iCompass at CheckThat! 2022: ARBERT and AraBERT for Arabic Checkworthy Tweet Identification

Bilel Taboubi[1], Mohamed Aziz Ben Nessir[1], and Hatem Haddad[1 [0000−0003−3599−7229]]

iCompass, 49 rue de Marseille, Tunis, Tunisia
{bennessir.mohamedaziz,bileltaboubi20,haddad.hatem}@gmail.com

Abstract. This paper provides a detailed overview of systems and its achieved results, which were produced as part of CLEF2022 - CheckThat! Lab Fighting the COVID-19 Infodemic and Fake News Detection. The task was carried out using a variety of techniques. We used transformers that achieved state-of-the art in many NLP tasks such as text classification.Transformers pre-trained models Arabic BERT, ARBERT, MARBERT, AraBERT, Arabic ALBERT and BERT base arabic were used and fine-tuned for the down-stream task in hand binary classification of Arabic tweets.According to the results, AraBERT had the highest 0.462 F1 score on the test set of subtask 1A and ArBERT had the best F1 score 0.557 on the test set of subtask 1C.

Keywords: GRU · ARBERT · AraBERT · Arabic.

1 Introduction

The spread of fake news misinformation is increasing in a huge number and almost turning to be unlimited due the increase of social media users and platforms allowing anyone these days can create and join and share articles and information in social medias platforms pretending to be a news agency or a popular person and this is causing serious problems to society, partly due to the fact that more and more people only read headlines or highlights of news assuming that everything is reliable, instead of carefully analysing whether it can contain distorted or false information Harmful Speech is particularly widespread in online communication due to users' anonymity and the lack of hate speech detection tools on social media platforms. Consequently, Harmful speech detection has determined a growing interest in using Machine/Deep Learning techniques to address this issue [11]. The increase of social media users conducted to a uncontrollable amount of information shared daily, making it impossible to coverd by manual fact checking sites where organizations and researchers began to move for a creation of automated systems with an aim to solve the mess caused by these misinformation. This paper focus on Subtask 1A and 1C in Arabic from CheckThat, a lab contest with various tasks for competitors **??**. This year,

the lab offerend the following three main tasks: Detecting Check-Worthy Claims (Task 1), Previously Fact-Checked Claims (Task 2), and Fake News Detection (Task 3).Task 1 was divided into four subtasks while the rest Tasks has two subtasks.Subtask 1A was provided in six different languages (Arabic, Bulgarian, Dutch, English, Spanish and Turkish) while Subtask 1C was in the same languages as Subtask 1A excluding Spanish.

2 Tasks Definition:

TASK 1 is presenting a supervised text classification problem aiming to classify tweets into categories based on their content for the purpose of developing an automated system to identify worth checking tweets out of unworthy ones. Subtask 1C and Subtask 1A offered in Arabic language aims at classifying real-world tweets into defined binary labels. Tweets has been labelled as worth checking or not for Subtask 1A and harmful or not for Subtask 1C.

3 Related Work

3.1 Accenture team approaches at CheckThat!

The winner team from the recent years contest CheckThat! Lab 2020 [1] and 2021 [2] proposed a solution that introduces two models: BERT and RoBERTa adding a mean-pooling and a dropout layers on their top before classification layer. This team scored 1st position in both the Arabic language for two consecutive years with 0.7 p@30 with AraBERTv1.0 for 2020 trained on augmented data and 0.658 MAP in 2021 year and scored 1st for English language in only 2020 for Subtask 1A.They used data augmentation in particular, they generated synthetic training data using lexical substitution to create additional synthetic examples for the positive class and used machine translation, translating Arabic data to English and then to Arabic again.

3.2 Deep Learning Approaches for Covid19 Fake News Detection

A work was done on Contraint@AAAI 2021 Covid-19 Fake news detection dataset, evaluating Deep learning approches using supervised algorithms for text classification based on Convolutional Neural Networks [3]. (CNN), Long Short Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT). Dataset was preprocessed by the following steps, Removal of HTML tags, Convert Accented Characters to ASCII, Expand contractions, Removal of Special Characters, Noise Removal, Normalization, Stemming, and Stop-words Removal. All Transformers based models outperformed basic models with a difference of 3-4% in accuarcy . The best accuracy was reached using language model pretraining on BERT 98.41%.

3.3 Context-Aware Misinformation Detection

Experiments was done by IEEEAcces [4] using the open source Fake News Corpus dataset available on Github, the dataset has been used for determining the veracity of news articles. Text Preprocessing techniques applied on news article transforms the text to UTF-8, removes stop words and punctuation, lemmatize the sentences to get them back to their root form and transforms the text to lowercase. Many deep learning architecture were applied such as LSTM, GRU, CNN with different word embeddings techniques Word2Vec, FastText and GloVe. where the best results was obtained using a Recurrent Convolutional Neural Network based architecture.

4 Data Description

4.1 Subtask 1A: Check-worthiness of tweets

Dataset Statistics The dataset for CLEF Subtask 1A contains 3439 tweets written in Arabic dialect divided on train, development and test sets as shown in Table , labelled with binary labels, 1 for worthy claims with a percentage of 61.4 and 0 for unworthy for the rest 38.6% of the data.

Type	Train	Dev	Test	Total
Worthy claims	962	100	266	1328
Unworthy claims	1551	135	425	2111

Table 1. Task 1A dataset statistics..

4.2 Subtask 1C: Harmful tweet detection

Dataset Statistics The provided training dataset of the CLEF Subtask 1C harmful tweet detection is about 5k tweets, labelled with the 2 categories Normal and Harmful. 81% of the tweets are Normal and 19% are Harmful as shown in Table 2.

Type	Train	Dev.	Dev. Test	Total
Harmful	678	60	189	927
Normal	2946	276	805	4027

Table 2. Task 1C dataset statistics..

The dataset is highly unbalanced so we downsampled the Normal tweets by 65% making it roughly 1.5 times the size of the harmful tweets.

5 Data preparation

We experimented with various preprocessing techniques, such as removing emojis, normalizing hashtags, removing Latin characters, removing URLs, data normalization, deleting tashkeel and the letter madda from texts, as well as duplicates etc. The best results were given on the raw unpreprocessed data for each of subtasks 1A and 1C.

6 Pre-trained Models

Different pre-trained models were used in order to achieve the best results when fine-tuning it in a multi-task fashion.

6.1 AraBERT

AraBERT (V2) [5], is a BERT based model for Modern Standard Arabic Language understanding, trained on 70M sentences from several public Arabic datasets and news websites. It was fine-tuned on 3 tasks: Sequence Classification, Named Entity Recognition and Question Answering. It was reported to achieve state-of-the-art performances even on Arabic dialects after fine-tuning by (Abu Farha and Magdy, 2020).

6.2 Bert base Arabic

The Arabic BERT model [6] was trained on 8.2 billion words using the Arabic version of OSCAR, Recent dump of Arabic Wikipedia and other Arabic resources which sum up to 95GB of text which was filtered using Common Crawl. 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.

6.3 ARBERT

ARBERT [10] is also a Bert based model trained on 61GB of Modern Standard Arabic text (6.5B tokens) gathered from books, news articles, crawled data and Wikipedia.

6.4 MARBERT

MARBERT [10] is a large-scale pretrained language model using the BERT base's architecture. MARBERT is trained on on 128 GB of tweets from various Arabic dialects containing at least 3 Arabic words. With very light preprocessing the tweets were almost kept at their initial state to retain a faithful representation of the naturally occurring text.

7 Results

7.1 Subtask 1A: Check-worthiness of tweets

Pre-trained models AraBERT and BERT base Arabic were trained and finetuned with the following architecture:

- Input layer
- Bert model
- A gated recurrent unit with 128 untis and 0.3 probability for dropout.
- Dense layer with 50 units and Relu activation function
- A dropout layer with 0.1 probability.
- Dense layer with a Sigmoid activation function and one unit

The average training time of a model is around 6 minutes. Best results achieved by each pre-trained model is presented in the table 3 where they got trained on the train set,validated on development set and tested with test set.

Type	F1	Accuracy	Precision	Recall
AraBERT	0.590	0.536	0.453	0.844
BERT base Arabic	0.576	0.672	0.601	0.554

Table 3. Task 1A Pre-trained models results on dev-test set.

The submitted model was AraBERT, trained with a 10 epochs, 2e-5 learning rate for Adam optimizer, a sequence length of 150, 32 batch size and binary cross entropy loss function.The model achieved F1_score 0.590 on the dev-test set,and 0.462 on the submission test set to get rank 3 in Subtask 1A Arabic leaderboard as shown in the table 4.

Participants (userid/team-name)	Subtask	F1 (postive class)
elfsong	Subtask-1A-Checkworthy-Arabic	0.628
mkutlu	Subtask-1A-Checkworthy-Arabic	0.495
HatemHaddad	Subtask-1A-Checkworthy-Arabic	0.462

Table 4. Top 3 on Subtask 1A Arabic leaderboard

7.2 Subtask 1C: Harmful tweet detection

All the models were finetuned with :

- A gated recurrent unit with 256 untis and 0.5 dropout.
- A gated recurrent unit with 128 untis and 0.4 dropout.
- A gated recurrent unit with 64 untis and 0.3 dropout.
- 1-dimensional convolution neural network with 64 units and a kernel size of 3.
- A 0.3 dropout layer.
- A layer to concatenate Global Average Pooling 1D and Global Maximum Pooling 1D of the previous output.
- A 0.05 dropout layer.
- A final dense layer with a Sigmoid activation function and one unit.

All of the models results are presented in table 5.

Type	F1	Accuracy	Precision	Recall
ARBERT	0.775	0.905	0.857	0.707
AraBERT	0.750	0.890	0.867	0.661
MARBERT	0.7	0.885	0.703	0.696

Table 5. Task 1C Pre-trained models Dev results.

the best results were achived with ARBERT, The submitted model was trained with a total of 16 epochs. The first 4 epochs were only used to warm up the GRU layers, we froze ARBERT and trained them with a learning rate of 1e-4 and then and for the rest 12 epochs we unfroze ARBERT and used a learning rate of 1e-5. For both parts we used Adam optimizer, a batch size of 64 and a binary cross entropy loss function. The model achieved an F1 score of 0.557 on the test set and got rank 1 among the subtask participants are shown in the table 6.

Participants (userid/team-name)	Subtask	F1 (postive class)
HatemHaddad	Subtask-1C-Harmful-Arabic	0.557
mkutlu	Subtask-1C-Harmful-Arabic	0.268
random-baseline	Subtask-1C-Harmful-Arabic	0.118

Table 6. Top 3 on Subtask 1C Arabic leaderboard

8 Discussion

8.1 Subtask 1A: Check-worthiness of tweets

BERT base Arabic and AraBERT choice for this subtask was based on recent studies.However Arabert overperformed BERT base Arabic and reached the best

results since this last was trained with more vocabulary, a corpus with a large vocabulary more than 8.6B words. Both F1 scores attained by models are low and that is due the imbalance presented in the data plus an assemblance between worthy and unworthy tweets text from semantic side.

8.2 Subtask 1C: Harmful tweet detection

Different language models were used in this work. However, ARBERT achieved the best results. This was the case because it was pre-trained on modern standard arabic text from tweets with little no normalization therefore works better for our case. In addition, the data imbalance further illustrated in figure 1 decreased the model performance causing it to easily overfit on the training dataset.

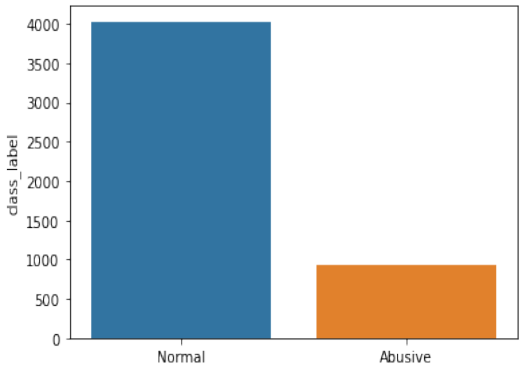

Fig. 1. Subtask 1c harmful speech statistics.

9 Conclusion

In this paper, we demonstrated the performance of gated recurrent unit for each fo the subtasks Harmful tweet detection and Check-worthiness of tweets by fine-tuning the pre-trained models ARBERT and AraBERT. Despite the small sized annotated data, the model achieved satisfactory results.

With respect to the models, further work should explore meta-learning, Focal loss, semi-supervised learning.

As for the data, further work should focus on the exploring other augmentation and resampleing strategies as well as collectiong more harmful tweets for Subtask1C, and feature extracting such accounts types, as number of likes, number of shares from tweet links provided within the data for more distinguishability between the worthy and unworthy claims.

References

1. Evan Williams, Paul Rodrigues, Valerie Novak: Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. Working Notes of CLEF (2020) https://doi.org/https://doi.org/10.48550/arXiv.2009.02431
2. Evan Williams, Paul Rodrigues, Valerie Novak: Accenture at CheckThat! 2020: Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation. Working Notes of CLEF (2021) https://doi.org/https://doi.org/10.48550/arXiv.2107.05684
3. Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, Raviraj Joshi: Evaluating Deep Learning Approaches for Covid19 Fake News Detection.(2021) https://doi.org/arXiv:2101.04012v2
4. VLAD-IULIAN ILIE , CIPRIAN-OCTAVIAN TRUICA , ELENA-SIMONA APOSTOL , ADRIAN PASCHKE: Context-Aware Misinformation Detection:A Benchmark of Deep Learning Architectures Using Word Embeddings.(2021) https://doi.org/https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9634064
5. AEl Moubtahij, H., Hajar A., El Bachir T.: AraBERT transformer model for Arabic comments and reviews analysis. Int J Artif Intell **11**(1), 79-387 (2022)https://doi.org/arXiv:2003.00104v4
6. Safaya Ali, Abdullatif Moutasem, Yuret Deniz: KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media (2020) https://doi.org/https://www.aclweb.org/anthology/2020.semeval-1.271
7. Nakov, Preslav and Barrón-Cedeño, Alberto and Da San Martino, Giovanni and Alam, Firoj and Struß, Julia Maria and Mandl, Thomas and Míguez, Rubén and Caselli, Tommaso and Kutlu, Mucahid and Zaghouani, Wajdi and Li, Chengkai and Shaar, Shaden and Shahi, Gautam Kishore and Mubarak, Hamdy and Nikolov, Alex and Babulkov, Nikolay and Kartal, Yavuz Selim and Beltrán, Javier: The CLEF-2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection,**Advances in Information Retrieval**, 416-428 (2022)
8. Nakov, Preslav and Barrón-Cedeño, Alberto and Da San Martino, Giovanni and Alam, Firoj and Struß, Julia Maria and Mandl, Thomas and Míguez, Rubén and Caselli, Tommaso and Kutlu, Mucahid and Zaghouani, Wajdi and Li, Chengkai and Shaar, Shaden and Shahi, Gautam Kishore and Mubarak, Hamdy and Nikolov, Alex and Babulkov, Nikolay and Kartal, Yavuz Selim and Beltrán, Javier and Wiegand, Michael and Siegel, Melanie and Köhler, Juliane : Overview of the CLEF-2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection (2022)
9. Nakov, Preslav and Barrón-Cedeño, Alberto and Da San Martino, Giovanni and Alam, Firoj and Míguez, Rubén and Caselli, Tommaso and Kutlu, Mucahid and Zaghouani, Wajdi and Li, Chengkai and Shaar, Shaden and Mubarak, Hamdy and Nikolov, Alex and Kartal, Yavuz Selim and Beltrán, Javier : Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets (2022)
10. Abdul-Mageed, Muhammad and Elmadany : ARBERT & MARBERT: deep bidirectional transformers for Arabic (2020)
11. Schmidt, Anna and Wiegand, Michael : A Survey on Hate Speech Detection using Natural Language Processing **Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media**,1–10 (2017)