Mirela at CheckThat! 2024: Check-Worthiness of Tweets with Multilingual Embeddings and Adversarial Training

Notebook for the CheckThat! Lab at CLEF 2024

Mirela Dryankova[1,*], Dimitar Dimitrov[1], Ivan Koychev[1] and Preslav Nakov[2]

[1] *Sofia University "St. Kliment Ohridski", Bulgaria*
[2] *Mohamed bin Zayed University of Artificial Intelligence, UAE*

Abstract

Accurately assessing the credibility and significance of texts is crucial in today's digital age where misinformation and disinformation abound, especially in social media. In this paper, we propose an approach for check-worthiness of tweets that integrates adversarial learning techniques to optimize classification accuracy and language identification simultaneously. We conduct fine-tuning of DistilBERT-multilingual and XLM-RoBERTa-base for English, Dutch, Spanish, and Arabic to allow the models to adapt to the intricacies of different languages. Furthermore, we introduce an adversarial training approach to enhance the performance of multilingual sentence transformers, ensuring their effectiveness across linguistic contexts. The proposed approach ranks 4th in Dutch, 11th in Arabic, and 16th in English with an F_1-score (positive class) of 0.65, 0.48, and 0.66, respectively.

Keywords

Check-worthiness, Misinformation, Disinformation, Social Media, Multilingual Classification, Sentence Transformers

1. Introduction

The fast development of social media platforms in recent years has greatly urged information dissemination. This advancement allows society to stay up-to-date with emerging news and follow the latest trends, fostering a more informed and connected global community. For instance, people can access real-time events, participate in online discussions and seminars, and engage with content from all over the world, encouraging public sharing of opinions. Thus, social media have become one of the main communication channels for information dissemination and consumption, and nowadays many people rely on them as their primary source of news [1]. However, the ease with which information can be shared and the often unchecked nature of user-generated content has led to the wide and rapid spread of false or misleading information, which can have negative societal consequences [2].

This dual-edged situation highlights the need for effective fact-checking mechanisms to distinguish between reliable and dubious sources. The CheckThat! Lab Task 1 [3] at the Conference and Labs of the Evaluation Forum (CLEF) 2024 focuses on developing models that automatically determine the tweets' worthiness. This task is designed to assist fact-checkers by identifying tweets that contain potentially false claims and have a significant impact if left unchecked, thus streamlining the fact-checking process and helping to mitigate the spread of misinformation [4].

The task is focused on four languages - English, Spanish, Arabic, and Dutch, where data was collected from Twitter. We participated in the check-worthiness sub-task with a focus on Dutch, English, and Arabic. For the submission phase, we proposed a multi-language text classification strategy emphasizing the incorporation of language adversarial learning into the training process of sentence transformers for Arabic and Dutch, and run BERT-base-uncased for English.

Our approach mainly focuses on using a pre-trained DistilBERT-multilingual model, which is lighter and faster at inference time, while also requiring a smaller computational training budget [5]. For

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
mireladryankova959@gmail.com (M. Dryankova); mitko.bg.ss@gmail.com (D. Dimitrov); koychev@fmi.uni-sofia.bg (I. Koychev); preslav.nakov@mbzuai.ac.ae (P. Nakov)

further experiments, we fine-tuned XLM-RoBERTa-base - a transformer-based multilingual masked language model pre-trained on text in 100 languages, which obtains state-of-the-art performance on cross-lingual classification, sequence labeling and question answering [6].

The mentioned methodologies above are essential to automate the fact-checking process and address misinformation in diverse linguistic contexts.

2. Related Work

Traditionally, fact-checking has been a manual process, relying heavily on human effort and some of the leaders in the field are FactCheck.org[1], Snopes[2], PolitiFact[3], and FullFact[4]. This meticulous work is very time-consuming and labor-intensive so the need to automate it emerged. Automated fact-checking appeared as an approach where methods of Natural Language Processing (NLP) and Machine Learning (ML) are used to assist experts in making these decisions [4]. One of the earlier efforts in this direction is the ClaimBuster [7], an end-to-end system that uses machine learning, natural language processing, and database query techniques to aid in the process of fact-checking [8].

In their study [7], the authors address the check-worthiness task by comparing traditional models (Random Forest and SVM) with transformer-based models BERT and XLM-RoBERTa. The evaluation shows that transformer models (BERT-multilingual and XLM-RoBERTa-base) outperform the SVM and Random Forest in Dutch and English languages [7], but the results for Spanish are better using Random Forest.

In another study [9], Fraunhofer SIT take first place for CLEF-2023 CheckThat! Task 1A and second place for CLEF-2023 CheckThat! Task 1B. To determine whether a claim in a tweet that contains both a snippet of text and an image is worth fact-checking [9], they combine BERT with an OCR analysis. To determine whether a text snippet from a political debate should be assessed for check-worthiness, the team run multiple experiments. The best approach for this task is an ensemble classification scheme centered on Model Souping [9].

Another interesting approach [10], compares GPT models with BERT models and uses zero-shot, few-shot, and fine-tuning techniques in the context of check-worthiness problem. As a result, the participants managed to outperform CheckThat! Lab 2022 Task 1 winning model, as fine-tuning DeBERTa v3 base.

Additionally, other methods have been explored, using Word2Vec [11], as well as many participants apply different machine learning methods such as k-nearest neighbors [12] and Gradient boosting [13].

All of the above methodologies are specifically focused on addressing the check-worthiness task for English.

3. Methodology

3.1. Data

The dataset used for the check-worthiness task is given by the organizers of the CheckThat! Lab. The train data is provided in four languages - English, Spanish, Arabic, and Dutch while the test datasets are in English, Arabic, and Dutch. The number of rows in English and Spanish given for training is relatively higher than in Arabic and Dutch. The datasets contain text, id and link of the tweet as well as class label ("Yes" / "No") whether or not the text can be fact-checked.

Table 1 displays the distributions of the given datasets. As can be seen, the dataset suffers from class imbalance [9]. It can be concluded that within each split the total number of "No" labels is relatively higher than "Yes" labels.

[1] http://www.factcheck.org/
[2] http://www.snopes.com/fact-check/
[3] http://www.politifact.com/
[4] http://fullfact.org/

Table 1
Class label distribution for Train, Dev, Dev-Test, and Test datasets.

Class label	Train	Dev	Dev-Test	Test
Arabic				
Yes	2,243	411	377	218
No	5,090	682	123	392
Total	7,333	1,093	500	610
Dutch				
Yes	405	102	316	397
No	590	150	350	603
Total	995	252	666	1,000
English				
Yes	5,413	238	108	88
No	17,088	794	210	253
Total	22,501	1,032	318	341
Spanish				
Yes	3,122	704	509	-
No	16,826	4,296	4,491	-
Total	19,948	5,000	5,000	-

3.2. Models

We implemented a multilingual sentence classification system, designed to classify sentences across multiple languages and ensure that the learned representations are independent of the language of input sentences.

We introduced two classes of models Sentence Transformer, representing a basic sentence embedding model, and Sentence Transformer Adversarial, extending the original model with an adversarial training. The key difference lies in incorporating an additional language classification in Sentence Transformer Adversarial architecture, enabling language prediction of the input sentence in an adversarial manner. Both models follow pre-trained transformer-based architectures, specifically DistilBERT-multilingual and XLM-RoBERTa-base.

The model configuration is set to output both attention and hidden states. The architecture (Figure 1) includes a fully connected neural network to refine the representations for classification tasks. During training, we employ cross-entropy loss for classification tasks and incorporate linear scheduling of learning rates to stabilize training and improve convergence.

For evaluation, the key metric used is F_1-score with respect to the positive class as proposed by the organizers of the CheckThat! Lab. Furthermore, accuracy, precision, and recall with respect to the positive and negative classes are also shown in this paper for a better understanding of model performance.

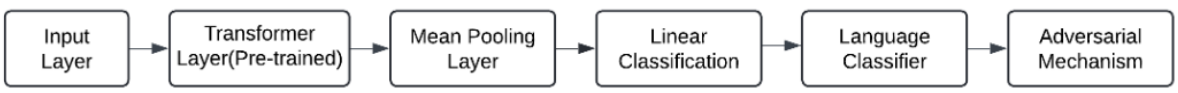

Figure 1: Model Pipeline.

Table 2
Performance metrics for different languages. **Bold** indicates positive class F_1-score. Underline indicates the best F_1-score.

Class label	Model	Accuracy	Precision	Recall	F_1- Score
		Arabic			
Yes	DistilBERT-multilingual	51.97	39.07	61.47	**47.77**
No			68.54	46.68	55.54
Yes	XLM-RoBERTa base	51.80	41.08	80.28	**54.35**
No			76.63	35.97	48.96
		Dutch			
Yes	DistilBERT-multilingual	71.50	63.39	66.75	**65.03**
No			77.32	74.63	75.95
Yes	XLM-RoBERTa base	72.80	63.56	73.80	**68.29**
No			80.71	72.14	76.18
		English			
Yes	DistilBERT-multilingual	86.22	79.71	62.50	**70.06**
No			87.87	94.47	91.05
Yes	XLM-RoBERTa base	86.22	78.08	64.77	**70.81**
No			88.43	93.68	90.98
Yes	BERT-base-uncased	84.46	75.12	57.95	**65.80**
No			86.49	93.68	89.94

3.3. Experiments

All models are trained on Google Colab Pro's T4 GPU. The T4 GPU offers significant computational capabilities, 16 GB of memory, and CUDA cores, which are crucial for our experiments.

DistilBERT-multilingual is trained on 5 epochs, while XLM-RoBERTa-base on 3 epochs due to higher computational demands. The training and validation data are processed in batches of 32 and 16, respectively. Additional hyperparameters include learning rate 2e-5, Adam optimizer 1e-8, and dropout level at 0.1.

For the official competition submission, we provided a Multilingual DistilBERT model, which demonstrated promising results in the fact-checking task, particularly for Dutch. Initially, our approach focused exclusively on English using BERT-base-uncased. Later, we upgraded our methodology with multilingual model. Due to time constraints, we ran only the Multilingual DistilBERT model for Arabic and Dutch, and BERT-base-uncased model for English and submitted the results. Following this initial phase, we expanded our approach by incorporating the XLM-RoBERTa-base model. After releasing gold labels once the submission period ended, we evaluated all the experiments and reported the statistics. The output results on the test set from all models are presented in Table 2.

We can conclude that our original submission model achieved the highest F_1-score over the positive class for Dutch. For all languages, XLM-RoBERTa base outperforms DistilBERT-multilingual and has the greatest increase in F_1-score for correctly identifying the positive class in the Arabic dataset. Moreover, XLM-RoBERTa-base model generally provides a better balance between precision and recall, making it slightly more reliable for identifying check-worthy tweets across multiple languages.

The official results and ranking from the competition submission are presented in Table 3:

Table 3
Overall ranking on the test set of Task 1: Check-worthiness of tweets

Language	Model	F_1 (positive class)	Rank
Arabic	DistilBERT-multilingual	0.478	11^{th} / 14^{th}
Dutch	DistilBERT-multilingual	0.65	4^{th} / 16^{th}
English	BERT-base-uncased	0.658	16^{th} / 27^{th}

4. Conclusion

In the following paper, we presented our experiments and insights gained from the check-worthiness task at CheckThat! 2024. Our methodology employs state-of-the-art transformer models, enhanced with adversarial training techniques, to improve fact-checking accuracy across multiple languages. By incorporating language classification into the training process, we ensure the models are capable of handling diverse linguistic inputs effectively. The proposed approach achieves 4^{th} place in Dutch with an F_1-score (positive class) of 0.65 and only 11^{th} in Arabic, and 16^{th} in English with an F_1-score (positive class) of 0.48, and 0.66, respectively. Overall, our study reveals that XLM-RoBERTa-base definitely outperforms DistilBERT-multilingual for all languages, regarding F_1-score over the positive class. The superior performance of XLM-RoBERTa-base can be attributed to its larger size and more intricate model architecture which is capable of capturing complex linguistic patterns more effectively. Moreover, XLM-RoBERTa-base benefits from extensive multilingual pretraining on a diverse corpus, enhancing its ability to generalize across different languages and understand diverse linguistic patterns. We can also conclude that our approach achieves the best results in Dutch due to the equal distribution of "Yes" and "No" class labels in train dataset. In all other languages the negative labels outnumber the positive labels by more than two times.

Further experiments can be conducted using a larger version of the discussed models or exploring bigger hyperparameter space, which can potentially lead to better results. Due to resource constraints, large transformer model architectures were not used in this research. Moreover, the current model can be expanded by incorporating additional contextual features, enhancing its capability to capture additional information from the input text and improve check-worthiness detection performance.

Acknowledgments

The work is partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

References

[1] A. Perrin, Social media usage, Pew research center (2015) 52–68.
[2] J. Vladika, F. Matthes, Scientific Fact-Checking: A survey of resources and approaches (2023).
[3] M. Hasanain, R. Suwaileh, S. Weering, C. Li, T. Caselli, W. Zaghouani, A. Barrón-Cedeño, P. Nakov, F. Alam, Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content (2024).
[4] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, J. Beltrán, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates, PNotes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF '2021, Bucharest, Romania (online) (2021).
[5] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, Hugging Face (2020).

[6] A. Conneau, K. Khandelwal, V. C. Naman Goyal, F. G. Guillaume Wenzek, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale v1 (2019).
[7] P. Tarannum, M. A. Hasan, F. Alam, S. R. H. Noori, Z-index at CheckThat! lab 2022: Check-worthiness identification on tweet text (2022).
[8] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, M. Tremayne, ClaimBuster: the first-ever end-to-end fact-checking system (2017).
[9] R. A. Frick, I. Vogel, J.-E. Choi, Fraunhofer SIT at CheckThat! 2023: Enhancing the detection of multimodal and multigenre check-worthiness using optical character recognition and model souping (2023).
[10] M. Sawiński1, K. Węcel1, E. Księżniak, M. Stróżyna1, W. Lewoniewski, P. Stolarski, W. Abramowicz, OpenFact at CheckThat! 2023: Head-to-head GPT vs. BERT - a comparative study of transformers language models for the detection of check-worthy claims (2023).
[11] M. Z. Ullah, An ML model for predicting information check-worthiness using a variety of features (2018).
[12] B. Ghanem, M. Montes-y-G´omez, F. Rangel, P. Rosso, UPV-INAOE - check that: preliminary approach for checking worthiness of claims. in: Working notes of CLEF 2018 - conference and labs of the Evaluation Forum, Avignon, France (2018).
[13] K. Yasser, M. Kutlu, T. Elsayed, bigIR at CLEF 2018: Detection and verification of check-worthy political claims (2018).