SCUoL at CheckThat! 2022: Fake News Detection Using Transformer-Based Models

Saud Althabiti [a,b], Mohammad Ammar Alsalka [c] and Eric Atwell [d]

[a] *University of Leeds, scssal@leeds.ac.uk, Leeds, United Kingdom*
[b] *King Abdulaziz University, salthabiti@kau.edu.sa, Jeddah, Saudi Arabia*
[c] *University of Leeds, m.a.alsalka@leeds.ac.uk, Leeds, United Kingdom*
[d] *University of Leeds, e.s.atwell@leeds.ac.uk, Leeds, United Kingdom*

Abstract

The fifth edition of the "CheckThat! Lab" is one of the 2022 Conference and Labs of the Evaluation Forum (CLEF) and aims to evaluate advances supporting three factuality-related tasks, covering several languages. Our team (SCUoL) participated in task 3A, which concentrates on multi-class fake news detection of English news articles. This paper describes our approach, including several experiments exploring different machine learning and transformer-based models. Furthermore, we employed an additional dataset to support our proposed model. During the validation results phase, the experiments highlight the best performing machine learning classifier, which achieved cross-validation scores of over 60% for the LinearSVC compared to the pre-trained BERT model that exceeds other models in this task. While in the testing results, we obtained an F1 of approximately 0.305 compared to the other participants' average F1 of 0.252.

Keywords

Fake News Detection, Misinformation, Misleading Information, CLEF 2022[1], CheckThat! Lab

1. Introduction

Fake news has evolved into a severe threat which may cause increased political, financial and societal losses making detection of such news significant because, unlike traditional newspapers, online information could mislead a broader range of communities. In addition, it may include both facts and parts of incorrect contents in one statement, which can be challenging to identify. Therefore, various research studies have developed models to find a beneficial solution to tackle this difficulty.

The CheckThat! Lab [1]–[5] provides annual competitions divided into three categories, namely, check-worthiness estimation, verified claim retrieval, and fake news detection (FND). In this experiment, we participated in the third task [6]. The performed task aims to detect fake news articles written in the English language and their topical domains. We focused only on the first part (Task 3A) in this experiment. This subtask provided a dataset including almost 1,300 articles divided into training and development datasets to determine whether each article's claim is false, partially false, true, or other [7]–[9]. Therefore, the main objective of this experiment is to classify real-world news articles into predefined categories. Each text has been labelled with a specific rating; hence, this is a supervised text classification problem aiming to rank upcoming news based on the article's content.

In this paper, we firstly discuss some of the related topics and previous studies in section 2. Then, we analyze the datasets and describe the methodology in sections 3 and 4, respectively. The following section discusses the results obtained after training the model, and finally, we conclude and suggest future work in the last section.

[1] CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

2. Related Work

Defining claims' credibility is a research problem that has drawn considerable attention in past years, and various studies have been developing methods to overcome this issue [9]–[14]. Since this topic has become trending in many languages, several surveys have attempted to review the various suggested techniques and practical approaches systematically. For instance, the study by Shu and Liu [15] details fake news detection methods in five categories: linguistics, topic-agnostic, knowledge-based, traditional machine learning, and hybrid approaches. Furthermore, they examined how these methods could interlink to be used jointly.

In another study by [12], [16], they provided a descriptive tutorial that reevaluated FND techniques and methods from four different viewpoints. The first viewpoint evaluates authenticity by extracting the facts and comparing them with knowledge. An additional perspective is apprehending the writing style of given news since manipulators who aim to distribute fake news usually spread distorted messages intended to persuade others [16]. Furthermore, the propagation of the spread of information is the third perspective. To explain, the route of widespread news messages forms a network that could hold indications for early fake news detection. Finally, a source-based method is another employed idea. This method mainly relies on the source of a particular post, such as the original news authors, the publishers who conveyed that news, or the person who shared the posted news [17].

The CheckThat! lab is part of the conference and labs of the evaluation forum (CLEF) [4], [7]. It has provided contests since 2018 [1] and attempts to assess competitors' systems each year related to factuality in different languages. It is divided into three challenges [4]: the first one is to predict which tweets are worth fact-checking [18]. Our team (SCUoL) participated in this competition last year and achieved the third-best result among eight other participating team [18], [19]. The second challenge is to decide whether a posted claim can be verified or not [20]. The last task is task 3, which aims to predict the veracity of a news article [7].

3. Datasets

3.1. Dataset provided from the competition

In this competition, the data was collected from 2010 to 2022 with more than one topic, such as elections, COVID-19 etc. [21]. The provided English dataset is about 1,300 articles for the training data with four main features: public id, the text, the title, and the rating or class-label of each article. In addition, the testing dataset includes more than 600 English articles with similar features except for the rating, which our proposed model aims to predict. The text provides the most important features that can help solve this multi-classification problem and determine the veracity of an article [22]. Therefore, we decided to use the article text only to be analyzed along with the rating. There are four different classes in the provided dataset (false, partially false, true, and other). We initially represent each category as a number since the proposed transformer model only accepts numerical classes.

3.2. External datasets

Team NoFake, the winner of last year's competition, used additional datasets [23], which increased their model performance. Therefore, we used an external dataset called the Fakenews Classification Datasets[2] from a Kaggle competition in our experiments. It contains more than 21,000 factual articles and over 23,000 fake articles.

4. Methods

4.1. Text pre-processing and machine learning models

[2] https://www.kaggle.com/datasets/liberoliber/onion-notonion-datasets

We conducted several experiments using various methods, including traditional machine learning and transformer-based models. Initially, the training datasets were divided into features and labels. In order for machine learning (ML) algorithms to be able to make predictions, all words included within each article's text have to be transformed into vectors. In this experiment, we used a statistical measure called TF-IDF. This measure stands for (Term Frequency - Inverse Document Frequency) and aims to evaluate how relevant each word is to a specific document in several documents [24]. The repeated words that may show in most or all documents, such as the words that, who, and which, will have a lower weight because these kinds of words will not add any valuable information for our predictions. The importance of each word will be determined by multiplying TF by IDF. In addition, we set some hyperparameters in the used vectorizer to minimize the unnecessary words, such as the 'stop words'. Then, we unified labels that indicate the same meaning and converted them into numerical features. For example, False, false, and untrue are represented as "1", and we also applied this to other labels. The last step in the pre-processing is splitting the data into training and testing with approximate ratios of 70% and 30%, respectively.

After that, we used four frequently used machine learning algorithms in classification problems: Random Forest Classifier (RFC), Linear Support Vector Classifier (SVC), Multinomial Naive Bayes (MNB), and Logistic Regression (LR). Then, we examined these models on the split training dataset and estimated the average prediction score over five folds of cross-validation.

4.2. Transformer-based models

In addition to the traditional machine learning algorithms, we aimed to utilize transformer-based models. To simplify the process of using such models, we employed an NLP library called "Simple Transformers", which includes multiple models, each intended to perform a particular task. For example, the "ClassificationModel" is designed to implement binary and multi-class text classification tasks. It can also implement other tasks, such as named entity recognition, multi-label text classification, question answering, language generation, and more tasks. The applied simple-transformer-based model contains a classification layer on top of the chosen transformer model. This layer has four output neurons corresponding to each class (true, false, partially false, or other). After creating the model, we specified the selected pre-trained model types and architectures based on the following supported models:

- **BERT**: The acronym indicates Bidirectional Encoder Representations from Transformers. The BERT model differs from other language models because it is open-source and developed to pre-train deep bidirectional language representations exclusively by using a plain unsupervised text [25].
- **XLNet** is a generalised autoregressive pretraining model. It combines a bidirectional context and avoids independent predictions to overcome the limitations of BERT based on its autoregressive formulation [26].
- **RoBERTa:** A Robustly Optimised BERT Pretraining Approach is built based on BERT with a modification on the hyperparameters; for instance, it is trained for a more extended time on bigger batches and learning rates [27].
- **DistilBERT** is another pre-trained transformer model. However, unlike previous models, this model is a distilled version that aims to reduce the size of a BERT model by 40 per cent and make it faster by 60 per cent while having more than 95 per cent of its language understanding abilities [28].

5. Results and Discussion

This section illustrates and discusses the results of the various experiments we conducted in this competition. As described in section 4, we investigated multiple ML algorithms (RFC, SVC, MNB, and LR) and transformer-based models (BERT, XLNet, RoBERTa, and DistilBERT). Firstly, we assessed the average prediction scores over five cross-validation folds. We only used the accuracy metric in the first experiment to choose the best performing ML classifier and then compare this selected classifier with transformer-based models based on the macro F1 score. The initial results indicate that the SVC usually outperforms other traditional ML models, as exhibited in Figure 1 with an average accuracy score of 0.61 compared to LR, MNB, and RFC with 0.57, 0.52, and 0.52, respectively. After that, we employed a simple transformer multi-classification model and fine-tuned it on both the learning rate and the number of epochs hyperparameters. Since the BERT model has been proven to provide state-of-the-art results in many studies, we employed it as the simple transformer model type. We conducted over 20 experiments, using BERT-large-cased with a learning rate ranging from 1.00E-4 to 1.00E-6 and several epochs ranging from 5 to 25. We concluded that the best combination of these hyperparameters was using the 1.00E-5 learning rate with five epochs. Then, we examined the other transformer-based models using these parameters. However, the comparison shows that the used combination delivers better scores with the BERT model than other transformer-based models, as presented in Table 1.

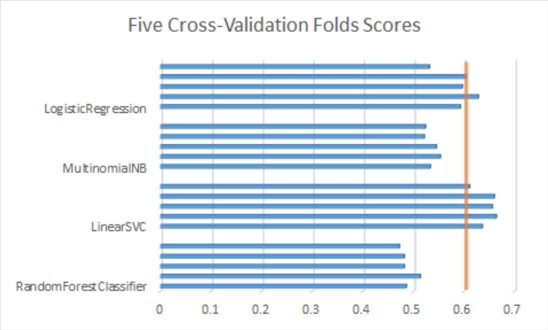

Figure 1: Applying cross validation score function on the four ML models

Table 1
The highest scores received from the employed models on the validation set

Model Name	Acc.	F1 Score
LinearSVC	0.62	0.51
bert-base-cased	0.43	0.46
bert-large-cased	**0.63**	**0.53**
xlnet-large-cased	0.53	0.38
roberta.large	0.59	0.47
distilbert-base-cased	0.58	0.45

In addition, we attempt to enhance the model using additional datasets. We used the Fakenews Classification Datasets described in subsection 3.2 with different numbers of samples. We only used 500 fake news and 500 real news in the first attempt and combined them with the provided dataset. Accordingly, we increased the amount of news to 1000 and 2000 samples in two different tries. Nevertheless, the outcomes still indicate that the previous fine-tuned model using only the provided dataset from the contest outperformed all other attempts. The reason here could be that the additional dataset only includes binary labels (false or real), while this task aims to classify a multiclassification problem. As a result, we decided to train the final model with only the provided dataset from the competition so that the model can only see similar data in the same format. The testing results released on the leaderboard for this task show that our team achieved 0.305 on the F1 measurement. Our score nearly reached the highest, which is 0.339, and it is higher than the average F1 of other participants at 0.252.

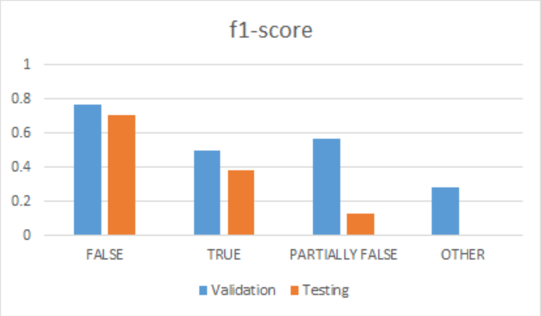

Figure 2: F1 scores for evaluating the model on the validation and testing sets

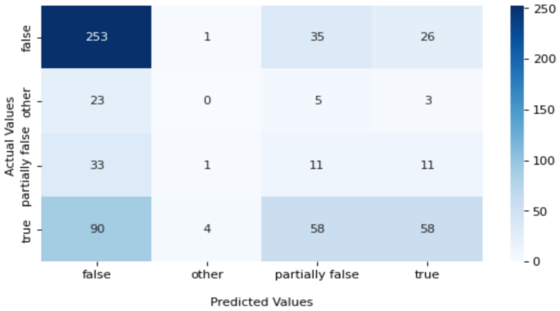

Figure 3: Confusion matrix on the testing set

For evaluation, we used a classification report and confusion matrix to assess our model during the validation phase. We also use these matrices to evaluate the submitted predictions on the released gold labels. Our observations show that our model performed better during the validation phase than in the testing, as illustrated in Figure 2. The received validation F1 scores are 0.77, 0.5, 0.57, and 0.28 for the labels 'false', 'true', 'partially false', and 'other', respectively. However, on the testing set, the model mispredicted most of the 'partially false' once, and none of the 'other' labels was predicted correctly, although it behaved close enough when predicting the 'false' and 'true' articles. Moreover, we observed from the presented confusion matrix in Figure 3 that the model mostly predicted the label 'false' with about 65% of the total predictions. In contrast, the positive labels are much fewer than the 'predicted-as-false' labels. One possible reason is that the labelling criteria in the training set differ from the last released testing set.

6. Conclusion and Future Work

Obtaining reliable information is considered an essential factor in our daily lives, especially when it comes to reading news. Due to the extensive number of published online articles, many developers have investigated and developed various models to tackle infodemic. This paper describes our system and participation in the "CLEF CheckThat! Lab" Task-3A competition. We examined an English dataset labelled as whether a particular article is 'true', 'false', 'partially false' and 'other'. We investigated four ML algorithms and pre-trained transformers to solve this multi-classification problem. Additionally, we attempted to use an external dataset from Kaggle to help improve the model. However, the additional dataset did not increase the performance, even though we used a different number of samples in each attempt. Finally, our findings from over 30 experiments show that the BERT model outperforms other models. The obtained testing results on the leaderboard indicate that we got an F1 of around 0.305, which slightly differs from the highest participant's score with only about 0.03. In future work, we recommend that finding an additional dataset with a similar format may help improve the model. Also, using an ensemble method, which considers both rule-based and deep learning methods, could significantly enhance the proposed system.

7. Acknowledgements

We would like to acknowledge and give our warmest thanks to the Ministry of Education in Saudi Arabia, King Abdulaziz University, and University of Leeds for their continuous support.

8. References

[1] P. Nakov *et al.*, “Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims,” in *International conference of the cross-language evaluation forum for european languages*, 2018, pp. 372–387.

[2] T. Elsayed *et al.*, “Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims,” 2019, doi: 10.1007/978-3-030-28577-7_25.

[3] A. Barrón-Cedeño *et al.*, “CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media,” 2020, doi: 10.1007/978-3-030-45442-5_65.

[4] P. Nakov *et al.*, “The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news,” in *European Conference on Information Retrieval*, 2021, pp. 639–649.

[5] P. Nakov *et al.*, “Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection,” 2022.

[6] J. K¨ohler, M. Shahi, Gautam Kishore Struß, Julia Maria Wiegand, M. Siegel, T. Mandl, and M. Sch¨utz, “Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection,” 2022.

[7] G. K. Shahi, J. M. Struß, and T. Mandl, “Overview of the CLEF-2021 CheckThat! Lab: Task 3 on fake news detection,” 2021.

[8] P. Nakov *et al.*, “The CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection,” in *European Conference on Information Retrieval*, 2022, pp. 416–428.

[9] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, “An exploratory study of covid-19 misinformation on twitter,” *Online Soc. networks media*, vol. 22, p. 100104, 2021.

[10] R. Mouty and A. Gazdar, “Survey on Steps of Truth Detection on Arabic Tweets,” 2018, doi:

10.1109/NCG.2018.8593060.

[11] G. Jardaneh, H. Abdelhaq, M. Buzz, and D. Johnson, “Classifying Arabic tweets based on credibility using content and user features,” 2019, doi: 10.1109/JEEIT.2019.8717386.

[12] X. Zhou and R. Zafarani, “Fake News: a survey of research, Detection Methods, and Opportunities,” *ACM Comput. Surv.*, pp. 1–40, 2018.

[13] S. Gaonkar, S. Itagi, R. Chalippatt, A. Gaonkar, S. Aswale, and P. Shetgaonkar, “Detection of Online Fake News : A Survey,” 2019, doi: 10.1109/ViTECoN.2019.8899556.

[14] G. K. Shahi and D. Nandini, “FakeCovid--A multilingual cross-domain fact check news dataset for COVID-19,” *arXiv Prepr. arXiv2006.11343*, 2020.

[15] K. Shu and H. Liu, “Detecting Fake News on Social Media,” *Synth. Lect. Data Min. Knowl. Discov.*, 2019, doi: 10.2200/s00926ed1v01y201906dmk018.

[16] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” *ACM Comput. Surv.*, 2020, doi: 10.1145/3395046.

[17] N. Y. Hassan, W. H. Gomaa, G. A. Khoriba, and M. H. Haggag, “Credibility detection in twitter using word n-gram analysis and supervised machine learning techniques,” *Int. Food Agribus. Manag. Rev.*, 2020, doi: 10.22434/IFAMR2019.0020.

[18] S. Shaar *et al.*, “Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates,” *Work. Notes CLEF*, 2021.

[19] S. Althabiti, M. Alsalka, and E. Atwell, “SCUoL at CheckThat! 2021: An AraBERT model for check-worthiness of Arabic tweets,” 2021.

[20] S. Shaar *et al.*, “Overview of the CLEF-2021 CheckThat! Lab Task 2 on detecting previously fact-checked claims in tweets and political debates,” in *CEUR Workshop Proceedings*, 2021, vol. 2936, pp. 393–405.

[21] G. K. Shahi, “Amused: An annotation framework of multi-modal social media data,” *arXiv Prepr. arXiv2010.00502*, 2020.

[22] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, “Content based fake news detection using knowledge graphs,” in *International semantic web conference*, 2018, pp. 669–683.

[23] S. Kumari, “NoFake at CheckThat! 2021: fake news detection using BERT,” *arXiv Prepr. arXiv2108.05419*, 2021.

[24] K. Poddar and K. S. Umadevi, “Comparison of various machine learning models for accurate detection of fake news,” in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2019, vol. 1, pp. 1–5.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv Prepr. arXiv1810.04805*, 2018.

[26] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[27] Y. Liu *et al.*, “Roberta: A robustly optimized bert pretraining approach,” *arXiv Prepr. arXiv1907.11692*, 2019.

[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv Prepr. arXiv1910.01108*, 2019.