DamascusTeam at NLP4IF2021: Fighting the Arabic COVID-19 Infodemic on Twitter Using AraBERT

Ahmad Hussein[1], Nada Ghneim[2], and Ammar Joukhadar[1]

[1]Faculty of Information Technology Engineering, Damascus University, Damascus, Syria
ahmadhussein.ah7gmail.com, ajoukhadar@el-ixir.com
[2]Faculty of Informatics&Communication Engineering, Arab International University, Damascus, Syria
n-ghneim@aiu.edu.sy

Abstract

In the modern era of computing, the news ecosystem has transformed from old traditional print media to social media outlets. Social media platforms allow us to consume news much faster, with less restricted editing results in the spread of infodemic misinformation at an incredible pace and scale. Consequently, the research on the infodemic of the post's misinformation is becoming more important than ever before. In this paper, we present our approach using AraBERT (Transformer-based Model for Arabic Language Understanding) to predict 7 binary properties of an Arabic tweet about COVID-19. To train our classification models, we use the dataset provided by NLP4IF 2021. We ranked 5th in the Fighting the COVID-19 Infodemic task results with an F1 of 0.664.

1 Introduction

In the past few years, various social media platforms such as Twitter, Facebook, Instagram, etc. have become very popular since they facilitate the easy acquisition of information and provide a quick platform for information sharing (Vicario et al., 2016; Kumar et al., 2018). The work presented in this paper primarily focuses on Twitter. Twitter is a micro-blogging web service with over 330 million Active Twitter Users per month, and has gained popularity as a major news source and information dissemination agent over the last years. Twitter provides the ground information and helps in reaching out to people in need, thus it plays an important role in aiding crisis management teams as the researchers have shown (Ntalla et al., 2015). The availability of unauthentic data on social media platforms has gained massive attention among researchers and become a hot-spot for sharing misinformation (Gorrell et al., 2019; Vosoughi et al., 2017). Infodemic misinformation has been an important issue due to its tremendous negative impact (Gorrell et al., 2019; Vosoughi et al., 2017; Zhou et al., 2018), it has increased attention among researchers, journalists, politicians and the general public. In the context of writing style, misinformation is written or published with the intent to mislead the people and to damage the image of an agency, entity, person, either for financial or political benefits (Zhou et al., 2018; Ghosh et al., 2018; Ruchansky et al., 2017; Shu et al., 2020).

This paper is organized as follows: Section 2 describes the related work in this domain; Section 3 gives our methodology in detail; Section 4 discusses the evaluation of our proposed solution and finally, the last section gives the conclusion and describes future works.

2 Related Works

There are various techniques used to solve the problem of infodemic misinformation on Online Social Media, especially in English content. This section briefly summarizes the work in this field. Allcott et al. (2017) have focused on a quantitative report to understand the impact of misinformation on social media in the 2016 U.S. Presidential General Election and its effect upon U.S. voters.

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 93–98
June 6, 2021. ©2021 Association for Computational Linguistics

Authors have investigated the authentic and unauthentic URLs related to misinformation from the BuzzFeed dataset. Shu et al. (2019) have investigated a way for robotization process through hashtag recurrence. Authors have also presented a comprehensive review of detecting misinformation on social media, false news classifications on psychology and social concepts, and existing algorithms from a data mining perspective. Ghosh et al. (2018) have investigated the impact of web-based social networking on political decisions. Quantity research (Zhou et al., 2018; Allcott et al., 2017; Zubiaga et al., 2018) has been done in the context of detecting political-news-based articles. Authors have investigated the effect of various political gatherings related to the discussion of any misinformation as agenda. Authors have also explored the Twitter-based data of six Venezuelan government officials with a specific end goal to investigate bot collaboration. Their discoveries recommend that political bots in Venezuela tend to imitate individuals from political gatherings or basic natives. In one of the studies, Zhou et al. (2018) have investigated the ability of social media to aggregate the judgments of a large community of users. In their further investigation, they have explained machine learning approaches with the end goal to develop a better rumors detection. They have investigated the difficulties for the spread of rumors, rumors classification, and deception for the advancement of such frameworks. They have also investigated the utilization of such useful strategies towards creating fascinating structures that can help individuals in settling on choices towards evaluating the integrity of data gathered from various social media platforms. In one of the studies, Jwa et al. (2019) have explored the approach towards automatic misinformation detection. They have used Bidirectional Encoder Representations from Transformers model (BERT) model to detect misinformation by analyzing the relationship between the headline and the body text of the news story. Their results improve the 0.14 F-score over existing state-of-the-art models. Williams et al. (2020) utilized BERT and RoBERTa models to identify claims in social media text a professional fact-checker should review. For the English language, they fine-tuned a RoBERTa model and added an extra mean pooling layer and a dropout layer to enhance generalizability to unseen text. For the Arabic language, they fine-tuned Arabic-language BERT models and demonstrate the use of back-translation to amplify the minority class and balance the dataset. Hussein et al. (2020) presented their approach to analyze the worthiness of Arabic information on Twitter. To train the classification model, they annotated for worthiness a dataset of 5000 Arabic tweets -corresponding to 4 high impact news events of 2020 around the world, in addition to a dataset of 1500 tweets provided by CLEF 2020. They proposed two models to classify the worthiness of Arabic tweets: BI-LSTM model, and a CNN-LSTM model. Results show that BI-LSTM model can extract better the worthiness of tweets.

3 Methodology

In this section, we will present our methodology by explaining the different steps of building the models, we use the same architecture for building them: Data Set, Data Preprocessing, AraBERT System Architecture, and Model Training.

3.1 Data Set

We used a dataset of 2556 tweets provided by NLP4IF 2021 (Shaar et al., 2021), which includes tweets about COVID-19. The dataset includes besides the tweet text and the tweet Id. Each tweet annotates with binary properties about COVID-19: whether it contains a verifiable claim (Q1), whether it appears to contain false information (Q2), whether it may be of interest to the general public (Q3), whether it is harmful (Q4), whether it needs to verification (Q5), whether it is harmful to society (Q6) and whether it requires attention of government entities (Q7). Each question has a Yes/No (binary) annotation. However, the answers to Q2, Q3, Q4 and Q5 are all "nan" if the answer to Q1 is No. Table 1 shows the statistics of the class labels for each property in the dataset.

Classifier	**Yes**	**No**	**Not Sure**
Q1	1926	610	0
Q2	376	1545	635
Q3	1895	22	639
Q4	351	1566	639
Q5	936	990	630
Q6	2075	459	0
Q7	2208	328	0

Table 1: Dataset with the class labels

3.2 Data Preprocessing

Tweets have certain special features, i.e., emojis, emoticons, hashtags and user mentions, coupled with typical web constructs, such as email addresses and URLs, and other noisy sources, such as phone numbers, percentages, money amounts, time, date, and generic numbers. In this work, a set of pre-processing procedures, which has been tailored to translate tweets into a more conventional form sentences, is adopted. Most of the noisy entities are normalized because their particular instances generally do not contribute to the identification of the class within a sentence. Regarding date, email addresses, money amounts, numbers, percentages, phone numbers and time, this process is performed by using the ekphrasis tool[1] (Baziotis et al., 2017), which enables to individuate regular expressions and replace them with normalized forms.

3.3 AraBERT System Architecture

Among modern language modeling architectures, AraBERT (Antoun et al., 2020) is one of the most popular for Arabic language. Its generalization capability is such that it can be adapted to different down-stream tasks according to different needs, be it NER or relation extraction, question answering or sentiment analysis. The core of the architecture is trained on particularly large text corpora and, consequently, the parameters of the most internal layers of the architecture are frozen. The outermost layers are instead those that adapt to the task and on which the so-called fine-tuning is performed. An overview is shown in Figure 1.

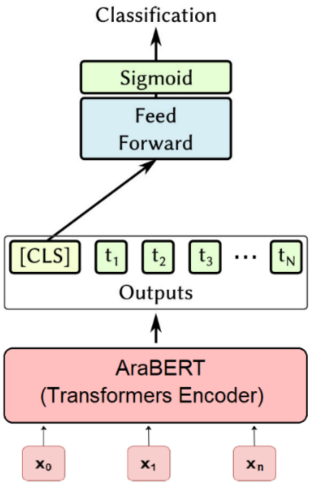

Figure 1: AraBERT architecture overview.

Going into details, one can distinguish two main architectures of AraBERT, the base and the large. The architectures differ mainly in four fundamental aspects: the number of hidden layers in the transformer encoder, also known as transformer blocks (12 vs. 24), the number of attention heads, also known as self-attention (Vaswani et al., 2017) (12 vs. 16), the hidden size of the feed-forward networks (768 vs. 1024) and finally the maximum sequence length parameter (512 vs. 1024), i.e., the maximum accepted input vector size. In this work, the base architecture is used, and the corresponding hyper-parameters are reported in Table 2.

Hyperparameter	**Value**
Attention heads	12
Batch size	6
Epochs	10
Gradient accumulation steps	16
Hidden size	768
Hidden layers	12
Learning rate	0.00002
Maximum sequence length	128
Parameters	136 M

Table 2: Hyper-parameters of the model

In addition, the AraBERT architecture employs two special tokens: [SEP] for segment separation and [CLS] for classification, used as the first input token for any classifier, representing the whole sequence and from which an output vector of the same size as the hidden size H is derived. Hence, the output of the transformers, i.e., the final hidden state of this first token used as input, can be denoted as a vector $C \in R^{H}$. The vector C is used as input of the final fully-connected classification layer. Given the parameter matrix $W \in R^{KxH}$ of the classification layer, where K is the number of

[1] https://github.com/cbaziotis/ekphrasis

categories, the probability of each category P can be calculated by the softmax function as:

$$P = softmax(CW^T)$$

3.4 Model Training

The whole classification model has been trained in two steps, involving firstly the pre-training of the AraBERT language model and then the fine-tuning of the outermost classification layer. The AraBERTv0.2-base (Antoun et al., 2020) is pre-trained on five corpora: OSCAR unshuffled and filtered, Arabic Wikipedia dump, the 1.5B words, Arabic corpus, the OSIAN corpus and Assafir news articles with a final corpus size equal to about 77 GB. The cased version was chosen, being more suitable for the proposed pre-processing method. The fine-tuning of the model was performed by using labeled tweets comprising the training set provided for the shared task. In particular, the fully connected classification layer was learned accordingly. During training, the loss function used was categorical cross-entropy. For this study, the hyper-parameters used are shown in Table 1. The maximum sequence length was reduced to 128, due to the short length of the tweets.

4 Evaluation and Results

To validate the results, we used the NLP4IF tweets dataset. The training and testing sets contain 90% and 10% of total samples, respectively. We split the training data set into 90% for training and 10% for validation.

In this section, we will introduce the different evaluation experiments of our implemented model on the test data. In Table 3, we present the accuracy, precision, recall, F1-score of each evaluation experiment on the test dataset.

Results show that our model can detect if the tweet is “harmfull to society” or “requires attention of government entities” with high accuracy (90% and 92% respectively), if the tweet “may be of interest to the general public” or “contains false information” with a very good accuracy (84% and 86% respectively), and if the tweet is “Harmfull”, “needs verification”, or “Verifiable” with fairly good accuracy (76%, 75%, and 74% respectively).

Evaluation Experiment	Recall	Precision	F1-score	Accuracy
Q1	73%	75%	70%	74%
Q2	87%	87%	87%	86%
Q3	83%	84%	84%	84%
Q4	76%	76%	76%	76%
Q5	74%	76%	71%	75%
Q6	91%	90%	90%	90%
Q7	93%	92%	90%	92%

Table 3: The evaluation results of our models on the test data.

In Table 4, we represent the evaluation results of our implementation models, which was conducted by the organizers based on our submitted predicted labels for the blind test set.

Recall	Precision	F1-score	Accuracy
67.7%	78.7%	66.4%	67.7%

Table 4: The evaluation results of our models on the blind test data.

5 Conclusions

The objective of this work was the introduction of an effective approach based on the AraBERT language model for fighting Tweets COVID-19 Infodemic. It was arranged in the form of a two-step pipeline, where the first step involved a series of pre-processing procedures to transform Twitter jargon, including emojis and emoticons, into plain text, and the second step exploited a version of AraBERT, which was pre-trained on plain text, to fine-tune and classify the tweets with respect to their Label.

Future work will be directed to investigate the specific contributions of each pre-processing procedure, as well as other settings associated with the tuning, so as to further characterize the language model for the purposes of COVID-19 Infodemic. Finally, the proposed approach will also be tested and assessed with respect to other datasets, languages and social media sources, such as Facebook posts, in order to further estimate its applicability and generalizability.

References

Hadeer Ahmed, Issa Traore, Sherif Saad. 2017. Detection of online fake news using N-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*. Springer, Cham, pages 127–138.

Hunt Allcott, and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. In *Journal of Economic Perspectives*. 31 (2): 211-36.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference.* pages 11-16.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*. San Diego, CA, USA, 16–17 June 2016; pages 747–754.

Alessandro Bondielli, and F. Marcelloni. 2019. A survey on fake news and rumour detection techniques. In *Inf. Sci. 497*. pages 38-55.

Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol. 2020. CLEF 2020 Working Notes. In *CEUR Workshop Proceedings, CEUR-WS.org*.

Souvick Ghosh, and Chirag Shah. 2018. Towards automatic fake news classification. *Proceedings of the Association for Information Science and Technology*. 55(1): pages 805–807.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854.

Maram Hasanain, Fatima Haouari, Reem Suwaileh, Zien Sheikh Ali,Bayan Hamdan, Tamer Elsayed , Alberto Barrón-Cedeňo, Giovanni Da San Martino , and Preslav Nakov. 2020. Overview of CheckThat! 2020 Arabic: Automatic Identi-fication and Verification of Claims in Social Media. In *arXiv:2007.07997*.

Ahmad Hussein, Abdulkarim Hussein, Nada Ghneim, and Ammar Joukhadar. 2020. DamascusTeam at CheckThat! 2020: Check worthiness on Twitter with hybrid CNN and RNN models. In *Cappellato et al. (2020).*

Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT). In *Applied Sciences (Switzerland), 9(19), [4062]*.

Srijan Kumar, and Neil Shah.2018. False Information on Web and Social Media: A Survey. In *arXiv:arXiv-1804*.

Athanasia Ntalla, and Stavros T. Ponis. (2015). Twitter as an instrument for crisis response: The Typhoon Haiyan case study. In *The 12th International Conference on Information Systems for Crisis Response and Management.*

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. Association for Computing Machinery, New York, NY, USA, pages 797–806.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov,and Anna Feldman. 2021. Findings of the {NLP4IF}-2021 Shared Task on Fighting the {COVID}-19 Infodemic and Censorship Detection. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda.*

Jieun Shin, Lian Jian, Kevin Driscoll, and Franois Bar. 2018. The diffusion of misinformation on social media. *Comput. Hum. Behav*. 83, C (June 2018), pages 278–287.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. In *Big Data. 8. 171-188*. 10.1089/big.2020.0062.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *KDD 2019 - Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 395-405.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016.

The spreading of misinformation online. In *Proceedings of the National Academy of Sciences Jan 2016*, 113 (3) pages 554-559; DOI: 10.1073/pnas.1517441113.

Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. 2017. Rumor Gauge: Predicting the Veracity of Rumors on Twitter. In *ACM Trans. Knowl. Discov*. Data 11, 4, Article 50 (August 2017), 36 pages.

Evan Williams, Paul Rodrigues, and Valerie Novak. 2020. Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. In *Cappellato et al. (2020).*

Xinyi Zhou, and Reza Zafarani. 2018. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. In *arXiv:arXiv-1812.*

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media.In *ACM Computing Surveys.* 51(2): pages 1–36.