VTU_BGM at CheckThat! 2022: An Autoregressive Encoding Model for Detecting Check-worthy Claims

Sanjana Kavatagi[1], Rashmi Rachh[2] and Madhura Mulimani[3]

[1]Dept of Computer Science and Engineering, VTU Belagavi, Karnataka, India
[2]Dept of Computer Science and Engineering, VTU Belagavi, Karnataka, India
[3]Dept of Computer Science and Engineering, VTU Belagavi, Karnataka, India

Abstract

Monitoring social media content has recently captured the attention of the artificial intelligence community due to the widespread use of social media platforms. To encourage research in this direction, the 5th edition of the CheckThat! lab was organized as part of CLEF-2022, on Fighthing the infodemic and fake news detection. In this paper, we present the results of the language model developed by our team VTU BGM at CheckThat! 2022 for determining the check-worthiness of tweets and verifying factual claims as part of the competition. In the proposed model we have used autoregressive model XLNet for feature extraction and Support Vector Machine for the classification of tweets. Our team positioned at 11th and 7th place for Subtask 1A and Subtask 1B, respectively.

Keywords

Fake news, Auto-regressive, Embedding, SVM

1. Introduction

With the increased availability of internet, social media has grown in popularity. With more content being easily shared on social media, vulnerabilities such as fake news and hate speech have emerged [1, 2, 3]. Fake news is defined as misinformation or disinformation that is widely disseminated as news on various platforms. Fake news has the capacity to misguide the people and its impact on the society is definitive [4]. Recent proof for this is, during the elections of West Bengal fake news and hate speech spread on social media resulted in serious clashes between the political parties. These types of issues are common during elections and natural disasters like COVID-19.

As a counter measure to prevent this type of content being spread on social media, people have tried different approaches. The traditional method is to verify the fake content by manually, however these methods are time consuming and inefficient. Further, a team of AI researchers have taken keen interest in identifying this fake news and save people from the adverse effects of fake news. This aims to automate the identification of the fake news and thus saves humans' time and effort to identify. Fake news on social media covers various areas like politics, finance, religion and natural disasters. The most recent and ongoing pandemic is COVID-19. During

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

kawatagi.sanjana@gmail.com (S. Kavatagi); rashmirachh@vtu.ac.in (R. Rachh); madhurasm@gmail.com (M. Mulimani)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3180/paper-42.pdf

the outbreak of the COVID-19 pandemic, people relied on or used the social media across the world [5]. The social media platforms such as Facebook, Twitter, Instagram, Youtube, etc., have become the major source of disinformation [6]. To promote research to fight against the fake news, the organizers of CLEF group shared 5th version of their lab-Check That! 2022 on Fighting the COVID-19 Infodemic and Fake News Detection [7]. The competition is conducted in three shared tasks which were aiming at fighting the misinformation or disinformation being spread on social media regarding the political debates in the news. The shared task was conducted in seven languages namely Arabic,Bulgarian, Dutch, English, German, Spanish and Turkish.

Task 1 focuses on Identifying relevant claims in tweets. Task 2 concentrates on detecting previously fact-checked claims and using a set of previously fact-checked claims to find out whether the claim has been previously fact-checked was supposed to be determined. Task 3 aims to detect the fake news. The text of a news article was given the main claim made in the article is to determine whether its true, partially true, false or others. This task was offered as a mono-lingual task for English and cross-lingual task for German and English. The idea of the latter is to use the English data and the concept of transfer learning to build classification model for the German language as well. Our team took part in Subtask 1A and Subtask 1B of Task 1: Identifying relevant claims in tweet. This task comprises of four subtasks. Subtask 1A is to detect the check-worthiness of tweets, given a tweet predict whether it is worth fact-checking or not. This task is defined with binary labels 0 and 1 indicating yes or no respectively. This is a classification task run on 6 languages and we have submitted language model for English. Subtask 1B is about verifiable factual claims detection. Given a tweet, we were supposed to predict whether it contains a verifiable factual claim. This is also a binary classification task with labels 0 and 1 indicating Yes or No respectively. This Subtask was run on 5 different languages and our team submitted a model for English language.

The remainder of the article is organised as follows: Section 2 provides an overview of previous work, Section 3 gives details about the dataset, while, Section 4 discusses our proposed model, Section 5 articulates insights about our findings, and Section 6 provides a conclusion and future work.

2. Literature Review

We provide the related work in two subsections which are relevant to our two subtasks, viz check-worthiness of tweets and verifiable factual claims detection.

2.1. Check worthiness of tweets

There are various methods for detecting the check-worthiness of tweets. Claim check-worthiness detection, the first step in automatic fact checking, is a binary classification problem that involves identifying whether a piece of text makes a proclamation about the world that can be checked [8]. Several studies have been conducted in an attempt to develop mechanisms for the extraction of check-worthy statements. Claim check-worthiness further involves (i) Claimrank [9] which employs a diverse set of features derived from individual sentences and the context of the situation. (ii) Claimbuster which combines various techniques like parts of speech tags (POS tags) , term frequency and inverse document frequency(TF-IDF) on

Table 1
Dataset description for Subtask 1A

	Checkworthy	**Non-checkworthy**	**Total**
Train Dataset	1675	447	2122
Validation Dataset	151	44	195
Test dataset	110	39	149

support vector machine(SVM) to generate the labels for each of the claims [10]. In the CLEF 2019 evaluation lab on check worthiness detection [11], most of the participants used neural networks for the classification/checking the worthiness of the claims. In CLEF CheckThat! 2020 and CLEF CheckThat! 2021 evaluation labs, the participated teams have used various embedding techniques of transformer models along with other neural networks successfully [12, 13].

2.2. Verifiable factual claim detection

Many approaches have been employed by researchers to tackle the concerns of misinformation or disinformation being spread [14, 15], in such approaches verifying the veracity of the claims is an important task. The process of comparing the veracity of a claim to relevant evidence is known as fact-checking [16]. The fact-checking task was previously conducted manually by journalists [17]. As the internet has become a burgeoning source of provocative comments from misleading news reports, party leaders, media speculation, and other domains, the need for developing an automated method for describing the veracity of claims has grown [18]. This challenge is addressed by two categories of computational methods database-based and AI-based. The central emphasis of AI-based systems is on features and patterns that can be used to anticipate the veracity of claims using machine learning techniques [19, 20, 21]. Database-based approaches, assuming sufficient relevant data, use knowledge bases to verify claims [22].

3. Dataset Distribution

We used datasets from the CLEF 2022 CheckThat! Lab for Fighting the COVID-19 infodemic and fake news detection [23]. The datasets for all three tasks(mentioned in the preceding section) have been provided by the organisers. We have participated in task 1 competition and submitted the models for Subtask 1A: Check-worthiness of tweets and Subtask 1B: Verifiable factual claims detection. The organizers of the task have provided the datasets in 6 different languages including English. We have used the dataset provided by the organizers in English for model building. The task providers have released data in three different phases such as training, validation, and testing datasets for subtasks 1A and 1B. The datasets contains the fields: topic, tweet id, tweet_url, tweet_text, and class_label. Tables 1 and 2 contain dataset details for subtasks 1A and 1B, respectively [23].

Table 2
Dataset description for Subtask 1B

	Verifiable factual claim	Non-verifiable factual claim	Total
Train Dataset	1202	2122	3324
Validation Dataset	112	195	307
Test Dataset	102	149	251

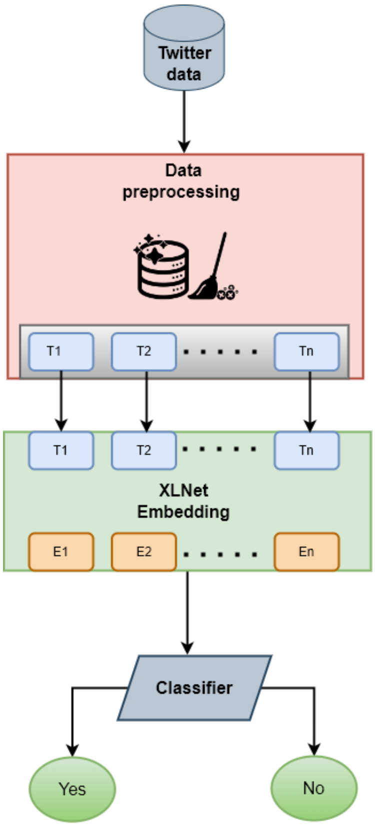

Figure 1: Architecture of the proposed model

4. Methodology

Figure 1 depicts the proposed model's architecture. As shown in the figure, we pre-process the data before passing it to an auto-regressive and auto-encoding model for feature extraction, followed by the addition of a classifier. Further subsections provide a detailed description of the architecture.

Table 3
Top performing model for Subtask-1A

Participants (userid/team-name)	Subtask	F1 (postive class)	Rank
Asavchev	Subtask-1A-Checkworthy-English	0.698	1
nicuBuliga	Subtask-1A-Checkworthy-English	0.667	2
Team_PoliMi-FlatEarthers	Subtask-1A-Checkworthy-English	0.626	3
Mkutlu	Subtask-1A-Checkworthy-English	0.561	4
fraunhofersit_checkthat22	Subtask-1A-Checkworthy-English	0.552	5
VTU_BGM	**Subtask-1A-Checkworthy-English**	**0.482**	**11**

4.1. Pre-processing

The data provided by the organizers is not in the format required by the model. As a result, some preliminary processing is carried out in order to prepare it for transmission to our model. Hyperlinks, white-spaces, special characters, and numbers were removed from the tweet text as they do not contribute to determining the worthiness or veracity of the claim. Furthermore, the information obtained from social media is not grammatically correct. As a result, lemmatization is used to convert it into meaningful statements by bringing it back to the base words. All of the words were converted to lower case to remove the redundant terms. The NLTK toolkit from the Python library [24] was used to complete all of the pre-processing steps. This preprocessed data is then fed into the tokenizer, which turns all tweets into tokens. Padding and masking were used to manage variable-length sentences. This preprocessed data acts as input to the feature extraction model.

4.2. Feature Extraction

According to the previous literature review, XLnet [25] works as the most effective method for extracting corpus features that can help us identify offensive content.We implemented the XLNet embedding approach to extract features. XLNet is a transformer-based generalised autoregressive and autoencoding approach [26]. It is a pre-trained method that can learn bidirectional contexts by maximizing the expected likelihood across all permutations of the factorization order. XLNet has 12 layers, 12 attention heads and 768 hidden layers and works based upon encoder and decoder approach. It contains [CLS] and [SEP] tokens at the end, and in our model, we have used embeddings generated by [CLS] token that provides full sentence embeddings.

4.3. Classifier

We used traditional machine learning algorithms such as Support Vector Machine (SVM)[27] for classification of checkworthy claims and verifiable factual claims since we know it produces promising results based on the literature review. Subtasks 1A and 1B are both binary classification problems. 10-fold cross validation is done on the SVM. The tweet data is divided into 10 smaller sets and each data point is allotted to one of the subsets of almost equal size. When the method is applied to the training data set, an individual model for each of these subsets is built.

Table 4
Top performing model for Subtask-1B

Participants (userid/team-name)	Subtask	Accuracy	Rank
Team_PoliMi-FlatEarthers	Subtask-1B-Claim-English	0.761	1
manansuri	Subtask-1B-Claim-English	0.749	2
Team_NLP&IR@UNED	Subtask-1B-Claim-English	0.725	3
asavchev	Subtask-1B-Claim-English	0.713	4
nicuBuliga	Subtask-1B-Claim-English	0.709	5
VTU_BGM	**Subtask-1B-Claim-English**	**0.709**	**7**

The average of the results of all model evaluations is then used to calculate the cross-validation value. The resulting model serves as a test set for calculating performance metrics such as accuracy.

5. Result

The teams were ranked in the competition based on the F1 score measure for the positive class in Subtask 1A and the accuracy for Subtask 1B. Among the participating teams, our language model was ranked 11^{th} for Checking the Worthiness of Claims and 7^{th} for Verifying Factual Claims. The results of the top five ranked teams, as well as our team, are shown in Table 3 and Table 4 below. Our model performance is highlighted in bold letters.

6. Conclusion and future enhancement

Our team has presented a language model at CLEF 2022 CheckThat! Fighting the COVID-19 Infodemic and Fake News Detection for two subtasks: predicting the worthiness of claims and verifying the factuality of the claims. We used XLNet embedding techniques in the proposed language model for its autoregressive and autoencoding properties and SVM classifier for tweet classification. This work can further be extended by including regional languages and code-mixed texts.

References

[1] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management 57 (2020) 102025.

[2] S. Kavatagi, R. Rachh, A context aware embedding for the detection of hate speech in social media networks, in: 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), IEEE, 2021, pp. 1–4.

[3] S. Biradar, S. Saumya, A. Chauhan, Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 2470–2475.

[4] S. Biradar, S. Saumya, A. Chauhan, Combating the infodemic: Covid-19 induced fake news recognition in social media networks, Complex & Intelligent Systems (2022) 1–13.

[5] M. K. Elhadad, K. F. Li, F. Gebali, Detecting misleading information on covid-19, Ieee Access 8 (2020) 165201–165215.

[6] B. Al-Ahmad, A. Al-Zoubi, R. Abu Khurma, I. Aljarah, An evolutionary fake news detection method for covid-19 pandemic information, Symmetry 13 (2021) 1091.

[7] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.

[8] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, Digital Threats: Research and Practice 2 (2021) 1–16.

[9] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, Claimrank: Detecting check-worthy claims in arabic and english, arXiv preprint arXiv:1804.07587 (2018).

[10] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, et al., Claimbuster: The first-ever end-to-end fact-checking system, Proceedings of the VLDB Endowment 10 (2017) 1945–1948.

[11] T. Elsayed, P. Nakov, A. Barrón-Cedeno, M. Hasanain, R. Suwaileh, G. D. San Martino, P. Atanasova, Overview of the clef-2019 checkthat! lab: automatic identification and verification of claims, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 301–321.

[12] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeno, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, et al., Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media, in: CLEF (Working Notes), 2020.

[13] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, et al., Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates, in: CLEF (Working Notes), 2021.

[14] P. Nakov, F. Alam, S. Shaar, G. Da San Martino, Y. Zhang, Covid-19 in bulgarian social media: Factuality, harmfulness, propaganda, and framing, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, pp. 997–1009.

[15] S. Biradar, S. Saumya, Iiitdwd@ tamilnlp-acl2022: Transformer-based approach to classify abusive content in dravidian code-mixed text, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 2022, pp. 100–104.

[16] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, arXiv preprint arXiv:2004.14974 (2020).

[17] D. Graves, Understanding the promise and limits of automated fact-checking (2018).

[18] F. T. Alkhawaldeh, Factual or non-factual claim: Verifying claims, International Journal of Advanced Studies in Computers, Science and Engineering 9 (2020) 1–8.

[19] Z. Jin, J. Cao, Y. Zhang, J. Luo, News verification by exploiting conflicting social viewpoints in microblogs, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[20] J. Ma, W. Gao, K.-F. Wong, Detect rumors in microblog posts using propagation structure via kernel learning, Association for Computational Linguistics, 2017.

[21] A. Zubiaga, M. Liakata, R. Procter, Exploiting context for rumour detection in social media, in: International conference on social informatics, Springer, 2017, pp. 109–123.

[22] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, A. Flammini, Computational fact checking from knowledge networks, PloS one 10 (2015) e0128193.

[23] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF ’2022, Bologna, Italy, 2022.

[24] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O’Reilly Media, Inc.", 2009.

[25] A. Gautam, V. Venktesh, S. Masud, Fake news detection system using xlnet model with topic distributions: Constraint@ aaai2021 shared task, in: International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation, Springer, 2021, pp. 189–200.

[26] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[27] S. Ratan, S. Sinha, S. Singh, Svm for hate speech and offensive content detection (2021).