders and analysts who focus on the Arab world that plays an essential role in the international policies and global economy.

Another implication is the possibility of using the BDs dataset as a benchmark for future NLP studies in Bahraini dialects, especially when considering the paucity of such dialects resources and studies.

Chapter 5 - Ensemble Learning

5.1 Introduction

Part of the work included in this chapter was accepted to be published in IEEE Xplore as a conference paper entitled “Ensemble learning for sentiment analysis of translation-based textual data”. This chapter describes the detailed steps of achieving another objective of this research that is to enhance the performance of our proposed LSTM model by incorporating it in the ensemble learning technique. The following subsections describe the detailed steps of achieving this objective.

5.2 Proposed LSTM Model and Ensemble Learning

Ensemble learning is a technique used in machine learning that combines several base learners of a specific learning task to generate a model with more generalized prediction than the individual learners (Ortiz *et al*., 2020). One of the ensemble techniques is stacking. Stacking is an ensemble learning method represented by independent models known as base learners plus meta learners. The meta learner can be any learning algorithm trained to make predictions based on the base learners' predictions (Wolpert, 1992), as shown in figure 5.1. In this research, our proposed LSTM model denoted as LSTM-1 in table 5.1, and two other LSTM models were employed as base learners. At the same time, the meta-learner is a decision tree (DT) whose parameters were kept at their default values. DT is a machine learning algorithm where the instances or data points are classified based on the values of the features when a set of cases is given (Kumar and Verma, 2012). It is composed of inner nodes and leaf nodes. The inner nodes represent the decision threshold, and the leaves represent the prediction (Liang *et al*., 2019). The performance of DT as a meta-learner was compared with it as a single learner to check the robustness of DT meta-learner.

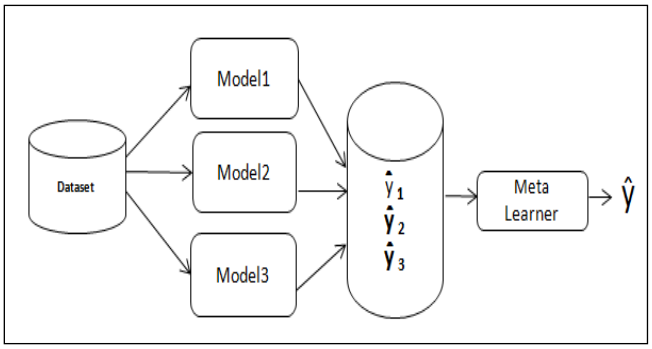


Figure 5.1 Stacking ensemble

Table 5.1 Architectures of LSTM models of stacking ensemble

Model	Number of Layers	Layer Number	nodes/ Layer	Dropout	Recurrent dropout	Activation function	Optimizer	Learning rate	Epochs	Batch size
LSTM-1	2	1	10	0.1	0.5	ReLU	Adamax	0.01	20	80
		2	10							
LSTM-2	2	1	5							
		2	2							
LSTM-3	2	1	8							
		2	4							

Each instance of LSTM models was trained on the training data and saved with a unique name. In the training stage of the meta-learner, all saved models were loaded and evaluated on testing data. The prediction results were used as input to the DT meta-learner. It is worth mentioning that each base learner will output two predictions for each instance of the test dataset. Since we have ten thousand instances of the dataset, two predictions, and three sub models, the sub models resulting three arrays that have the shape [10000,2]. These arrays were combined into a three-dimensional array having the shape of [10000, 3, 2]. The combination was applied using the dstack () function provided by the NumPy library. The three-dimensional array was reshaped to [10000, 6] because the three models give two predictions (3*2) per instance. The resulting array of a shape [10000, 6] was used in the meta-learner training process, DT. Knowing that transforming the shape of the array was applied using the reshape () function provided by the NumPy library.

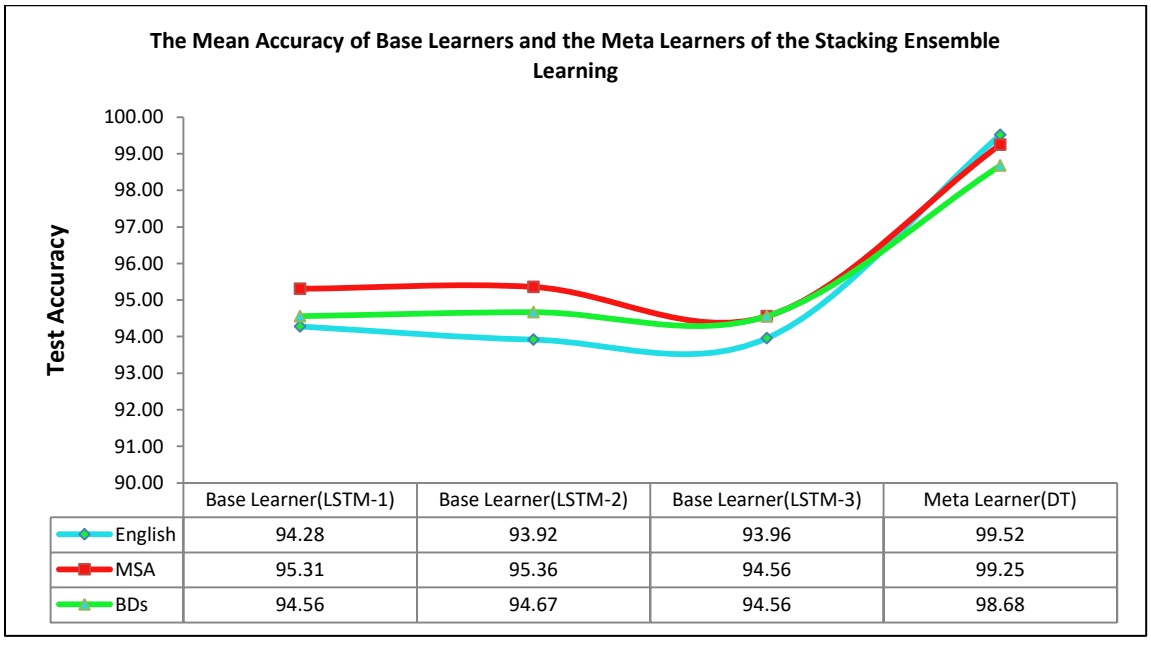
All members of the ensemble technique were evaluated using the mean accuracy of k-fold cross-validation with k-fold =10.

5.3 Experiments and Results

Many studies have utilized ensemble learning to improve the prediction task on various and multi datasets, but very few of them considered the ensemble technique on a translated datasets. Two experiments were conducted to evaluate the performance of the ensemble learning technique. The first was represented by forming the members of the stacking ensemble that consists of three LSTM models with different architectures, one of which is our proposed standalone model and a DT as a meta- learner. In contrast, the second experiment presents the DT as a single learner without incorporating it into an

ensemble learning process. This single DT learner was trained and tested to classify the reviews in the three datasets: English, MSA, and BDs, where the features representation was implemented using TF-IDF vectorizer with n-gram_range = (1.2). TF-IDF vectorizer converts the raw text to a matrix of TF-IDF features. TF-IDF is a term used to represent the weight of the word in the document (Joachims, 1996 ; n-gram_range determines the lower and upper value of the range of n-gram that should be extracted. In both experiments, k-fold cross-validation at k=10 was applied.

Figure 5.2 presents the experimental results of the stacking ensemble learning on the three datasets of English, MSA, and BDs. It can be seen from figure 5.2 that the meta-learner achieved the best results over the constituent ones for all datasets. This finding broadly supports the work of other studies in this area. An increase in the mean accuracy score was detected in the performance of the meta-learner when comparing it in BDs to MSA and English datasets. For example, the mean accuracy obtained by the meta-learner on BDs was 98.68%, which increased and rose to 99.25% on MSA, and 99.52% on the English dataset.



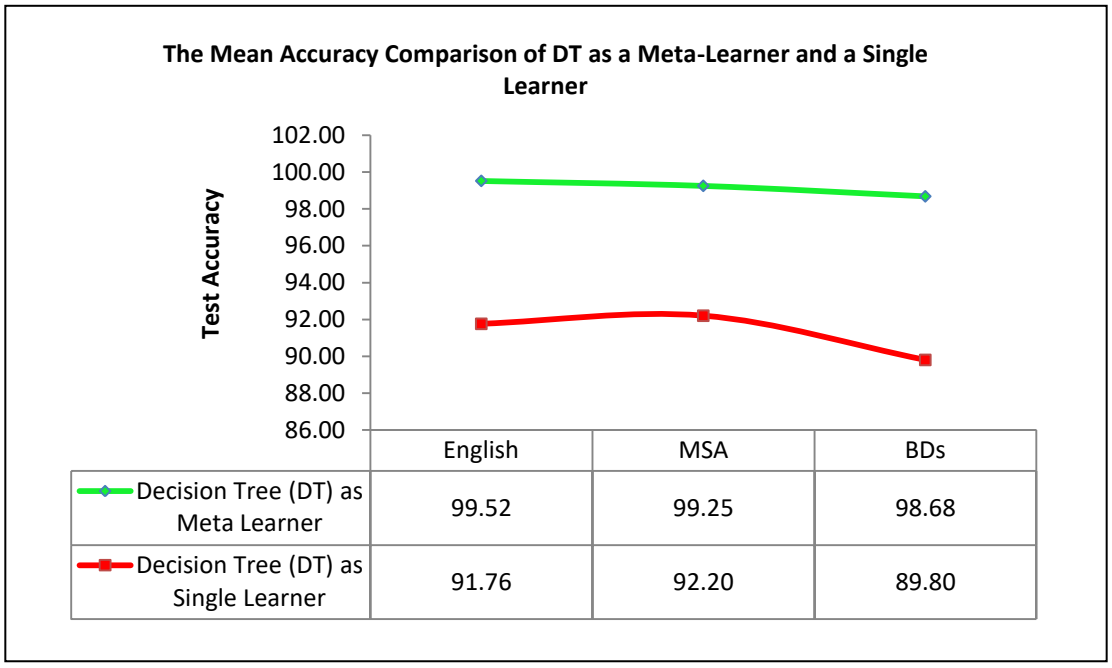
	Base Learner(LSTM-1)	Base Learner(LSTM-2)	Base Learner(LSTM-3)	Meta Learner(DT)
English	94.28	93.92	93.96	99.52
MSA	95.31	95.36	94.56	99.25
BDs	94.56	94.67	94.56	98.68

Figure 5.2 The mean accuracy of base-Learners and the meta-learners of the stacking ensemble learning

More investigation on the obtained results revealed that the difference between the source language dataset, English, and the final target language, BDs, was 0.84%, which is

considered a slight difference. What emerges from the results reported here is that the translation approach did not affect the sentiment analysis process of textual data, despite the different linguistic features of each language.

By looking at figure 5.3, it is clear that DT as a meta-learner outperforms its achievement as a single. A possible explanation for this might be due to the base learner, which is LSTM that can learn complex sequential patterns of the input dataset and also model the input text characterized by long-term dependencies.



	English	MSA	BDs
Decision Tree (DT) as Meta Learner	99.52	99.25	98.68
Decision Tree (DT) as Single Learner	91.76	92.20	89.80

Figure 5.3 The mean accuracy comparison of DT as a meta-Learner and a single learner

The slight difference in the obtained results of DT meta-learner of 0.27% between the English and MSA datasets and 0.57% between the MSA and BDs may indicate that translation of textual data; either machine or manual had maintained the sentiment features specifically with BDs dataset. Also, these results draw out attention to the importance of considering ensemble learning as an efficient tool for enhancing the sentiment analysis process. However, more research on this topic needs to be undertaken, for example, the effect of the number of base learners.

To show the enhancement in the classification process that was achieved by the stacking ensemble technique utilized in this research, table 5.2 shows the improvement magnitude

in the obtained accuracy by the standalone LSTM model and the one obtained by the ensemble method.

Table 5.2 Improvement value in classification of standalone LSTM model and ensemble meta-learner

Dataset	Our LSTM model	Ensemble Meta-Learner (DT)	Improvement Value
English	97.04	99.52	2.48
MSA	97.01	99.25	2.24
BDs	96.72	98.68	1.96

It can be noticed that the best classification improvement occurred in the English dataset with a value of 2.48%, followed by MSA, where the improvement value was 2.24%, while it was 1.96% in BDs. This can be justified by the morphology richness of the Arabic language (Guellil, Azouaou & Mendoza, 2019). Added to that, it is polysemic. For example, the word "أكبر", which is pronounced as “Akbar” in Arabic that means "larger" in English, may be used as a person's noun or as a comparative adjective.

Chapter 6 – Transfer Learning

6.1 Introduction

Parts of the work included in this chapter were previously published in (Omran, T.M., Sharif, B.T., Grosan, C. and Li, Y., 2022. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, p.102106).

This chapter presents the methods and experiments of achieving one more objective of this research that is analyzing the sentiment of small dataset by exploiting the knowledge gained from phase 1.

6.2 The Pre -Trained Model and Transfer Learning

This section presents the transfer learning process by exploiting the information acquired by the learner of BDs product reviews in Phase 1 in classifying another Bahraini dialects dataset that covers movie comments. Usually, in machine learning, the testing data and training data belong to the same domain, but in some other cases, the training data is very low or difficult to collect, which necessitates the availability of a good learner on a source that can be used in a target with low data (Weiss, Khoshgoftaar & Wang, 2016)

A model with good performance on the source task is one of the principal factors for transfer learning (Sarkar, Bali & Ghosh, 2018). In this research, the transfer learning process was for feature representation, which was conducted by creating a pre-trained LSTM model on BDs products reviews, where training is accelerated (Brownlee, 2020) and the divergence of the domains is minimized (Sarkar, Bali & Ghosh, 2018). The feature representation transfer occurs via saving the layers of our proposed trained LSTM model of Phase 1 and loading it as a pre-trained model in Phase 2 when classifying another small dataset covering different domains like movies comments.

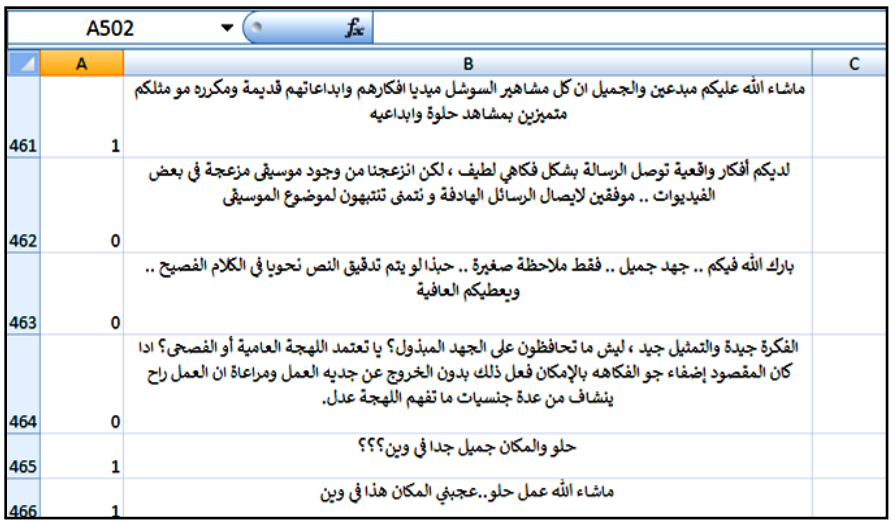
6.3 Dataset Design and Preparation

The following are the detailed steps of creating the target movie comments dataset that will be classified using the pre-trained model.

1. A number of movies comments were collected from different internet sites like “https://www.youtube.com/watch?v=y03-cdJPBH8” (Accessed on February 5, 2020) that includes comments on Arabic movies. About seven URLs contain movie

comments, where the comments were extracted from. Each URL includes comments about one movie. Other comments were collected from some accounts on Instagram like https://www.instagram.com/p/CNsBk0GDcoI/ Accessed on February 8, 2020). These comments were in Bahraini dialects. The collection process occurred by utilizing a tool called "YouTube comment extractor" provided by seobots.io, where the URL of the site that contains the movie comments was paste , and the number of the comments that was specified by the user were scraped. The collected comments were exported and downloaded to a csv file. The csv file was modified to contain two columns, one for comments' labels while the second was for the text of the comments.

2. The collected comments which were saved to the csv file passed through a manual process of cleaning and validation, where some comments which contain no sentiment were excluded and the misspelled words were corrected. This process of validating and cleaning resulting in 500 comments. It is worth noting that these comments were about short movies shed lights on some social, ethical, and educational issues in Bahrain society, such as divorce, visiting sick people, the impact of social media, and distance learning.
3. The 500 comments which were saved to the csv file were uncategorized, so their categorization or annotation process was assigned to two Bahraini dialects native speakers who are highly qualified persons in computer science and electrical engineering.
4. The annotators were asked to label the comments with 1 or 0 according to the included sentiment in the text, label 1 for positive sentiment comments and 0 for negative ones. The final label for the comments that the annotators disagreed on was decided by the author of this thesis. Figure 6.1 shows part of the collected and annotated movies' comments.



	A	B	C
461	1	ماشاء الله عليكم مبدعين والجميل ان كل مشاهير السوشل ميديا افكارهم وابداعاتهم قديمة ومكرره مو مثلكم متميزين بمشاهد حلوة وابداعيه	
462	0	لديكم أفكار واقعية توصل الرسالة بشكل فكاهي لطيف ، لكن انزعجنا من وجود موسيقى مزعجة في بعض الفيديوات .. موفقين لايصال الرسائل الهادفة و نتمنى تنتبهون لموضوع الموسيقى	
463	0	بارك الله فيكم .. جهد جميل .. فقط ملاحظة صغيرة .. حبذا لو يتم تدقيق النص نحويا في الكلام الفصيح .. وبعطيكم العافية	
464	0	الفكرة جيدة والتمثيل جيد ، ليش ما تحافظون على الجهد المبذول؟ يا تعتمد اللهجة العامية أو الفصحى؟ اذا كان المقصود إضفاء جو الفكاهه بالإمكان فعل ذلك بدون الخروج عن جديه العمل ومراعاة ان العمل راح ينشاف من عدة جنسيات ما تفهم اللهجة عدل.	
465	1	حلو والمكان جميل جدا في وين؟؟؟	
466	1	ماشاء الله عمل حلو..عجبني المكان هذا في وين	

Figure 6.1 A snippet of movies' comments

5. The annotation process classified the 500 comments into 140 negative and 360 positive comments.
6. After completing the annotation process, a Cohen's kappa was applied to get a reliable annotation. The Cohen's Kappa is a statistic used for measuring the value of agreement between inter-raters on a nominal scale. It is a function of the proportion of expected agreement and observed one (Warrens, 2015).
7. The obtained kappa value was 0.71, which is considered good.

6.4 Dataset Preprocessing

The movie comment dataset preprocessing steps were the exact ones applied to product reviews in BDs. The preprocessing steps involved applying the double swap augmentation technique, removing non-alphanumeric characters, text tokenization, and removing the stop words.

6.5 Pre-trained LSTM model Creation

The pre-trained LSTM model was created by saving our proposed LSTM model using the *save ()* function of the h5py library. This pre-trained model was then loaded using the *load_model* function of Keras *(Brownlee, 2019d)* to classify the movie comment dataset, which represents the target domain of the transfer learning process. The pre-trained LSTM model was trained on 20 epochs where the model achieved an appropriate fitting; a batch size equals to 80, and compiled using an Adamax optimizer.

6.6 Experiments and Results

One of the particular issues in reviewing the literature is the absence of transfer learning studies in Bahraini dialects. The purpose of Phase 2 experiments is to evaluate the transfer learning process by evaluating the pre-trained LSTM model by investing the extracted features from Phase 1 Bahraini dialects product reviews to classify the polarities of 1000 movie comments obtained by augmenting the 500 ones. The pre-trained model was created by saving our proposed LSTM model.

The pre-trained LSTM model was uploaded and trained on the movie's comments. Some experiments were conducted to evaluate the pre-trained model using the accuracy, F1 score, and AUC metrics. A part of them was conducted using a train-validate-test split and multiple runs, where 75% of the dataset was for training, while the 25% was split into two parts, a quarter of it 6.26% as validation and the remaining part, which is 18.74% as testing. The other parts were implemented using 3, 5, and 10 folds cross-validation with multiple runs.

Figure 6.2 compares the LSTM testing accuracy, F1 score, and AUC on the 1,000 movie comments dataset.

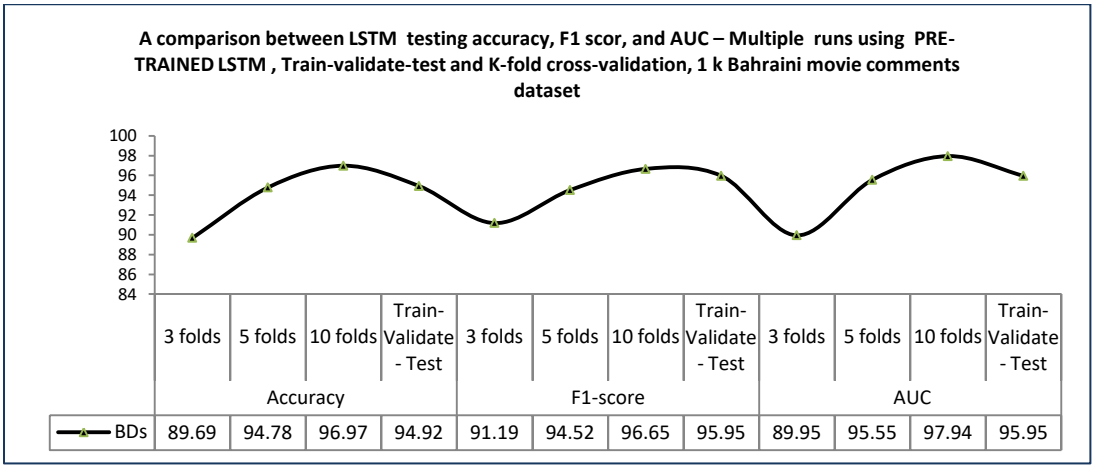


Figure 6.2 A comparison between pre-trained LSTM testing accuracy – Multiple runs using train-validate-test and k-fold cross-validation, 1000 Bahraini movie comments dataset

Looking at figure 6.2, it is clear that LSTM pre-trained model performance improved when moving from 3 folds to 10 folds and decreased when using the train-validate-test split.

One interesting finding is the F1 score, which was 96.65%, which can be viewed as a good result, taking into account that the dataset is imbalanced, composed of 720 positive and 280 negative comments. In addition, it can be inferred that FP and FN predicted instances were very few.

Comparing our pre-trained LSTM model performance on BDs, shown in figure 6.2, with our proposed LSTM model performance on BDs, shown in figure 4.10, a similarity in the two models' performance can be detected. The fact of the intersection may explain this similarity in extracted features between the source domain, products, and the target dataset's movie comments.

Although the dataset size was 1000 comments, considered minor, the obtained results were encouraging, indicating that the transfer learning process using the pre-trained LSTM model was successful; especially the LSTM learning method is greedy for data.

What is striking about the figures in figure 6.2 is that the best performance of the pre-trained model was achieved when using ten folds in all metrics: accuracy, F1 score, and AUC, where it was 96.97%, 96.65%, and 97.94%, respectively. These results can be explained by the fact that the pre-trained model was unbiased towards the training data of the source domain, in addition to its generalization ability. An implication of these results is the possibility of using our pre-trained model in classifying another dataset that covers another domain characterized by limited training data because of the expenses or difficulty of collecting or labelling data, where the input feature space is similar to our source dataset. In other words, the target dataset should have a high-level common domain with our source domain called marginal distribution (Weiss, Khoshgoftaar & Wang, 2016); for example, food review and book review are subdomains belonging to a review domain. The pre-trained LSTM model can be more successful when some conditions are considered, such as the similarity of conditional properties distribution between the source and target domains (Weiss, Khoshgoftaar & Wang, 2016); for example, the data label should be the same. If the labels are positive and negative in the source dataset, they should be so in the target dataset. Besides, the dataset balance in labelling should be taken into consideration (Weiss, Khoshgoftaar & Wang, 2016).

Chapter 7- Conclusion and Future Work

7.1 Introduction

Part of the work included in this chapter was previously published in (Omran, T.M., Sharif, B.T., Grosan, C. and Li, Y., 2022. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering*, p.102106).

This research set out to address two of the Arabic NLP challenges. One is the scarcity of standard and colloquial Arabic datasets, specifically the Bahraini one, by creating a BDs dataset utilizing the translation approach, where English Amazon product reviews were translated by machine to MSA, which translated manually to BDs. The other is the lack of SA studies of Bahraini dialects, tackled by sentiment analysis of the created datasets using a deep learning approach incorporated with the stacking ensemble learning technique.

The following subsections shed light on each challenge, how it was tackled, and its implication.

7.1.1 Dataset Design

The first challenge of creating the dataset includes sub-challenges such as specifying the appropriate domain of the dataset that will be created and the mechanism of collecting data, creating the most effective tool for data collection, besides finding collaborative participants who appreciate the importance of the research and respond within the specified period—added to that, collecting the data and its validation.

Creating a dataset is difficult, especially with resource rarity, the difficulty of collecting contents, and the labelling process. It is expensive and requires excellent collaboration. So providing a dataset will save many expenses, efforts, and time, representing necessary and facilitating factors in conducting any research. Accordingly, this research significantly contributes to NLP by providing two BDs' datasets. The first one is composed of 5000 reviews covering the products' domain. This dataset was generated by applying a translation approach that takes advantage of the availability of rich resource languages such as English, where Amazon product reviews were translated through machine to modern standard Arabic (MSA).

In contrast, manual translation was applied to translate the resulting standard Arabic reviews to Bahraini dialects (BDs). The manual translation necessitates the creation of customized forms with a specific structure to ease the conversion and validation process of MSA reviews to BDs. The other dataset is movie comments that require a reliable labelling process.

The best-obtained results of 96.72% accuracy, 97.93% F1 score, and 98.46% AUC on BDs revealed something about the nature of the BDs dataset, such as good preprocessing and good validation of the translation process. Accordingly, it can be inferred that the BDs dataset is qualified and reliable for future use by NLP researchers.

7.1.2 Multilingual Dataset Sentiment Analysis LSTM model

The other challenge of this research is the lack of SA studies of Bahraini dialects, which was tackled by sentiment analysis of the created datasets using the LSTM deep learning algorithm incorporated with the stacking ensemble learning technique. The findings indicate that the LSTM model has achieved the best performance when applying k-fold cross-validation at k=10 on the English dataset, followed by the MSA and BDs. In contrast, the highest accuracy has been obtained on MSA, followed by English and BDs datasets when using a train-validate-test split.

The obtained results in this research such as 97.04%, 97.91%, and 98.79%in accuracy, F1 score, and AUC, respectively, on the English dataset, 97.01% in accuracy, 97.69% in F1 score, and 98.67% in AUC on MSA, 96.72%, 97.93%, and 98.46% in accuracy, F1 score, and AUC, respectively on BDs, encourage the use of our proposed model as a reference or benchmark for the work of future researches. Also, the highest model performance in the AUC metric, which was 98.79% in English, 98.67% on MSA, and 98.46% in the BDs dataset, explains the high ability of the model to distinguish between the positive and negative instances, which indirectly reflect the correctness and appropriateness of the words in the translated reviews per class, precisely the ones of BDs.

The slight difference in the obtained results by the proposed LSTM model in classifying the multilingual datasets indicates its success in performing SA in a poor source dialect by taking advantage of the translation approach from the rich source language and

building an independent language model. Additionally, the difference in the F1 score between BDs and the English and MSA datasets, which was 0.02% and 0.24%, was very slight. This can be explained by the fact that the terms and words used in constructing the BDs dataset were carefully and accurately chosen, reflecting the validity and authenticity of the created dataset of Bahraini dialects.

This research also showed that the stacking ensemble technique had significantly enhanced the model achievement across the three datasets, English, MSA, and BDs.

7.1.3 Transfer Learning Pre-Trained LSTM model

One more objective of this research was employing a transfer learning technique to exploit the gained knowledge of analyzing the product reviews in Bahraini dialects to analyze another dataset of the same dialect that covers different domains and is characterized by small size by creating a pre-trained LSTM model.

The standalone and the pre-trained LSTM models of analyzing the Bahrain dialects datasets were evaluated using two data splits, train-validate-test and k-folds. The best results were obtained using ten folds and running the model 10 times.

Comparing our pre-trained LSTM model performance on BDs with our proposed LSTM model performance shows a similarity in the two models' performance. This similarity may explain the intersection of extracted features between the source domain, which is products and the target dataset's movie comments.

Although the dataset size was 1,000 comments, considered minor, the obtained results were encouraging, indicating that the transfer learning process using the pre-trained LSTM model was successful, especially since the LSTM learning method is greedy for data.

The best performance of the pre-trained model that was achieved when using ten folds in all metrics: accuracy, F1 score, and AUC, where it was 96.97%, 96.65%, and 97.94%, respectively, indicates that the pre-trained model was unbiased towards the training data of the source domain, in addition to its generalization ability. This solves the problem of obtaining a training data feature space that matches the test data prediction distribution by

providing a good learner. An implication is the possibility of using this pre-trained model in classifying another dataset that covers another domain.

7.2 Discussion

The results mentioned above, either of the multilingual dataset sentiment analysis LSTM model or the pre-trained one, using different metrics such as accuracy, F1 score, and AUC, reflect success in the proposed models' performance with limited availability of Bahraini dialects datasets despite the data greedy characteristic of deep learning models. The greedy characteristic of deep learning models for data necessitates the availability of abundant data, which was a challenge that imposed the creation of artificial data using augmentation techniques, specifically the swap one. The scarcity of Bahraini dialects' resources prevented the usage of other data augmentation methods, such as random insertion and synonym replacement. The scarcity of the SA analysis model on BDs datasets represented another limitation in comparing this research results with a benchmark one. Also, the scarcity of the pre-trained models, such as Bidirectional Encoder Representation from Transformers (BERT) on BDs, represented a limitation for the transfer learning process.

Besides, this research results consider some of the legal and ethical issues, such as the bias and the transparency of both the dataset and model performance which were revealed through the documentation process of detailed steps of datasets design and creation and the model configuration and evaluation.

Another ethical issue considered here was dataset privacy, where the participants of writing the reviews in BDs were anonymous. Additionally, the dataset was undertaken per the research ethics regulations of Brunel University London and was approved by the university ethics committee.

All this documentation and clarification allow the stakeholders to ask more questions which, when answered, will contribute to the future development of model designing, especially human-based. The documentation process also will help give the users a general idea about the model's suitability for their specific purposes.

7.3 Summary

The obtained results of this research have important implications for developing more studies in Arabic NLP that consider the SA of sequential textual data, specifically in Bahraini dialects, not only for the researchers but for all stakeholders and analysts who focus on the Arab world that plays an essential role in the international policies and global economy. The NLP researchers who can cite this work in their related works, and stakeholders such as 1- people who are interested in linguistic studies to compare dialects and standard languages, 2- technology companies who tend to use abundant data to satisfy the needs of some parties of business and money world, 3- the people of who work in commercial fields by developing the marketing strategy, promoting the customer relationship, and improving the product which in turn contribute in purchasing new products.

7.4 Future Work

There are a few ways by which this research could be further extended:

1. Conducting other classification problems rather than binary ones for different types of applications such as live interviews, where multimodality SA, where visual and audio alongside the text-based SA can be carried out or bimodal SA, where a combination of two modalities can be applied: speech and text, speech and image, or image and text which helps in predicting the individual sentiment more comprehensively.

2. Creating a larger dataset covering different domains such as books, restaurants, hotels, travelling, and healthcare from various sources, as the availability of more than one dataset that covers more domains can help in compensating the scarcity of resources and studies on other domains through utilizing the transfer learning. Additionally, various datasets can serve NLP researchers in different fields, such as medical, commercial, educational, and nutrition.

3. Creating a sentiment lexicon covering the previously mentioned areas that can be used in an unsupervised learning process.

4. Implementing the Bidirectional Encoder Representation from Transformers (BERT) for word embedding. By BERT, every word in the document has different

embeddings in different sentences due to the context-aware-embedding, where 15% of input words are masked, and the model is asked to predict the missing words in an unsupervised manner, where missing words are learned based on attention to the forthcoming and previous words. BERT can be used in transfer learning by adding a new layer on top of a pre-trained one and training the model for another task. This is useful when unlabelled data are available due to BERT's masked language modelling technique (Ghojogh, Ghodsi, 2020).

References

1. Abdullah, M., Hadzikadicy, M. and Shaikhz, S., 2018, December. SEDAT: Sentiment and Emotion Detection in Arabic Text Using CNN-LSTM Deep Learning. *In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 835-840). IEEE.
2. Abdul-Latif, E. 2016, "Bassiouney, Reem (2009) Arabic Sociolinguistics. Edinburgh University Press: Edinburgh; pp. 311 i-xiv", *International Journal of Arabic Linguistics*, vol. 2, no. 2, pp. 154-158.
3. Abdul-Mageed, M., Alhuzali, H. and Elaraby, M., 2018, May." You tweet what you speak: A city-level dataset of Arabic dialects". *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018*).
4. Abo, M.E.M., Shah, N.A.K., Balakrishnan, V., Kamal, M., Abdelaziz, A. and Haruna, K., 2019, April. SSA-SDA: Subjectivity and Sentiment Analysis of Sudanese Dialect Arabic. *In 2019 International Conference on Computer and Information Sciences (ICCIS*) (pp. 1-5). IEEE.
5. Agüero-Torales, M.M., Salas, J.I.A. and López-Herrera, A.G. (2021) 'Deep learning and multilingual sentiment analysis on social media data: An overview', *Applied Soft Computing*, , pp. 107373.
6. Ahmad, M., Aftab, S., Muhammad, S.S. and Ahmad, S. (2017) 'Machine learning techniques for sentiment analysis: A review', *Int.J.Multidiscip.Sci.Eng*, 8(3), pp. 27.
7. Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B. and Rehman, A., 2017. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl*, *8*(6), p.424.
8. Al Sallab, A., Hajj, H., Badaro, G., Baly, R., El-Hajj, W. and Shaban, K., 2015, July. Deep learning models for sentiment analysis in Arabic. In *Proceedings of the second workshop on Arabic natural language processing* (pp. 9-17).
9. Alahmary, R.M., Al-Dossari, H.Z. and Emam, A.Z., 2019, January.

Sentiment analysis of Saudi dialect using deep learning techniques. In *2019 International Conference on Electronics, Information, and Communication (ICEIC)* (pp. 1-6). IEEE.

10. Alayba, A.M., Palade, V., England, M. and Iqbal, R., 2018, August. A combined CNN and LSTM model for arabic sentiment analysis. In *International cross-domain conference for machine learning and knowledge extraction* (pp. 179-191). Springer, Cham..
11. Al-Azani, S. and El-Alfy, E.S., 2018, March. Emojis-based sentiment classification of Arabic microblogs using deep recurrent neural networks. In *2018 International Conference on Computing Sciences and Engineering (ICCSE)* (pp. 1-6). IEEE.
12. Albiladpress.com. 2016. *اللهجات الشعبية في البحرين وجذورها التاريخية*. [online] Available at: <http://albiladpress.com/article350009-4.html> [Accessed 19 October 2021].
13. Alessia, D., Ferri, F., Grifoni, P. and Guzzo, T., 2015. Approaches, tools and applications for
    sentiment analysis implementation. *International Journal of Computer Applications*, 125(3)
14. Algburi, M.A., Mustapha, A., Mostafa, S.A. and Saringatb, M.Z. (2019) *Comparative Analysis for Arabic Sentiment Classification.* Springer, pp. 271.
15. Al-Hashedi, A., Al-Fuhaidi, B., Mohsen, A.M., Ali, Y., Gamal Al-Kaf, H.A., Al-Sorori, W. & Maqtary, N. 2022, "Ensemble Classifiers for Arabic Sentiment Analysis of Social Network (Twitter Data) towards COVID-19-Related Conspiracy Theories", *Applied Computational Intelligence and Soft Computing*, vol. 2022.
16. Alnawas, A. and Arıcı, N., 2018. The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: A literature review. *Politeknik Dergisi*, 21(2), pp.461-470.
17. Al-Sallab, A., Baly, R., Hajj, H., Shaban, K.B., El-Hajj, W. and Badaro, G., 2017. Aroma: A recursive deep learning model for opinion mining in arabic

as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *16*(4), pp.1-20.

18. Alsarsour, I., Mohamed, E., Suwaileh, R. and Elsayed, T., 2018, May. Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

19. Alsayat, A. 2022, "Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model", *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 2499-2511.

20. Alsayat, A. and Elmitwally, N., 2020. A comprehensive study for Arabic sentiment analysis (Challenges and Applications). *Egyptian Informatics Journal*, *21*(1), pp.7-12

21. Alshuaibi, A.S., Mohd Shamsudin, F. and Alshuaibi, M.S.I., 2015. Internet misuse at work in Jordan: Challenges and implications.

22. Altaher, A., 2017. Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting. *International Journal of Advanced and Applied Sciences*, *4*(8), pp.43-49.

23. Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A. and Al-Ohali, Y., 2017. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, *117*, pp.63-72.

24. Ameen, F., 2020. *لسانيات (6) جناحا اللسان البحريني* [online] Akhbar-alkhaleej.com. Available at: <http://www.akhbar-alkhaleej.com/news/article/1210862> [Accessed 20 October 2021].

25. Arora, A., Candel, A., Lanford, J., LeDell, E. and Parmar, V., 2015. Deep learning with h2o. *H2O. ai*, *Mountain View*, *587*.

26. Attia, M., Samih, Y., Elkahky, A. and Kallmeyer, L., 2018, May. Multilingual multi-class sentiment classification using convolutional neural networks. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

27. Badaro, G., Baly, R., Hajj, H., El-Hajj, W., Shaban, K.B., Habash, N., Al-Sallab, A. and Hamdi, A. (2019) 'A survey of opinion mining in Arabic: a comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations', *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3), pp. 1-52.

28. Baly, R., Badaro, G., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., El-Hajj, W., Habash, N. and Shaban, K., 2017b, April. A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models. In *Proceedings of the third Arabic natural language processing workshop* (pp. 110-118).

29. Baly, R., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., Shaban, K.B. and El-Hajj, W., 2017a. Comparative evaluation of sentiment analysis methods across Arabic dialects. *Procedia Computer Science*, *117*, pp.266-273.

30. Becker, W., Wehrmann, J., Cagnini, H.E. and Barros, R.C., 2017, May. An efficient deep neural architecture for multilingual sentiment analysis in twitter. In *The Thirtieth International Flairs Conference*.

31. Birjali, M., Kasri, M. and Beni-Hssane, A., 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, p.107134.

32. Boudad, N., Faizi, R., Thami, R.O.H. and Chiheb, R., 2018. Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal*, **9**(4), pp. 2479-2490.

33. Brașoveanu, A.M. and Andonie, R., 2020, September. Visualizing transformers for nlp: a brief survey. In *2020 24th International Conference Information Visualisation (IV)* (pp. 270-279). IEEE.

34. Brownlee, J. 2016, "How to grid search hyperparameters for deep learning models in python with keras". Available at: *https://machinelearningmastery.com/grid-search-hyperparameters-deep-*

learning-models-python-keras, (Accessed: 3 June 2021).

35. Brownlee, J. 2017a. *Long Short-term Memory Networks with Python.* V1.7 ed.: Jason Brownlee.
36. Brownlee, J. 2017b**, "***How to Use Word Embedding Layers for Deep Learning with Keras, Machine Learning Mastery***.** Available at: https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/ (Accessed: 10 April 2021).
37. Brownlee, J. 2017c**,"***Stacked Long Short-Term Memory Networks, Machine Learning Mastery*. Available at: https://machinelearningmastery.com/stacked-long-short-term-memory-networks/ (Accessed: 7 September 2021).
38. Brownlee, J. 2018a, "*Difference Between a Batch and an Epoch in a Neural Network, Machine Learning Mastery*. Available at: https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/ (Accessed: 20 May 2021).
39. Brownlee, J. 2018b, *How to Use ROC Curves and Precision-Recall Curves for Classification in Python*.
40. Brownlee, J. 2019a," *How to Control Neural Network Model Capacity With Nodes and Layers, Machine Learning Mastery*. Available at: https://machinelearningmastery.com/how-to-control-neural-network-model-capacity-with-nodes-and-layers/ (Accessed: 25 April 2021).
41. Brownlee, J. 2019b, "*A Gentle Introduction to the Rectified Linear Unit (ReLU), Machine Learning Mastery*. Available at: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ (Accessed: 25 April 2021).
42. Brownlee, J. 2019c," *How to Configure the Learning Rate When Training Deep Learning Neural Networks, Machine Learning Mastery*. Available at: https://machinelearningmastery.com/learning-rate-for-deep-learning-neural-networks/ (Accessed: 20 May 2021).
43. Brownlee, J. 2019d, " *How to Improve Performance With Transfer*

Learning for Deep Learning Neural Networks, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/how-to-improve-performance-with-transfer-learning-for-deep-learning-neural-networks/ (Accessed: 28 December 2021).

44. Brownlee, J., 2020. Jump-Start Training with Transfer Learning. In *Better Deep Learning*. pp. 221–221.
45. Budumam, N and Locascio, N. (2017). *Fundamentals of Deep Learning Designing Next-Generation Machine Intelligence Algorithms.* 1st ed. United States of America: OREILY. P 174-178.
46. Can, E.F., Ezen-Can, A. and Can, F., 2018. Multilingual sentiment analysis: An rnn-based framework for limited data. *arXiv preprint arXiv:1806.04511*.
47. Chandra, A. & Yao, X. 2006, "Evolving hybrid ensembles of learning machines for better generalisation", *Neurocomputing*, vol. 69, no. 7-9, pp. 686-700.
48. Chaudhari, V.V., Dhawale, C.A. and Misra, S., 2016.Sentiment Analysis Classification: A Brief.
49. Chen, H. and Ji, Y. (2019) 'Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation', *arXiv preprint arXiv:1909.04225*, .
50. Chicco, D. & Jurman, G. 2020, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", *BMC genomics*, vol. 21, no. 1, pp. 1-13.
51. Chinea, A., 2009, September. Understanding the Principles of Recursive Neural Networks: A Generative Approach to Tackle Model Complexity. In International Conference on Artificial Neural Networks (pp. 952-963). Springer, Berlin, Heidelberg.
52. Chollet, CB (ed.) 2018, Deep Learning with python, *Manning Publications Co*, Shelter Island.
53. Cliche, M., 2017. BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. *arXiv preprint arXiv:1704.06125*..

54. Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H.T., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V., El-Beltagy, S.R., El-Hajj, W. and Jarrar, M., 2021. A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, *64*(4), pp.72-81.

55. Dong, X.L. and De Melo, G., 2018, July. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2524-2534).

56. Elnagar, A., Einea, O. and Lulu, L., 2017, October. Comparative study of sentiment classification for automated translated Latin reviews into Arabic. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)* (pp. 443-448). IEEE.

57. Elnagar, A., Yagi, S.M., Nassif, A.B., Shahin, I. and Salloum, S.A. (2021) 'Systematic literature review of dialectal Arabic: identification and detection', *IEEE Access*, 9, pp. 31010-31042.

58. Fouad, M.M., Mahany, A., Aljohani, N., Abbasi, R.A. and Hassan, S. (2020) 'ArWordVec: efficient word embedding models for Arabic tweets', *Soft Computing*, 24(11), pp. 8061-8068.

59. Fouadi, H., El Moubtahij, H., Lamtougui, H., SATORI, K. and Yahyaouy, A. (2020) *Applications of Deep Learning in Arabic Sentiment Analysis: Research Perspective.* IEEE, pp. 1.

60. Fourure, D., Javaid, M.U., Posocco, N. & Tihon, S. 2021, "Anomaly detection: how to artificially increase your f1-score with a biased evaluation protocol", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*Springer, , pp. 3.

61. Gangula, R.R.R. and Mamidi, R., 2018, March. Impact of Translation on Sentiment Analysis: A Case-Study on Telugu Reviews. In *19th International Conference on Computational Linguistics and Intelligent Text Processing*.

62. Ghojogh, B. and Ghodsi, A., 2020. Attention mechanism, transformers,

BERT, and GPT: tutorial and survey.

63. Goldberg, Y. (2017) 'Neural network methods for natural language processing', *Synthesis Lectures on Human Language Technologies,* 10(1), pp. 1-309.
64. Goldberg, Y., 2016. A Primer on Neural Network Models for Natural Language Processing. J. Artif. Intell. Res.(JAIR), 57, pp.345-420.
65. *Google developers, 2020, Classification: Accuracy | Machine Learning Crash Course* (2021). Available at: https://developers.google.com/machine-learning/crash-course/classification/accuracy (Accessed: 4 September 2021).
66. Guellil, I., Azouaou, F. and Mendoza, M., 2019. Arabic sentiment analysis: studies, resources, and tools. *Social Network Analysis and Mining, 9(1), p.56.*
67. Guo, Y. and Xiao, M., 2012. Cross language text classification via subspace co-regularized multi-view learning. *arXiv preprint arXiv:1206.6481*.
68. Haeri, N. 2003, *Sacred language, ordinary people: Dilemmas of culture and politics in Egypt,* Springer.
69. Haje, U.A., Abdalla, M.H., Kurda, R.M.S. & Khalid, Z.M. 2022, "A New Model for Emotions Analysis in Social Network Text Using Ensemble Learning and Deep learning", *Academic Journal of Nawroz University*, vol. 11, no. 1, pp. 130-140.
70. Hajiali, M., 2020. Big data and sentiment analysis: A comprehensive and systematic literature review. *Concurrency and Computation: Practice and Experience*, *32*(14), p.e5671.
71. HAN, J. and Moraga, C., 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning, *International workshop on artificial neural networks* 1995, Springer, pp. 195-201.
72. Heikal, M., Torki, M. and El-Makky, N., 2018. Sentiment analysis of Arabic Tweets using deep learning. *Procedia Computer Science*, *142*, pp.114-122.

73. Itani, M., Roast, C. and Al-Khayatt, S., 2017. Developing resources for sentiment analysis of informal Arabic text in social media. *Procedia Computer Science*, *117*, pp.129-136.

74. Joachims, T., 1996. A probabilistic analysis of the rocchio algorithm with tfidf for text classification. In *14th International Conference on Machine Learning* (pp. 143-151).

75. *Kaggle* (2019). Available at: https://en.wikipedia.org/wiki/Kaggle (Accessed: 27 October 2020)

76. Kaur, J. and Saini, J.R., 2014. Emotion detection and sentiment analysis in text corpus: a differential study with informal and formal writing styles. *International Journal of Computer Application, ISSN*, pp.0975-8887.

77. Kazmaier, J. & van Vuuren, J.H. 2022, "The power of ensemble learning in sentiment analysis", *Expert Systems with Applications*, vol. 187, pp. 115819.

78. Kobayashi, S. (2018) 'Contextual augmentation: Data augmentation by words with paradigmatic relations', *arXiv preprint arXiv:1805.06201*, .

79. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

80. Kumar, R. & Verma, R. 2012, "Classification algorithms for data mining: A survey", *International Journal of Innovations in Engineering and Technology (IJIET)*, vol. 1, no. 2, pp. 7-14.

81. Liang, J., Qin, Z., Xiao, S., Ou, L. & Lin, X. 2019, "Efficient and secure decision tree classification for cloud-assisted online diagnosis services", *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 4, pp. 1632-1644.

82. Lin, S., Kung, Y. & Leu, F. 2022, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis", *Information Processing & Management*, vol. 59, no. 2, pp.

102872.

83. Lin, T., Wang, Y., Liu, X. and Qiu, X., 2022. A survey of transformers. *AI Open*.

84. Liu, R., Shi, Y., Ji, C. and Jia, M. (2019) 'A survey of sentiment analysis based on transfer learning', *IEEE Access*, 7, pp. 85401-85412.

85. Luo, S., Gu, Y., Yao, X. & Fan, W. 2021, "Research on Text Sentiment Analysis Based on Neural Network and Ensemble Learning.", *Rev.d'Intelligence Artif.*, vol. 35, no. 1, pp. 63-70.

86. Luque, F.M. (2019) 'Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis', *arXiv preprint arXiv:1909.11241*, .

87. *makcedward/nlpaug* (2021). Available at: https://github.com/makcedward/nlpaug/blob/master/example/quick_example.ipynb (Accessed: 6 April 2021).

88. Mdhaffar, S., Bougares, F., Esteve, Y. and Hadrich-Belguith, L., 2017, April. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Third Arabic Natural Language Processing Workshop (WANLP)* (pp. 55-61).

89. Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp.1093-1113.

90. Medrouk, L. and Pappa, A., 2017, November. Deep learning model for sentiment analysis in multi-lingual corpus. In *International Conference on Neural Information Processing* (pp. 205-212). Springer, Cham.

91. Mitchell, T.F., 1990. *Pronouncing Arabic* (Vol. 2). Oxford University Press, USA.

92. Mohammad, S.M., Salameh, M. and Kiritchenko, S., 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55, pp.95-130.

93. Mohammed, A. & Kora, R. 2021, "An effective ensemble deep learning framework for text classification", *Journal of King Saud University-Computer and Information Sciences*, .

94. Mohammed, A. and Kora, R. (2019) 'Deep learning approaches for Arabic sentiment analysis', *Social Network Analysis and Mining*, 9(1), pp. 1-12.

95. Munezero, M., Montero, C.S., Mozgovoy, M. and Sutinen, E. (2013) *Exploiting sentiment analysis to track emotions in students' learning diaries.* pp. 145.

96. Obeid, O., Salameh, M., Bouamor, H. and Habash, N., 2019, June. ADIDA: Automatic Dialect Identification for Arabic. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (pp. 6-11).

97. Omara, E., Mosa, M. and Ismail, N. (2019) *Emotion Analysis in Arabic Language Applying Transfer Learning.* IEEE, pp. 204.

98. Omran, T., Sharef, B., Grosan, C. and Li, Y., 2023, March. The Impact of Data Augmentation on Sentiment Analysis of Translated Textual Data. In *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)* (pp. 1-4). IEEE.

99. Ortiz, M., Scheidegger, F., Casas, M., Malossi, C. & Ayguadé, E. 2020, "Generating Efficient DNN-Ensembles with Evolutionary Computation", *arXiv preprint arXiv:2009.08698*.

100. Osama, M. and El-Beltagy, S.R. (2019) *A Transfer Learning Approach for Emotion Intensity Prediction in Microblog Text.* Springer, pp. 512.

101. Özbey, C., Dilekoğlu, B. & Açiksöz, S. 2021, "The Impact of Ensemble Learning in Sentiment Analysis under Domain Shift", *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*IEEE, , pp. 1.

102. Pawar, A.B., Jawale, M.A. and Kyatanavar, D.N. (2016) 'Fundamentals of sentiment analysis: concepts and methodology'*Sentiment analysis and ontology engineering* Springer, pp. 25-48.

103. Phi, M 2018,*Illustrated Guide to LSTM's and GRU's: A step by step explanation* (2020), Medium, Available at: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (Accessed: 30 August 2021).

104. Rahab, H., Zitouni, A. and Djoudi, M., 2017, September. Siaac: Sentiment

polarity identification on arabic algerian newspaper comments. In *Proceedings of the Computational Methods in Systems and Software* (pp. 139-149). Springer, Cham..

105. Rokach, L. 2010, *Pattern classification using ensemble methods*, World Scientific.
106. Ruangkanokmas, P., Achalakul, T. and Akkarajitsakul, K., 2016, January. Deep belief networks with feature selection for sentiment classification. In *2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)* (pp. 9-14). IEEE
107. Ruby, U. & Yendapalli, V. 2020, "Binary cross entropy with deep learning technique for image classification", *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 10.
108. Sarkar, D., Bali, R. and Ghosh, T., 2018. *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras*. Packt Publishing Ltd.
109. Setia, M. (2023) *Binary Cross Entropy aka Log Loss-the cost function used in logistic regression*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/11/binary-cross-entropy-aka-log-loss-the-cost-function-used-in-logistic-regression/ (Accessed: 14 July 2023).
110. Sharaf, A.M. and Atwell, E., 2012. QurAna: Corpus of the Quran annotated with Pronominal Anaphora. *LREC* 2012, Citeseer, pp. 130-137.
111. Sharami, J.P.R., Sarabestani, P.A. and Mirroshandel, S.A. (2020) 'Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus', *arXiv preprint arXiv:2004.05328*, .
112. Smagulova, K. and James, A.P. (2019) 'A survey on LSTM memristive neural network architectures and applications', *The European Physical Journal Special Topics*, 228(10), pp. 2313-2324.
113. Soliman, A.B., Eissa, K. and El-Beltagy, S.R., 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, *117*, pp.256-265.

114. Soufan, A., 2019, March. Deep learning for sentiment analysis of arabic text. In *Proceedings of the ArabWIC 6th Annual International Conference Research Track* (pp. 1-8).

115. Staudemeyer, R.C. and Morris, E.R. (2019) 'Understanding LSTM--a tutorial into Long Short-Term Memory Recurrent Neural Networks', *arXiv preprint arXiv:1909.09586*, .

116. Sun, X. and He, J. (2020) 'A novel approach to generate a large scale of supervised data for short text sentiment analysis', *Multimedia Tools and Applications*, 79(9), pp. 5439-5459.

117. Team, K. (2021) *Keras documentation: Adamax, Keras.io*. Available at: https://keras.io/api/optimizers/adamax/ (Accessed: 25 April 2021).

118. Team, K., *Keras documentation: Classification metrics based on True/False positives & negatives*. Available: https://keras.io/api/metrics/classification_metrics/ [2022, Sep 15,].

119. Thomas, A., 2019. *Coding The Deep Learning Revolution Ebook - Adventures In Machine Learning*. [online] Adventures in Machine Learning. Available at: <https://adventuresinmachinelearning.com/coding-deep-learning-ebook/> [Accessed 6 June 2020].

120. Tilmatine, M. 1999, "Substrat et convergences: le berbère et l´ arabe nord-africain", *EDNA, Estudios de dialectología norteafricana y andalusí*, vol. 4, pp. 99-119.

121. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

122. Versteegh, K., 2014. *Arabic language*. Edinburgh University Press.

123. Warrens, M.J. 2015, "Five ways to look at Cohen's kappa", *Journal of Psychology & Psychotherapy*, vol. 5, no. 4, pp. 1.

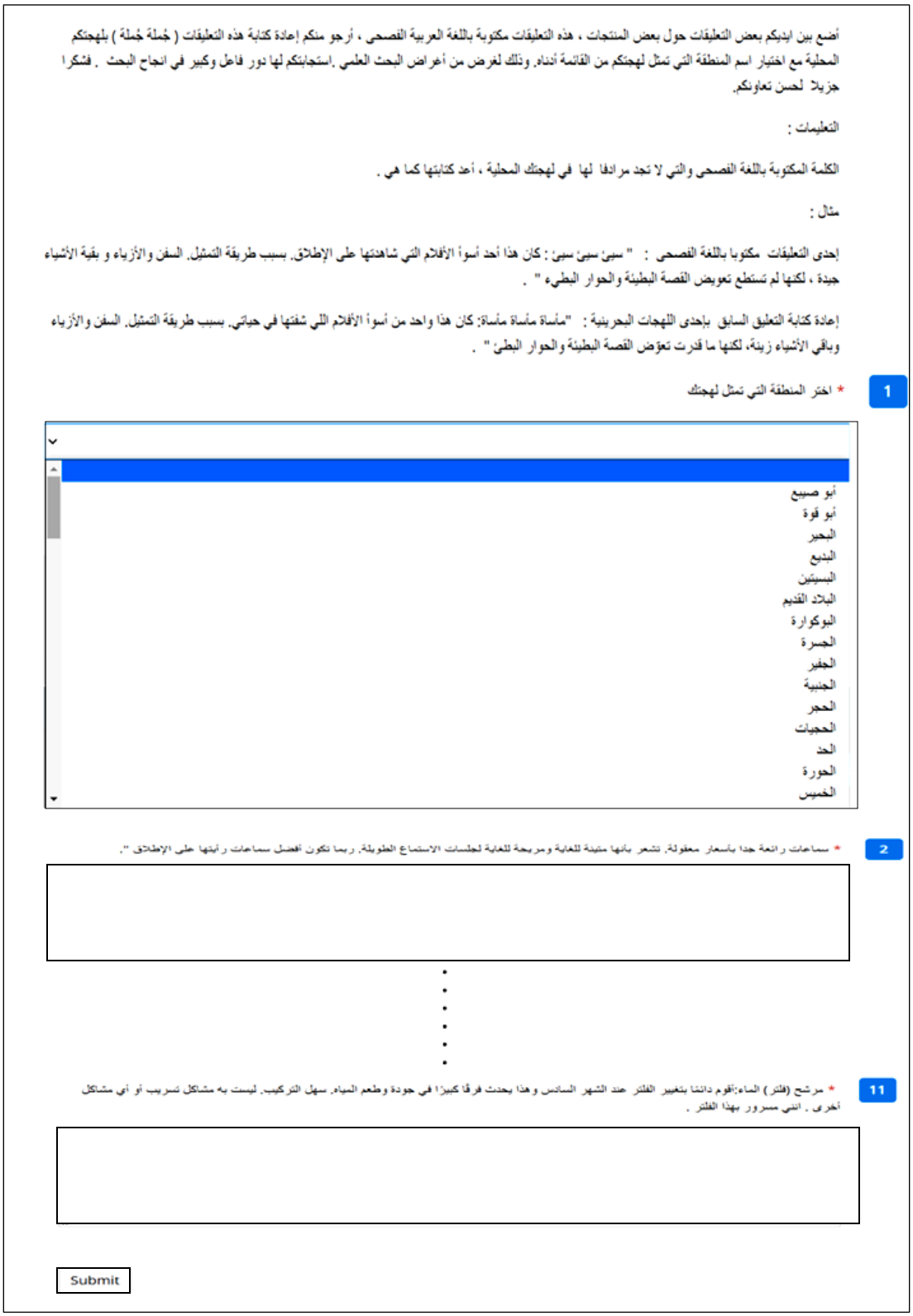
124. Wehrmann, J., Becker, W.E. and Barros, R.C., 2018, April. A multi-task neural network for multilingual sentiment classification and language detection on twitter. In *Proceedings of the 33rd Annual ACM Symposium*

on Applied Computing (pp. 1805-1812).

125. Wei, J. and Zou, K. (2019) 'Eda: Easy data augmentation techniques for boosting performance on text classification tasks', *arXiv preprint arXiv:1901.11196*,
126. Weiss, K., Khoshgoftaar, T.M. and Wang, D. (2016) 'A survey of transfer learning', *Journal of Big data*, 3(1), pp. 1-40.
127. Wolpert, D.H. 1992, "Stacked generalization", *Neural Networks*, vol. 5, no. 2, pp. 241-259.
128. Wong, T.T., 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), pp.2839-2846.
129. Zaidan, O.F. and Callison-Burch, C., 2014. Arabic dialect identification. *Computational Linguistics*, *40*(1), pp.171-202.
130. Zaman, M.F. & Hirose, H. 2011, "Classification performance of bagging and boosting type ensemble methods with small training sets", *New Generation Computing*, vol. 29, no. 3, pp. 277-292.
131. Zhang, L., Wang, S. and Liu, B. (2018) 'Deep learning for sentiment analysis: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), pp. e1253.
132. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y., 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Appendices

Appendix 1: An example of the distributed form that used for obtaining the corresponding Bahraini dialects of Modern Standard Arabic products reviews.



أضع بين ايديكم بعض التعليقات حول بعض المنتجات ، هذه التعليقات مكتوبة باللغة العربية الفصحى ، أرجو منكم إعادة كتابة هذه التعليقات (جُملة جُملة) بلهجتكم المحلية مع اختيار اسم المنطقة التي تمثل لهجتكم من القائمة أدناه. وذلك لغرض من أغراض البحث العلمي .استجابتكم لها دور فاعل وكبير في انجاح البحث . فشكرا جزيلا لحسن تعاونكم.

التعليمات :

الكلمة المكتوبة باللغة الفصحى والتي لا تجد مرادفا لها في لهجتك المحلية ، أعد كتابتها كما هي .

مثال :

إحدى التعليقات مكتوبا باللغة الفصحى : " سيئ سيئ سيئ : كان هذا أحد أسوأ الأفلام التي شاهدتها على الإطلاق. بسبب طريقة التمثيل. السفن والأزياء و بقية الأشياء جيدة ، لكنها لم تستطع تعويض القصة البطيئة والحوار البطيء " .

إعادة كتابة التعليق السابق بإحدى اللهجات البحرينية : "مأساة مأساة مأساة: كان هذا واحد من أسوأ الأفلام اللي شفتها في حياتي. بسبب طريقة التمثيل. السفن والأزياء وباقي الأشياء زينة، لكنها ما قدرت تعوّض القصة البطيئة والحوار البطئ " .

1 * اختر المنطقة التي تمثل لهجتك

أبو صبيع
أبو قوة
البحير
البديع
البسيتين
البلاد القديم
البوكوارة
الجسرة
الجفير
الجنبية
الحجر
الحجيات
الحد
الحورة
الخميس

2 * سماعات رائعة جدا بأسعار معقولة. تشعر بأنها متينة للغاية ومريحة للغاية لجلسات الاستماع الطويلة. ربما تكون أفضل سماعات رأيتها على الإطلاق ".

•
•
•
•
•
•

11 * مرشح (فلتر) الماء:أقوم دائما بتغيير الفلتر عند الشهر السادس وهذا يحدث فرقًا كبيرًا في جودة وطعم المياه. سهل التركيب. ليست به مشاكل تسريب أو أي مشاكل أخرى . انني مسرور بهذا الفلتر .

Submit

Appendix 2: Ethical approval letter

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

24 June 2020

LETTER OF CONFIRMATION

Applicant: Ms Thuraya Omran

Project Title: AN EFFICIENT DEEP LEARNING APPROACH FOR MULTI-LANGUAGES SENTIMENT ANALYSIS

Reference: 23598-NER-Jun/2020- 25852-1

Dear Ms Thuraya Omran

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has confirmed that on the basis of the information provided in your application, your project does not require ethical review.

Please note that:

- Approval to proceed with the study is granted providing that you do not carry out any research which concerns a human participant, their tissue and/or their data.
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must inform the Committee by submitting an appropriate Research Ethics Application. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,

Professor Hua Zhao

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

Appendix 3: List of the covered Bahraini towns and villages in obtaining the products reviews in Bahraini dialects and the corresponding statistics of obtained responses

Number	Name of Town/Village In Arabic	Name of Town/Village In English	Number of Responses (Filled forms)
1	أبو قوة	AbuQuwah	3
2	البديع	Al Budaiya	2
3	البلاد القديم	Al Bilad Al Qadeem	3
4	الجفير	Al Juffair	4
5	الحد	Al Hidd	4
6	الدراز	Al Diraz	13
7	الدير	Al Dair	9
8	الديه	Al Daih	4
9	الرفاع الشرقي	East Rifaa	18
10	السنابس	Al Sanabis	8
11	السهلة الشمالية	North Sehla	2
12	العكر	Al Eker	1
13	القرية	Al Qurrayah	1
14	الكورة	Al Kawarah	6
15	المالكية	Al Malikiyah	6
16	المحرق	Al Muharraq	8
17	المصلى	Al Musalla	2
18	المعامير	Al Ma’ameer	19
19	المنامة	Al Manama	27
20	النعيم	Al Naim	3
21	النويدرات	Al Nuwaidrat	264
22	أم الحصم	Umm AlHassam	4
23	بوري	Buri	5
24	توبلي	Tubli	4
25	جبلة حبشي	Jeblat Hebshi	1
26	جدحفص	Jidhafs	3
27	جدعلي	JidAli	5
28	دمستان	Damistan	1
29	رأس رمان	Ras Romman	10
30	سار	Saar	4
31	سترة	Sitra	21
32	سماهيج	Samaheej	4
33	سند	Sanad	4
34	عراد	Arad	3
35	كرانة	Karranah	10
36	كرزكان	Karzakkan	5
37	مدينة عيسى	Isa Town	7
38	مقابة	Maqabah	2

Appendix 4: Stopwords list for Bahraini dialect

هذولة	كنت	أو	ابه
هذوله	لان	اى	ابو
هذي	لاى	أى	احد
هل	لاي	اي	احدى
هم	لحد	أي	اذا
هنا	لدا	اياه	إذا
هنى	لقد	إيلي	اكون
هني	للي	بس	ال
هو	لما	به	الاخرى
هوه	لنا	بها	الأخرى
هي	له	بهذا	اللى
و	لها	تكون	اللي
وا	لهذي	تكونين	إللي
وانت	لو	جدي	المو
وانه	لي	حتى	النا
واني	لينا	حق	اله
وأني	ليه	دا	إله
وهذه	ليها	دلين	الى
وهي	ليها	دي	إلى
ويا	ليي	ديلاك	الي
وياها	مال	ذاك	إلي
وياهم	ماله	ذي	اليه
ويايي	مالها	ذيك	إليه
ويكون	مالهم	راح	ان
يكون	مايكون	شكلاة	أن
	مثل	شكلاه	انا
	من	عشان	أنا
	منها	علشان	آنا
	هادا	على	انت
	هادي	عليه	انك
	هاذا	عليهم	إنك
	هاذلين	عليي	انه
	هاذي	عندما	إنه
	هال	عنه	انها
	هاي	عنها	إنها
	هدا	فهذا	أنها
	هدة	في	اني
	هدي	فيها	أني
	هذه	كان	اهيا
	هذولا	كانت	او