Checker Hacker at CheckThat! 2024: Detecting Check-Worthy Claims and Analyzing Subjectivity with Transformers

Notebook for the CheckThat! Lab at CLEF 2024

Syeda DuaE Zehra[1,*], Kushal Chandani[1], Muhammad Khubaib[1], Ahmed Ali Aun Muhammed[1], Faisal Alvi[1] and Abdul Samad[1]

[1]*Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan.*

Abstract

This paper represents our approach on the CheckThat! 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 tasks we focused on the English dataset only. For task 1, after standard preprocessing, we used an ensemble approach 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. For task 1 our F1 score was 0.696 and our rank was 14th in the English leaderboard. For task 2 we augmented our data after standard pre-processing using Google AI Studio and it's gemini-1.0-pro-latest model and then used the transformer-based model RoBERTa and finetuned it on the augmented dataset. For task 2, our macro F1 score was 0.7081 and our rank was 4th in the English leaderboard.

Keywords

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

1. Introduction

The CLEF CheckThat! Lab [17] initiative is at the forefront of technological developments in automated fact-checking, aiming to combat misinformation in the digital age. Misinformation poses significant risks to public discourse and democratic processes, making the development of effective fact-checking tools crucial. In the 2024 edition [1][17], the Lab focuses on two key tasks, each addressing critical aspects of this challenge.

The first one concentrated on assessing the check-worthiness of claims made in tweets and other English texts. This involves identifying which statements require verification, thereby prioritizing efforts, as not all claims can be fact-checked due to resource constraints, and determining check-worthiness ensures that the most impactful misinformation is addressed promptly. The second task aimed to distinguish subjective opinions from objective facts in the sentences of the news articles, something essential for maintaining factual integrity and preventing spread of misinformation. By accurately identifying and separating opinions from facts, we can improve the reliability of news content and support informed public discourse. Unlike sentiment analysis, which would be focused on identifying emotional tones, subjectivity analysis actually aims to improve the working of Task 1 as it aims at discerning statements that may require verification (subjective) from those presenting factual information (objective). By categorizing claims, fact-checkers can prioritize rigorous scrutiny for

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

sz06571@st.habib.edu.pk (S. D. Zehra); kc07535@st.habib.edu.pk (K. Chandani); mk07218@st.habib.edu.pk (M. Khubaib); aa07600@st.habib.edu.pk (A. A. A. Muhammed); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

0009-0002-3207-1826 (S. D. Zehra); 0009-0002-0954-7742 (K. Chandani); 0009-0002-0699-4029 (M. Khubaib); 0009-0004-5910-6018 (A. A. A. Muhammed); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

subjective claims that may influence public opinion or require context evaluation, while focusing factual verification efforts on objective claims backed by evidence. Together, these tasks help in providing a comprehensive approach to validating information, preventing the spread of misinformation, and upholding the credibility of information sources.
Both tasks make use of binary classification and measure effectiveness through F1 scores, ensuring precise and efficient validation of information, and preventing the spread of misinformation.

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 [4][5]. 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 [6][8]. The second-place team, Fight for 4230, also used BERTweet but added a dropout layer and implemented data augmentation techniques [7][8]. 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 [9][10]. 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[11][12]. Additionally, recent research has shown that models like FACT-GPT, which use synthetic data generated by large language models for training, can closely match human judgment in identifying related claims, highlighting the potential for AI tools to enhance the fact-checking process [20].

2.2. Task 2

This task has only appeared in one previous edition CheckThat! 2023. The top submissions used many different models, the most used were BERT, RoBERTa, ChatGPT and GPT3. Team DWReCo [13][16] 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 [14][16] who used the XLMRoBERTa model and trained it on the multilingual dataset. Team Thesis Titan [15][16] achieved top positions in 4 languages. Their 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.

Similar to Team DWReCo's strategy, FACT-GPT utilized large language models (LLMs) to generate synthetic training data, enhancing the adaptability of models for specific tasks, which is crucial for claim matching in fact-checking contexts. Like the approach of using XLMRoBERTa for multilingual datasets, FACT-GPT demonstrated that fine-tuning language models on synthetic datasets could improve classification accuracy and reduce computational costs. Both FACT-GPT and the approaches in CheckThat! 2023 emphasize the importance of leveraging AI to assist and enhance human expertise in the fact-checking process. [20]

3. Task 1

3.1. Our Approach

The goal of Task 1 [2] 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

We were provided with three datasets: [18] the training dataset, the dev 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 of 0.80 on the test-dev dataset.**

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.

3.2. Results

Table 1: Performance metrics for Task 1 across different datasets

Task 1	dev Set	dev-test Set	**Test Set**
F1 scores	0.93	0.87	**0.696**

4. Task 2

4.1. Our Approach

The goal of Task 2 was to evaluate the Subjectivity of news articles and decide whether a sentence from the news article [3][19] was subjective or objective.

4.1.1. Data Preparation, Model Training and Evaluation

Our focus was on the English datasets: the training dataset, the dev 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. This method changed one word at random from the sentence and replaced it with its synonyms.

Our initial modeling was done using mDeBERTa 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 of 0.76.** This was achieved on the dataset that had been augmented using the WordNet model and NTLK.

4.1.2. Modifications Made For Final Approach

The current approach of changing the words with their synonyms at times did not portray the sentence correctly. We then decided to use 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. While modifying, we tried different models and even used the ensemble approach using models such as RoBERTa-base, mDeBERTa, 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

The probability calculated went through a probability threshold of 0.5, based on which we determined the final label ("Subj" or "Obj").

4.2. Results

Table 2: Performance Metrics for Task 2 across different datasets

Task 2	Dev Set	Dev-test Set	**Test Set**
MACRO F1	0.86	0.82	**0.708**
SUBJ F1	0.82	0.83	**0.54**

5. Analysis

We saw an overall drop in the scores of our model as it achieved high scores on the training set compared to the dev set, dev-test, and test set scores, which indicates potential overfitting. This means that the model did not perform well with new, unseen data.

In Task 1, the validation loss showed a slight increase, which potentially contributed to the model's underperformance on new data. This increase in validation loss indicates that the model might have started overfitting to the training data, thereby reducing its generalizability. As a result, when the model was applied to the test set, it did not produce equally good results. Additionally, the class imbalance in the dataset could have affected the model's performance. With fewer instances labeled as check-worthy compared to non-check-worthy, the model might have struggled to accurately identify the check-worthy instances, leading to a lower overall score. The preprocessing steps, while essential for cleaning and preparing the data, might not have fully addressed the inherent variability in the text, further complicating the model's ability to generalize well to unseen data.

Moreover in task 2, a reason for the low SUBJ F1 score on the test set suggests that the model had difficulty with the "SUBJ" class. One possible reason for this could be that the features used for identifying the "SUBJ" class may not be as strong or distinctive, or there might be more variability or noise in the "SUBJ" class in the test set compared to the training set. Another reason for the low SUBJ F1 could be the way we conducted our data augmentation. The approach that we used, created three similar sentences for each of the "Objective" label and five similar sentences for each of the "Subjective" label. Considering all the sentences might have been similar to the original one from which they were made, we might have experienced over-fitting as the features of those sentences might have been similar.

6. 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}.

In Task 2, the fine-tuned RoBERTa model proved better than the other models we had tested as it showed better performance in differentiating the subjective from objective statements on the devtest file, utilizing a refined approach with a lower learning rate (5×10^{-6}) and an increased number of epochs (12), ensuring thorough learning. However, the performance, as indicated by a macro F1 score of 0.7081 and an F1 score of 0.54 for the SUBJ class, suggests room for improvement. A deeper analysis reveals that the model struggled with the "SUBJ" class, possibly due to weaker feature representation or greater variability and noise in the test set. Another issue that might have been prevalent is of class imabalance which might have led to weaker SUBJ identification. Future work could focus on enhancing the feature set for this class and reducing noise through better data preprocessing and augmentation.

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.

Acknowledgments

The authors would like to acknowledge the support provided by the Office Of Research (OoR) at Habib University, Karachi, Pakistan for funding this project through internal research grant IRG-2235.

References

[1] Barrón-Cedeño, A. et al. (2024). The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. In: Goharian, N., et al. Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science, vol 14612. Springer, Cham. https://doi.org/10.1007/978-3-031-56069-9_62

[2] Hasanain, M., Suwaileh, R., Weering, S., Li, C., Caselli, T., Zaghouani, W., Barrón-Cedeño, A., Nakov, P., Alam, F.: Overview of the CLEF-2024 CheckThat! Lab Task 1 on Check-Worthiness Estimation of Multigenre Content. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[3] Struß, J. M., Ruggeri, F., Barrón-Cedeño, A., Alam, F., Dimitrov, D., Galassi, A., Siegel, M., Wiegand, M.: Overview of the CLEF-2024 CheckThat! Lab Task 2 on Subjectivity in News Articles. In: Faggioli, G., Ferro, N., Galuščáková, P., García Seco de Herrera, A. (eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[4] Williams, E., Rodrigues, P., Novak, V.: Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. In: Cappellato et al., CLEF 2020

[5] Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeño, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 English: Automatic Identification and Verification of Claims in Social Media. In: Cappellato et al., CLEF 2020.

[6] J. M.-R. Juan R. Martinez-Rico, L. Araujo, NLP&IR@UNED at CheckThat! 2021: Checkworthiness estimation and fake news detection using transformer models, 2021.

[7] X. Zhou, B. Wu, P. Fung, Fight for 4230 at CLEF CheckThat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness, 2021.

[8] Shaar, S., Hasanain, M., Hamdan, B., Ali, Z. S., Haouari, F., Nikolov, A., Kutlu, M., Kartal, Y. S., Alam, F., Da San Martino, G., Barrón-Cedeño, A., Miguez, R., Beltrán, J., Elsayed, T., Nakov, P.: Overview of the CLEF-2021 CheckThat! Lab: Task 1 on Check-Worthiness Estimation in Tweets and Political Debates. In: Cappellato et al., CLEF 2021, 369-392.

[9] S. Agrestia, A. S. Hashemianb, M. J. Carmanc, PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[10] 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.: Overview of the CLEF-2022 CheckThat! Lab: Task 1 on Identifying Relevant Claims in Tweets. In: Cappellato et al., CLEF 2022, 368-392.

[11] M. Sawiński, K. Wecel, E. Ksieżniak, M. Stróżyna, 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, in 2023.

[12] 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.: Overview of the CLEF-2023 CheckThat! Lab: Task 1 on Check-Worthiness in Multimodal and Multigenre Content. In: Cappellato et al., CLEF 2023, 219-235.

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

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

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

[16] 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.: Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity Detection. In: Cappellato et al., CLEF 2023, 236-249.

[17] Barrón-Cedeño, A., Alam, F., Struß, J. M., Nakov, P., Chakraborty, T., Elsayed, T., Przybyła, P., Caselli, T., Da San Martino, G., Haouari, F., Li, C., Piskorski, J., Ruggeri, F., Song, X., Suwaileh, R. (2024). Overview of the CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Soulier, L., Di Nunzio, G. M., Galuščáková, P., García Seco de Herrera, A., Faggioli, G., Ferro, N. (eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024).

[18] Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., et al. (2021). Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 611–649.

[19] Ruggeri, F., Antici, F., Galassi, A., Korre, K., Muti, A., Barrón-Cedeño, A. (2023). On the Definition of Prescriptive Annotation Guidelines for Language-Agnostic Subjectivity Detection. In: Text2Story@ECIR, CEUR Workshop Proceedings, vol. 3370, pp. 103–111. CEUR-WS.org.

[20] Choi, E. C., Ferrara, E.: FACT-GPT: Fact-Checking Augmentation via Claim Matching with LLMs. In: Companion Proceedings of the ACM on Web Conference 2024, 13 May 2024. https://doi.org/10.1145/3589335.3651504.