

iCompass at CheckThat! 2021: Identifying Check-Worthy Arabic Tweets

Oumayma Rjab¹, Hatem Haddad¹, Wassim Henia¹ and Chayma Fourati¹

¹*iCompass, 49 rue de Marseille, 1001, Tunis, Tunisia*

Abstract

A news reader has the ability to respond, express, and share opinions with others in a highly interactive and quick manner in this digital age of news consumption. Because of the limited capacity of large corporations and individuals to check news on the Internet, misinformation has found its way into our everyday lives. In this paper, we introduce the strategies used by the iCompass Team for the CLEF2021 CheckThat! Lab, Task 1, on Arabic. This shared task evaluated whether a claim in social media text should be professionally fact checked or not. Three Arabic versions of BERT were experimented and fine tuned for this task. For the final submission, an ensemble of Arabic BERT and AraBERT were used which made our team in the 4th place with 0,597 Mean Average Precision score.

Keywords

Fact-Checking, Social Media, Twitter, BERT

1. Introduction

With the COVID-19 pandemic situation, a rapid increase in social media usage was noticed. In measures, during 2020, 490 million new users joined indicating a more than 13% of increase [1]. This growth is mainly due to the impacts on day-to-day activities and information sharing about daily events. As a drawback of these exponential growths, the spread of false and harmful information has been quite frequent. As a result, a number of initiatives to fact-check claims of general interest and to confirm or to debunk them were conducted. Due to the time-consuming nature of manual fact-checking, automated methods have been suggested as a quicker alternative. Even with automated methods, it is impossible to fact-check every argument, and automated fact-checking systems are substantially less accurate than human experts. Furthermore, pre-filtering and prioritizing what should be sent to human fact-checkers is needed. The role of check-worthiness estimation, which is seen as a critical first phase in the general fact-checking pipeline, has steadily gained attention alongside the need to prioritize.

The remainder of the paper is organized as follows: Section 2 briefly presents the CheckThat! lab and describes the task. Section 3 describe the given data for the task. Section 4 presents the preprocessing techniques applied to the tweet. Section 5 presents the pre-trained BERT models used in this paper. Section 6 presents details of our approach for Task-1. Section 7 describes official results and ranking. Section 8 summarizes our work and gives future research directions.


CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ oumayma.rjab@etudiant-isi.utm.tn (O. Rjab); hatem@icompass.digital (H. Haddad);

wassim0henia@gmail.com (W. Henia); chayma@icompass.digital (C. Fourati)



© 2021 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)

2. Related Work

In a related area of study, the winning team Accenture Team presents their methods for the CLEF2020 CheckThat! Lab Task 1; assessed if a claim made in a social media text could be properly fact-checked. Thus, for the Arabic challenge they have fine-tuned BERT models in Arabic and illustrate the use of Back Translation to enhance and balance the Minority Class. The final models were fine-tuned by 2 epochs, $2e-05$ of a learning rate, Adam optimization, and a batch size of 32 after a grid search. they used a BERT sequence classification function from Huggingface and applied a linear layer above the pooled output. The output was then provided to a softmax function and tweets in each class of topic were ranked by the difference between the positive and the negative class probabilities. The best results were achieved with AraBERT v0.1 and Upsampling the data they a mAP score of 0.6232 [2].

3. Task Description

The CheckThat! 2021 [3] is intended to combat misinformation and disinformation in social media, political debates, and news by focusing on three major tasks that contribute to the fact-checking pipeline.

The first step in fact-checking systems is check-worthiness, which predicts which sentences should be prioritized for fact-checking. Each claim is ranked based on its importance to the topic. A check-worthy tweet is the one that includes a claim that is of interest to a wide audience. That is why ranking the tweet is critical since its large audience would aid in the spreading of the content. Our work focuses only on the Arabic subtask-1A [4] first because we are Arabic native speakers also because Arabic is a challenging language and we wanted to explore more about it.

4. Data Description

The dataset is provided by the CLEF2021 CheckThat! lab and publicly available on GitLab¹. It includes tweets on different topics annotated as either Check-Worthy or not, defining a binary classification task. The following presents two examples of a check-worthy and non-check-worthy tweets respectively:

- #عاجل نتنياهو: يسعدني أن أرى سفراء الإمارات والبحرين وعمان هنا وهذا مؤشر لا " لبصقة _القرن <https://t.co/hbdlhrTNZe>."
- اقسم بالله مش مبالغة بس شعور القهر والعجز احتلني وقلبي وجعني من كمية الدل "

¹https://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/tree/master/task1

والاهانة يلي سمعتها، الله يلعن كل واحد كان السبب او ساهم فيها ويلعن كل المطيعين وكل
 "رئيس عربي ساكت ##تسقط_صفقة_القرن ##لا_لبصقة_القرن

The English translation of the preceding tweets are as follow :

- #urgent|| Netanyahu: I am happy to see the ambassadors of the UAE, Bahrain and Oman here, and this is an indication not only of the future, but of the present. #No Spit of the Century. <https://t.co/hbdlhrTNZe>.
- I swear to God, it is not an exaggeration, but the feeling of oppression and powerlessness occupied me and my heart hurt from the amount of humiliation and insult that I heard. #Down with the Deal of the Century #No Spit of the Century

Additionally, the Train set has 3439 tweets for training; only 22.2% of them are check-worthy and 77.8% are not check-worthy. The test set has 600 tweets for evaluation and it contains only two topics which are different than the train set topics.

We can observe the highly imbalanced data distribution, and when we checked the check-worthiness per topic, we noticed that certain topics were closely balanced while others were not, as shown in Figure 1. For example, the topic CT20-AR-08 had 428 non-check-worthy and 20 check-worthy tweets.

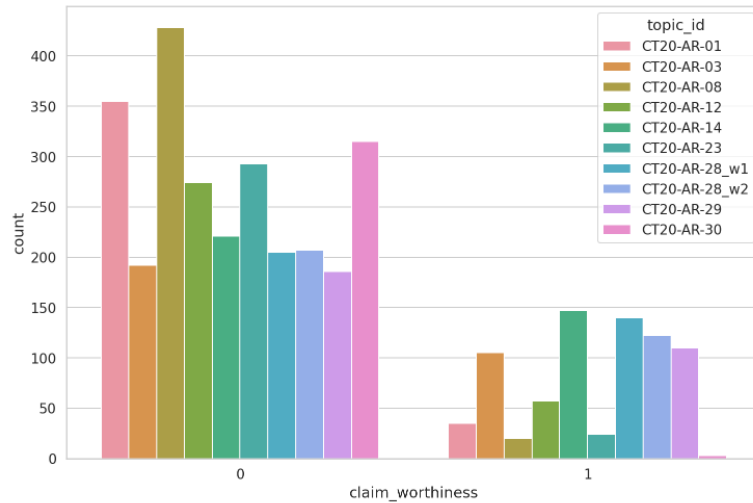


Figure 1: Check-worthiness per topic

5. Preprocessing

We used a variety of pre-processing methods on the raw tweet text. We began with deleting all English words from the text so we would only be focusing on Arabic. Second, we deleted all of the links in the tweet text, as well as all of the username tags. Third, we performed some data normalization, removing tashkeel and the letter madda from texts, as well as duplicates (ra2, alif,

waw, ha2, ya2, ta2) and replacing some characters to prevent mixing, as seen below in the Table 1.

Table 1

Arabic letters before and after preprocessing

Before	After
أ	ا
آ	ا
إ	ا
ة	ه
ى	ي

As a result, the following sentences present one tweet before and after preprocessing respectively:

- يا قدس وردُ الشوقِ باح وشدني وطمعتُ في نيل الغرام فصَدني .. ##القدس "عاصمة فلسطين الأبدية ##القدس ##المسجد الأقصى
- يا قدس ورد الشوق باح وشد ني وطمع ت في نيل الغرام فصد ني القدس _عاصمه _ "فلسطين _الأبدية القدس المسجد _الأقصى

6. Pre-trained models

Pretrained contextualized text representation models have shown effectiveness in making natural language recognizable by computers. Bidirectional Encoder Representations from Transformers (BERT) [5] is currently the most advanced model for language comprehension, outperforming previous versions and opening up new areas in the Natural Language Processing (NLP) sector. A similar study was recently completed for Arabic, that is getting popular. We used AraBERT and ArabicBERT. Added-on, we used the base version Arabic Albert.

- Arabic BERT [6]: Arabic BERT was pretrained on 8.2 Billion words from the Arabic version of OSCAR filtered from Common Crawl. The model includes Modern Standard Arabic and some dialectical Arabic. It was trained for 3M training steps with batch size of 128.
- AraBERT [7]: AraBERT is a pre-trained language model based on the BERT architecture. It was released using a pre-training dataset of 70 million sentences, corresponding to

24 GB of text covering news from different Arab media. To train the model the BERT base configuration was used with 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length and a total of 110M parameters.

- Arabic-ALBERT [8]: We used Arabic-ALBERT Base² trained with 7M training steps with a batch size of 64. The corpus and vocabulary collection include dialectical Arabic as well as Modern Standard Arabic.

7. Experiments

The datasets given included Twitter metadata fields, which we ignored. In our approach we only require use of the Tweet’s message text and the check-worthy target, which contains a binary label with a positive class denoting check-worthy tweets. We used Transformers in our approach where we fine-tuned the pretrained language models described in the previous section. While validating our models, we decided not to use Arabic BERT because of its low performance and we focused on improving the other two models.

To prevent our model from overfitting during training phase, we added a dropout layer to the model architecture. As a result, we were able to improve the loss plot as well as the average precision metric on our development dataset. Experiments included ones using preprocessing and others using the undersampling technique where we decreased the non-check-worthy class because of the huge imbalance.

As shown in Table 2, AraBERT performed **86.51%** MAP and **98%** Precision @30 on the development data with the following parameters : 3e-6 learning rate, 16 batch size , 150 sequence length and 6 epochs. However, Arabic BERT performed **86.51%** MAP and **98%** Precision @30 with the following parameters : 1e-6 learning rate, 32 batch size , 150 sequence length and 10 epochs. These results of Arabic Bert are obtained when training the model with an undersampled data. Since some topics were highly imbalanced as we mentioned before, we decreased the none check-worthy claims by 50% for the following topics: CT20-AR-30,CT20-AR-23, CT20-AR-12, CT20-AR-08, CT20-AR-01.

Table 2

Experiments Results on the Development dataset

Model	MAP	P@30
AraBERT	86.51%	98%
Arabic BERT	86.51%	98%
Ensemble	87.49	98%

8. Results and Discussion

Table 3 reviews the official results of iCompass system against the top three ranked systems.

The submitted results were due to an ensemble of two submissions. With ensembling the two models: AraBERT v01 with the highest MAP and the best Arabic BERT. The resulted model

²<https://github.com/KUIS-AI-Lab/Arabic-ALBERT>

Table 3

Official Results on Test set and ranking

Team	Rank	MAP
Accenture	1	0.658
bigIR	2	0.615
SCUoL	3	0.612
iCompass	4	0.597

performed 87.49% of MAP score. Hence, we decided to do the same thing and ensemble the test prediction, which did not produce the expected results. This can be due to:

- Test had different tweet content, in other words the topics weren't the same.
- Ensemble harmed the results because, while using the Arabic BERT, we used undersampled data and that achieved good results on the development dataset however while trying it on the test data when test-golds were released , it only achieved 0,581 MAP which confirmed that the test data distribution isn't the same as the development data. But, the fine-tuned AraBERT model that is trained using all the data achieved on test set 0,629 MAP which makes us conclude that the test set had a close distribution to the train set.

9. Conclusion

In this paper, we describe our system, for detecting check-worthy tweets, which we developed for the Arabic version of CLEF-2021 CheckThat! Subtask-1A. Our system is built using an ensemble of deep contextualized text representations, social context, and advanced pre-processing techniques. Our system was ranked fourth with a Mean Average Precision score of 0,597. Future work will include developing larger contextual system covering as many topics as possible. Furthermore, we intend to use meta-data and other modalities including photos and videos found in tweets for claim's check-worthiness.

References

- [1] S. Kemp, Whatsapp is the world's favorite social platform (and other facts) (2021).
- [2] 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. [arXiv:2009.02431](#).
- [3] P. Nakov, G. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. Shahi, J. Struß, T. Mandl, The CLEF-2021 Check-That! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News, 2021, pp. 639–649. doi:10.1007/978-3-030-72240-1_75.
- [4] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, M. K. Alex Nikolov, F. A. Yavuz Selim Kartal, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF '2021, Bucharest, Romania (online), 2021.

- [5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [6] 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://www.aclweb.org/anthology/2020.semeval-1.271>.
- [7] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, arXiv preprint arXiv:2003.00104 (2020).
- [8] A. Safaya, Arabic-albert, 2020. URL: <https://doi.org/10.5281/zenodo.4718724>. doi:10.5281/zenodo.4718724.