iCompass at NLP4IF-2021–Fighting the COVID-19 Infodemic

Wassim Henia
Oumayma Rjab
iCompass, Tunisia
{wassim.henia, oumayma.rjab}
@etudiant-isi.utm.tn

Hatem Haddad
Chayma Fourati
iCompass, Tunisia
{Hatem, Chayma}
@icompass.degital

Abstract

This paper provides a detailed overview of the system and its outcomes, which were produced as part of the NLP4IF Shared Task on Fighting the COVID-19 Infodemic at NAACL 2021. This task is accomplished using a variety of techniques. We used state-of-the-art contextualized text representation models that were fine-tuned for the down-stream task in hand. ARBERT, MARBERT,AraBERT, Arabic ALBERT and BERT-base-arabic were used. According to the results, BERT-base-arabic had the highest 0.748 F1 score on the test set.

1 Introduction

In recent years, there has been a massive increase in the number of people using social media (such as Facebook and Twitter) to share, post information, and voice their thoughts. The increasing number of users has resulted in the development of an enormous number of posts on Twitter. Although social media networks have enhanced information exchange, they have also created a space for anti-social and illegal activities such as spreading false information, rumors, and abuse. These anti-social behaviors intensify in a massive way during crisis cases, creating a toxic impact on society, either purposely or accidentally. The COVID-19 pandemic is one such situation that has impacted people's lives by locking them down to their houses and causing them to turn to social media. Since the beginning of the pandemic, false information concerning Covid-19 has circulated in a variety of languages, but the spread in Arabic is especially harmful due to a lack of quality reporting. For example, the tweet "مسائكم أخبار جميلة عن #كورونا 40 ثانية فقط م صاحب مبادرة تجميع العلماء لإيجاد علاج ضد #كرونا يعلنها على الهواء مباشرة أن فريق كامل من ضمنهم طبيب فرنسي اسمه “راولت” اكتشفوا أن دواء الملاريا هو الذي يعالج # كورونا_الجديد بنسبة 100% وتم تجربته على 40 مريّض #ترجمات_عبدالله_الخريّف" is translated as follows: "Good evening, good news, 40 seconds, the owner of the initiative to gather scientists to find a treatment against Corona announces on the air that an entire team, including a French doctor named "Raoult", discovered that the malaria treatment is the one that treats the new Corona, and it has been tried on 40 patients". This tweet contains false information that is harmful to the society and people believing it could be faced with real danger. Basically, we are not only fighting the coronavirus, but there is a war against infodemic which makes it crucial to identify this type of false information. For instance, the NLP4IF Task 2 is fighting the COVID-19 Infodemic by predicting several binary properties of a tweet about COVID-19 as follows: whether it is harmful, whether it contains a verifiable claim, whether it may be of interest to the general public, whether it appears to contain false information, whether it needs verification or/and requires attention. This is why we performed a multi-label classification using Arabic pretrained models including ALBERT Arabic (Lan et al., 2019), BERT-base-arabic (Devlin et al., 2018), AraBERT (Antoun et al., 2020), ARBERT(Abdul-Mageed et al., 2020), and MARBERT (Abdul-Mageed et al., 2020) with different hyper-parameters. The paper is structured as follows: Section 2 provides a concise description of the used dataset. Section 3 describes the used systems and the experimental setup to build models for Fighting the COVID-19 Infodemic. Section 4 presents the obtained results. Section 5 presents the official submission results. Finally, section 6 concludes and points to possible directions for future work.

2 Dataset description

The provided training dataset of the competition, fighting the COVID-19 Infodemic Arabic, consists of 2536 tweets and the development dataset con-

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 115–118
June 6, 2021. ©2021 Association for Computational Linguistics

sists of 520 tweets (Shaar et al., 2021). The data was labelled as yes/no questions answering seven questions:

1. Verifiable Factual Claim: Does the tweet contain a verifiable factual claim?
2. False Information: To what extent does the tweet appear to contain false information?
3. Interest to General Public: Will the tweet have an effect on or be of interest to the general public?
4. Harmfulness: To what extent is the tweet harmful to the society/person(s)/company(s)/product(s)?
5. Need of Verification: Do you think that a professional fact-checker should verify the claim in the tweet?
6. Harmful to Society: Is the tweet harmful for society and why?
7. Require attention: Do you think that this tweet should get the attention of government entities?

Questions 2,3,4 and 5 will be labelled as nan if the answer to the first question is no. The tweets are in Modern Standard Arabic (MSA) and no other Arabic dialect was observed. Data was preprocessed by removing emojis, URLs, punctuation, duplicated characters in a word, diacritics, and any non Arabic words.
We present an example sentence before and after preprocessing:

- Before preprocessing: #وزارة_الصحة : تلزم المشاركين من منسوبيها في حج 1442[1] باخذ لقاح كوفيد_19
- After preprocessing: وزاره_الصحه تلزم المشاركين من منسوبيها في حج 1442 باخذ لقاح كوفيد 19

3 System description

Pretrained contextualized text representation models have shown to perform effectively in order to make a natural language understandable by machines. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is, nowadays, the state-of-the-art model for language understanding, outperforming previous models and opening new perspectives in the Natural Language Processing (NLP) field. Recent similar work was conducted for Arabic which is increasingly gaining attention. In our work, we used three BERT Arabic variants: AraBERT (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2020), MARBERT (Abdul-Mageed et al., 2020) and Arabic BERT (Safaya et al., 2020). Added-on, we used the xlarge version Arabic Albert[2].

3.1 AraBERT

AraBERT (Antoun et al., 2020), was trained on 70 million sentences, equivalent to 24 GB of text, covering news in Arabic from different media sources. It achieved state-of-the-art performances on three Arabic tasks including Sentiment Analysis. Yet, the pre-training dataset was mostly in MSA and therefore can't handle dialectal Arabic as much as official Arabic.

3.2 ARBERT

ARBERT (Abdul-Mageed et al., 2020) is a large-scale pretrained language model using BERT base's architecture and focusing on MSA. It was trained on 61 GB of text gathered from books, news articles, crawled data and the Arabic Wikipedia. The vocabulary size was equal to 100k WordPieces which is the largest compared to AraBERT (60k for Arabic out of 64k) and mBERT (5k for Arabic out of 110k).

3.3 MARBERT

MARBERT, also by (Abdul-Mageed et al., 2020), is a large-scale pretrained language model using BERT base's architecture and focusing on the various Arabic dialects. It was trained on 128 GB of Arabic Tweets. The authors chose to keep the Tweets that have at least three Arabic words. Therefore, Tweets that have three or more Arabic words and some other non-Arabic words are kept. This is because dialects are often times mixed with other foreign languages. Hence, the vocabulary size is equal to 100k WordPieces. MARBERT enhances the language variety as it focuses on representing the previously underrepresented dialects and Arabic variants.

[1] https://t.co/6MEMHFMQj2

[2] https://github.com/KUIS-AI-Lab/Arabic-ALBERT

3.4 Arabic ALBERT

Arabic ALBERT[2] by (KUIS-AI-Lab) models were pretrained on 4.4 Billion words: Arabic version of OSCAR (unshuffled version of the corpus) filtered from Common Crawl and Recent dump of Arabic Wikipedia. Also, the corpus and vocabulary set are not restricted to MSA, but contain some dialectical Arabic too.

3.5 Arabic BERT

Arabic BERT (Safaya et al., 2020) is a set of BERT language models that consists of four models of different sizes trained using masked language modeling with whole word masking (Devlin et al., 2018). Using a corpus that consists of the unshuffled version of OSCAR data (Ortiz Suárez et al., 2020) and a recent data dump from Wikipedia, which sums up to 8.2B words, a vocabulary set of 32,000 Wordpieces was constructed. The final version of corpus contains some non-Arabic words inlines. The corpus and the vocabulary set are not restricted to MSA, they contain some dialectical (spoken) Arabic too, which boosted models performance in terms of data from social media platforms.

3.6 Fine-tuning

We use these pretrained language models and build upon them to obtain our final models. Other than outperforming previous techniques, huge amounts of unlabelled text have been used to train general purpose models. Fine-tuning them on much smaller annotated datasets achieves good results thanks to the knowledge gained during the pretraining phase, which is expensive especially in terms of computational power. Hence, given our relatively small dataset, we chose to fine-tune these pretrained models. The fine-tuning actually consists of adding an untrained layer of neurons on top of the pretrained model and only tweaking the weights of the last layers to adjust them to the new labelled dataset.
We chose to train our models on a Google Cloud GPU using Google Colaboratory. The average training time of one model is around 10 minutes. We experimented with Arabic ALBERT, Arabic BERT, AraBERT, ARBERT and MARBERT with different hyperparameters.
The final model that we used to make the submission is a model based on BERT-base-arabic, trained for 10 epochs with a learning rate of 5e-5, a batch size of 32 and max sequence length of 128.

4 Development dataset results

We have validated our models through the development dataset as mentioned in the data section. The results of all models were close but the BERT-base-arabic achieved the best results performing 78.27% F1 score. For reference, and to compare with other models, we also showcase the results obtained with ARBERT, AraBERT, and Arabic ALBERT in Table 1.

- The best ARBERT model was achieved using 2e-5 learning rate, 32 batch size, 10 epochs, 128 max length.
- The best MARBERT model was achieved using 6e-5 learning rate, 32 batch size, 10 epochs, 128 max length.
- The best AraBERT model was achieved using 4e-5 learning rate, 32 batch size, 10 epochs, 128 max length.
- The best ALBERT Arabic model was achieved using 2e-5 learning rate, 16 batch size, 8 epochs, 128 max length.

5 Official submission results

Table 1 presents the results obtained over development data for Fighting COVID-19 Infodemic. The result of all the models used are very close. However, bert-base-arabic outperformed all other models. This may be due to the pretrained data for bert-base-arabic. The final version has some non-Arabic words inlines. Also, the corpus of bert-base-arabic and vocabulary set are not restricted to MSA, they contain some dialectical Arabic too which can boost the model performance in terms of data from social media.
Table 2 reviews the official results of iCompass system against the top three ranked systems.
Table 3 presents the official results per class of iCompass system.

6 Conclusion

This paper describes the system built in the NLP4IF 2021 shared Task , along with comprehensive results. Various learning techniques have been investigated using five language models (Arabic ALBERT, AraBERT, ARBERT, MARBERT, and BERT-base-arabic) to accomplish the task of Fighting the COVID-19 Infodemic. The results show

Model	F1 Score	Precision	Recall
ARBERT	0.7734	0.8153	0.7502
MARBERT	0.7654	0.7879	0.7662
AraBERT	0.7635	0.8223	0.7403
ALBERT-Arabic	0.7603	0.8202	0.7399
BERT-base-Arabic	**0.7827**	**0.8255**	**0.7712**

Table 1: Models performances on the Dev dataset.

Team	Rank	F1 Score
R00	1	0.781
iCompass	**2**	**0.748**
HunterSpeechLab	3	0.741
advex	4	0.728

Table 2: Official Results on Test set and ranking as reported by the task organisers (Shaar et al., 2021).

Questions	F1 Score
Q1	0.797
Q2	0.746
Q3	0.881
Q4	0.796
Q5	0.544
Q6	0.885
Q7	0.585

Table 3: Official Results for each classifier as reported by the task organisers (Shaar et al., 2021).

that BERT-base-arabic outperforms all of the previously listed models in terms of overall performance, and was chosen for the final submission. Future work will include developing larger contextualized pretrained models and improving the current COVID-19 Infodemic Detection .

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media.

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*, NLP4IF@NAACL' 21, Online. Association for Computational Linguistics.