

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

Bilel Taboubi<sup>1</sup>, Mohamed Aziz Ben Nessir<sup>1</sup> and Hatem Haddad<sup>1</sup>

<sup>1</sup>*iCompass, Emeraude Palace, Rue du Lac Windermère, Les Berges du Lac, Tunis 1053*

## Abstract

This paper provides a detailed overview of systems and its achieved results, which were produced as part of CLEF2022 - Check- That! Lab Fighting the COVID-19 Infodemic and Fake News Detection. The task was carried out using transformers pre-trained models Arabic BERT, ARBERT, MARBERT, AraBERT, Arabic ALBERT and BERT base arabic. The models were fine-tuned for the down-stream task in hand, binary classification of Arabic tweets. According to the results, AraBERT attained the highest 0.462 F1 score on the test set of subtask 1A and ARBERT attained 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 and almost turning to be unlimited due to 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 harmful 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 [1]. The increase of social media users conducted to a uncontrollable amount of information shared daily, making it impossible to be covered 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 [2]. This year, the lab offered the following three main tasks: Detecting Check-Worthy Claims (Task 1), Fact Checking Claims (Task 2), and Fake News Detection (Task 3). Task 1 was divided into four subtasks and the rest of the tasks each contain two subtasks. Both Subtask 1A and 1C where

---


*CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy*

✉ [bileltaboubi20@gmail.com](mailto:bileltaboubi20@gmail.com) (B. Taboubi); [mohamedaziz.benessir@etudiant-isi.utm.tn](mailto:mohamedaziz.benessir@etudiant-isi.utm.tn) (M. A. B. Nessir); [haddad.hatem@gmail.com](mailto:haddad.hatem@gmail.com) (H. Haddad)

🆔 0000-0003-3599-7229 (H. Haddad)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

provided in six different languages (Arabic, Bulgarian, Dutch, English, Spanish and Turkish). Detecting Check-Worthy Claims (Task 1) presents a supervised text classification problem aiming to classify tweets into categories based on their content, the purpose is to develop an automated system to identify trust unworthy tweets.

## 2. Related Work

The winner team from the recent years contest CheckThat! Lab 2020 [3] and 2021 [4] proposed a solution using two models BERT and RoBERTa then adding a mean-pooling, a dropout layers and finally a classification layer.

They also 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 to translate Arabic data to English and then to Arabic again.

The paper [5] evaluates Deep learning approaches using supervised algorithms for text classification based on Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT) for Fake news detection. The dataset was preprocessed as following, 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 accuracy . The best accuracy was reached using language model pretraining on BERT 98.41%.

Experiments were done by IEEEAccess [6] using the open source Fake News Corpus dataset available on Github, the dataset has been used for determining a veracity of news articles. Text Preprocessing techniques were applied on news article to transform the text to UTF-8, remove stop words and punctuation, lemmatize the sentences to get them back to their root form and transform the text to lowercase. Many deep learning architectures were applied such as LSTM, GRU, CNN with different word embedding techniques Word2Vec, FastText and GloVe.

## 3. Data Description

### 3.1. Subtask 1A: Check-worthiness of tweets

#### 3.1.1. Dataset Statistics

The dataset for CLEF Subtask 1A contains 3439 tweets written in Arabic dialect, the data set had originally 4 parts, train, dev, dev test, and test but we reassembled it into 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.

### 3.2. Subtask 1C: Harmful tweet detection

#### 3.2.1. 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

**Table 1**

Task 1A dataset statistics.

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

19% are Harmful as shown in Table 2. Again this data set was also reassembled.

**Table 2**

Task 1C dataset statistics..

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

The dataset is highly unbalanced so we downsampled the Normal tweets, we tried multiple combinations and percentages but down sampling it to 65%, as in making it roughly 1.5 times the size of the harmful tweets gave the best results.

## 4. 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.

## 5. 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.

### 5.1. AraBERT

AraBERT (V2) [7], 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.

## 5.2. Bert base Arabic

The Arabic BERT model [8] 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.

## 5.3. ARBERT

ARBERT [9] 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.

## 5.4. MARBERT

MARBERT [9] 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.

# 6. Results

## 6.1. 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 units 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

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.

**Table 3**

Task 1A Pre-trained models results on dev 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

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 set, and 0.462 on the submission test set to get rank 3 in Subtask 1A Arabic leaderboard as shown in the table 4.

**Table 4**  
Top 3 on Subtask 1A Arabic leaderboard

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

## 6.2. 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.

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

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

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, the subtask participants are shown in the table 6.

**Table 6**

Top 3 on Subtask 1C Arabic leaderboard

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

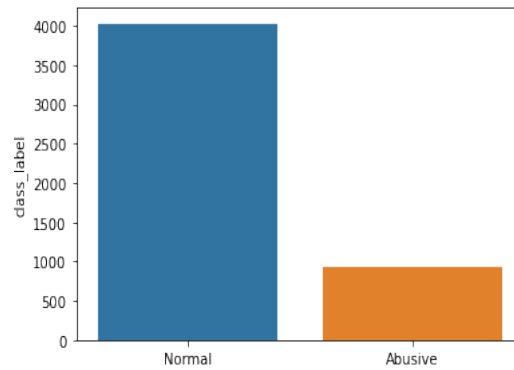
## 7. Discussion

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

### 7.2. 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.

**Figure 1:** Subtask 1c harmful speech statistics.

## 8. 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 resampling strategies as well as collectiong more harmful tweets for Subtask1C, and feature extracting features like account 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] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: <https://aclanthology.org/W17-1101>. doi:10.18653/v1/W17-1101.
- [2] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: N. Faggioli, Guglielmo and Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.
- [3] E. Williams, P. Rodrigues, V. Novak, Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, 2020. URL: <https://arxiv.org/abs/2009.02431>. doi:10.48550/ARXIV.2009.02431.
- [4] E. Williams, P. Rodrigues, S. Tran, Accenture at checkthat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, 2021. URL: <https://arxiv.org/abs/2107.05684>. doi:10.48550/ARXIV.2107.05684.
- [5] A. Wani, I. Joshi, S. Khandve, V. Wagh, R. Joshi, Evaluating deep learning approaches for covid19 fake news detection, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, 2021, pp. 153–163. URL: [https://doi.org/10.1007/978-3-030-73696-5\\_15](https://doi.org/10.1007/978-3-030-73696-5_15). doi:10.1007/978-3-030-73696-5\_15.
- [6] V.-I. Ilie, C.-O. Truică, E. S. Apostol, A. Paschke, Context-aware misinformation detection: A benchmark of deep learning architectures using word embeddings, IEEE Access PP (2021) 1–1. doi:10.1109/ACCESS.2021.3132502.
- [7] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, 2020. URL: <https://arxiv.org/abs/2003.00104>. doi:10.48550/ARXIV.2003.00104.
- [8] A. Safaya, M. Abdullatif, D. Yuret, KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona

(online), 2020, pp. 2054–2059. URL: <https://aclanthology.org/2020.semeval-1.271>. doi:10.18653/v1/2020.semeval-1.271.

- [9] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, Arbert & marbert: Deep bidirectional transformers for arabic (2021). URL: <https://arxiv.org/abs/2101.01785>. doi:10.48550/ARXIV.2101.01785.