

CERIST Natural Language Processing Challenge

March 29th, 2023.

Classifying Arabic covid-19 related tweets for fake news detection and sentiment analysis with BERT-based models

Rabia Bounaama^a, Mohammed El Amine Abderrahim^b

^a*Biomedical Engineering Laboratory, Tlemcen University, Algeria,*

^b*Laboratory of Arabic Natural Language Processing, Tlemcen University, Algeria*

Abstract

The present paper is about the participation of our team “techno” at CERIST Natural Language Processing Challenge. We used an available dataset for task1.c: Arabic sentiment analysis and fake news detection within covid-19. It comprises 4128 tweets for sentiment analysis task and 8661 tweets for fake news detection task. We used natural language processing tools with the combination of the most renowned pre-trained language models BERT (Bidirectional Encoder Representations from Transformers). The results shows the efficacy of pre-trained language models as we attained an accuracy of 0.93 for the sentiment analysis task and 0.90 for the fake news detection task.

Keywords: CERIST; BERT; Fake news detection; Sentiment analysis; Twitter

1. Introduction

Covid-19 has affected each individual’s daily life. On October 2022, over of 600 million cases have been reported around the world. Under the current disruption caused by covid-19, everyone is responsible to government measures aimed at preventing the spread of the virus, including distance learning, quarantine, staying at home and social distancing. Those decisions have results an amount of user’s opinion on social media. One of the most effective methods for analyzing these opinions is Sentiment Analysis (SA). SA is a

classification process that identifies the opinions and emotions of users through the written contents Hassonah et al., 2020.

Information on social media platforms may contain true and false facts. Unfortunately, false facts spread rapidly, Stahl, 2018. Fake news (FN) is the deliberate spread of misinformation via traditional news media or via social media, Stahl, 2018. The critical situation imposed by the covid-19 pandemic has increased people's attention particularly regards to the number of new cases and deaths. If this information is wrong, it can have a negative impact on people's mental health, like depression and anxiety disorder. For that reason, FN detection has become an urgent research area for the research community. FN detecting is to predict whether the news article is a FN piece or not Shu et al., 2017. In this context, there is pressing need for practical studies. To this end CERIST has proposed several tasks to solve. Our participation is related to opinion mining and SA task, precisely "task1.c: Arabic sentiment analysis and fake news detection within COVID-19"*.

To address SA and FN detection tasks several approaches can be used. Among the most popular that has shown better results is the transfer learning approach using BERT. The use of pre-trained BERT models allows us to fine-tune the model for the specific task without having to train the model from scratch. This involves training the model on a labelled dataset, Hadj Ameer & Aliane, 2021, provided by the challenge organizers. The model's biases and weights were adjusted to fit the requirements of either the SA or FN detection task. After training, the model was used to predict the labels on an unseen test dataset.

The remainder of this paper is organized as follow. In Section 2, we present our related work on SA and FN detection in comments about covid-19. Section 3 describes the experimental study and the obtained results. Finally, section 4 concludes the paper.

2. Related work

This related work is divided into two parts: Arabic SA related covid-19 and Arabic FN detection related covid-19.

Several studies have been conducted to address SA related to Covid-19 pandemic and different approaches and methods are applied. Among these approaches, deep learning is a commonly used tool especially for short texts such as tweets. The study given by Alhumoud, 2020 on Arabic SA applied deep learning to a corpus of 416,292 annotated tweets about covid-19. Authors found that the word 'corona' and 'virus' are dominant in their corpus. They used traditional Machine Learning (ML) such as Support Vector Machines (SVM), the new pre-trained transformer model "araBERT" and ensemble model. As results, they found that negative sentiment is dominant as opposed to the positive and neutral. They conclude that the proposed ensemble model outperforms SVM, deep learning models and AraBERT with 0.90 as accuracy.

Covid-19 had an important impact on government decision in all countries worldwide. For example, Saudi Arabia implemented a series of control measures in order to deal with the pandemic. The two following studies concern the SA of tweets in Saudi Arabia regarding distance learning during Covid-19 Aljabri et al., 2021, and regarding governmental preventive measure to contain Covid-19, Alhajji et al., 2020. In the study of Alhajji et al., 2020, authors used Naïve Bayes (NB) as ML model to train a corpus of 53,127 tweets containing hashtags related to seven public health measures imposed by the Saudi government. The measures are: Grand Mosque closure, Qatif lockdown, school and university closure, shopping malls, parks and restaurants closure, sports competition suspension, congregational and weekly Friday prayers suspension and nationwide curfew. The results of their analysis conclude that twitter users in Saudi Arabia support the applicable government measures except one measure: shopping malls, parks, and restaurant closure. For the

* <http://www.nlpchallenge.cerist.dz>

six other measures, results show that positive sentiments are more than negative ones especially for measure that pertain religious practice.

The research conducted by Aljabri et al., 2021 showed the acceptance rate of the distance learning enforcement in Saudi Arabia. Their sentiment classifier was built using well-known ML and feature extraction techniques. As best result, Logistic Regression (LR) with the use of unigrams and tf-idf approaches for feature extraction had an accuracy of 0.89. Their research concludes that positive opinions are more numerous than negative ones.

The work described in Ali, 2021 concerns e-learning to mitigate Covid-19. This author also used traditional ML like SVM, K Nearest Neighbour (KNN), Multinomial Naïve Bayes (MNB) and LR. The data extracted from twitter, via twitter APIs, are pre-processed and labelled depending on the emotional weights of eight emotions using word-emotion lexicon NRC (the National Research Council Canada). As best result, SVM classifier outperforms the others classifiers with 0.89 as accuracy.

For the Arabic language, there are few studies and datasets published during Covid-19 about FN detection and it hasn't achieved the level compared to other languages, especially English, Mahlous and Al-Laith, 2021. In the following, we present some recent automated systems for Arabic FN detection during Covid-19. The study of Mahlous and Al-Laith, 2021 introduces a public dataset within Covid-19 FN detection in Arabic tweets. The authors compare two built corpus: 2500 manually labelled tweets and 34529 automated labelled tweets from 7 million extracted tweets. By using classical ML, The results show that the automatically and manually labelled corpora reached 0.93 and 0.87 respectively as f1 score.

In Shishah, 2022, the author tested transformer models with a new strategy "JointBERT" in four FN dataset: Covid19Fakes, AraNews, Satirical News and ANS. A new technique in detecting Arabic FN is proposed that had an improvement reach to 10% comparing to baseline Arabic FN detection models. The author used his proposed method "JointBERT" for sequence labelling and encoding process, more details about the proposed model are in Shishah, 2021.

In Nassif et al., 2022, authors constructed two Arabic FN detection datasets related to Covid-19. The first dataset comprises 16,000 examples and was translated to English from "Kaggle", while the second dataset comprises 10,000 examples and was collected from Arabic websites and tweets. The authors developed eight transformer models, including Giga-BERT, RoBERTa-Base, AraBERT, Arabic-BERT, ARBERT, MarBERT, Araelectra and QaribBERT, as well as contextualized Arabic embedding models. The developed models performed much better on the Arabic dataset compared to the translated one. ARBERT pre-trained model achieved 0.98 as accuracy.

3. Experiments

In this section, we describe the followed steps for building our SA model, and FN detection model.

3.1. Pre-trained models description

Transformer is a simple network architecture proposed by Vaswani et al., 2017 which is based on the attention mechanism Bahdanau et al., 2014. The main idea behind the attention mechanism is to allow the encoder-decoder model to automatically search for relevant parts from the input source in order to predict a relevant target word for machine translation purposes by extending the vector length, Cho et al., 2014.

BERT, Devlin et al., 2018 is a language model. For text classification BERT-base model contains an encoder with 12 Transformer blocks. BERT takes as input a sequence of no more than 512 tokens and outputs the representation of sequence, Sun et al., 2019. The pre-trained BERT model becomes a state-of-the-art of several NLP tasks such as question answering and language inference, Devlin et al., 2018. We implemented

the pre-trained BERT-mini-arabic and BERT-base-arabic, Safaya et al., 2020 for SA task and FN detection task respectively from hugging face library[†]. The model was pre-trained on approximately 8.2 Billion words. The model is a combination of convolutional neural network (CNN) with BERT, Safaya et al., 2020.

3.2. Data pre-processing

We perform data distribution before preprocessing it. As shown in the table below (see table1), the training data for the SA task and the FN detection task has a balanced distribution of classes, which may impact the model’s accuracy performance.

We implement some known steps for preparing the training dataset for BERT models. Knowing that the BERT model relies heavily on the context of a word in a sentence, then, during data cleaning we try to preserve the semantic and syntactic structure of each sentence “tweet”. We used regular expressions to remove URLs and punctuations marks by using the python “String” module. We replaced certain punctuation marks in the text with space and we removed emoji symbols. We kept digits unchanged. Note that we did not find any duplicate tweets and the traditional pre-processing was implemented for both datasets. Finally data was ready for training stage using BERT pre-trained models.

Table 1. Datasets distribution

Task	Count values per each class
SA (sentiment analysis)	neutral sentiment 1600 positive sentiment 1571 negative sentiment 957
Total	4128 tweets
FN (fake news)	Can’t decide 5024 maybe 1471 yes 367 no 1799
Total	8661 tweets

As indicated in Table 1, the corpus of total tweets in the SA task consists of 38% neutral and positive sentiments. In the FN task, the “can’t decide” class accounts for 58%, while the “yes” and “no” classes make up 4% and 16% of the corpus, respectively. Predicting minority classes may be difficult due to the impact of imbalanced datasets on most ML algorithms, Zhang et al., 2020.

[†] <https://huggingface.co/>

3.3. Experimental fine-tuning

The fine-tuning process is an essential step in using pre-trained models like BERT. It involves adjusting the parameters of the pre-trained model to suit the specific task that it's being applied to. For the BERT models, in our case, we used the following fine-tuning parameters:

- Training batch size: We used a batch size of 10 during the training process. The batch size is the number of samples that are processed at once before the model's weights are updated.
- Learning rate: The learning rate determines the step size at which the optimizer updates the model's weights. For our models, we used a learning rate of $2e-5$.
- Optimizer: AdamW is the optimizer that we used to update the model's weights during the training process. AdamW is a variation of the Adam optimizer that's well-suited for fine-tuning pre-trained models like BERT.
- Train epochs: The number of training epochs refers to the number of times that the model is trained on the entire dataset. For our models, we used a total of 10 training epochs.
- Maximum sequence length: The maximum sequence length is the maximum number of tokens that can be fed into the BERT model at once. In our case, we used a maximum sequence length of 256.

By using these fine-tuning parameters, we were able to effectively fine-tune the pre-trained BERT models for our SA and FN detection tasks.

3.4. Results and discussion

To evaluate the performance of our models, we used the fine-tuned BERT models to predict the targets for both the SA task and FN detection task.

We can say that the use of the pre-trained language models BERT gives good results in the context of the SA and the FN detection.

As shown in the tables below (Table 2 and Table 3), the SA task result, in term of the three performance measures, outperforms the FN detection task. This result can be attributed to the balanced distribution of classes in the SA task dataset.

The FN detection dataset is unbalanced, which could lead to bias in the output and affect the overall performance of the model.

Table 2. Techno-team results for task1.c on validation dataset for sentiment analysis and fake news detection tasks

Task	Precision	Recall	F1-score
SA	0.94	0.93	0.94
FN	0.87	0.86	0.86

Table 3. Techno-team results for task1.c on test dataset for sentiment analysis and fake news detection tasks

Task	Accuracy	F1-score
SA (run1)	0.9332	0.9332
SA (run2)	0.9332	0.9331
SA (run3)	0.9381	0.9381
FN	0.9021	0.9016

4. Conclusion

In this work, we utilized natural language processing and pre-trained language models BERT to address two classification subtasks as part of CERIST Natural Language Processing Challenge shared tasks “task1.c”. We obtained 0.93 for f1-score on sentiment analysis task and a 0.90 for fake news detection task. These results indicated that this approach effectively solved the sentiment analysis task related to COVID-19. However, the limitations of the model are evident in the unbalanced dataset of the fake news detection task. To further improve our results, in future work, we plan to implement other learning approaches.

References

- Alhajji, M., Al Khalifah, A., Aljubran, M., & Alkhalifah, M., 2020. Sentiment analysis of tweets in Saudi Arabia regarding governmental preventive measures to contain COVID-19.
- Alhumoud, S., 2020. Arabic sentiment analysis using deep learning for covid-19 twitter data. *International Journal of Computer Science and Network Security*, 132-138.
- Ali, M. M., 2021. Arabic sentiment analysis about online learning to mitigate covid-19. *Journal of Intelligent Systems*, 30(1), 524-540.
- Aljabri, M., Chrouf, S. M. B., Alzahrani, N. A., Alghamdi, L., Alfahaid, R., Alqarawi, R., Alduhailan, N., 2021. Sentiment analysis of Arabic tweets regarding distance learning in Saudi Arabia during the COVID-19 pandemic. *Sensors*, 21(16), 5431.
- Bahdanau, D., Cho, K., & Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Devlin, J., Chang, M. W., Lee, K., outanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Hadj Ameer M.S, Aliane, H., 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset. *Procedia Computer Science* 189, 232-241.2021.
- Hassonah, M. A., Al-Sayyed, R., Rodan, A., Ala'M, A. Z., Aljarah, I., Faris, H., 2020. An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter. *Knowledge-Based Systems*, 192, 105353.
- Mahlous, A. R., & Al-Laith, A., 2021. Fake news detection in Arabic tweets during the COVID-19 pandemic. *International Journal of Advanced Computer Science and Applications*, 12(6), 778-788.
- Nassif, A. B., Elnagar, A., Elgendy, O., Afadar, Y., 2022. Arabic fake news detection based on deep contextualized embedding models. *Neural Computing and Applications*, 34(18), 16019-16032.
- Safaya, A., Abdullatif, M., Yuret, D., 2020. Kuisail at semeval-2020 task 12: BERT-CNN for offensive speech identification in social media. *arXiv preprint arXiv:2007.13184*.
- Shishah, W., 2022. JointBERT for Detecting Arabic Fake News. *IEEE Access*, 10, 71951-71960.
- Shishah, W., 2021. Fake news detection using BERT model with joint learning. *Arabian Journal for Science and Engineering*, 46(9), 9115-9127.
- Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1), 22-36.
- Stahl, K., 2018. Fake news detection in social media. *California State University Stanislaus*, 6, 4-15.
- Sun, C., Qiu, X., Xu, Y., & Huang, X., 2019. How to fine-tune BERT for text classification?. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18* (pp. 194-206). Springer International Publishing.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., Artzi, Y., 2020. Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv:2006.05987*.