The Impact of Data Augmentation on Sentiment Analysis of Translated Textual Data

Thuraya Omran
Department of Computer Science
Brunel University London
Uxbridge, United Kingdom
Thuraya.Omran@brunel.ac.uk

Baraa Sharef
Department of Information Technology
Ahlia University
Manama, Kingdom of Bahrain
bsharif@ahlia.edu.bh

Crina Grosan
Division of Applied Technologies for Clinical Care
King's College London
Strand, United Kingdom
crina.grosan@kcl.ac.uk

Yongmin Li
Department of Computer Science
Brunel University London
Uxbridge, United Kingdom
Yongmin.Li@brunel.ac.uk

***Abstract*—Sentiment analysis is an application of natural language processing that requires an abundance of data that may not be achieved sometimes for some reason. Data augmentation is one technique that deals with the lack of data by creating synthetic training data without adding new ones. It boosts model performance, especially with deep learning ones. Despite its influential role in boosting the model performance, it attracted very little attention from the researchers of the Arabic NLP community, specifically with scarce language resources such as Arabic and its dialects. In this study, one of the augmentation techniques called random swap was applied with LSTM deep learning model to classify three parallel datasets. The three parallel datasets are Bahraini dialects, Modern Standard Arabic and English. The results show an improvement in the LSTM model by 14.06%, 12.57%, and 11.04% on Bahraini dialects, Modern Standard Arabic, and English datasets, respectively, when applying the augmentation technique over that of no application.**

Keywords—Data augmentation, LSTM, translation-based, Modern standard Arabic, Bahraini dialects.

I. Introduction

Sentiment analysis (SA) is a way of mathematically studying peoples' impressions, and emotions about entities such as organizations, individuals, and events [1], [2] whether these impressions were positive or not [3]. SA is a significant area of interest in natural language processing (NLP), which requires the availability of resources such as lexicons or corpora.

Dataset creation is not easy; its collection is difficult, expensive, and time-consuming. According to [4], the manual creation of the dataset gives accurate results. However, its construction leads to small datasets unsuitable for machine-learning tasks. Data augmentation is one solution that deals with the lack of data. It is a technique of increasing the training data by creating diverse and artificial ones by methods of transformation [5]. It contributes to boosting data [6] and model performance, especially with deep learning-based models; those need a large training data size to overcome problems like data sparsity and overfitting [7].

There are different methods of data augmentation, some of which are carried out at the data space level that is applied to raw data. In contrast, others are conducted at the feature space level and applied to the preprocessed data. The transformation methods on data space can be applied at the character level, word level, sentence level or document level. In contrast, the transformation method on feature space can be applied by interpolation such as SMOTE or mix-up interpolation [5].

Four powerful data augmentation methods consisted of easy data augmentation (EDA) techniques of [8], such as random deletion, random insertion, random swap, and synonym replacement.

Data augmentation is broadly used in speech [9], [10], and computer vision [11] when training efficient deep learning models. However, it attracted very little attention from the Arabic NLP scholarly community despite the spread of the Arabic language [12]. This little attention may be due to the absence of a general rule of data augmentation that, when applied, helps in preserving the meaning of opinion and its label. In this study, an exploration of one of the data augmentation techniques was adopted on the English dataset of Amazon product reviews that was translated by Google translate to modern standard Arabic (MSA), which in turn was translated manually to Bahraini dialects.

The primary aim of this study is to investigate the effect of an augmentation technique on the SA of textual data obtained using a translation approach and evaluate its effect on LSTM performance.

In this study, a random swap augmentation technique was applied to our dataset. Random swap is considered one method of noise induction at the word level [5] .

The experimental work presented here provides one of the first investigations to explore and compare the effect of augmentation techniques on the translated datasets. The obtained results can be used as a benchmark for interested researchers in the same area.

This paper is divided into 5 sections. Section 2 shed lights on related work; section 3 describes the methods of this study, section 4 reports and discusses the results, while conclusion and future works were covered in section 5.

II. Related Work

Data augmentation is one technique that copes with the lack of data and boosts the data and model performance.

In a shared task to detect the polarity of the sentiment of Spanish tweets, two techniques of data augmentation were used by[6] to cope with data scarcity: 1-machine back translation from Spanish to other four pivot languages (English, Arabic, Portuguese, and French) and 200 languages, and back to Spanish. The translation process was done using Google's cloud translation API services. 2- Instance cross-over: a novel technique tried by [6]inspired by genetic algorithm cross-over operation. This technique creates a new instance by combining two halves, one half from the original instance while the other represents half of another. The instance cross-over technique augmentation showed feasible results, despite using a logistic regression algorithm. At the same time, the back translation of the four pivot languages (English, Arabic, Portuguese, and French) does not improve model performance.

Similarly, a novel technique called contextual augmentation was proposed by [13], where a sentence was supposed to stay invariant, even if words in it are replaced with other words that have paradigmatic relation with it. This technique provides a wide range of words to replace the word in the original text with predicted words from its context, utilizing bi-directional convolutional (CNN) or recurrent network (RNN) in multi-tasks of classification.

Along the same lines, three data augmentation methods were suggested by [14] for extracting sentiments at the sentence level from low resources Persian dataset. These methods were: synonym data augmentation, extra data augmentation to increase the small number of instances, and translation data augmentation.

A similar idea has been supported broadly by those who proposed multi-granularity data augmentation techniques, including sentence level, phrase level, and word level, that enhanced their hybrid neural network-based model, which was evaluated on a corpus of Chinese news headlines and a dataset of Chinese online comments.

At word level data augmentation, four parts were considered by [7]. The four parts were: 1- Substituting words with their synonyms using a thesaurus, where Stanford parser was used to parse sentences and substitute certain words with their synonyms. 2- Substituting word2vec where similar words were searched for in the semantic space to enrich the substitution process.3- Translating the sentence by adding some words that have no meaning to either side of the sentence left or right, which results 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, the focus was on two attributes of contemporary Chinese, adverbial phrases and attribute head. More details regarding these attributes in [7]. At sentence level data augmentation, the focus was on deleting the sentence that did not affect the sentiment of the text.

A similar point was made in the [15] study of addressing the problem of the explainability lack of the sentiment classifier. Two methods of data augmentation were proposed by [15] for creating more training instances. One of these methods is based on using a list of predefined words of sentiment as external knowledge, while the other method is based on adversarial instances. The augmentation methods of [15] were tested using LSTM and CNN classifiers and three benchmark datasets (MR, IMDB, and SST).

In the same vein, data augmentation by shuffling technique was adopted by [16] 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 data augmentation techniques with deep learning models due to their impact on models' performance and achievement. In this study, a random swap augmentation was applied.

III. Study Methodology

The methodology of this study is represented in two folds. The first fold included dataset creation and preprocessing, and the second fold included the sentiment analysis process of the dataset using the LSTM deep learning model.

The dataset was a parallel one composed of English reviews of Amazon products and their corresponding ones in modern standard Arabic (MSA) and Bahraini dialects (BDs). The dataset was created by translating 5k reviews to modern standard Arabic using machine translation via the Google translate web page. The MSA reviews were then manually translated to BDs with the participation of native speakers.

The random swap augmentation method at the word level was applied in this study to take advantage of data augmentation techniques. This method was specifically selected due to the restrictions of the Bahraini dialects (BDs) dataset. Random swap randomly chooses any pair of words in the sentence and interchanges their positions, which is done *n* times [8].

Python NLP augmentation library provides various textual augmenters at the character level, word level, and sentence level [17]. 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. In this study, a random swap was applied two times. The first time, it was applied with a minimum number of words equal to one and a maximum of 10. While in the second time of applying the word swap, the minimum number of words was 1, and the maximum number of words was 5. These numbers were the best ones that fit our dataset. It is worth mentioning that each augmented review was labeled with the same label as its corresponding original review.

Fig. 1 shows the steps of obtaining 10k reviews from the original 5k reviews by repeating the random swap method twice.

Table 1 shows an example of an English review and its corresponding ones in MSA and BDs in its initial form and that after applying the swap augmentation technique two times.

IV. Study Results

This study examined the impact of data augmentation on the LSTM model performance in classifying the translated datasets composed of English and its corresponding ones in MSA and BDs.

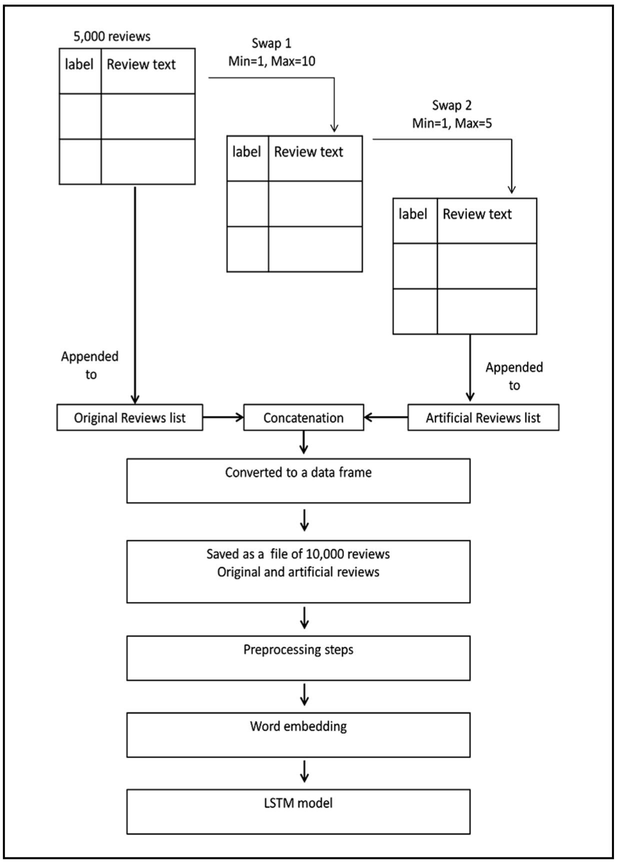


Fig. 1. Steps of obtaining 10k reviews from 5k reviews using random swap augmentation technique

All experiments were implemented twice on the English, MSA, and BDs datasets. The first time on 5k reviews, the second time on 10k reviews. All experiments on all datasets were conducted using 1- cross-validation split at k=3, 5, and 10. 2- Train-validate-test split with a ratio of 75% for training and 25 for both of validation and test part. 3- Learning rate = 0.01.

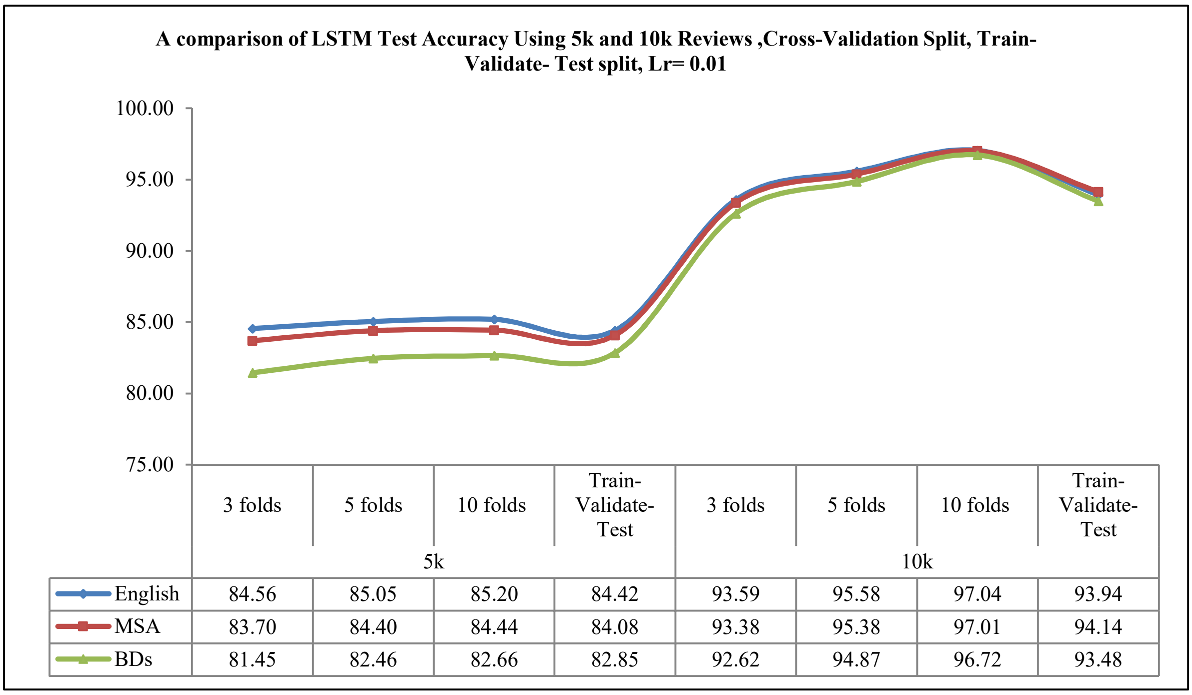
A comparison of the results in Fig. 2 of the LSTM model performance before and after applying the swap augmentation technique reveals an outperformance in the model achievement when adopting the augmentation technique. What stands out from the data in Fig. 2 is the marked improvement in the LSTM performance on 10k reviews that were obtained by adopting the swap augmentation technique in case of 10 folds cross-validation by 11.84%, 12.57%, and 14.06% on English, MSA, and BDs datasets, where it was 85.20% and became 97.04% on English dataset, 88.44% and became 97.01% on MSA, and 82.66% and became 96.72% on BDs. In contrast, the random swap augmentation technique boosted the LSTM model performance by 9.52%, 10.06%, and 10.63% on English, MSA, and BDs datasets, respectively, when using the train-validate-test split, where the accuracy rose from 84.42% to 93.94% on English dataset, and rose from 84.08% to 94.14% on MSA, while it rose from 82.85% to 93.48% on BDs.

There are some possible explanations for these results. Firstly, the model generalizability got better with data augmentation that satisfies the greedy characteristic of data of the LSTM. Secondly, data augmentation worked as a regularizer for the LSTM model, reducing the model overfitting. These results support the idea that data augmentation effectively boosts the classifier performance in data paucity.

A closer inspection of the data table in Fig.2 shows a slight difference between the obtained test accuracy of the LSTM performance on English, MSA, and BDs datasets in the case of 10 folds cross-validation and 10k reviews. For example, the accuracy difference between English and MSA was 0.03%, 0.29% between MSA and BDs, where it was 97.04% on English, 97.01% on MSA and 96.72% on BDs datasets. These results may be explained by pertaining text sentiment in addition to label preservation despite the text translation and transformation.

TABLE I. AN EXAMPLE OF APPLYING TWO TIMES RANDOM SWAP AUGMENTATION TECHNIQUE FOR PARALLEL DATA SET OF ENGLISH, MSA, AND BDS.

Review Text in	Initial Form of The Review	First Swap	Second Swap
English	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.	Boring: I bought 3 books because I read the review they and interesting look funny and. But the after pages first, I knew that this a was mistake. very Is repetitive and some of the really content for this ridiculous age and time.	Boring: I bought 3 books because I read the review they interesting and look funny and. But the after pages first I, knew that this a was. mistake very Is and repetitive some of the really content for this ridiculous and age time.
MSA	مملة : اشتريت 3 كتب لأنني قرأت المراجعة وهي تبدو ممتعة ومضحكة. لكن بعد الصفحات الأولى ، علمت أن هذا كان خطأ. مكرر جدا وبعض المحتويات سخيفة حقا لهذا العصر والزمان.	: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.	: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.
BDs	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 وااجد و محتواها تااااااااافه بالنسبة لهالزمن	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 وااجد محتواها تااااااااافه بالنسبة و لهالزمن	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 وااجد محتواها تااااااااافه بالنسبة و لهالزمن



	5k				10k			
	3 folds	5 folds	10 folds	Train-Validate-Test	3 folds	5 folds	10 folds	Train-Validate-Test
English	84.56	85.05	85.20	84.42	93.59	95.58	97.04	93.94
MSA	83.70	84.40	84.44	84.08	93.38	95.38	97.01	94.14
BDs	81.45	82.46	82.66	82.85	92.62	94.87	96.72	93.48

Fig. 2. A comparison of LSTM Test Accuracy Using 5k and 10k Reviews ,Cross-Validation Split, Train-Validate- Test split, Lr= 0.01

V. Conclusion and Future Work

In this study, the aim was to investigate the effect of the random swap data augmentation technique on the performance of the LSTM model in classifying translation-based datasets.

This study has shown that the random swap augmentation technique boosted the translated data and model performance. On the other hand, it indirectly reveals the authenticity of the resulting datasets from the translation approach.

This study has been one of the first attempts to examine the data augmentation technique on the LSTM model in classifying a parallel dataset of English, MSA, and BDs.

Regarding research methods, some limitations must be acknowledged, such as the limited resources that restrict the applied augmentation method in the random swap only. It would be interesting to repeat the experiments here using synonym replacement that necessitates the creation of the BD thesaurus.

References

[1] L. Zhang, S. Wang and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 8, *(4),* pp. e1253, 2018.

[2] M. Munezero *et al*, "Exploiting sentiment analysis to track emotions in students' learning diaries," in *Proceedings of the 13th Koli Calling International Conference on Computing Education Research,* 2013, .

[3] A. B. Pawar, M. A. Jawale and D. N. Kyatanavar, "Fundamentals of sentiment analysis: Concepts and methodology," in *Sentiment Analysis and Ontology Engineering*Anonymous 2016, .

[4] I. Guellil, F. Azouaou and M. Mendoza, "Arabic sentiment analysis: studies, resources, and tools," *Social Network Analysis and Mining,* vol. 9, *(1),* pp. 1-17, 2019.

[5] M. Bayer, M. Kaufhold and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys,* 2021.

[6] F. M. Luque, "Atalaya at TASS 2019: Data augmentation and robust embeddings for sentiment analysis," *arXiv Preprint arXiv:1909.11241,* 2019.

[7] X. Sun and J. He, "A novel approach to generate a large scale of supervised data for short text sentiment analysis," *Multimedia Tools Appl,* vol. 79, *(9),* pp. 5439-5459, 2020.

[8] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv Preprint arXiv:1901.11196,* 2019.

[9] X. Cui, V. Goel and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, *(9),* pp. 1469-1477, 2015.

[10] T. Ko *et al*, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association,* 2015, .

[11] P. Y. Simard *et al*, "Transformation invariance in pattern recognition—tangent distance and tangent propagation," in *Neural Networks: Tricks of the Trade*Anonymous 1998, .

[12] T. Omran, B. T. Sharef and C. Grosan, "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, 2021.

[13] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," *arXiv Preprint arXiv:1805.06201,* 2018.

[14] J. P. R. Sharami, P. A. Sarabestani and S. A. Mirroshandel, "Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus," *arXiv Preprint arXiv:2004.05328,* 2020.

[15] H. Chen and Y. Ji, "Improving the explainability of neural sentiment classifiers via data augmentation," *arXiv Preprint arXiv:1909.04225,* 2019.

[16] A. Mohammed and R. Kora, "Deep learning approaches for Arabic sentiment analysis," *Social Network Analysis and Mining,* vol. 9, *(1),* pp. 1-12, 2019.

[17] Makcedward, "Nlpaug/quick_example.ipynb at master · Makcedward/NLPAUG," GitHub, 19-Sep-2020. [Online]. Available: https://github.com/makcedward/nlpaug/blob/master/example/quick_example.ipynb. [Accessed: 06-Oct-2022].