CNN BASED CLASSIFICATION SYSTEM FOR COVID-19 VACCINE RELATED TWEETS

Mehmet Hanifi IŞIK[1], Saed ALQARALEH[2]

ABSTRACT

Nowadays, a huge amount of information is shared through social media. This information could be opinions, feelings, complaints, requests, etc. Also, the Internet becomes a space where peoples feel free to share their thought with others. For example, currently due to the coronavirus, social media has an extremely large number of opinions, news about almost all the aspects of coronavirus. Such information is very useful either for governments or even societies. However, the extremely huge number of available information leads to the use of a very minimum amount of such useful information.

In this paper, the Deep Convolutional Neural Networks (ConvNet) have been integrated to build an efficient and automated classification system for the Covid-19 vaccine-related tweets. The system can help to get the most relevant information about the people's opinions related to the vaccine.

Several experiments using manually annotated 15,000 tweets about the vaccine have been conducted, and it has been observed that the developed system has the ability to obtain quite good performance.

Keywords: *Sentiment Analysis, Natural Language Processing, Convolutional Neural Networks, Text Classification, Covid-19 Vaccine.*

[1] Undergraduate Student, Hasan Kalyoncu University, mhanifi.isik@std.hku.edu.tr

[2] Corresponding Author, Assist. Prof. Dr., Hasan Kalyoncu University, saed.alqaraleh@hku.edu.tr

1 INTRODUCTION

In today's modern world, the Internet has an exclusive impact on people's lives, where websites such as social media became extremely popular and people are expressing their thoughts and opinions about every topic through such websites. Twitter is one of the most used and visited microblogging websites that people share tweets about various topics such as politics, entertainment, gaming, music, sports, etc. According to (internetlivestats.com, 2021), every second, on average, around 6,000 tweets are published, i.e., over 350,000 tweets per minute, 500 million tweets per day. For this reason, Twitter is an ideal platform to find out what people think about all kinds of topics.

On the other hand, Coronavirus disease (COVID-19) is one of the most dangerous diseases of the 21st century. It has adversely affected all aspects of our daily life and overall humanity. All humanity is hoping that COVID-19 will end as soon as possible and that we will return to our normal life. Recently a number of vaccines are developed and approved such as Pfizer/BioNTech, Moderna, AstraZeneca/Oxford, Sinovac, Sputnik. However, governments, ministries of health, and healthcare corporations are suffering from the opinions of people about the vaccines and the lack of accurate information about how can encourage people to take the vaccine. Such information can be obtained using sentiment analysis systems.

Sentiment Analysis is a branch of Natural Language Processing (NLP) and it is a task of identifying sentiment polarities expressed in text, typically positive, neutral, or negative (Shin et al., 2016), recently there are several approaches used for sentiment analysis. Generally, machine learning and lexicon-based approaches such as SVM, Naive Bayes, Maximum Entropy have been widely used by researchers to implement efficient SA systems. Recently, several researchers investigated the possibility of using deep learning, such as the Convolutional Neutral Networks (CNN), Recursive Neutral Networks (RNN), and Long Short-Term Memory Networks (LSTM) for sentiment analysis.

In this paper, we investigated the possibility of using pre-trained CNN models to build a SA system that can efficiently classify COVİD-19 vaccine-related tweets. Also, two of the state-of-the-art word embedding, i.e., BERT and XLNet were investigated to find the most suitable one for the proposed system. In other words, this paper aims to integrate transformers-based methods into Convolutional Neural Networks (CNN) to realize an efficient and noteworthy system with high accuracy.

This paper is organized as follows sections: Section 2 presented the related works and researches, Section 3 introduced the components of the SA model including feature extraction and classification, and in section 4 the experiments and results were explained. Finally, conclusions are summarized in the last section.

2 RELATED WORKS

Currently, sentiment analysis is one of the fastest-growing research areas in Natural Language Processing. According to (Mäntylä et al., 2018), just after 2004, more than 7,000 papers have been published and studied the sentimental analysis. It is expected that this number will increase in the coming years and more papers and projects will be published in the field of sentiment analysis. Also, most of the sentiment analyses in prior research were performed using machine learning, lexicon-based, and hybrid approaches. In the following, some of these researchers are summarized.

In (Dong et al., 2020), the BERT-CNN model has been used for SA using 5,000 positives and 5,000 negatives cell phone review data. Also, the performance BERT-CNN model was compared with the CNN and BERT when used individually. The experimental results showed that BERT-CNN improved the performance of BERT and CNN by a factor of 14.4% and 17.4%, respectively (Dong et al., 2020).

(Pipalia et al., 2020) uses multiple pre-trained language models with LSTM model. This study was done using the IMDB reviews dataset where 25000 samples is used for training and 25000 for testing. Overall, XLNet outperformed all other studied models and achieved the highest accuracy (96.2%).

The Turkish sentiment analysis using some ensemble learning approaches (AdaBoost Classifier, Random Forest, and Gradient Boosting Classifier) was presented in (Alqaraleh, 2020). The results of the paper have shown that 1) Random Forest outperformed others, and 2) preprocessing plays a critical role in such system and can significantly improve the accuracy.

In (Ahmad et al., 2017), both machine learning and lexicon-based approaches were used. Support Vector Machines (SVM) was also used as learning algorithm. In result, the accuracy of the proposed approach outperformed the baseline classifier and the average of its accuracy is 85.4%.

The sentiment analysis of financial news using an unsupervised approach was studied in (Yadav et al., 2020) using a financial news dataset. Samples were preprocessed first, then Turney's pattern, POS tags, and noun-verb combinations were used. As a result, Turney's approach has achieved a 75% accuracy, and between 76% - 79% accuracy was achieved by the hybrid approach. Overall, the Hybrid approach obtained the best accuracy.

(Kaliyar et al., 2021) mainly worked on integrating BERT with a deep convolutional network for Fake news detection in social media. They worked on the real-world fake news dataset which is consists of 20800 news. By using the BERT, a 92.7% accuracy was obtained using the CNN and 97.55% with LSTM. On the other hand, a new version of BERT(FakeBERT) was proposed and was able to achieve more accurate result with an accuracy of 98.90%.

Twitter Sentiment Analysis during COVID19 was presented by (Dubey, 2020). In this paper, tweets that from twelve countries have been used to analyze how people are thinking of COVID19. Also, the NRC Emotion lexicon was applied for analyzing the contents of tweets. Results showed that most positive sentiments were reflected from the tweets of Belgians, while the most negative sentiments are from the tweets of China. Overall, tweets were classified into the basis of sentiments (positive and negative) and also categorizes into 8 emotions (fear, joy, anticipation, anger, disgust, sadness, surprise, trust) (Dubey,2020).

(Sosa,2017) also proposes to combine CNN and LSTM to improve the performance of sentiment analysis. They combined CNN and LSTM to build two models CNN-LSTM and LSTM-CNN. Overall, the LSTM-CNN model achieved the highest accuracy, i.e., 75.2% while the CNN-LSTM model obtained a 69.7% accuracy.

In summary, a comparative analysis for some state-of-the-art sentiment analysis approaches is shown in Table 1.

Table 1. Comparative analysis for some state-of-the-art sentiment analysis approaches.

Reference	**Year**	**Approach**	**Dataset**	**Techniques**	**Accuracy (%)**
(Dong et al.)	2020	Supervised	Cell phone review	BERT-CNN	%14-%17 (F1 Score)
(Pipalia et al.)	2020	Unsupervised	IMDB	LSTM	%86.6
				BERT	%93.6
				RoBERTa	%94.2
				XLNet	%96.2
				T5	%94.6
				DistilBERT	%92.4
(Alqaraleh)	2020	Supervised	Movie review	GBC	%86,16
				AdaBoost	%86.23
				RF	%87,64
(Ahmad et al.)	2017	Hybrid	Twitter Tweets	LMS	%85.4
				SVM	%86,7
(Yadav et al.)	2020	Unsupervised	Financial news	Turney's P	%75.00
				Hybrid	%76.00
				Noun-Verb	%79.00
(Kaliyar et al.)	2021	Unsupervised	Real world fake news	BERT-LSTM	97.55%
				BERT-CNN	92.70%
				FakeBERT	98.90%.

3.PROPOSED APPROACH

In This section, the main used approaches, i.e., BERT-CNN and XLNet-CNN, and the main components of building an efficient sentimental analysis system are explained. Mainly, samples(dataset) are collected first, then these samples are preprocessed, next one of the state-of-the-art words embedding (BERT or XLNet) is used to convert the text into numerical vectors. Finally, A CNN model is used as the system classifier.

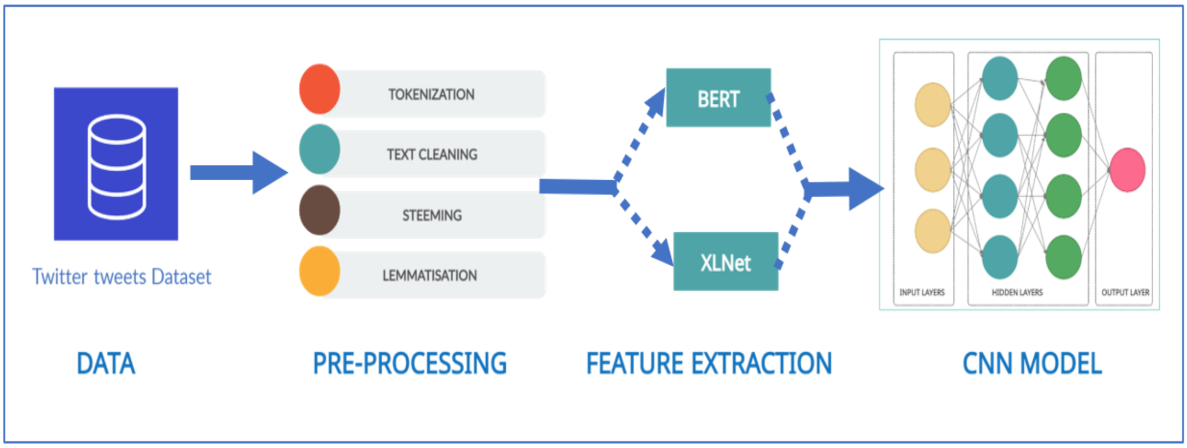

Figure 1. The architecture of the proposed approach. Note that the dashed arrow refers to the possibility of using either BERT OR XLNet for feature extraction.

3.1 Data Gathering

The availability of dataset is the main key for building and classification system. Having enough and quality samples give us the chance to proceed in building the required system. In this paper, we have used 15,000 tweets about COVID-19 vaccine that are manually annotated and classified as relevant or not relevant.

3.2 Preprocessing

One of the first steps of preprocessing is to determine which columns of datasets are required and which are not required. The used samples have 13 features such as author, tweet text, date of publication, retweet, etc. However, for building the needed system, the useless features were removed and only 3 features that are shown in Table 2, are used.

Table 2. The features of the used 15,000 COVID-19 vaccine's tweets.

Attribute	Type	Description
Id	Int	Unique identifier for tweets
Text	String	UTF-8 text of the Tweet
Class	String	Evaluated tweet 1 (relevant) or 0 (not relevant)

Next step, is to improve the quality of the dataset (the used samples), briefly the main preprocessing step that have been performed in this work are:

Eliminating Duplicate Samples

After data selection, the second step is data cleaning (Eliminating Duplicate Samples). Before this step, our dataset consisted of 15000 tweets, where 286 tweets found to be duplicate. Hence, after removing these tweets, 14714 tweets are kept for the following steps.

Text Cleaning

In this step, we firstly tokenized the tweets, i.e., splitting sentences(samples) into words. After tokenization, the test of samples was cleaned by removing URL, multiple spaces, single words, ASCII and special characters, punctuations, and stop words. Lastly, both stemming and lemmatizing were applied on the output of the previous sub-step.

3.3 Feature Extraction

In general, feature extraction from text refer to the process of converting the text into numerical values. In this paper the performance of two state-of-the-art BERT and XLNet were investigated. The details of these two approaches are summarized below.

3.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a new language representation that was introduced by Google in 2018. BERT mainly is a deep bidirectional representation that used unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2018). BERT is a multilingual model which supports 104 languages such as Turkish, Dutch, Italian, Persian. Also, BERT has achieved state-of-the-art performance in various natural language processing (NLP) tasks. Two main implementations of BERT are frequently used, they are:

BERT$_{\text{BASE}}$ which is consists of 12 layers, 768 hidden dimensions, and 12 attention heads (in transformer) with the total number of parameters, 110M.

BERT$_{\text{LARGE}}$: which is consists of 24 layers, 1024 hidden dimensions, and 16 attention heads (in transformer) with the total number of parameters, 340M.

In this paper, we have used the **BERT$_{\text{BASE}}$** as a decoder, i.e., transforming each sample into a numerical vector. As shown in Figure 2, the BERT tokenizer was used to split tweet text into a list of tokens that are added into the BERT vocabularies. BERT has special tokens which are [CLS] that is added to the beginning, and [SEP] which refers to the ending. Finally, BERT assigns an identifies (id) to each word in the current input(sample).

3.3.2 XLNet

XLNet is a bidirectional transformer that was introduced in 2019. XLNet is motivated by enabling autoregressive (AR) and autoencoding (AE) language models to learn bidirectional contexts while avoiding their limitations and the parameters of the transformer encoder are learned using permutation language modeling, which allows the model to capture general knowledge and representation (Yang et al., 2019). And also, XLNet achieves state-of-the-art performance on various natural language processing (NLP) tasks.

Mainly, XLNet consists of two model settings $\text{XLNet}_{\text{BASE}}$ and $\text{XLNet}_{\text{LARGE}}$. $\text{XLNet}_{\text{BASE}}$ consists of 12 layers, 768 hidden dimensions, and 12 attention heads, while $\text{XLNet}_{\text{LARGE}}$ is consisted of 24 layers, 1024 hidden dimensions, and 16 attention heads. In this paper, we used $\text{XLNet}_{\text{BASE}}$ and its performance was compared with the $\text{BERT}_{\text{BASE}}$.

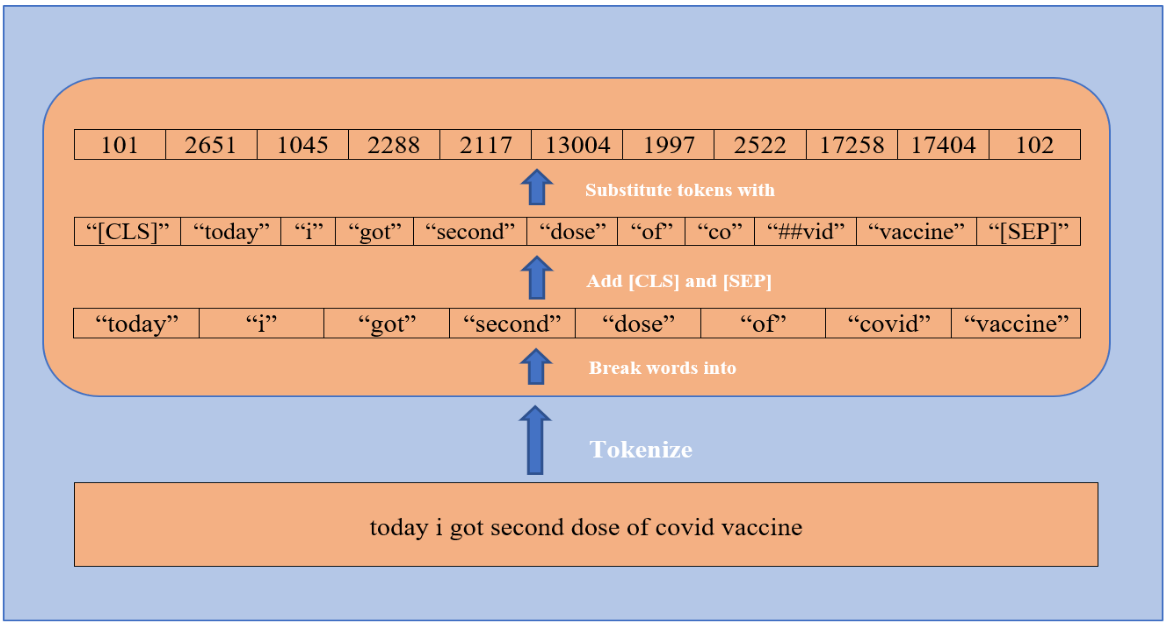

Figure 2. Example of transforming a sample into a numerical vector using BERT.

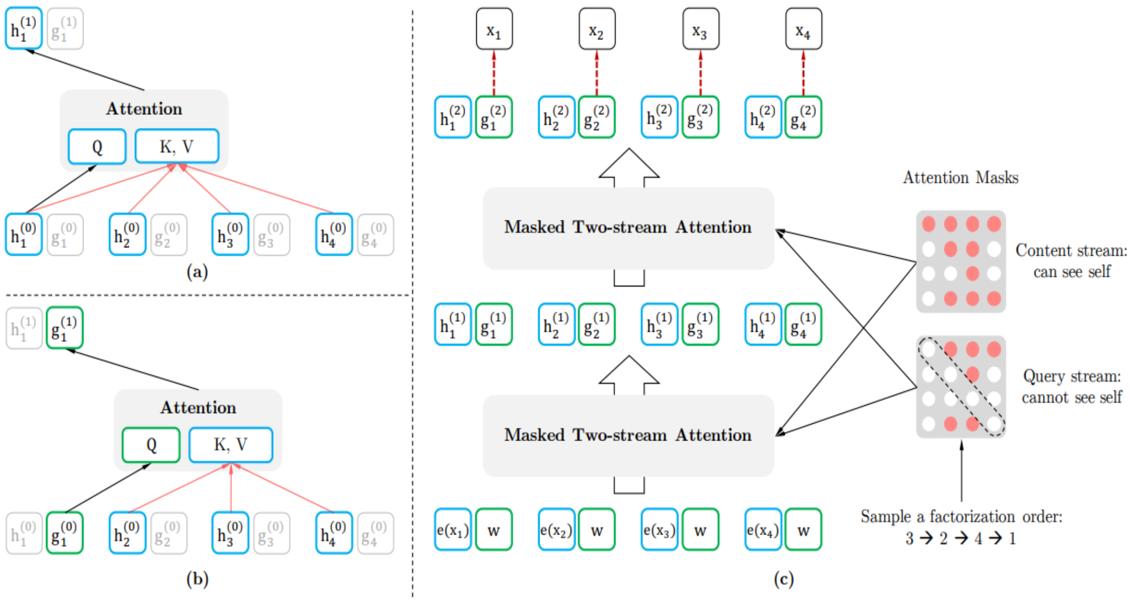

Figure 3. The architecture of XLNet two-stream self-attention for target-aware representations, where (a) is the content stream attention, which is the same as standard self-attention. (b) is the query stream attention, (c) is the overview of the permutation language modelling tanning with two-stream attention (Yang et al., 2019).

3.4 Classification

Deep and Neural learning models such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Auto encoders, Recurrent Neural Networks (RNN), and Reinforcement Learning were able to obtained state-of-the-art performance in many fields including sentiment analysis. In this paper the performance of CNN is investigated, and specifically the architecture proposed by Yoon Kim (Kim,2014) was selected. This model mainly consists of three main parts, i.e., a set of convolutional layers, pooling layer, and fully connected layer. Briefly, the convolution layer works on passing the input through some filters (feature maps) to obtain more efficient representatives. Secondly, the pooling layer operates upon each feature map independently to build new less dimensionality representatives. Related to the output layer, which is also known as the Fully-connected layer, is used to generate an output (predict the class of the current input). The architecture of the used CNN is shown below (Figure 3).

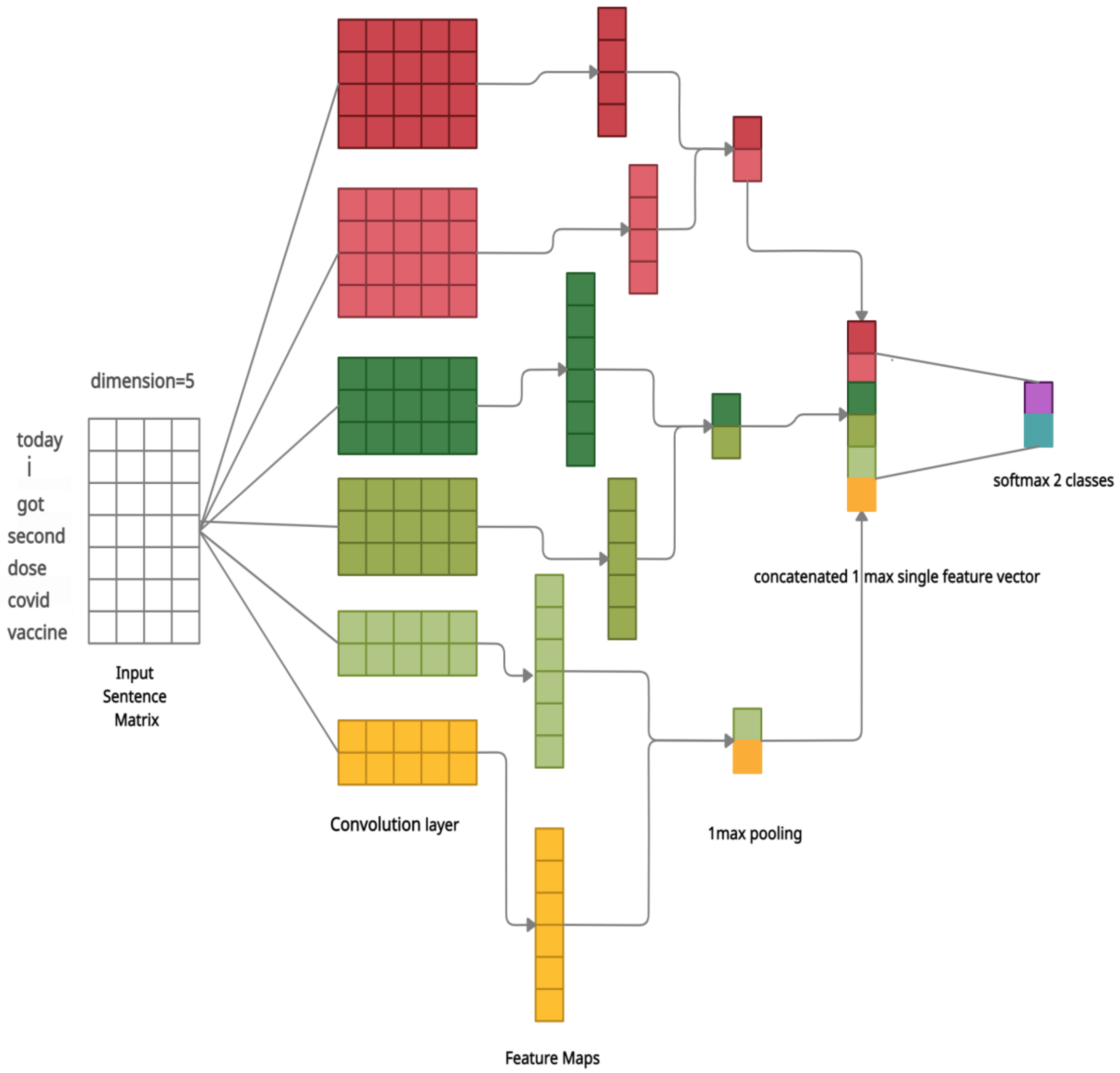

Figure 3. Architecture of the used CNN model (Kim,2014).

4. EXPERIMENTS

In this section, the performance of CNN and BERT, and XLNet word embedding was investigated (called BERT-CNN and XLNet-CNN). Also, the performance was compared with one of the state-of-the-art SA models, which we refer to it as LSTM-CNN that was presented in (Sosa, 2020).

As mentioned before, sentiment analysis has some main steps that start by improving the quality of the used dataset(samples) by preprocessing steps like removing duplicate and unnecessary information. Then, this dataset is used to train both BERT and XLnet and obtain the sample's numerical representation.

4.1 Evaluation Setups

Python and some of its libraries were used to implement the studied models. For example, the preprocessing steps were performed using the NLTK library, TensorFlow library is used in building the CNN classifiers, where the CNN main parameters are: Embedding number =100, embedding dimensions =768, and the number of kernels = 3, and the specifications of the experimental environment are shown in Table 3.

Table 3 Development environment specifications.

Memory	16GB
Processor	Intel Core i5 8300H
Development Tool	Python 3.9
Libraries	TensorFlow, Transformers

Related to the datasets, 15000 tweets collected from the COVID-19 vaccine dataset from Kaggle were used (Yadav,2020) These tweets were manually evaluated and classified one by one as relevant or irrelevant by 3 judges. Then, the class of each tweet is specified by the majority voting. To ensure the quality of the obtained results, four sub-datasets were constructed from the 15000 tweets. The details of the four constructed datasets are shown in table 4, where the first three sub-dataset are balanced, i.e., the number of samples for the two classes is same. The last sub-dataset has all 15000 tweets.

Table 4. The details of the four constructed datasets.

Datasets	Relevant Tweet	Not Relevant Tweet	Total Tweet
1st	1500	1500	3000
2nd	3000	3000	6000
3rd	4500	4500	9000
4th	10168	4832	15000

4.2 Experimental Results

In the first experiment, we have investigated and analyzed the performance of two transformer methods, i.e., BERT and XLNet. These two methods were integrated with the CNN model presented in (Kim,2014), which we refer to as BERT-CNN and XLNet-CNN respectively. As shown in Table 5, both transformer methods achieved a quite good performance. However, XLNet-CNN was able to slightly achieve higher results. All results reported by accuracy, precision, recall, and F1 score are shown in Table 5.

Table 5. Accuracy, Precision, Recall, F1 Score, and AUC of the two constructed models BERT-CNN and XLNet-CNN.

Dataset	BERT-CNN					XLNet-CNN				
	Accuracy	Precision	Recall	F1 Score	AUC	Accuracy	Precision	Recall	F1 Score	AUC
1st	0. 94	0. 94	0. 94	0. 93	0. 97	0. 95	0. 95	0. 94	0. 94	0. 98
2nd	0. 93	0. 93	0. 93	0. 93	0. 97	0. 94	0. 94	0.93	0.93	0.97
3rd	0.91	0.91	0.91	0.91	0.96	0.94	0.94	0.94	0.94	0.97
4th	0. 90	0. 91	0. 90	0. 91	0. 97	0.93	0.92	0.94	0.93	0.95

In the second experiment, to demonstrate the performance of BERT-CNN and XLNet-CNN models, they were also compared with the model of (Sosa,2017) which known as LSTM-CNN. Briefly, (Sosa,2017) has tested the effect of using combined two models, i.e, CNN, and LSTM. In the first part LSTM was the first model and CNN are the consecutive one, and vice versa is applied in the second part. As shown in (Sosa,2017), LSTM-CNN achieved the best results. Based on this, we have compared the two models that we have proposed (BERT-CNN and XLNet-CNN) with the LSTM-CNN (Sosa,2017). The results are shown in Figure 4.

As shown in Figure 4, both BERT-CNN and XLNet-CNN models have significantly outperformed LSTM-CNN, however, we believe that cab be improved by using the larger dataset(s), i.e., increasing the number of training samples.

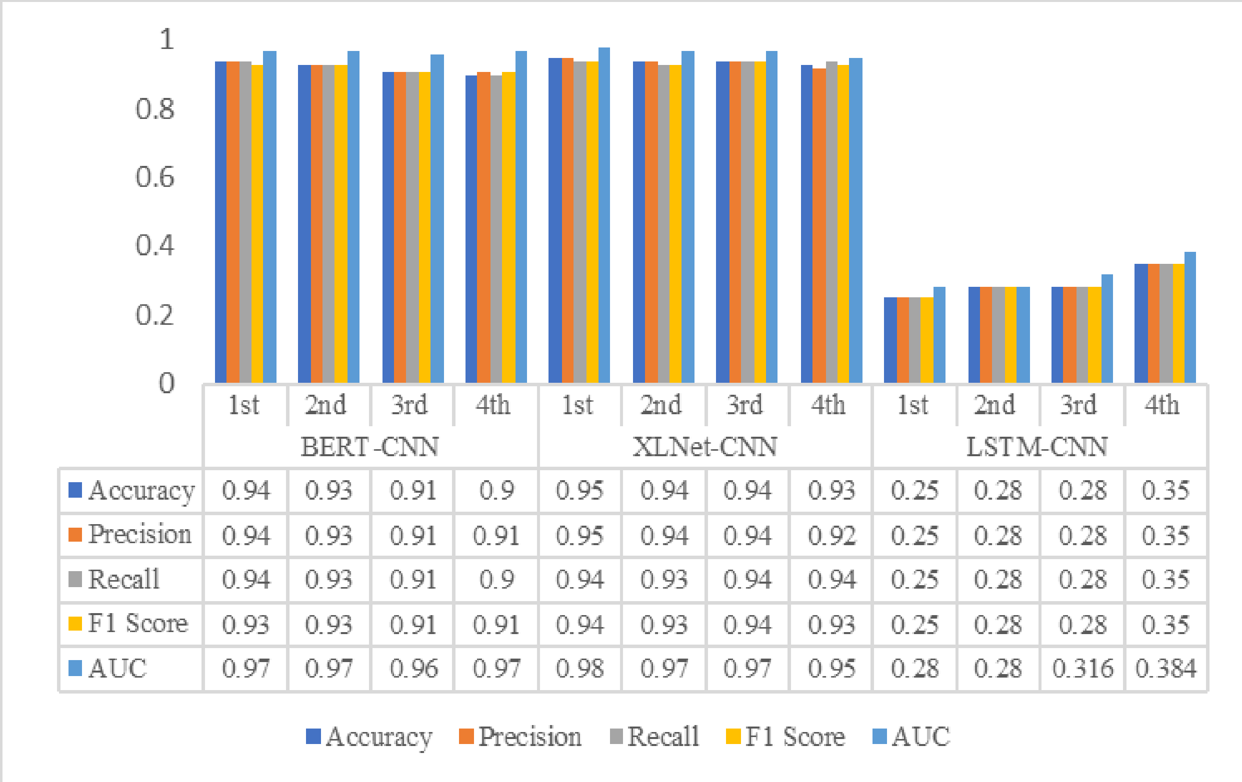

Figure 4. Accuracy, Precision, Recall, F1 score, and AUC for BERT-CNN, XLNet-CNN, and LSTM-CNN models.

5. CONCLUSIONS

In this paper, the performance of two state-of-the-art transformer-based methods which are BERT and XLnet while integrated with CNN is investigated (the models are known as BERT -CNN, and XLNet-CNN respectively). Overall, XLNet has slightly outperformed the BERT.

The mentioned models were also compared with one of SA's state-of-the-art model LSTM-CNN and similar to the results of the first experiment, the XLNet-CNN, which is a combination XLNet word embedding combined with the CNN model of (Kim,2014) was able to obtain the best accuracy, precision, recall, F1 score, and AUC for all used datasets.

As future work, this work can be further improved by investigating the possibility of integrating state-of-the-art transformer-based methods with more advanced deep learning methods in order to build a more efficient and multi-objective sentiment analysis system.

REFERENCES

'Twitter Usage Statistics - internet Live Stats. (2021). Twitter Usage Statistics. Retrieved May 8, 2021, from https://www.internetlivestats.com/twitter-statistics/

Shin, Bonggun, Timothy Lee, and Jinho D. Choi (2016), "Lexicon integrated cnn models with attention for sentiment analysis." arXiv preprint arXiv:1610.06272.

Mäntylä, Mika V., Daniel Graziotin, and Miikka Kuutila (2016),"The evolution of sentiment analysis—A review of research topics, venues, and top cited papers." Computer Science Review

J. Dong, F. He, Y. Guo and H. Zhang (2020) "A Commodity Review Sentiment Analysis Based on BERT-CNN Model," 5th International Conference on Computer and Communication Systems (ICCCS), Shanghai, China, 2020, pp. 143-147, doi: 10.1109/ICCCS49078.2020.9118434.- 1/04/2021

K. Pipalia, R. Bhadja and M. Shukla (2020), "Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis," 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020, pp. 411-415, doi: 10.1109/SMART50582.2020.9337081.

Alqaraleh, S. (2020). Turkish Sentiment Analysis System via Ensemble Learning. European Journal of Science and Technology, (Special Issue), 122-129.- 1/04/2021

Ahmad, Munir & Aftab, Shabib & Ali, Iftikhar & Hameed, Noureen. (2017). Hybrid Tools and Techniques for Sentiment Analysis: A Review. International Journal of Multidisciplinary Sciences and Engineering. 8. 31-38.- 1/04/2021

Yadav, A., Jha, C. K., Sharan, A., & Vaish, V. (2020). Sentiment analysis of financial news using unsupervised approach. Procedia Computer Science, 167, 589-598.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765-11788.

Dubey, Akash. (2020). Twitter Sentiment Analysis during COVID19 Outbreak. SSRN Electronic Journal. 10.2139/ssrn.3572023.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237.

Yadav, R. (2020). COVID-19 Vaccine Tweets. https://www.kaggle.com/ritesh2000/covid19-vaccine-tweets

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/d14-1181

Sosa, P. M. (2017). Twitter sentiment analysis using combined LSTM-CNN models. *Eprint Arxiv*, 1-9.