

Trio Titans at Check That! 2024 : Check Worthiness Estimation

Sai Geetha M^{1,*†}, Prarthna M^{1,†}, Chiranjeev Prasanna V V^{1,†} and Saritha M^{1,†,^}

¹ Department Of Computer Science Engineering, Sri Sivasubramaniya Nadar College Of Engineering, Kalavakkam, Tamil Nadu - 110.

Abstract

We present an overview of CheckThat! Lab's 2024 task 1, part of CLEF-2024, focused on determining the check-worthiness of text items. Utilizing various BERT models, including DistilBERT, RoBERTa, and ALBERT, we developed a text classification model using simple transformer frameworks. The model was trained on a preprocessed dataset and evaluated using a pre-processed development test set. Our final model, based on RoBERTa, achieved an accuracy of 0.86. This paper details our approach and experiments with different machine learning and transformer-based models.

Keywords

Check worthiness Estimation, Simple Transformer, Classification Model, DistilBERT , ALBERT , RoBERTa

Introduction


We undertook task 1 of CLEF -2024 CheckThat![\[11\]](#)[\[12\]](#) in which we had to verify the worthiness of the comments. We first preprocessed the data, removed the stopwords , punctuations and lowercase conversion and fed the data to the Bert model first and then switched to different models of BERT that were more optimized and then finally we used the RoBERTa which proved to show a higher accuracy and faster run speeds compared to the other models. The CheckThat! task-1 focuses on the worthiness of the comments. In the modern world where everyone uses social medias with ease of access and in seconds they can communicate with the entire world, it is necessary to uphold ethics and good practices. This helps the upcoming generation to concentrate only on the worthy messages and texts in social media and other platforms so that they won't be distracted by other useless information. This CheckThat! task focuses on classifying whether a given comment is worthy enough to be considered to check with the facts (i.e) is it related to any political events showcasing some statistics based data or just some inaccurate use of data just to express the person's point of view on social media. As young people it is our duty to make sure that the upcoming generation gets fended only on the good and correct information rather than inaccurate and biased information explaining someone's political or other interest.

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

*Corresponding author.

†These authors contributed equally.

^Mentor

 prarthna2310607@ssn.edu.in (Prarthna M); saigeetha2310537@ssn.edu.in (Sai Geetha M); chiranjeevprasanna2310132@ssn.edu.in (Chiranjeev Prasanna V V); sarithamadhesh@ssn.edu.in (Saritha M)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Related Works

In this section, we explore papers related to our methodology involving RoBERTa [1-5] and BERT [6,7] models, along with preprocessing techniques [8-10] like stop words removal, punctuation removal, and lowercase conversion of the texts. These studies include text classification papers from CheckThat! Lab and various other topics. They illustrate how effective preprocessing enhances the performance of RoBERTa and BERT models in tasks such as fact-checking, spam detection, short text classification, and more diverse applications in natural language processing.

Let's delve into these papers to understand their contributions and methodologies. The Accenture Team participated in the CLEF2020 CheckThat! Lab, Task 1, using BERT and RoBERTa models to identify and prioritize claims in social media text for professional fact-checking. For the English task, they fine-tuned a RoBERTa model with additional mean pooling and dropout layers to enhance generalizability, achieving 1st place in the English track [\[1\]](#).

In 2023, a study introduced a RoBERTa-based bi-directional Recurrent Neural Network model for spam detection on social networks. The RoBERTa model enhances the performance of a stacked BLSTM network by learning contextualized word representations. Experimental results showed that the RoBERTa-BLSTM model outperformed common models with accuracies of 98.15%, 94.41%, and 99.74% on Twitter, YouTube, and SMS datasets, respectively [\[2\]](#).

Published in 2021, a paper introduces a short text classification method using RoBERTa. RoBERTa generates semantic text vectors, significantly enhancing classification accuracy. Experiments on the THUCNews dataset demonstrate that RoBERTa achieves an accuracy of 94.64%, surpassing TextRCNN by 4.53% and Bert+RCNN by 0.62%, highlighting its effectiveness in semantic representation learning for short text contexts [\[3\]](#).

The paper provides an overview of team AI Rational's approach in CheckThat! 2022 for Task 1A in English, focusing on classifying COVID-19 tweets for check-worthiness detection. The team achieved first place among 13 teams, utilizing transformer models BERT, DistilBERT, and RoBERTa, where RoBERTa achieved the highest accuracy of 0.84, while both DistilBERT and BERT reached 0.81 [\[4\]](#).

In 2021, a comparative study evaluated BERT, RoBERTa, and Electra for fact verification. RoBERTa achieved the highest accuracy and F1-Score, both at 95.4% and 95.3% respectively, using an epoch value of 5 and batch size of 32, followed by BERT with 94.3% accuracy and F1-Score using epoch 10 and batch size 32 [\[5\]](#).

Also in 2021, a paper introduces a BERT-based model for classifying and categorizing fake news domains, leveraging fact-checked articles to achieve macro F1 scores of 83.76% for Task 3A and 85.55% for Task 3B by merging classes and using additional training data [\[6\]](#).

Addressing the spread of fake news through electronic media, another 2021 paper proposes a BERT-based model for early-stage fake news detection. By analyzing article content and stance, the model achieves an accuracy of 95.32% on a real-world dataset, surpassing previous methods [\[7\]](#).

Published in 2020, a paper investigates how different text preprocessing methods affect text classification (TC) accuracy. It finds that stop word removal significantly improves accuracy across most datasets, with converting uppercase to lowercase letters particularly effective in one dataset [\[8\]](#).

A Thesis explores the impact of preprocessing techniques and learning algorithms on classification accuracy for a customer feedback dataset, demonstrating that removing stop words and performing spelling correction significantly increase accuracy across classifiers [\[9\]](#).

Introducing CovBERT in 2021, a BERT-based pre-trained language model automates the literature review process for COVID-19 biomedical publications. CovBERT, trained on COV-Dat-20, shows significant improvements in classification accuracy, using preprocessing techniques like stop words removal and text conversion to lowercase [\[10\]](#).

2. Dataset

The task provides Three different Datasets. It contains texts that are in English as well as some political debates.

Training Dataset

This Dataset has three columns with 22501 rows. The columns given are Sentence id, Text, class label. The count for class label 'no' is 17088 and 'yes' is 5413, refer to table 1.

Development Test Dataset

This Dataset has three columns with 318 rows. The columns given are Sentence id, Text, class label. The count for class label 'no' is 210 and 'yes' is 108, refer to table 2.

Test Dataset - final

This Dataset has three columns with 341 rows. The columns given are Sentence id, Text, class label. The count for class label 'no' is 284 and 'yes' is 57, refer to table 3.

Table 1 – Training Dataset

Class label	counts
NO	17088
YES	5413

Table 2 – Development Test Dataset

Class label	counts
NO	210
YES	108

Table 3 – Final Test Dataset

Class label	counts
NO	284
YES	57

3. Methodology

The given Dataset consists of text that are in English as well as some political debates. The Dataset is first divided into training, development and test partitions from which we initially consider only the training and development dataset. Improving the quality of the training set is very important when building models since it makes raw data acceptable for accurate learning. We improve the quality by using various techniques for preprocessing like lowercase conversion, remove punctuation and stop words. This process is applied to both development datasets and training datasets. As the given dataset contains a mix of punctuations and stopwords, we first convert the text to lowercase as the model treats “Words” and “words” as different tokens. This is done to ensure uniformity. Then the next step is to remove all the punctuations to reduce the noise and stop words using the stopword corpus imported from the Natural Language Toolkit NLTK library in python ensuring the focus remains on only the meaningful words. Then the class labels ‘Yes’ and ‘No’ were mapped to binary values ‘1’ and ‘0’ respectively which simplifies the classification task during model training. Since the task involves binary classification we made use of ‘simpletransformers’ library to train a RoBERTa model, the model which performs effectively in Natural Language Understanding tasks and the number of epochs used were 1. We preprocess the input data with each training iteration. Also, the cached or stored evaluation features from previous runs are not used during the evaluation phase. We write the output file in a new file so that it does not conflict with the original datasets. While training the model, we first check if the GPU is available and utilize it. RoBERTa is built on BERT’s architecture but the main difference is that it uses more data with longer running time. DistilBERT is a compressed model of BERT which uses the process of knowledge distillation, in which the smaller model is made to run like a larger model. At first, we made use of the DistilBERT model but the accuracy we got was very low. Then we made use of ALBERT which provided the lowest accuracy. ALBERT is the model of BERT that makes use of less parameters to make the model work effectively and faster with usage of less memory but that too didn’t work out. This made us continue our task with the RoBERTa model. Then we calculate the accuracy, F1 score, precision, recall and the data preprocessing is done to the development dataset in the same way as that of the training and development datasets. Ensuring consistency in preprocessing is essential to maintain the coherence between the datasets. Once the development test data is preprocessed the text data is input into the trained model. Then each input in the dataset is utilized by the model to generate predictions for each instance. The model will output the class labels indicating the likelihood of each instance being checkworthy or non-checkworthy and then the output file is generated which contains the predictions of the test dataset. The system flow diagram is shown in Figure 1.

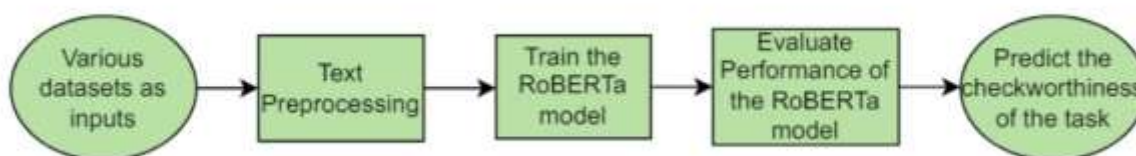


Figure 1 – system flow diagram

4. Experimental Result

We conducted a series of experiments to evaluate the performance of various transformer-based models on a test_dataset from the official Checkthat! site , consisting of English statements and political debates. Our primary objective was to determine the most effective model for the binary classification task of identifying checkworthy statements. As discussed in methodology , we tried using different BERT models like DistilBERT , RoBERTa and ALBERT with the help of classification models. RoBERTa produced higher accuracy of 86% than any other Bert

model. Hence we generated the class label of final test data for the submission, using this RoBERTa model. This is because the DistilBERT model has very few parameters as compared to other models. This can especially happen if the task requires a high degree of understanding, as in this case whether the given text is checkworthy or not. The low performance of the ALBERT model may be due to shared parameter techniques or smaller capacity as compared to RoBERTa. The observation of low accuracy with DistilBERT and other BERT models highlights the importance of model selection in achieving satisfactory performance. Let's look into variation in the accuracy, precision, F1 score and recall produced by different bert models. These metrics are calculated using the dev test data. The various metrics are discussed in table 4 and the graph for F1 score is shown in Figure 2.

Table 4 - Evaluation Metrics

MODEL	ACCURACY	PRECISION	F1 SCORE	RECALL
DistilBERT	0.84	0.91	0.73	0.61
RoBERTa	0.86	0.94	0.75	0.62
ALBERT	0.70	0.77	0.31	0.19

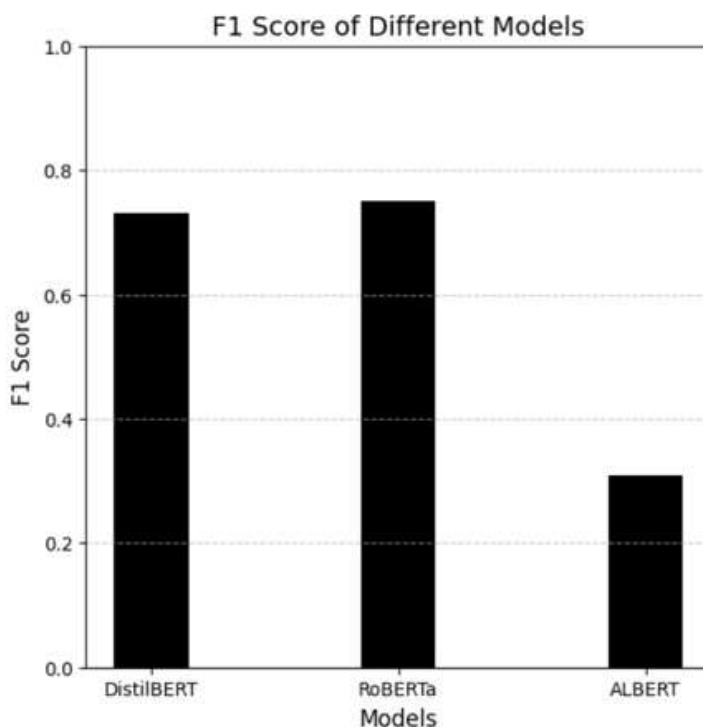


Figure 2 - F1 score of different model

5. Conclusion and Future Works

This paper describes our participation in CheckThat! task 1. This task involves identifying check-worthiness in texts whether a particular sentence is 'Yes' or 'No'. We used three transformer models and compared the results and finally chose the RoBERTa model. The Zobtained test results from the leaderboard indicate that the first team got the F1 score of 0.802 and our team got the F1 score of 0.600 and we find that the difference between them is 0.202. This indicates that our experiments with these three models are not enough and we have to explore various other methods and also have a look at deep learning techniques.

For future work, we aim to enlarge the dataset as a more extensive dataset improves the model's ability to perform better on unseen data. Then we can expand it to many languages to enhance the model's versatility and apply them in various platforms so that the model can handle diverse texts. Additionally we plan to implement various other transformer models and advanced deep learning models so that the texts can be analyzed in depth. These measures can improve the accuracy of our model which helps in achieving better results

6. Citations

In the context of check-worthiness estimation, the overviews provided in the CLEF – Check That! 2024 [13] and CLEF – Check That! Task -1 [12] papers serve as foundational references. These papers summarize the objectives and methodologies of Task 1 within the CLEF framework. Additional contributions from various participants in Task 1 have been documented in subsequent works [14-22].

7. Acknowledgments

We would like to thank the contributors – CLEF'24 of the dataset available at [<https://checkthat.gitlab.io>] for providing the data used in this research. We would like to give our acknowledgements and our sincere thanks to our mentors M.Saritha and Krithika Swaminathan who contributed their time and efforts and we also extend our heartfelt thanks to Sri Sivasubramaniya Nadar College of Engineering, for letting us know about this opportunity.

8. References

- [1] Williams E, Rodrigues P, Novak V. Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. arXiv preprint arXiv:2009.02431. 2020 Sep 5.
- [2] Ghanem R, Erbay H, Bakour K. Contents-based spam detection on social networks using RoBERTa embedding and stacked BLSTM. SN Computer Science. 2023 May 6;4(4):380.
- [3] Guo Z, Zhu L, Han L. Research on short text classification based on roberta-textrcnn. In 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI) 2021 Sep 17 (pp. 845-849). IEEE.
- [4] Savchev A. AI Rational at CheckThat!-2022: Using transformer models for tweet classification. In CLEF (Working Notes) 2022 (pp. 656-659).
- [5] Naseer M, Asvial M, Sari RF. An empirical comparison of bert, roberta, and electra for fact verification. In 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC) 2021 Apr 13 (pp. 241-246). IEEE
- [6] Kumari S. NoFake at CheckThat! 2021: fake news detection using BERT. arXiv preprint arXiv:2108.05419. 2021 Aug 11.

- [7] Karande H, Walambe R, Benjamin V, Kotecha K, Raghu TS. Stance detection with BERT embeddings for credibility analysis of information on social media. *PeerJ Computer Science*. 2021 Apr 14;7:e467.
- [8] HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*. 2020 May 1;15(5):e0232525.
- [9] Grancharova M, Jangefalk M. Comparative study of the combined performance of learning algorithms and preprocessing techniques for text classification.
- [10] Khadhraoui M, Bellaaj H, Ammar MB, Hamam H, Jmaiel M. Survey of BERT-base models for scientific text classification: COVID-19 case study. *Applied Sciences*. 2022 Mar 11;12(6):2891.
- [11] Barrón-Cedeño A, Alam F, Chakraborty T, Elsayed T, Nakov P, Przybyła P, Struß JM, Haouari F, Hasanain M, Ruggeri F, Song X. The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. In *European Conference on Information Retrieval 2024 Mar 23* (pp. 449-458). Cham: Springer Nature Switzerland.
- [12] Maram Hasanain, Reem Suwaileh, Sanne Weering, Chengkai Li, Tommaso Caselli, Wajdi Zaghouni, Alberto Barrón-Cedeño, Preslav Nakov, and Firoj Alam. Overview of the CLEF-2024 CheckThat! lab task 1 on checkworthiness estimation of multigenre content.
- [13] Alberto Barrón-Cedeño, Firoj Alam, Julia Maria Struß, Preslav Nakov, Tanmoy Chakraborty, Tamer Elsayed, Piotr Przybyła, Tommaso Caselli, Giovanni Da San Martino, Fatima Haouari, Chengkai Li, Jakub Piskorski, Federico Ruggeri, Xingyi Song, and Reem Suwaileh. Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness. In Lorraine Goeriot, Philippe Mulhem, Georges Qu'énoc, Didier Schwab, Laure Soulier, Giorgio Maria Di Nunzio, Petra Galušćáková, Alba García Seco de Herrera, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, 2024.
- [14] B. Bharathi, D. Dilsha Singh, and K. Harinishree. Aqua Wave at CheckThat! 2024: Check-worthiness estimation.
- [15] Kushal Chandani and Dua E Zehra Syeda. Checker Hacker at CheckThat! 2024: Ensemble models for check-worthy tweet identification. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*.
- [16] Md. Sajid Alam Chowdhury, Anik Mahmud Shanto, Mostak Mahmud Chowdhury, Hasan Murad, and Uday Das. Fired from NLP at CheckThat! 2024: Estimating the check-worthiness of tweets using a fine-tuned transformer-based approach.
- [17] Mirela Dryankova¹, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. Mirela at CheckThat! 2024: Check-worthiness of tweets with multilingual embeddings and adversarial training.
- [18] Sanjai Balajee Kannan Giridharan, Sanjhit Sounderrajan, B Bharathi, and Nilu R. Salim. SSN-NLP at CheckThat! 2024: Assessing the checkworthiness of tweets and debate excerpts using traditional machine learning and transformer models.
- [19] Yufeng Li, Rrubaa Panchendrarajan, and Arkaitz Zubiaga. FactFinders at CheckThat! 2024: Refining check-worthy statement detection with LLMs through data pruning.
- [20] NA. Team Artists at CheckThat! 2024: Text-based binary classification for check-worthiness detection.

[21] Abdullah Al Mamun Sardar, Kaniz Fatema, and Mohammad Ashraf Islam. JUNLP at CheckThat! 2024: Enhancing check-worthiness and subjectivity detection through model optimization.

[22] Symom Hossain Shohan, Ashraf Islam Paran, Md. Sajjad Hossain, Jawad Hossain, and Mohammed Moshiul Hoque. SemanticCuetSync at CheckThat! 2024: Finetuning transformer models for checkworthy tweet identification.