Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Original Article

Data Augmentation for Sorani Kurdish News Headline Classification Using Back-Translation and Deep Learning Model

Soran Badawi [a*]

[a]Language Center, Charmo University, KRG, Chamchamal, Kurdistan, Iraq

Submitted: 23 May2023
Revised: 20 June2023
Accepted: 29 June 2023

*** Corresponding Author:**
soran.sedeeq@charmouniversity.

Keywords: Data augmentation, Deep learning, Machine learning, Kurdish language, Text classification.

How to cite this paper: S. Badawi, "Data Augmentation For Sorani Kurdish News Headline Classification Using Back-Translation And Deep Learning Model", *KJAR*, vol. 8, no. 1, pp. 27–34, Jun. 2023, doi: 10.24017/science/2023.1.4

Abstract: With the increased volume of news articles and headlines being generated, it is becoming more difficult for individuals to keep up with the latest developments and find relevant news articles in the Kurdish language. To address this issue, this paper proposes a novel data augmentation approach for improving the performance of Sorani Kurdish news headline classification using back-translation and a proposed deep learning Bidirectional Long Short-Term Memory (BiLSTM) model. The approach involves generating synthetic training data by translating Sorani Kurdish headlines into a target language in this context, English, and back-translating them into the Kurdish language, resulting in an augmented dataset. The proposed BiLSTM model is trained on the augmented data and compared with baseline models support vector machines (SVM) and Naïve Bayes and trained on the original data. The experimental results demonstrate that the proposed BiLSTM model outperforms the baseline and other existing models, such as SVM and Naïve Bayes, achieving state-of-the-art performance on the Sorani Kurdish news headline classification task by scoring %94 for the Kurdish documents classification-4007 dataset and %89 for the Kurdish news dataset headlines dataset, using five folds method which +2 higher than non-augmented texts. The findings suggest that the combination of back-translation and a proposed BiLSTM model is a promising approach for data augmentation in less-resourced languages, contributing to the advancement of natural language processing in under-resourced languages. Moreover, having a Sorani Kurdish news headline classification model can improve Sorani Kurdish speakers' access to news and information. With the classification model, they can easily and quickly search for news articles that interest them based on their preferred categories, such as politics, sports, or entertainment.

1. Introduction

In recent years, digital information has been shared significantly more through social media platforms [1]. Over time, there has been rapid development in Natural Language Processing (NLP). However, it is crucial to note that the accuracy of NLP models heavily relies on the quality and quantity of data utilized for training. However, collecting sufficient labelled data for under-resource languages is challenging due to the need for more resources and funding[2]. Augmenting data refers to creating synthetic data based on existing data. The synthetic data is typically designed to include minor variations from the original dataset while requiring consistent predictions. Synthetic data can represent distant examples and combinations that are extremely difficult to infer the class of a piece of text without them [3]. Studies have shown that deep neural networks are highly effective when trained with Data Augmentation since data augmentation is most used to prevent overfitting [4,5]. The deep neural network, without regularization or augmentation, is more prone to learning spurious correlations and memorizing patterns that are hard to detect by humans [6]. Overfitting can be mitigated by shuffling specific forms of language with data augmentation. For the model to perform well with noisy data, it does so by learning abstractions of information with a higher chance of generalization [7].

The Kurdish language is one of these low-resource languages, spoken by approximately 30 million people worldwide, mainly in Iran, Iraq, Turkey, and Syria [8]. Despite the increasing interest in Kurdish NLP, the limited availability of labelled datasets hinders the development of accurate and reliable models for text classification tasks such as news headline classification.

This paper proposes an approach that chains the use of back-translation and deep learning models to improve the performance of Kurdish news headline classification. The evaluation of the proposed approach involves comparing it with a baseline model trained on the original data. Further, the study propose a novel Bidirectional Long Short-Term Memory (BiLSTM) architecture to enhance the performance of the deep learning model. The proposed approach can contribute to the advancement of NLP in low-resource languages such as Kurdish and can be extended to other similar tasks in under-resourced languages. Developing a robust news headline classification system for Kurdish is crucial for improving information access and promoting the Kurdish language and culture. It will help to improve the dissemination of Kurdish news and enable individuals to stay informed about current events in their native language. It is worth noting that limited research exists on data augmentation for less-resourced languages, such as Kurdish. This paper aims to bridge this gap by showcasing various data augmentation methods and utilizing Kurdish corpora.

The remainder of this paper is structured this way. Section two presents the related works. Section three provides information about the datasets utilized in this study and explains the data augmenting methods and the architecture of the proposed method. Section four provides the results obtained by the proposed text classification model and the machine learning approaches. In section five, I discuss the results achieved from each method. The last section of this paper will be the conclusion.

2. Related work

Modern computer vision uses data augmentation extensively to enhance models' performance. By incorporating data augmentation, models are exposed to a broader range of diverse examples, enabling them to learn and adapt to various non-semantic variations in the input. Contrary to this, data augmentation in NLP is less prevalent. However, previous studies have explored and presented several approaches to tackle this limitation and promote the effective use of data augmentation in NLP tasks. A popular NLP strategy for text modification is Easy Data Augmentation (EDA), introduced by Wei *et al.* EDA is a set of straightforward and widely applicable techniques that include synonym replacement, random insertion, random swapping, and random deletion. However, this method has certain limitations. It often introduces incorrect sentence grammar or alters the text's original meaning [9].

Moreover, data augmentation has been used in various NLP tasks to address the challenge of limited-labelled data. In the field of machine translation, one study used French to English translations [10]. They employed data augmentation techniques that involved translating context and passage pairs to and from another language, which served as a means of paraphrasing the questions and contexts. Furthermore, several studies have investigated operations on input sequences based on words. For instance, Fadaee*et al.*[11] proposed a novel data augmentation technique that addresses low-frequency words by creating new sentence pairs that include infrequent terms in artificially generated contexts. Through experiments in simulated low-resource settings, their approach demonstrates improved translation quality [11]. In a study by Sennrich*et al.*[12], data augmentation was used to improve text classification performance for low-resource languages. They used back-translation and synonym replacement as data augmentation techniques and achieved significant improvements in accuracy on low-resource datasets [12]. Sabty*et al.*[13] implemented data augmentation techniques for Named Entity Recognition (NER) in Code-Switching (CS) data. They proposed several practical and easily implementable data augmentation techniques to enhance Arabic NER performance, particularly on CS data. These techniques include word embedding substitution, a modified version of the EDA technique, and back-translation [13]. In addition to this, Alsayady*et al.*[14] the application of data augmentation techniques in the context of Arabic automatic speech recognition (ASR) using deep end-to-end learning.

They apply data augmentation on the original corpus by incorporating noise adaptation, pitch-shifting, and speed transformation techniques. The experimental results demonstrated that data augmentation improves automatic speech recognition in Arabic [14]. In the scope of the Persian language, Nasiri *et al.*[15] experimented with data augmentation techniques and the ParsBERT pre-training model to address the challenge of limited and insufficient annotated datasets in the Persian stance detection task. The findings of their study demonstrate that by utilizing data augmentation methods, content-based representation of the data, and the ParsBERT model, they achieved superior performance in identifying the stance of news towards specific claims compared to previous approaches [15]. Finally, Nazarizadeh and Sayyadpour[16] employed a novel approach combining Group Deep Learning and Data Augmentation to classify emotions in Persian. By applying data augmentation and utilizing the group deep learning approach for classification, they achieved an impressive accuracy of 96.5% in Persian sentiment analysis [16].

Kurdish is a low-resource language, and as such, there have been limited studies on text classification for the Kurdish language. However, some previous studies have explored various approaches to tackle the challenges of performing text classification in the Kurdish language. Awlla and Veisi [17] proposed a deep learning approach for analyzing Kurdish text. By manually labeling Kurdish texts, they obtained an accuracy of 67% employing Word2Vec and Long Short-Term Memory (LSTM). Unfortunately, the sample size of the study was small [17]. Moreover, Badawi [18]proposed deep learning-based Bidirectional Encoder Representations from Transformers (BERT) to perform medical text classification and achieved an F1-score of 92% [18].

3. Materials and Methods

3.1. *Datasets*

The selection of an appropriate corpus is crucial for the text classification of low-resourced languages. Undoubtedly. A dataset's application domain will invariably significantly impact the final prediction. In this study, a total of two datasets were used. Kurdish documents classification-4007 (KDC-4007)[19] is the first dataset. The simplicity and well-documentation of this dataset are its most impressive features. The dataset contains Kurdish texts written in the Sorani dialect and can be used for various text classification studies. A total of 4,007 text files were included in the dataset, and they were divided into eight categories: sports, religious, arts, economics, education, social, styles, and health. The corpus comprised 4,007 texts divided into 500 categories, as shown in figure 1. The dataset has 24,817 unique words [19].

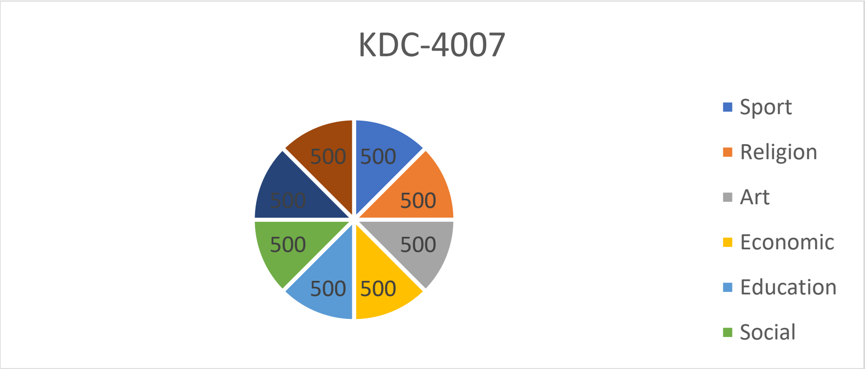

Figure 1: The number of text samples in KDC-4007.

Kurdish news dataset headlines (KNDH)is the second dataset. Various Kurdish news websites' headlines are gathered in this dataset. The dataset used in this study includes 50,000 news headlines, with an equal distribution of 10,000 headlines among five categories: social, sport, health, economic, and technology, as illustrated in figure 2. Although the number of samples is equal for each category, the percentage ratio of the channels used to collect the headlines is different. A total of 34 different channels were utilized to gather the headlines for each category, including 8 channels for economics, 14 for health, 18 for science, 15 for social, and 5 for sport [20] as shown in figure 2.

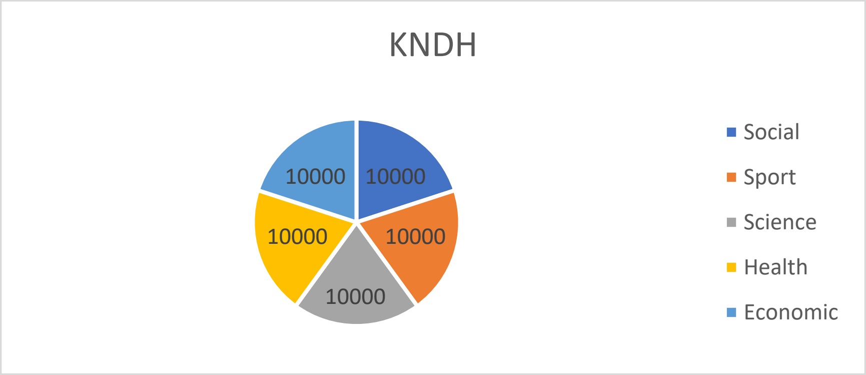

Figure 2: The number of text samples in KNDH

3.2. *Data Preprocessing*

Whenever a task involves NLP, preprocessing is the first step. This study carried out the preprocessing process in five steps on each text inside KDC-4007 dataset since the KDC-4007 dataset contains only raw texts and preprocessing was not performed on it, as shown in figure 3. For the first and second steps, I used a toolkit called kurdishlanguage processing toolkit for standardizing and normalizing Kurdish text [21]. This step involved removing white-space space and correcting misspelled words. Next, I removed punctuation, single characters, ineffective digits, and single characters from the sentence. I finished preprocessing by cleaning the datasets from stop-words as shown in table 1.

Table 1: Examples of preprocessing steps.

Steps	Raw Text	English Translation of Raw Text	Clean Text	Translation of Clean Texts
Normalization	فلیمەنوێیەکە٦خەڵاتی بەدەستهێنا	The new movie won six trophies	خەڵات6فلیمەنوێیەکە یبەدەستهێنا	The new movie won six trophies
Standardization	ئە مرو روزیکی خوشە	Today is a good day	ئەمڕۆ ڕۆژێکی خۆشە	Today is a good day
Unwanted Character Removal	فیلمەکە، (The Birth Of A Nation) پارەیەکی زۆری تێچوو!	The movie (The Birth Of A Nation) cost a lot of money!	فیلمەکەپارەیەکی زۆری تێچوو	The movie cost a lot of money
Stop-Word Removal	ریکلامەکەی لەیوتیوب لەملیۆنێک تەماشاکاری هەبووە	His ad had a million hits on YouTube	ریکلامیوتیوب ملیۆنێک تەماشاکاری هەبووە	Ad had a million hits on YouTube

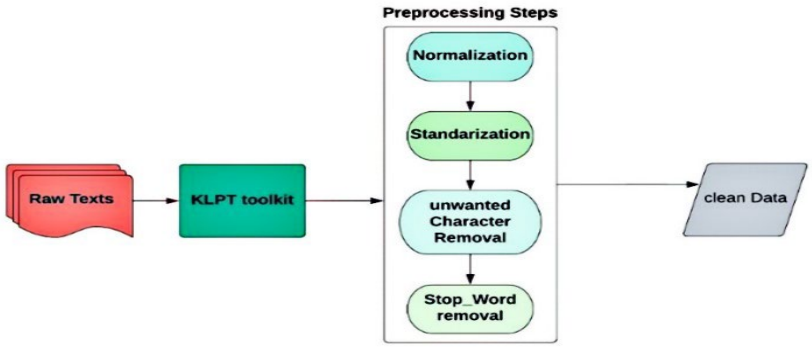

Figure 3: The stages of preprocessing

3.3. *Data Augmentation*

In back-translation, text is translated from one language into another, then translated back to the original language. This approach leverages the semantic invariances observed in supervised translation datasets to generate semantic invariances for augmentation purposes. Furthermore, back-translation ensures consistency in the back-translations in order to train unsupervised translation models [3]. In this approach, data is enhanced through data noising. Notably, text data augmentation has limitations with regard to grammatical structure, vocabulary, and pragmatic meanings. The order and structure of words are rarely vital since I am looking for multiple classes. Thus, examining common words used in particular polarities is necessary. Additionally, translations from dictionaries like Google Translator (GT) can be used to generate labels. GT is used to translate each sentence in the training data into a middle language, in this case English. Subsequently, the original Kurdish text is retranslated [11]. During this process, some words may change to synonyms, some words may change positions, and other unpredictable changes can occur, but the overall polarity and sentiment remain unaltered. Implementing this approach can result in a two-fold increase in the provided datasets. An example of translation data augmentation is shown in figure 4.

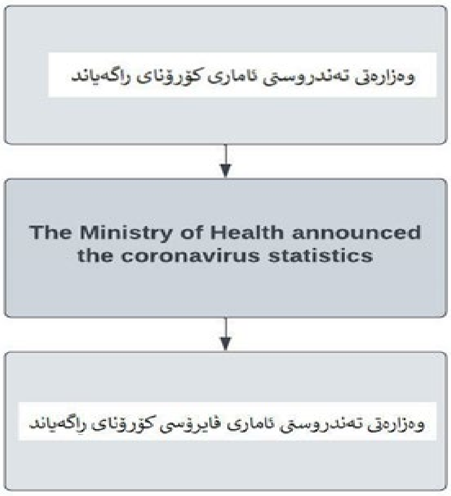
وەزارەتی تەندروستی ئاماری کۆرۆنای راگەیاند
The Ministry of Health announced the coronavirus statistics
وەزارەتی تەندروستی ئاماری ڤایرۆسی کۆرۆنای ڕاگەیاند

Figure 4: Back translation of a text sample in the datasets

3.4. *Baseline Classifiers*

Popular baseline patterns generally evaluate new models. This study employs two machine-learning algorithms. These include a naive Bayes classifier and a support vector machine [22]. In the baseline support vector machine (SVM) model, text data is classified into predefined categories by using a linear kernel to implement a SVM classifier [23]. The term frequency-inverse document frequency (TF-IDF) is used to transform the text data into a feature matrix. For each term in the text corpus, the vectorizer produces an inverse document frequency value TF-IDF. TF-IDF measures the importance of words within a corpus. Naive Bayes is a probabilistic classification algorithm that leverages Bayes' theorem. It is referred to as "naive" because it assumes feature independence, which is frequently unrealistic in real-world scenarios.

3.5. *Bilstm Architecture*

Recent years have seen an increase in the utilization of deep learning for natural language processing [24], as they do not require the assistance of humans [25]. As well as supervised methods, they are also capable of classifying datasets in unsupervised methods. In BiLSTM architecture, BiLSTM layers provide multidimensional long short-term memory processing of sequential data such as time series or text data. It consists of two Long-Short-Term Memory (LSTM) layers that process the input sequence forwards and backwards respectively, and concatenate their output at each time step. By capturing both the past and the future context of an input sequence, the network is able to perform tasks such as sentiment analysis, named entity recognition, and machine translation more effectively.LSTM is a deep learning architecture because it contains a storage unit for recent data [26].

Recurrent neural networks are employed in the proposed deep learning architecture. These networks are called Bidirectional LSTM. The proposed model consists of six additional layers in addition to the two initial layers' input and embedding the layers. First, The NVIDIA CuDNN-backed Bidirectional CuDNNLSTM implementation offers fast LSTM implementation utilizing GPUs (with TensorFlow as the backend). Second, in max-pooling 1D, all input channels are computed the maximum for default input dimensional tensor, where the input channels and strides are one. Thirdly, the dropout regularization [27] prevents overfitting by randomly dropping 20% of the input. Fourth, there is a dense layer of neural connections that are regular and deeply connected. I was able to define Y, W, and b a output, weight, and bias. Next, I apply the following equation to the input (x):

$$y = \sum_{i=1}^{n} x_i w_i + b \tag{1}$$

As an activation function, I use the ReLU [28] Hence, the final output would be:

$$y = \sum_{i=1}^{n} x_i w_i + b \tag{2}$$

A total of 1000 neurons are included in the model. Fifth, the dropout layer has been re-created with the same configuration as before. Finally, the activation function SoftMax is introduced in a new dense layer.Click or tap here to enter text.The design of the proposed BILSTM model is shown in figure 5.

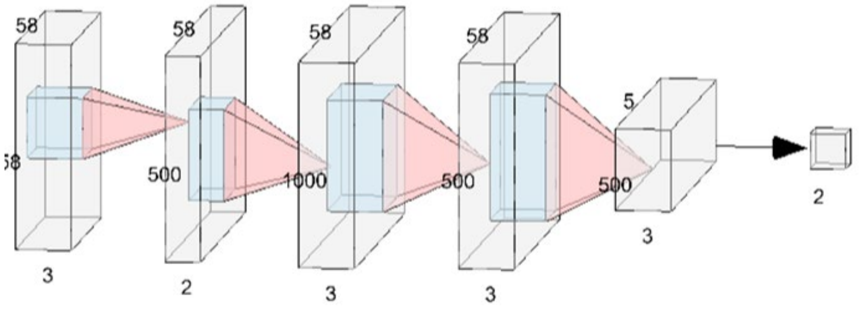

Figure 5: Proposed BiLSTM architecture.

4. Results

In the experiment section, the proposed BiLSTM, Naive Bayes and SVM classifiers were evaluated on two Kurdish news headline datasets: KDC-4007 and KNDH. For the proposed BiLSTM classifier, I utilized the Keras deep learning framework with embedding layers, bidirectional LSTM layers, global max pooling layers, and two dense layers with dropout regularization. I trained the model using Adam optimizer with a categorical cross-entropy loss function. As for the Naive Bayes classifier, the model was trained using the scikit-learn machine learning library with a multinomial naive Bayes algorithm. For the SVM classifier, the model was trained using the scikit-learn machine learning library with a linear support vector machine algorithm. Using the fit method, the model was trained and evaluated using the score method.

Five fold cross-validation was used, which split the data into five equal parts, with each fold being used once as the test set and the remaining four folds used as the training set. The performance of the classifiers was evaluated using the F1-score. The results were reported in tables 2, 3 and 4, where the F1-score of each classifier was calculated for each fold and dataset separately.

Finally, the results were analyzed and compared to determine the best classifier for Kurdish news headline classification. For the baseline classifiers. The results are shown in tables 2,3 and 4.

Table 1: F1-score results for BiLSTM

Title	K=1		K=2		K=3		K=4		K=5	
	Au*	Un	Au	Un	Au	Un	Au	Un	Au	Un
KDC-4007	93.4	91.2	93.8	92.8	93.5	93.8	94	92.5	93.8	92.4
KNDH	88.4	87.3	85.6	88.3	86.8	86.4	86.9	86.7	89	87.3

*Aug: refers to augmented texts: Un: refers to unaugmented texts

Table 3: F1-score results for Naïve Bayes

Title	K=1		K=2		K=3		K=4		K=5	
	Au	Un	Au	Un	Au	Un	Au	Un	Au	Un
KDC-4007	91.2	90.4	91.5	91.5	91.7	93.4	92.1	92.6	91.8	92.8
KNDH	85.3	86	85.6	86.3	85.7	86.8	89.3	88.1	86	87.4

Table 4: F1-score results for SVM

Title	K=1		K=2		K=3		K=4		K=5	
	Au	Un	Au	Un	Au	Un	Au	Un	Au	Un
KDC-4007	92.2	90.1	92.4	91.6	92.5	91.1	92.8	92.3	92.7	92.6
KNDH	86.3	85.1	86.6	86.1	86.7	86.2	86.9	85.4	87	86.6

5. Discussion

In the given tables, the results of three different models, i.e., BiLSTM, Naive Bayes, and SVM, have been presented for two different datasets, i.e., KDC-4007 and KNDH for augmented and non-augmented texts. Each table shows the F1-score of the models for five different values of K, where K represents the number of folds used in the cross-validation process.

Starting with the results of the proposed BiLSTM model, it can be seen that it consistently outperforms the other two models in both datasets. For the KDC-4007 dataset, the proposed BiLSTM provides the highest F1-score of 94% at K=4, with a range of 93.4% to 94%. Similarly, for the KNDH dataset, our model achieves the highest F1-score of 89% at K=5, and the F1-score ranges from 88.4% to 89%. These results indicate that our proposed model is successful in accurately classifying the Kurdish news headlines and can outperform existing models such as Naive Bayes and SVM. I hypothesize that the SDC-4007 was preprocessed and even labelled manually. While the KNDH data is significantly higher, they were labelled and even preprocessed automatically.

Moving on to the results of the Naive Bayes model, it can be observed that it performs better than the SVM model, but it lags behind the proposed BiLSTM model. For the KDC-4007 dataset, the Naive Bayes model achieves the highest F1-score of 92.1% at K=4, and the F1-score ranges from 91.2% to 92.1%. For the KNDH dataset, the Naive Bayes model achieves the highest F1-score of 86% at K=5, and the F1-score ranges from 85.3% to 86%. Based on these results, Naive Bayes may be a good alternative to SVM, but its effectiveness may not be as high as the BiLSTM model proposed for classifying Kurdish news headlines. Finally, considering the results of the SVM model, it can be observed that it performs the worst among the three models for both datasets. For the KDC-4007 dataset, the SVM model achieves the highest F1-score of 92.8% at K=4, and the F1-score ranges from 92.2% to 92.8%. For the KNDH dataset, the SVM model achieves the highest F1-score of 87% at K=5, and the F1-score ranges from 86.3% to 87%. These results indicate that the SVM model may not be the best choice for classifying Kurdish news headlines, and the proposed BiLSTM model can provide significantly better results.

In summary, the results presented in the tables show that the proposed BiLSTM model can outperform the existing models such as Naive Bayes and SVM in accurately classifying Kurdish news headlines. The proposed model outclassed the other machine learning classifiers scoring +2 higher than

SVM and + 3 higher than Naïve Bayes for the first dataset. Similarly, the proposed BiLSTM model outperforms the baseline classifiers for the second dataset. The BiLSTM model scores are +2 higher than SVM and + 4 higher than Naïve Bayes. The results demonstrate that using augmented data significantly improved the performance by comparing raw data with the proposed model. BiLSTM trained using augmented texts, performs best with unseen texts. By using augmentation, the model is potentially more resilient to unfamiliar data owing to the variation introduced by the augmentation. It is worth noting that F1-score has increased in both datasets when the texts underwent the augmentation process for the proposed BiLStm model as shown in figures 6 and 7. Whilst the F1-score slightly increases in SVM and Naïve Bayes.

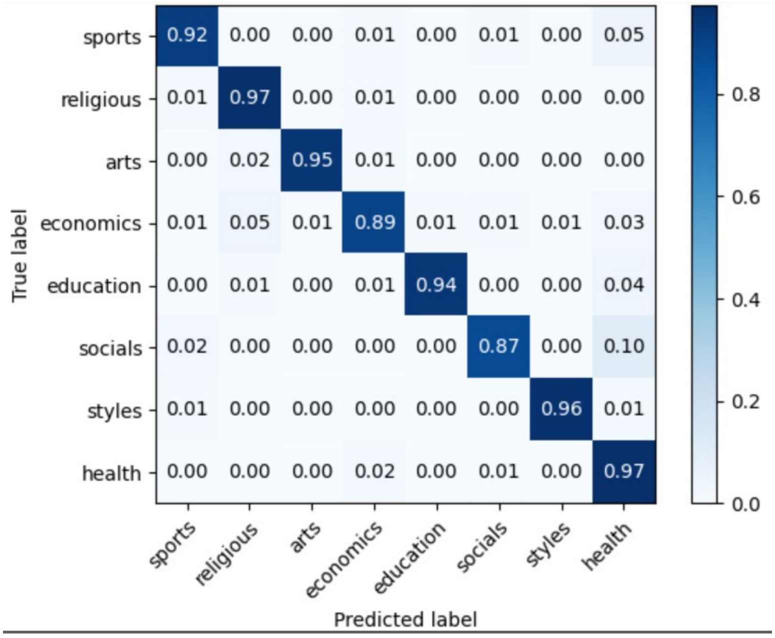

Figure 6: Confusion Matrix for the augmented texts in KDC-4007 Dataset**.**

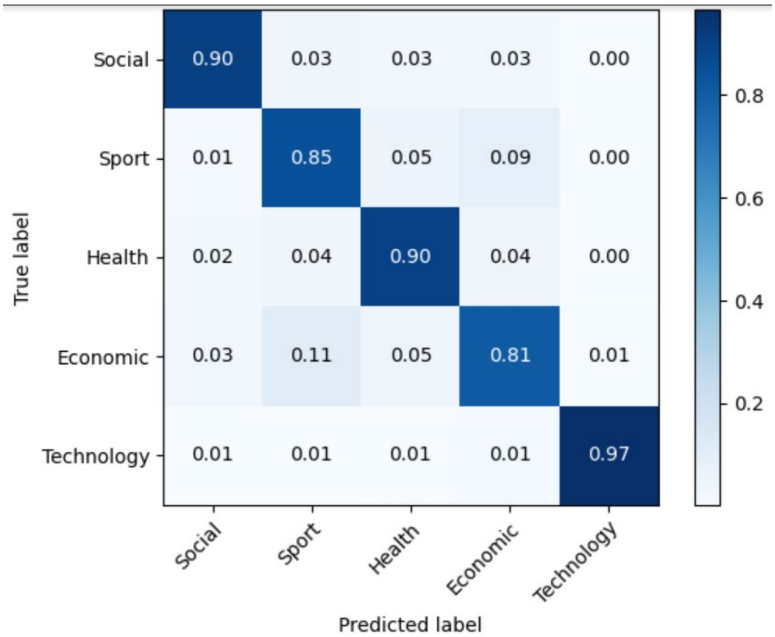

Figure 7: Confusion Matrix for the augmented texts in KNDH.

In this study, I sought to evaluate the performance of our proposed BiLSTM model in Kurdish text classification. To further validate the effectiveness of our proposed model, I compared it with the BERT-based model developed by in the previous study [18]. The same dataset was and experimental setup as in our proposed model to ensure a fair comparison between the two models. The evaluation metric used was F1-score as shown in table 4.

Table 3: F1-scores of BiLSTM and Multilingual Bert.

Dataset	Multilingual Bert-Based	Proposed BiLSTM
Kurdish Medical Dataset	0.92	0.93

The results showed that our proposed BiLSTM model outperformed the BERT-based model, with an F1 score of 93 compared to the BERT-based model's score of 92. This result indicates that our proposed BiLSTM model can effectively classify Kurdish text with high F1-score and outperforms the BERT-based model in this task. It is worth noting that while our proposed model achieved higher performance compared to the BERT-based model, there may still be room for improvement. Further experimentation and optimization may be necessary to achieve even higher performance in Kurdish text classification. Nevertheless, our proposed BiLSTM model has shown promise as an effective approach for this task.

6. Conclusions

This paper proposes augmenting the input data and then using a deep learning model on the augmented data for Kurdish text classification. I augment by applying back translation. Essentially, I was able to support the initial idea and provide a solution to a lack of annotated corpora in the field of Kurdish text classification. As a result of the augmentation, the deep learning model can efficiently generalize to unseen and new words. In this study, I proposed a BiLSTM-based model for Kurdish news headline classification and compared its performance with Naive Bayes and SVM classifiers. I also conducted experiments using k-fold cross-validation on two datasets: KDC-4007 and KNDH. The results showed that our proposed BiLSTM model outperformed the Naive Bayes and SVM classifiers on both datasets in terms of F1-score. Specifically, the BiLSTM model attained an average F1-score of 93.5% on KDC-4007 and 88.7% on KNDH, while Naive Bayes obtained an average F1-score of 91.6% and 85.6%, and SVM achieved an average F1-score of 92.5% and 86.5%, respectively. I also conducted experiments to test the robustness of our BiLSTM model using k-fold cross-validation. The results showed that the model consistently performed well across different folds on both datasets, with an average F1-score of 93.5% on KDC-4007 and 88.7% on KNDH. Overall, our proposed BiLSTM model demonstrates its effectiveness in classifying Kurdish news headlines, outperforming existing classifiers such as Naive Bayes and SVM. For future studies, I suggest the implementation of other argumentation techniques such as random insertion, random swap, and random deletion as this has been proven to be effective in other languages such as English and French. This study has important implications for the development of automated tools for analyzing Kurdish text data, and could potentially contribute to improving information management and decision-making processes in various domains.

Data availability: All authors have contributed equally.

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: The authors did not receive support from any organization for the submitted work.

Reference

[1] B. R. Chakravarthi *et al.*, "Detecting abusive comments at a fine-grained level in a low-resource language," Natural Language Processing Journal, vol. 3, p. 100006, Jun. 2023, doi: 10.1016/j.nlp.2023.100006.

[2] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," Oct. 2020, doi:10.48550/arXiv.2010.12309.

[3] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text Data Augmentation for Deep Learning," *J Big Data*, vol. 8, no. 1, p. 101, Dec. 2021, doi: 10.1186/s40537-021-00492-0.

[4] M. Varasteh and A. Kazemi, "Using ParsBert on Augmented Data for Persian News Classification," in 2021 7th International Conference on Web Research (ICWR), IEEE, May 2021, pp. 78–81. doi: 10.1109/ICWR51868.2021.9443119.

[5] J. P. R. Sharami, P. A. Sarabestani, and S. A. Mirroshandel, "DeepSentiPers: Novel Deep Learning Models Trained Over Proposed Augmented Persian Sentiment Corpus," Apr. 2020, [Online].doi:10.48550/arXiv.2004.05328.

[6] A. Karimi, L. Rossi, and A. Prati, "AEDA: An Easier Data Augmentation Technique for Text Classification," Aug. 2021,doi:10.48550/arXiv.2108.13230.

[7] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A Survey on Data Augmentation for Text Classification," ACM Comput Surv, vol. 55, no. 7, pp. 1–39, Jul. 2023, doi: 10.1145/3544558.

[8] S. Badawi, "Transformer-Based Neural Network Machine Translation Model for the Kurdish Sorani Dialect," UHD Journal of Science and Technology, vol. 7, no. 1, pp. 15–21, Jan. 2023, doi: 10.21928/uhdjst.v7n1y2023.pp15-21.

[9] D. T. Vu, G. Yu, C. Lee, and J. Kim, "Text Data Augmentation for the Korean Language," Applied Sciences, vol. 12, no. 7, p. 3425, Mar. 2022, doi: 10.3390/app12073425.

[10] A. W. Yu et al., "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension," Apr. 2018, [Online]. doi:10.48550/arXiv.1804.09541.

[11] M. Fadaee, A. Bisazza, and C. Monz, "Data Augmentation for Low-Resource Neural Machine Translation," in Proceedings of the 55th Annual Meeting of the Association forComputational Linguistics (Volume 2: Short Papers), Stroudsburg, PA, USA: Association for Computational Linguistics, May 2017, pp. 567–573. doi: 10.18653/v1/P17-2090.

[12] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2016, pp. 86–96. doi: 10.18653/v1/P16-1009.

[13] C. Sabty, I. Omar, F. Wasfalla, M. Islam, and S. Abdennadher, "Data Augmentation Techniques on Arabic Data for Named Entity Recognition," Procedia Comput Sci, vol. 189, pp. 292–299, Jul. 2021, doi: 10.1016/j.procs.2021.05.092.

[14] H. Alsayadi, A. Abdelhamid, I. Hegazy, and Z. Taha, "Data Augmentation for Arabic Speech Recognition Based on End-to-End Deep Learning," International Journal of Intelligent Computing and Information Sciences, vol. 21, no. 2, pp. 50–64, Jul. 2021, doi: 10.21608/ijicis.2021.73581.1086.

[15] H. Nasiri and M. Analoui, "Persian Stance Detection with Transfer Learning and Data Augmentation," in 2022 27th International Computer Conference, Computer Society of Iran (CSICC), IEEE, Feb. 2022, pp. 1–5. doi:10.1109/CSICC55295.2022.9780479.

[16] A. Nazarizadeh, T. Banirostam, and M. Sayyadpour, "Using Group Deep Learning and Data Augmentation in Persian Sentiment Analysis," in 2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), IEEE,pp. 1–5. Dec. 2022, ,doi: 10.1109/ICSPIS56952.2022.10044052.

[17] K. Awlla and H. Veisi, "Central Kurdish Sentiment Analysis Using Deep Learning," Journal of University of Anbar for Pure Science, vol. 16, no. 2, pp. 119–130, Dec. 2022, doi: 10.37652/juaps.2022.176501.

[18] S. S. Badawi, "Using Multilingual Bidirectional Encoder Representations from Transformers on Medical Corpus for Kurdish Text Classification," ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY, vol. 11, no. 1, pp. 10–15, Jan. 2023, doi: 10.14500/aro.11088.

[19] T. A. Rashid, A. M. Mustafa, and A. M. Saeed, "Automatic Kurdish Text Classification Using KDC 4007 Dataset," in International Conference on Emerging Intelligent Data and Web Technologies, May 2017, doi:10.1007/978-3-319-59463-719.

[20] S. Badawi, A. M. Saeed, S. A. Ahmed, P. A. Abdalla, and D. A. Hassan, "Kurdish News Dataset Headlines (KNDH) through multiclass classification," Data Brief, vol. 48, p. 109120, Jun. 2023, doi: 10.1016/j.dib.2023.109120.

[21] S. Ahmadi, "KLPT – Kurdish Language Processing Toolkit," in NLPOSS, Nov. 2020, doi:10.18653/v1/2020.nlposs-1.11.

[22] Y.-M. Li and T.-Y. Li, "Deriving market intelligence from microblogs," Decis Support Syst, vol. 55, no. 1, pp. 206–217, Apr. 2013, doi: 10.1016/j.dss.2013.01.023.

[23] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," Natural Language Processing Journal, vol. 2, p. 100003, Mar. 2023, doi: 10.1016/j.nlp.2022.100003.

[24] R. Collobert, J. Weston, J. Com, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," 2011.doi: 10.48550/arXiv.1103.0398

[25] P. Vateekul and T. Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," in 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, pp. 1–6, Jul. 2016, doi: 10.1109/JCSSE.2016.7748849.

[26] G. Liu, X. Huang, X. Liu, and A. Yang, "A Novel Aspect-based Sentiment Analysis Network Model Based on Multilingual Hierarchy in Online Social Network," Comput J, vol. 63, no. 3, pp. 410–424, Mar. 2020, doi: 10.1093/comjnl/bxz031.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskeverand R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929-1958, 2014. [Online]. Available: https://dl.acm.org/doi/10.5555/2627435.2670313

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks,"Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings,pp. 315-323, Jun. 2011