PTUK-HULAT at ArAIEval Shared Task: Fine-tuned Distilbert to Predict Disinformative Tweets

Areej Jaber
Palestine Technical University - Khadoorie
a.jabir@ptuk.edu.ps

Paloma Martínez
Computer Science Department
Universidad Carlos III de Madrid
pmf@inf.uc3m.es

Abstract

Disinformation involves the dissemination of incomplete, inaccurate, or misleading information; it has the objective, goal, or purpose of deliberately or intentionally lying to others about the truth. The spread of disinformative information on social media has serious implications, and it causes concern among internet users in different aspects. Automatic classification models are required to detect disinformative posts on social media, especially on Twitter. In this article, DistilBERT multilingual model was fine-tuned to classify tweets either as dis-informative or not dis-informative in Subtask 2A of the ArAIEval shared task. The system outperformed the baseline and achieved F1 micro 87% and F1 macro 80%. Our system ranked 11 compared with all participants.

1 Introduction

Nowadays, social media has advanced to the point that it can compete with traditional media. The freedom of user participation could have negative consequences Dhiman (2023). Disinformation is one of the side effects of this intentionally aiming to mislead the truth that could affect negatively people in many fields like politics, and health, among others.

Spreading fake news can lead to misunderstanding, harm individuals or groups, damage reputation, or even influence public opinion and decision-making, Nasery et al. (2023). Thus, automatic detection of these kinds of data is a very important issue. For a while, it seemed so easy to detect disinformation data by domain experts or fact-checkers, but with daily huge propagation data in social media, more resources are needed to automate and speed up the process of detection of this kind of information.

Arabic language is one of the languages spoken in the world, with 422 million people including native and non-native speakers. It is the official language in 22 countries with at least 30 distinct dialects. However, there are three categories: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) Kadaoui et al. (2023). CA is the original Arabic language that has been used for over 1,500 years, and it is usually used in most Arabic religious texts. MSA is one of the official languages of the United Nations and is widely used in todays Arabic newspapers, letters, and formal meetings, which are also focused on by researchers. DA is spoken Arabic used in informal daily communication.

With the recent advanced improvements in the natural language processing (NLP) field and the evolution of large language modeling which is based on transformers architecture Wolf et al. (2020) the development of Arabic language solutions in the NLP field has evolved. To mention some previous efforts devoted to creating Arabic datasets to train and test systems, the Arabic fact-checking and stance detection corpus Baly et al. (2018) contains 422 claims: 219 false claims from Verify [1], and 203 true claims from Reuters. All these claims were made about the war in Syria and related Middle East political issues. Alkhair et al. (2019) describes an Arabic corpus of 342 rumors and 3,000 no rumors about death personalities.

Automatic disinformation classification is the most straightforward way of disinformation analysis. Some previous work has been developed in the Arabic language such as Harrag and Djahli (2022) that explored convolutional neural networks (CNNs) for fact-checking using Baly et al. (2018) to evaluate the proposal obtaining an accuracy averaged from 0.886 to 0.898. A more recent work Nassif et al. (2022) evaluated a transformer-based classifier to recognize fake news using Arabic word embeddings. Authors reported a performance accuracy of 98% using models such as

[1] https://verify-sy.com/

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 525–529

QuariBert-Bse and Arabic-BERT among others.

The remainder of this article is organized as follows: The task definition is described in Section 2. Section 3 describes the data set that was used, then an explanation of the baselines and the proposed system are given in Section 4. System evaluation is introduced in Section 5. Finally, conclusions are given in Section 7.

2 Task Definition

Text classification is a machine-learning process that assigns a document to one or more predetermined categories based on its content Abdulghani and Abdullah (2022). It is a key problem in NLP, with applications ranging from sentiment analysis to email routing to offensive language detection, spam filtering, and language identification.

Disinformation classification is a form of text classification that is normally identified as binary classification Mu et al. (2022). Given a set of labeled contexts that will be modeled as a feature f, the task aims to predict whether f is disinformative or not.

$$g(f) = \begin{cases} 1, & \text{if f is dis-informative text} \\ 0, & \text{if f is not dis-informative text} \end{cases}$$

where **g** is the function we want to learn from the available data. The combination of the features to obtain **g** can be done manually or automatically.

3 Data

The organizers of the ArAIEval shared task Hasanain et al. (2023) released a data set that aims to categorize a tweet whether it is a disinformative or not. These shared tasks represent continuous works on the Arabic language after Alam et al. (2022).

The data set is extracted from the Twitter website by Twarc package Mubarak et al. (2023). These tweets were extracted by using the word corona in Arabic in February and March 2020. Each sample in the data sets is composed of three fields, the ID which represents the sample identifier, the text field which includes the tweet text, and the label which represents the annotated label for the text either "**disinfo**" or "**no-disinfo**". Table 1 illustrates a set of examples of the provided data.

Three separate data sets were released in two phases. The training and the developing data set were released in the first phase, containing 14,147 and 2,115 samples respectively. Then the test data set was released in the second phase with 3729 samples. Table 2 shows the stats of the data provided by the organizers and it is clear that the data sets are imbalanced.

4 System

Baseline: In order to familiarize the participants with the task, the organizers provided two baselines in the code repository, random and majority baselines.

Proposed System: Pre-trained transformer-based architectures have recently proven to be particularly efficient at language modeling and understanding when trained on a large enough corpus. Bidirectional Encoder Representations from Transformers (BERT) Vaswani et al. (2017) is one of these models that gains the attention of the researchers due to its ability to predict words considering left and right context sides.

Two model sizes are released for BERT as modeling language goals, both of them depend on encoder architecture, BERT$_{large}$ and BERT$_{base}$. The main difference between them is the number of encoders. BERT$_{base}$ consists of a stack of 12 encoders, on the other hand, BERT$_{large}$ consists of a stack of 24 encoders. In addition, they differ in the number of hidden units (768, 1,024) and attention heads (12,16).

Despite the notable results of pre-trained BERT models, it has a drawback which makes it very slow due to its parameter numbers Han et al. (2021). So, the distillation process, which is known as a compression technique in which a small model (the student) is trained to mimic the behavior of a bigger model (the teacher) or an ensemble of models Gou et al. (2021) is produced to deal with this issue.

Based on the current resources for NLP, 90% of the worlds population speaks languages that do not benefit from recent language technologies due to the lack of resources Joshi et al. (2020). Arabic NLP is among these languages that still need more interest to make it mature Bourahouat et al. (2023).

Cross-language transfer is considered the main technique used for addressing the lack of resources in the target language, in which higher resource language models are adapted to the low resource language. The cross-lingual transfer could be achieved in two ways, the first one is by using a trained single high-resource language model, or

ID	Text	Label
0	"الله يلعن ابو الساعة اللي عرفنا فيها كورونا اقسم بالله ماناقص الا الطعوس والبدو يعرفونها	no-disinfo
1	حفل زفاف في القاهرة ...الشعب المصري هو اللي بجيب الجلطة لفيروس كورونا	disinfo
2	البقاء في المنزل يقينا من كورونا حفظ الله بلادنا من كل شر كلنامسؤول	no-disinfo

Table 1: Examples of the ArAIEval dataset.

Data set	Disinfo	No-disinfo	Total
Training	2,656	11,491	14,147
Development	397	1,718	2,115
Test	876	2,853	3,729

Table 2: Description of training, development, and test data sets.

the second way is by using multiple languages with varying amounts of resources. The idea behind these strategies is that the lower-resource language benefits from the model's learning of language invariant features from a huge amount of data in the high-resource language.

Thus, to overcome the low resources problems for the Arabic language and the slowness of the BERT model, in our proposed model we used a distilled multilingual version of BERT which was released by Sanh et al. (2019).

DistilBERT is a multilingual model that is trained in 104 different languages including the Arabic language from the Wikipedia website. Thus, the Distilbert model has 6 layers, 768 dimensions, and 12 heads, totalizing 134M parameters. Table 3 illustrates the main differences between BERT_$base$ and DistilBERT. As shown, the model was able to reduce the size of a BERT model by 40% while retraining 97% of its language understanding capabilities and being 60% faster.

In the following sub-sections, the description of two phases, Development, and evaluation, will be described in detail.

4.1 System Development Phase

In this phase, the organizers of the ArAIEval shared task Hasanain et al. (2023) released training and development datasets. First, fundamental cleaning and preprocessing were performed on both data sets to improve their quality. Hence, white spaces, punctuation marks, hashtags, URLs, special characters, and hyperlinks were removed from the texts, and the null values were dropped. Therefore, the final samples for the experiments were 14126, and 2110 for the training and developing data sets respectively.

For each sample, the labeling is converted to either 1 to represent "no-disinfo" or 0 to represent "disinfo".

As known, input IDs' (which encode the words of the text to sequences of numbers) and attention mask (to tell the model which numbers of input_ids to pay attention to or to ignore) vectors should be generated from the DistilBERT tokenizer for each sample. During the fine-tuning, the training data set was used for optimization and model parameters. On the other hand, the developing data set was used as an evaluation data set to validate the results of the model updates independent of the data it is trained on.

The training arguments were adjusted before running the experiment, the learning rate was 2e-5, and the number of epochs was 2.

4.2 Final Evaluation Phase

When the testing data set was released by the organizers, the same preprocessing and cleaning processes were done on the test samples. Then, the data set was fed to the generated model after converted it into real numbers from the previous phase to get the predictions and submitted to the task portal. After the submission was closed, the organizers published the golden standard for the test data set for analysis of the errors. Figure 1 shows the whole pipeline during the two phases.

5 Results

The system performance was evaluated by using F1 macro, and F1 micro; micro and macro averages are aggregation methods for the F1 score, a metric that is used to measure the performance of classification machine learning models.

F1 score is calculated per class, which means that if you want to calculate the overall F1 score for a dataset with more than one class you will need to aggregate in some way. Micro F1 score is the normal F1 formula but calculated using the total number of True Positives (TP), False Positives

	BERT	DistilBERT
Parameters (millions)	base: 110	base:66
Training Time (days)	8 X V100 X 12	4 times less than BERT
Performance	Outperforms state-of-the-art	3% degradation from BERT

Table 3: Comparison between BERT_$base$ and distilBERT

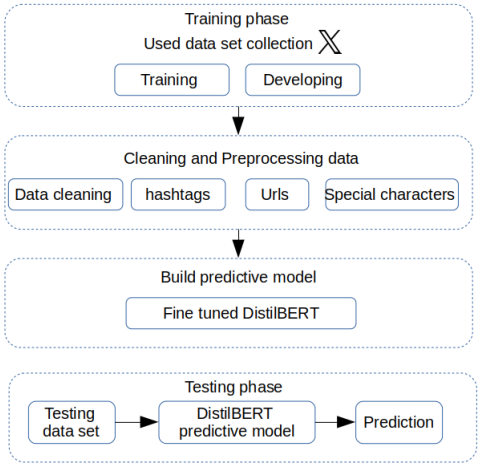

Figure 1: Overview of the proposed approach to predict the dis-informative tweets.

(FP), and False Negatives (FN), instead of individually for each class.

The formula for the micro F1 score is therefore:

$$Micro\,F_1 = \frac{TP}{TP + \frac{1}{2} \times (FP + FN)} \quad (1)$$

The Macro F1 score is the unweighted mean of the F1 scores calculated per class. It is the simplest aggregation for the F1 score. The formula for the macro F1 score is therefore:

$$Macro\,F_1 = \frac{sum(F1\,scores)}{number\,of\,classes} \quad (2)$$

In the training phase, the system achieved 81% F1 micro and 72% F1 macro. The system achieved 87% F1 micro and 80% F1 macro on the testing data set. To get further, the F1 score was computed per class; for the "disinfo" class the system achieved 68% , and "no-disinfo" class, the proposed model achieved 92% f1 score.

6 Discussion

In this work, a distilled multilingual version of BERT was fine-tuned to predict disinformative tweets that are extracted from the social media website Twitter. As shown in Table 4, the system outperformed the baselines in the two phases, training and evaluating phases.

Data set	F1 micro	F1 macro
our result (testing)	0.8675	0.7992
majority-baseline	0.7651	0.4335
random-baseline	0.5154	0.4764

Table 4: The performance of the proposed system compared with the baselines.

The data set which are provided by the organizers is imbalanced and this affects the results as shown in the result section. The proposed system failed to predict 494 samples in total, 346 samples related to "disinfo" class which is the minority class in the data set. On the other hand, 148 samples that were labeled with "no-disinfo" were predicted false from the data set.

Based on our in-depth failure analysis, we found that the system failed to predict correctly the examples containing English words in Arabic letters such as " فولو" which means "follow" in English. Another reason of failure is that some users repeat some characters in some words to express their emotions such as "طفللل".

7 Conclusion

In this work, we described our proposed system to classify Arabic tweets as either disinformative or not. Distilbert's multilingual model was fine-tuned on the task dataset. The system overcomes the baselines and achieves F1 micro 87% and F1 macro 80% on the testing data set.

The Arabic language is the official language of 22 countries and it is spoken by over 422 million people, but more efforts are needed to get benefits of the recent NLP technologies. Writing a foreign language in Arabic letter should be taken into account to improve the proposed model, in addition to using repeated characters to express emotions.

8 Acknowledgments

This work has been supported by Palestine Technical University - Khadoori (Palestine) and Madrid Regional Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17) and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation.

References

Farah A Abdulghani and Nada AZ Abdullah. 2022. A survey on arabic text classification using deep and machine learning algorithms. *Iraqi Journal of Science*, pages 409--419.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop, WANLP@EMNLP 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 8, 2022*, pages 108--118. Association for Computational Linguistics.

Maysoon Alkhair, Karima Meftouh, Kamel Smaïli, and Nouha Othman. 2019. An arabic corpus of fake news: Collection, analysis and classification. In *Arabic Language Processing: From Theory to Practice: 7th International Conference, ICALP 2019, Nancy, France, October 16--17, 2019, Proceedings 7*, pages 292--302. Springer.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. *arXiv preprint arXiv:1804.08012*.

GHIZLANE Bourahouat, MANAR ABOUREZQ, and NAJIMA DAOUDI. 2023. Systematic review of the arabic natural language processing: Challenges, techniques and new trends. *Journal of Theoretical and Applied Information Technology*, 101(3).

Dr Bharat Dhiman. 2023. Ethical issues and challenges in social media: A current scenario. *Available at SSRN 4406610*.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789--1819.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225--250.

Fouzi Harrag and Mohamed Khalil Djahli. 2022. Arabic fake news detection: A fact checking based deep learning approach. 21(4).

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abdelhakim Freihat. 2023. Araieval shared task: Persuasion techniques and disinformation detection in arabic text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Karima Kadaoui, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. *arXiv preprint arXiv:2308.03051*.

Yida Mu, Pu Niu, and Nikolaos Aletras. 2022. Identifying and characterizing active citizens who refute misinformation in social media. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 401--410.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and reasoning of deleted tweets before they are posted. *arXiv preprint arXiv:2305.04927*.

Mona Nasery, Ofir Turel, and Yufei Yuan. 2023. Combating fake news on social media: A framework, review, and future opportunities. *Communications of the Association for Information Systems*, 53(1):9.

Ali Bou Nassif, Ashraf Elnagar, Omar Elgendy, and Yaman Afadar. 2022. Arabic fake news detection based on deep contextualized embedding models. *Neural Computing and Applications*, 34(18):16019--16032.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38--45.