CheckThat! 2024: Automated Detection of Check-Worthy Claims and Subjectivity Analysis Using Pre-Trained Transformers

Syeda Dua E Zehra[1], Ahmed Ali Aun Mohammad[2], Kushal Chandani[3] and Muhammad Khubaib[4]

CS-355: Introduction to Large Language Models, Habib University, Karachi, Pakistan

Abstract

The CheckThat! Lab is a challenging lab designed to address the issue of disinformation. We participated in CheckThat! Lab Task 1 which focuses on identifying check-worthy claims in various forms of media, and Task 2 which targets the detection of subjective viewpoints in news articles. For both the tasks we focused only on the English dataset. For task 1, after standard preprocessing, we used the Ensembler Technique where we combined two models, namely Bert-Base-Uncased and XLM-Roberta-Base in order to finetune and to find the average probabilities to determine a unified ensemble probability. We achieved 14th position in the English leaderboard of task 1. For task 2 we augmented our data after standard pre-processing. We used the transformer-based model RoBERTa and finetuned it on the augmented dataset. We achieved 4th position in the English leaderboard for task 2.

Keywords

CLEF CheckThat!, fact-checking, transformer models, binary classification

1. Introduction

The CLEF CheckThat! initiative is at the cutting edge of technological developments in automated fact-checking, aimed at combating misinformation in the digital age. As we enter the 2024 edition, our efforts are focused on two key tasks: the first involves assessing the check-worthiness of claims made in tweets and other texts in Arabic, English, and Spanish, determining which statements require verification to prioritize efforts against potentially harmful misinformation. The second task explores the nuances of news reporting across multiple languages, aiming to distinguish subjective opinions from objective facts, essential for maintaining factual integrity and educating the public on the differences between opinion and fact-based reporting. Both tasks utilize binary classification and measure effectiveness through F1 scores, ensuring precise and efficient validation of information, reinforcing our toolkit against the proliferation of misinformation and fostering a healthier, more truthful digital environment.

[†] These authors contributed equally.

2. Literature Review

2.1. Task 1

In recent years, the CLEF CheckThat! competition has showcased innovative approaches to claim detection. Top teams have consistently relied on transformer-based models to enhance their systems. Accenture, the top-ranked team in 2020, utilized a RoBERTa-based model, incorporating mean pooling and dropout layers to improve generalization and reduce overfitting [2]. This strategy helped them achieve strong performance over baseline models.

In 2021, NLP&IR@UNED explored several pre-trained transformer models, discovering that BERTweet was the most effective on the development set. BERTweet, trained on 850 million English tweets and 23 million COVID-19-specific tweets, excelled at identifying check-worthy claims [3]. The second-place team, Fight for 4230, also used BERTweet but added a dropout layer and implemented data augmentation techniques [3]. In the following year, PoliMi-FlatEarthers stood out by fine-tuning GPT-3 for Task 1B. They combined deep learning with domain-specific customization to accurately classify check-worthy claims [5]. Finally, in 2023, OpenFact leveraged a fine-tuned GPT-3 model, utilizing a rich, annotated dataset of sentences from political debates and speeches. Their data-centric approach, tailored specifically for fact-checking, helped them outperform other submissions[7].

2.2. Task 2

This task has only appeared in one previous edition that was of CheckThat! 2023. The top submissions used many different models, the most used were BERT, RoBERTa, ChatGPT and GPT3. Team DWReCo [9] got the best score in the English category. There approach involved augmenting the dataset using GPT and then trained on RoBERTa. Two other teams also went with a data augmenting approach. The overall best score on the multilingual dataset was achieved by Team NN [10] who used the XLMRoBERTa model and trained it on the multilingual dataset. Team Thesis Titan [11] achieved top positions in 4 languages. There approach was to train the mdeberta model finetuned for each specific languages seperately allowing them to achieve those scores. Many other teams also tried an ensemble approach and got decent results.

3. Our Approach

3.1. Task 1

The goal of Task 1 was to evaluate the necessity of fact-checking claims in tweets and transcriptions. This typically requires either the expertise of professional fact-checkers or answers to several auxiliary questions by human annotators.

3.1.1. Data Preparation, Model Training and Evaluation

The datasets were initially provided in zip files. To access the data, we utilized direct drive links. We began with three datasets: the Training Dataset, the Validation Dataset, and the Test-Dev Dataset. Later, we received the fourth dataset, the main Test Dataset which was unlabeled. Our initial modeling used the following parameters with the Bert-base-uncased model:

- Batch size: 8 for both training and validation
- Learning rate: 2×10^{-5}
- Number of epochs: 3

After training, we used the model to process the Test-dev dataset. The procedure involved:

1. Tokenizing the text entries.
2. Feeding the tokenized data into the model.
3. Converting the output logits to probabilities using a sigmoid function.
4. Classifying each entry as "Yes" or "No" based on a probability threshold of 0.5.
5. Collecting these classifications and their corresponding 'Sentence_id' into a list for comparison with the original labels.

The approach achieved an **F1 score around 0.80.**

3.1.2. Modifications Made For Final Approach

To improve results, we experimented with various models like Alberta, Roberta-base, XLM-Roberta, and Electra. The most significant improvement was observed with **XLM-Roberta-base and Bert-base-uncased**. We then implemented an Ensemble approach with these two models using the following training configurations:

- Batch size: 16 for both training and validation
- Learning rate: 5×10^{-5}
- Number of epochs: 5
- Weight Decay: 0.005

Both trained models were evaluated on the Test-dev dataset. Each text data point from the test dataset was processed by both models, and their predictions were averaged to form a single ensemble probability. This probability determined the final label ("Yes" or "No"), which was collected along with the text's unique identifier into a list.

3.1.3. Results

Task 1	Training Set	Validation Set	Dev-test Set	**Test Set**
F1 scores	0.92	0.93	0.87	**0.696**

3.2. Task 2

The goal of Task 2 was to evaluate the Subjectivity of news articles and decide whether an extract from the article was subjective or objective.

3.2.1. Data Preparation, Model Training and Evaluation

Our focus was on the English Dataset. The datasets were initially provided to us in TSV files. We began with three datasets: the Training Dataset, the Validation Dataset, and the Test-Dev Dataset. We used data augmentation to enhance our dataset as the Train dataset was very small and the model was not able to learn and effectively. We initially tried to augment the data using WordNet model and the NTLK library but this did not prove effective. This was because this method changed some word at random from the sentence and replaced it with it's synonym which at times did not portray the sentence correctly. After some guidance from our supervisor, we used the Gemini Api using GOOGLE AI STUDIO and it's 'gemini-1.0-pro-latest' model and augmented our data. The approach used in this case was to create three similar sentences for each of the "Objective" label and five similar sentences for each of the "Subjective" label. This allowed us to have a more balanced dataset and allowed the model to have a better learning. We then imported the dataset called "data", which has been uploaded on GitHub as well.

Our initial modeling was done using MdBerta and we used the following parameters:

- Batch size: 16 for both training and validation
- Learning rate: 5×10^{-5}
- Number of epochs: 6
- Warmup steps: 100
- Weight decay: 0.01

After training, we used the model to process the Test-dev dataset. The procedure involved:

1. Processing the data and tokenizing the text entries.
2. Feeding the tokenized data into the model.
3. Converting the output logits to probabilities.
4. Classifying each entry as "Subj" or "Obj" using Sigmoid and Argmax.
5. Collecting these classifications into a list for comparison with the original labels.

The approach achieved an **F1 score around 0.76.**

3.2.2. Modifications Made For Final Approach

While modifying, we tried different models and even used the ensemble approach using models such as Roberta-base, MdBerta, Roberta-xlm, and Bert-base, but the best results were achieved using Roberta-base alone, hence we used that for our final submission using the following training configurations:

- Batch size: 64 for both training and validation

- Learning rate: 5×10^{-6}
- Number of epochs: 12
- Warmup steps: 100
- Weight decay: 0.01

This probability determined the final label ("Subj" or "Obj"), which was collected along into a list.

3.2.3. Results

Task 2	Training Set	Dev-test Set	**Test Set**
MACRO F1	0.96	0.82	**0.708**
SUBJ F1	0.95	0.83	**0.54**

3.2.4. Analysis

Leaving Blank as Indicated

3.2.5. Conclusion

In conclusion, our detailed exploration in the CheckThat! Lab 2024 challenge demonstrated the significant capabilities of transformer-based models in tasks of check-worthiness detection and subjectivity analysis. For Task 1, the ensemble method combining XLM-Roberta and Bert-base-uncased models effectively navigated the complexities of identifying check-worthy claims. By using a strategic ensemble of predictions and applying a robust training regimen involving multiple epochs (up to 5) and a learning rate of 5×10^{-5}, the system achieved a commendable balance between precision and generalization. In Task 2, the fine-tuned RoBERTa model excelled in differentiating subjective from objective statements, utilizing a refined approach with a lower learning rate (5×10^{-6}) and an increased number of epochs (12), ensuring thorough learning. Data augmentation played a crucial role here, bolstering the dataset and thereby enhancing the model's ability to handle nuanced textual variations. While these results were promising, they also suggest potential areas for further refinement to optimize performance, particularly in handling more complex misinformation scenarios. These efforts exemplify the essential role of adaptive, transformer-based architectures in leveraging deep learning for critical media literacy tasks in a multilingual context.

3.2.6. Acknowledgements

The text will be provided later

References

[1] "CheckThat!" Checkthat.gitlab.io, https://checkthat.gitlab.io/clef2024/

[2] Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeño, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Martino, G. D. S., & Nakov, P. (2020). Overview of CheckThat! 2020 English: Automatic Identification and Verification of Claims in Social Media. In Proceedings of the CLEF 2020. Thessaloniki, Greece. Retrieved from https://ceur-ws.org/Vol-2696/paper_265.pdf

[3] Shaar, S., Hasanain, M., Hamdan, B., Sheikh Ali, Z., Haouari, F., Nikolov, A., Kutlu, M., Kartal, Y. S., Alam, F., Da San Martino, G., Barrón-Cedeño, A., Míguez, R., Beltrán, J., Elsayed, T., & Nakov, P. (2021). Overview of the CLEF-2021 CheckThat! Lab Task 1 on Check-Worthiness Estimation in Tweets and Political Debates. In Proceedings of the CLEF 2021. Retreived from https://ceur-ws.org/Vol-2936/paper-28.pdf

[4] Shaar, S., Haouari, F., Mansour, W., Hasanain, M., Babulkov, N., Alam, F., Da San Martino, G., Elsayed, T., & Nakov, P. (2021). Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates. In Proceedings of the CLEF 2021. Retreived from https://ceur-ws.org/Vol-2936/paper-29.pdf

[5] Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Mubarak, H., Nikolov, A., & Kartal, Y. S. (2022). Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets. In Proceedings of the CLEF 2022. Retreived from https://ceur-ws.org/Vol-3180/paper-28.pdf

[6] Nakov, P., Da San Martino, G., Alam, F., Shaar, S., Mubarak, H., & Babulkov, N. (2022). Overview of the CLEF-2022 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims. In Proceedings of the CLEF 2022. Retreived from https://ceur-ws.org/Vol-3180/paper-29.pdf

[7] Alam, F., Barrón-Cedeño, A., Cheema, G. S., Shahi, G. K., Hakimov, S., Hasanain, M., Li, C., Míguez, R., Mubarak, H., Zaghouani, W., & Nakov, P. (2023). Overview of the CLEF-2023 CheckThat! Lab Task 1 on Check-Worthiness of Multimodal and Multigenre Content. In Proceedings of the CLEF 2023. Retreived from https://ceur-ws.org/Vol-3497/paper-019.pdf

[8] Galassi, A., Ruggeri, F., Barrón-Cedeño, A., Alam, F., Caselli, T., Kutlu, M., Struß, J. M., Antici, F., Hasanain, M., Köhler, J., Korre, K., Leistra, F., Muti, A., Siegel, M., Türkmen, M. D., Wiegand, M., & Zaghouani, W. (2023). Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity in News Articles. In Proceedings of the CLEF 2023. Retreived from https://ceur-ws.org/Vol-3497/paper-020.pdf

[9] I. B. Schlicht, L. Khellaf, D. Altiok, Dwreco at CheckThat! 2023: Enhancing subjectivity detection through style-based data sampling

[10] K. Dey, P. Tarannum, M. A. Hasan, S. R. H. Noori, Nn at CheckThat! 2023: Subjectivity in news articles classification with transformer based models

[11] F. Leistra, T. Caselli, Thesis titan at CheckThat! 2023: Language-specific fine-tuning of mdebertav3 for subjectivity detection