Using Multilingual Bidirectional Encoder Representations from Transformers on Medical Corpus for Kurdish Text Classification

Soran S. Badawi

Charmo Center for Scientific Research and Consulting – Language and Linguistic Center, Charmo University
Chamchamal, Sulaimani, Kurdistan region - F.R. Iraq

***Abstract*—Technology has dominated a huge part of human life. Furthermore, technology users use language continuously to express feelings and sentiments about things. The science behind identifying human attitudes toward a particular product, service, or topic is one of the most active fields of research, and it is called sentiment analysis. While the English language is making real progress in sentiment analysis daily, other less-resourced languages, such as Kurdish, still suffer from fundamental issues and challenges in Natural Language Processing (NLP). This paper experiments with the recently published medical corpus using the classical machine learning method and the latest deep learning tool in NLP and Bidirectional Encoder Representations from Transformers (BERT). We evaluated the findings of both machine learning and deep learning. The outcome indicates that BERT outperforms all the machine learning classifiers by scoring (92%) in accuracy, which is by two points higher than machine learning classifiers.**

***Index Terms*—Bidirectional Encoder Representations from Transformers, Deep learning, Machine learning, Natural language processing, Sentiment analysis, Transformers.**

I. Introduction

The text classification method in natural language processing (NLP) is one of the approaches of identifying the emotions in text. The field has gained more popularity since the emergence of social platforms such as Twitter and Facebook (Hoang, Bihorac, and Rouces, 2019). It has been tackled very well in the English Language. Conversely, the work done in the Kurdish language remains in its infancy; thus, more cooperation and contributions are required from research communities to offer a mature sentiment analysis system in Kurdish. The previous Kurdish sentiment analysis works mostly centered on classical machine learning classifiers. In general, these classical methods are considered to be super-fast and simple. Due to feature engineering, their performance firmly hangs on the feature selection before training.

Later on, deep learning was developed as a popular alternative to traditional machine learning methods because of its excellence in NLP tasks like text classifications (Collobert, et al., 2011). The main idea of deep learning algorithms is the automated extraction of representations from data (LeCun, et al., 2015).

The previous methods, for instance, a bag-of-words (BOW) and a Term Frequency Inverse Document Frequency (TF-IDF) approach, statistically represent word frequency in documents. Therefore, they could not recognize the relationship between different keywords in a document as inherently statistical methods. Consequently, Word Embedding methods emerged to help solve this issue by representing words as mathematical vectors in a multidimensional space. Usually, these vectors provide vital information about the associations between words. Numerous studies on word embedding proved that pre-trained word embedding models, such as word2vec (Mikolov, et al., 2013) and GloVe (Pennington, Socher, and Manning, 2014) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin, 2018), can significantly enhance text classification and other NLP tasks.

BERT is based on a multi-layer bidirectional transformer (Vaswani, et al., 2017). BERT is pre-trained on a large corpus of multilingual data in a self-supervised pattern, which implies that it was only pre-trained on the raw text independence of humans, labeling them in any way. It uses an automatic process to generate inputs and labels from those texts. Google Search Team has pre-trained BERT with 12 layers and 768 hidden dimensions per single token. Bert's total parameter equals 110 million parameters (Devlin, 2018).

In this study, we use BERT to classify Kurdish texts. We begin by comparing the BERT's performance with traditional machine learning methods that have been extensively utilized in earlier publications. The rest of this paper is structured in this way. We examine the literature on classifying Kurdish texts in the next part. Then, we apply a customized BERT-Multilingual model to the medical corpus and compare the outcomes with text categorization methods based on machine learning. The conclusion of this study will be included in the final section.

ARO-The Scientific Journal of Koya University
Vol. XI, No. 1 (2023), Article ID: ARO.11088. 6 pages
Doi: 10.14500/aro.11088
Received: 10 October 2022; Accepted: 29 December 2022
Regular research paper: Published: 15 January 2023.
Corresponding author's e-mail: Soran.sedeeq@charmouniversity.org

II. Related Works

Kurdish language has more than 30 million speakers around the globe and is categorized as one of the less-resourced languages, particularly in the field of NLP (Esmaili, 2012). Unlike English, mountainous works have been done in different areas of NLP; the sentiment analysis process in Kurdish is still in its early stages. So far, only one research study has been carried out in this direction. Two Kurdish researchers, Salam Abdulla and Mzhda Hiwa Hama, carried out the work. Their work entitled "Sentiment Analyses for Kurdish Social Network Texts using Naive Bayes Classifier' (2015)". Their data contained 15k tweets containing positive and negative labels, distributed half for each tag. The result was achieved using Naïve Bayes (0.66) (Abdulla and Hama, 2015). In addition to this, the corpus is not available online. Moreover, their corpus was not trained in deep learning tools.

III. BERT Architecture

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a transformers model pre-trained on a large corpus of multilingual data in a self-supervised fashion, which means that it was pre-trained on the raw texts only, independence of humans labeling them in any way with an automatic process to generate inputs, and labels from those texts. Google Search Team has pre-trained BERT with 12 layers and 768 hidden dimensions per single token. Bert's total parameter equals 110 million parameters (Devlin, 2018). The architecture of the model is displayed in Fig. 1.

Moreover, BERT requires its input token sequence to have a specific format. The first token of every sequence should be assigned as (CLS) (classification token), and there should be a (SEP) token (separation token) after every sentence to achieve the same format (Ling, 2020). To construct an input representation for a token, a sum is applied to its token, segment, and position embeddings. An illustration of this construction is shown in Fig. 2.

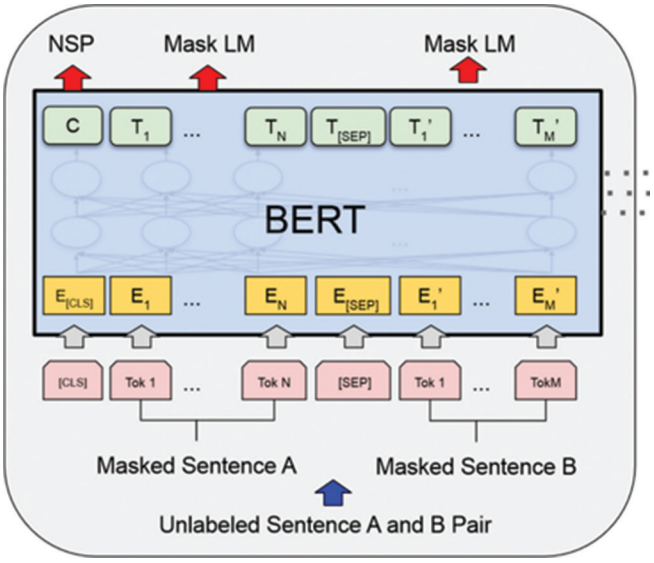

Fig. 1. The architecture of the model.

BERT is widely implemented in text classification for other resourceful languages. Since it gives high accuracy compared to traditional machine learning algorithms due to having many layers, as shown in Fig. 3.

As illustrated in Fig. 3, the first is called Input, the input layer. This layer accepts the initial word embedding and delivers it into BERT. The second part is *BERT* which is the pre-trained BERT model. The output of this part is the final word embedding of each input token. The last part is predict. In this part, the hidden representation is passed to a dense layer followed by $softmax$. The $softmax$ is applied along the dimension of the sequence. Mathematically, the $softmax$ function takes input as a vector of K real numbers and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers. The output is the probabilities of each label (Ling, 2020)

IV. Methodology

The dataset used in this paper contains social media comments written in the Kurdish language. Because many users used Arabic-supported keyboards built-in on their mobile devices when commenting on videos, posts, and pictures. For example, in the Arabic letter "ۆ" is absent. If users intend to type words that have the letter "ۆ," they write "خوش" instead of "خۆش." This dictation error causes the machine learning classifiers, BOW and Bert, which use the tokenization process to understand those two words when they are a single word.

Pre-processing is an essential phase of machine learning and deep learning. The pre-processing helps the classifiers to provide better results. From a morphological perspective, Kurdish is a language with numerous attachments, such as Arabic and Persian. Knowing the extension might give information about the pronouns, the plurality, and the location pre-positionally (Cieliebak, et al., 2017). Thus, pre-processing would be challenging since the language has progressed in NLP. Luckily, we could use the python KLPT toolkit developed by Ahmadi (2020). The libraries on KLPT helped us with normalization, standardization, and tokenization. It is essential to know that there are no special libraries to point out stop-words in the Kurdish language. We created the stop-word lists and implemented them on the corpus. The dataset includes a collection of raw comments from Kurdish social media users. The initial cleaning of the dataset included the removal of URLs, not-Kurdish alphabetical, emojis, and numbers. However, there are no mentions of using software or a library to accomplish this. We used the KLPT library to perform the following process;

1. Normalization for unifying dialects and scripts based on different encodings
   For example
   Unnormalized texts - " دکتور کیان ٤٥ روزه نەشتە رکە ریم کردوە"
   Normalized text "دکتور کیان 45 روزه نەشتەرکەریم کردوە"
2. Standardization – When given a normalized text, it returns standardized Kurdish text according to Sorani's recommendations.

http://dx.doi.org/10.14500/aro.11088

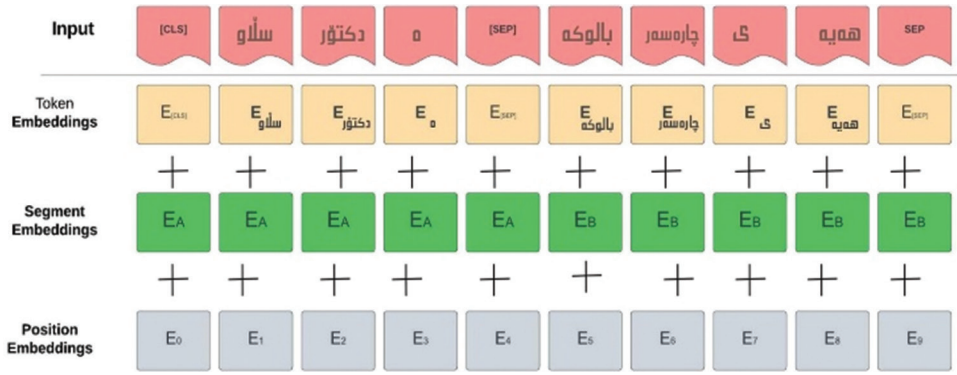

Fig. 2. Input representation of BERT.

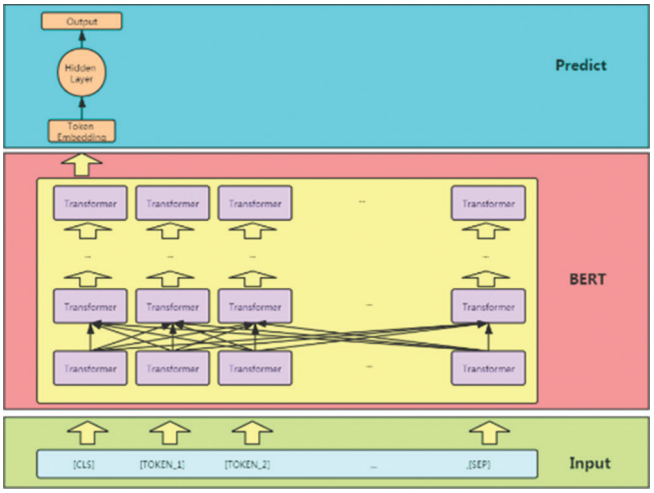

Fig. 3. BERT layers for text classification.

For instance

Unstandardized text - "دکتور کیان 45 روزە نەشتەرکەریم کردوە"

Standardized text- "دکتور گیان 45 رۆژە نەشتەرگەریم کردووە"

3. Tokenization refers to slicing sentences into words and putting them in a list.

   For example, the tokenization of the example above is ['دکتور', '_ـگیان_', '45', 'رۆژە', 'نەشتەرگەریم', 'کردووە']

Luckily, the library's documentation would help us construct a scientific set of vocabulary, which would be crucial to the machine learning classifiers when using BOW and for Bert to find the tokens of each language in the pre-trained model.

We implement HuggingFace, a BERT-multilingual tokenizer, and its model on the input data. Our data are tokenized before being used for training by fine-tuning the BERT model. The process of fine-tuning Bert begins with stacking multilingual-BERT with five multilingual-BERT layers. We add a dense layer with a softmax activation function in the next step. A binary cross-entropy loss function was also used to minimize the errors while training our models (Zahera, et al., 2019). Compared to the original BERT, the fine-tuned model requires much less time to train. Furthermore, fine-tuning BERT assists us with training a model to good performance on a much smaller amount of training data. Finally, this simple fine-tuning procedure (typically adding one fully connected layer on top of BERT and training for a few epochs) was shown to achieve a state of the art results with minimal task-specific adjustments for a wide variety of tasks: Classification, language inference, semantic similarity, question answering, etc. Rather than implementing custom and sometimes obscure architecture shown to work well on a specific task, simply fine-tuning BERT is shown to be a better (or at least equal) alternative. We reduce the max length to 128 for the BERT tokenizer, the batch size to 8, as shown in Fig. 4, and the training epochs to 3.

We experiment with decision tree and support vector machine (SVM) classifiers, multinomial stochastic gradient descent (SGD), k-nearest neighbor (kNN), SVM, Random Forest, and Logistic Regression. We compare their results with BERT, as shown in Table I. We also use a Count Vectorizer with a mixture of unigrams, bi-grams, and tri-gram representations of words for our machine-learning methods, because of the nature of our corpus. Usually, medical texts contain many keywords with low frequency (e.g., a disease's name or a medicine's name). The disease or medicine names have been transliterated into Kurdish language, as displayed in Fig. 5. These words contribute heavily to the classification since they are often less frequent; their importance would be lost in a BOW approach because of their low frequency, which only counts the word frequency in documents. Furthermore, we notice that the majority of medical texts include words such as "دکتور، عیادە، ڤیتامین، مرهەم، لیزەر، فیلەر، پیست".

Data-splitting is an essential step. The method used for splitting the data predominantly affects the model's outcome. Since our dataset is not very big, we used holdout to split our data. In the first phase, we used (the 80% train-20% test) technique. Moreover, we split the trainset using the Holdout technique to create a validation set. Having a validation set is vital, particularly in the case of deep learning.

V. Results and Discussion

This work's database comprises 6756 samples distributed between two labels (medical and non-medical). The number of medical texts equals 3076, while the number of non-medical texts is 3680 (Saeed, 2022), , as shown in Fig. 6.

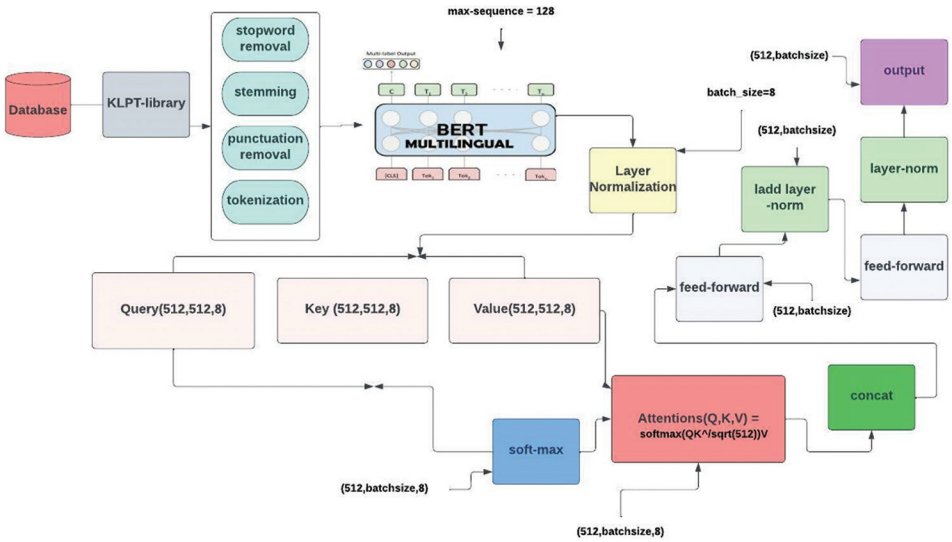

Fig. 4. The design of the full fine-tuned (proposed) model.

من سالیک عەمەلایات کراوم لەکاتەوە بیستم هیشک دەبیت	medical
سلاو من دە مو جاو م خالی قاوەی یە جی باشە بیزە حمە ت	medical
سە رکە وتوبیت	notmedical
رجیم لە رە مە زان زە رە ر نینە	medical
سڵاو دکتۆرگیان منبرجم زۆر دەوەرێ	medical
دەستت خۆش بێت دکتۆرە	medical
دکتۆر من بیستم زور جورە	medical
من لیوم زۆر ووشک ابیت بە لام لیتر او اخۆمە وە رۆزانە ج بکە م دایم جە وری اکە م سودی نیبە	medical
سلاو دکتورە بی زەحمت مرهمی اکزیمای دەست جی باشە	medical
ئەی دەست خۆش سوپاس	notmedical
چارەسەری گوشتی پشتی چاو بەچی ئە کریت پێلوی چاو	medical
سڵاو دکتور سەرم توک دەوەرێنێ	medical
رای بە رێزت بۆ بۆتوکس و فیلە ر زیانی زۆرترت یاخود سود	medical
بیستی دم جاوم سورە زو دە سوتی وشکیش جیبکە م تە مە نم سال	medical
وشک بونی بیست دەست خورانی زور	medical
دکتورە فیس بوک ناونیشان	medical
سلاو دکتۆر شوێنی زیپکە چۆن لادەجێت	medical
سڵاو پستی دەسم وشک ئەبی خورانی زۆرە چۆن چارەسەرئەکری	medical

Fig. 5. Sample texts in the corpus.

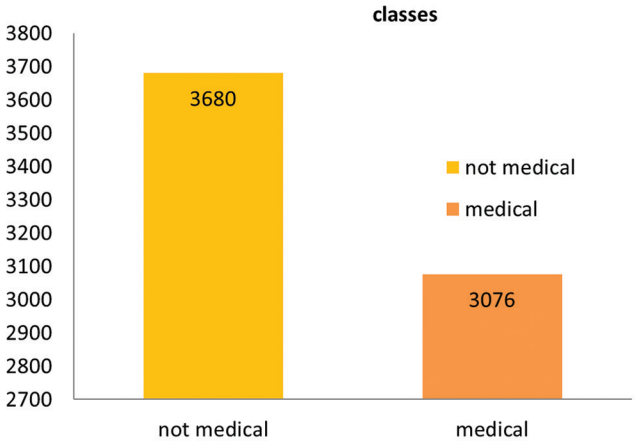

Fig. 6. The amount of data in the corpus.

We separate (4324 texts) for training and validation (1081 texts) and keep the rest (1352 texts) for testing. A shuffling technique was implemented to block the model from learning the specific order of words and inputs and provide a more realistic result. Table I shows the scores achieved from training on the classifiers.

The corpus works well with machine learning and deep learning classifiers, except for KNN, with the lowest score of 0.57. In terms of precision, the multinomial classifier outperforms other classical classifiers with a score of 0.91. Moreover, the second-best classifiers are SGD and

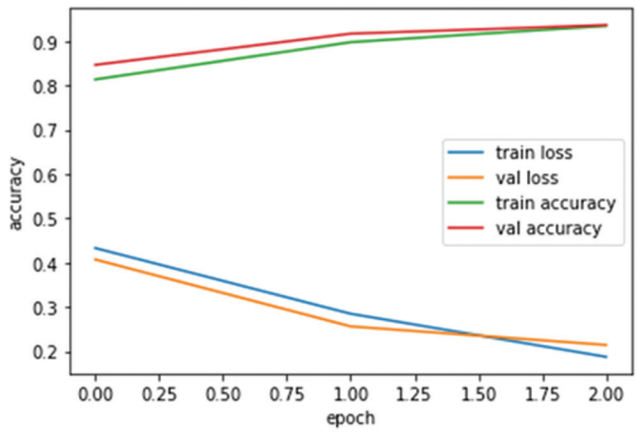

Fig. 7. The accuracy PER 2 epochs.

TABLE 1
THE SCORE OF (PRECISION, RECALL, ACCURACY, AND F1_SCORE) METRICS QF EACH CLASSIFIER

Classifier	Precision	Recall	Accuracy	F1_score
Multinomial	0.91	0.90	0.90	0.90
SGD	0.90	0.89	0.89	0.88
Decision tree	0.85	0.85	0.85	0.85
Random forest	0.90	0.89	0.89	0.89
SVM	0.82	0.82	0.82	0.82
KNN	0.33	0.57	0.57	0.42
Logistic regression	0.89	0.88	0.88	0.87
Bert multilingual	0.92	0.92	0.92	0.92

SVM: Support vector machines, kNN: k-nearest neighbor, SGD: Stochastic gradient descent

random forest, scoring 0.90. The rest of the classifiers stay between 0.80 and 0.89. Similarly, multinomial yields a more significant result than other ML classifiers for recall which is 0.90. SGD and random forest come second by scoring one point lower than multinomial.

Regarding deep learning, BERT outperforms the classical classifier by scoring the highest point in all measurements, which is 0.92. the model was merely trained for ten epochs. It is essential to state that training the model on a higher number helps the model achieve a score closer to 100% for all measurements, as shown in Fig. 7. Furthermore, the validation loss decreases significantly, guaranteeing a higher accuracy score.

Training loss is a metric to evaluate how well a deep-learning model fits the training data. On the other hand, validation loss specifies how well a deep learning model performs when evaluated against validation data. Moreover, the validation and training accuracy measure the model's overfitting. Overfitting refers to a statistical modeling error that occurs when a function is too closely aligned with a limited set of data points. Thus, the model is only helpful for its initial data set and not any other data sets. Since the gap between training loss and validation loss is too narrow, as illustrated in Fig. 7, it indicates that our model is fully optimized and has zero overfitting cases.

In the next phase, We attempt to compare our model with another state-of-the-art model in the Kurdish language. Unfortunately, this matter has yet to be tackled by researchers in the language. Therefore, we sought other language models which are close to the English language. We discovered a Bert-based model in the persian language known as ParsBert. Moreover, we compared our fine-tuned BERT model with the latest ParsBert model (Farahani, et al., 2021). ParBert is a recent state-of-the-art model developed for persian languages. The pre-trained model is used for numerous tasks such as text classification, question-answering, and named entity recognition. We trained our dataset on the pre-trained model; the results are displayed (Table II).

Overall, it can be noted that our fine-tuned model works slightly better than ParsBert, particularly in the case of precision and F1_score. The main reason behind this is that the Persian language is close to the Kurdish language, particularly in the sense of having nearly similar alphabetical letters and many standard vocabularies that exited in our corpus, as shown in Table III. Naturally, these similarities helped PrasBert find the tokens for most of the lexicons in our corpus, which ultimately yielded high results for ParsBert.

It is worth noting that having such a model is crucial to the Kurdish language. Even though this is the first time, the Kurdish language is introduced to a pre-trained model like BERT. The model outperformed the state-of-the-art by fine-tuning and using widespread softmax activation. The outcome achieved can add another source for the Kurdish language and prevent it from being labeled a less-resourced language. Moreover, the model can be utilized on Kurdish clinical websites or social media pages to separate medical and nonmedical questions. They can answer medical questions and serve their users, guaranteeing more customers. This is because our model can recognize non-medical text with high accuracy.

TABLE II
The Score of (Precision, Recall, Accuracy, and F1_Score) Metrics QF Multilingual and Parsbert

Model	Precision	Recall	Accuracy	F1_score
BERT-multilingual	92	92	92	92
ParsBert	91	92	92	91

TABLE III
Sample of Common Words between Kurdish and Persian

Kurdish	Persian	English meaning
دکتور	دکتر	Doctor
(مو(قژ	مو	Hair
سەر	سر	Head
دەست	دست	Hand
عیلاج	عیلاج	Treatment
گۆشت	گوشت	Meat

VI. Conclusion

In this paper, we experimented with the recently published medical corpus for the Kurdish language using machine learning and deep learning (BERT) to classify texts. We removed the stop words and irrelevant texts in the pre-processing stage. We compared the performances of the deep learning method with the conventional machine learning classifiers. Our experiments indicate that the BERT-multilingual model achieved higher accuracy of 0.92 in the text classification task and showed at least +0.2 improvement over the traditional machine learning methods. For future work, we suggest using augmentation techniques by lemmatizing and giving the stem of the keywords in the input data, as this yielded higher results in other languages.

References

Abdulla, S. and Hama, M. H., 2015. Sentiment analyses for kurdish social network texts using naive bayes classifier. *Journal of University of Human Development*, 1(4), pp. 393-397.

Ahmadi, S., 2020. *KLPT-Kurdish Language Processing Toolkit*., In: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pp. 72-84.

Cieliebak, M., Deriu, J.M., Egger, D. and Uzdilli, F., 2017. *A Twitter Corpus and Benchmark Resources for German Sentiment Analysis*. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Medias. pp. 45-51.

Collobert, R., Weston, J., Bottu, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P., 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, pp. 2493-2537.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. Available from: http://arxiv.org/abs/1810.04805

Esmaili, K., 2012. Challenges in Kurdish text processing. arXiv preprint arXiv:1212.0074.

Farahani, M., Gharachorloo, M., Farahani, M. and Manthouri, M., 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6), pp. 3831-3847.

Hoang, M., Bihorac, O.A. and Rouces, J., 2019. *Aspect-Based Sentiment Analysis Using Bert.* In: Proceedings of the 22nd Nordic Conference on Computational Linguistics. pp. 187-196.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp. 436-444.

Ling, J., 2020. *Coronavirus Public Sentiment Analysis with BERT Deep Learning*. Dalarna University, Sweden.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In: Proc. Advances in Neural Information Processing Systems. 26, pp.3111–3119.

Pennington, J., Socher, R. and Manning, C., 2014. *Glove: Global vectors for word representation.* In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

Saeed, A.M., Hussein, S. R., Ali, C.M. and Rashid, T. A., 2022. Medical dataset classification for Kurdish short text over social media. *Data Brief*, 42, p.108089.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L, Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems*, Vol. 30. NeurIPS Proceedings.

Zahera, H. M. Elgendy, I., Jalota, R. and Sherif, M.A., 2019. *Fine-tuned BERT Model for Multi-Label Tweets Classification*. The Real Estate Company, Mumbai. pp. 1-7.