International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET218366

# A Deep Learning Based Approach for Fake News Detection

**Srishti Sharma, Vaishali Kalra**
Department of Computer Science and Engineering, The NorthCap University, Gurugram, India

**Article Info**
Volume 8, Issue 3
Page Number: 01-05

**Publication Issue :**
May-June-2021

**Article History**
Accepted : 10 June 2021
Published: 15 June 2021

## ABSTRACT

Owing to the rapid explosion of social media platforms in the past decade, we spread and consume information via the internet at an expeditious rate. It has caused an alarming proliferation of fake news on social networks. The global nature of social networks has facilitated international blowout of fake news. Fake news has proven to increase political polarization and partisan conflict. Fake news is also found to be more rampant on social media than mainstream media. The evil of fake news is garnering a lot of attention and research effort. In this work, we have tried to handle the spread of fake news via tweets. We have performed fake news classification by employing user characteristics as well as tweet text. Thus, trying to provide a holistic solution for fake news detection. For classifying user characteristics, we have used the XGBoost algorithm which is an ensemble of decision trees utilising the boosting method. Further to correctly classify the tweet text we used various natural language processing techniques to preprocess the tweets and then applied a sequential neural network and state-of-the-art BERT transformer to classify the tweets. The models have then been evaluated and compared with various baseline models to show that our approach effectively tackles this problem.

**Keywords :** Fake News, Transfer Learning, Classification, Transformers, Gradient Boosting, Text Classification, Twitter.

## I. INTRODUCTION

The increasing proliferation of social media platforms is considered to be the key reason behind the dissemination of fake news at an unprecedented scale. This computerized data age has produced new outlets for content makers to fabricate imaginary articles to build readership or as part of psychological warfare, financial and political gain. News of questionable credibility breaks the genuineness equilibrium of the news network. The developing use of algorithms in robotized news circulation and creation has made it easy and inexpensive to provide news online at a fast pace. Gartner analysis predicts that "By 2022, the overwhelming majority of individuals in developed economies will devour more false knowledge than real information [1]". Social media is a perilous weapon when mishandled, abused or invaded. One of the biggest challenges is that in social networks, the influence of spread and impact of content sharing happens so quickly that contorted, incorrect or misleading information obtains a colossal probability

of causing significant negative societal effects, in practically no time, for millions of users. It is palpable that one of the utmost well-known fake news was considerably more rapidly diffused on a popular social networking giant than the most renowned credible standard newsflash at the time of ballots in USA in 2016 [2]. Social media platforms enable sharing, commenting and talking about news in seconds. For instance, 62% people in the United States admitted to getting newscasts via social networks in 2016, whereas in 2012, only 49% testified to obtaining updates via social media. It has also been revealed that social networking sites on the internet now outflank TV as the significant news source [3]. The issue of tackling and controlling the explosion of fake news needs immediate attention. Any endeavours to mislead or troll in the cyberspace through fake news or misleading content sources are now considered grave matters with supreme adequacy and warrants genuine efforts from security scientists. As such, there is a dire need to come up with a fake news detection and filtering system. This is of utmost importance that such systems are built, as they can help both news readers and tech companies alike. Since, the dynamic nature and varying styles of fake news are a great hurdle, the objective is to propose a Fake News Detection system that comprehensively considers the user characteristics, content and social context. This hybrid approach should yield us a robust and effective system to combat the fake news epidemic efficiently at early stages of its propagation. Through this work, we outline a system for fake news detection that makes use of a tool to detect and eliminate counterfeit sites amongst the results returned by search giants or news applications. This tool can be downloaded by the user and, then, be supplemented to the user's browser or any application that the user is making use of to acquire news feeds.

Social media has proven to be a powerful place to spread false news. In 2015, Chen et al. [4] highlighted the need for automated assistive tools to verify and evaluate the authenticity of news. They highlighted that with the continuous propagation of the internet and social media channels the distinction between user created content and mainstream media content is becoming hazy and difficult to stop.

Mi et al. [5] used Stacked Auto-encoder (SAE) to detect spam news and a multi-layer neural network to examine performance. The SAE outperforms several popular ML algorithms like Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and Artificial Neural Networks. Rubin et al. [6] discussed three kinds of fake news: a) serious fabrications, (b) large-scale hoaxes, and c) humorous fakes (news satire, parody, game shows), each was compared to certified genuine reportage, and their benefits and disadvantages as a corpus for content investigation and predictive modelling were evaluated. Conroy et al. [7] asserted that crossover approaches that combine language knowledge and machine learning with network-based social knowledge can be used to test the validity of a statement.

Zhao et al. [8] designed a rumor detection approach based on a detector that searches for the question patterns like "Is this true?", "Really?", and "What?". Perform clustering of such related posts, and do the ranking of those clusters based on statistical features of truly containing questioned factual statements. Jin et al. [9] detected rumor tweets by corresponding them with tested rumor articles and classified them as rumored or non-rumored tweets. For analysis they collected the data on presidential elections from a news website and applied the five feature extraction algorithms for the analysis and compared them to find the most appropriate algorithm for detecting the rumored tweets. They applied TF-IDF, Word2Vev, Doc2Vec, BM25 and Lexicon matching algorithm out of which TF-IDF outperformed the others.

Castellini et al. provided a system for detecting bogus Twitter profiles using artificial neural networks [10].

An anomaly detector was developed using a denoising autoencoder and a semi-supervised learning approach, and it was tested on benchmark datasets. Duong et al. [11] gave a provenance-aware approach along with the text to improve the accuracy of the system in comparison to the other popular systems. The provenance means considering the related articles for detection in case if any citation is also given in the article. The proposed model uses recurrent neural networks to consolidate the provenance data and the content of the post to intensify the precision of counterfeit news discovery system. Thota et al. [12] implemented a dense deep learning neural network with TF-IDF, Word2Vec, and BoW, and the TF-IDF approach outperformed the other approaches for identifying bogus newsflashes.

In [13], Vicario, Quattrociocchi, Scala, and Zollo introduced a system for timely discovery of polarizing content on the internet, taking into account user's polarization and confirmation bias, to predict possible potential targets vulnerable to disinformation with 77% accuracy. User's behavior and related characteristics on social media were took into consideration to build a FND classifier.

Yang et al. [14] investigated the unsupervised identification of fake news. The authenticity of the broadcast and the dependability of handlers were viewed as hidden random variables, and a probabilistic graphical model was developed to record the total generative range. Without any labelled results, an operative collapsed Gibbs sampling process was planned to determine both the validity of the bulletin and the credibility of the handlers. The activities of handlers on social networks were used to classify their attitudes toward the validity of news.

## II. METHODS AND MATERIAL

In this section, we outline the particulars of the proposed model for Fake News Detection. The proposed model for Fake News Detection works in 3 phases.

Dataset Used and Pre-processing

In this work, we utilize the dataset provided by Shu et. al. in FakeNewsNet. They've created a repository of news content from fact-checking websites like PolitiFact and GossipCop, that assign rating to news bulletins on a scale of 0 (fake) to 10 (real) to categorize news bulletins. We amass from this news bulletins having ratings of less than 5 as fake news stories. The dataset provided had tweet ids and their corresponding labels of either True or False. We used the python library tweepy to extract tweet features by utilizing the tweet ids. We collected a total of 27,910 tweets out of which 14,854 tweets had the label 'True' while 13,056 had the label 'False'. Since we are using tweets the data was highly unstructured and noisy. We performed the following pre-processing steps:

1. Converted emojis to a textual description. This was done using the emoji library in python. For eg. The emoticon 😂 was converted to face_with_tears_of_joy. All occurrences of emojis were replaced in this manner.
2. Any mentions of the following were normalized using the ekphrasis python library: 'url', 'email', 'percent', 'money', 'phone', 'user', 'time', 'url', 'date', 'number'.
3. All occurrences of the following were annotated: 'hashtag', 'allcaps', 'elongated', 'repeated', 'emphasis', 'censored'.
4. All hashtags were unpacked. For eg. #BreakingNews was unpacked to 'breaking news'. 5. All contractions were unpacked. For eg. 'can't' unpacked to 'can not'.
5. Word Segmentation was carried out using the Viterbi algorithm which in turn is adapted from [15]. For eg. "smallandinsignificant" segmented to "small and insignificant".

6. Spell correction was also performed using Peter Norvig's spell-corrector [16]. Lemmatization is not executed and punctuation marks are not eliminated as pre-trained embeddings are always used. No stop-word is removed for fluency.

Classification of Tweets Based on Twitter User Characteristics

For this approach we use the XGBoost classifier. It is an ensemble technique i.e., it uses a group of decision trees. It combines a group of weak learners to create a strong learner. XGBoost is a boosting based ensemble classifier. In boosting, classifiers are constructed successively by taking the error of the preceding step into consideration for the next step. It gives weights to every data point, when a data point is misclassified, its weight is amplified so that the consequent classifiers try to rectify the incorrect classifications.

Classification of Tweets Based on Tweet Text

Language typical pre-training has proven beneficial in learning widespread linguistic depictions. Hence, we will use transformers for Classification of tweets. For language pre-training, Bidirectional Encoder Representations from Transformers (BERT) has attained remarkable consequences in several language understanding errands. Pre-trained models on huge corpora are advantageous for text categorization and natural language processing responsibilities, that can circumvent training a model from scratch. BERT is constructed on top of a multi-layer bidirectional Transformer [17] and trained on plain text for masked word estimation and subsequent sentence estimation. BERT uses Transformer, an attention machinery that absorbs contextual relationships amongst words or sub-words. In its vanilla arrangement, Transformer comprises of two distinct machineries — an encoder to read the lingsutic input and a decoder to produce an estimation.

Implementation

For the first task i.e., tweet classification on the basis of user and tweet characteristics, we choose 14 attributes of a particular tweet namely: 'No. Of Favourites', 'No. Of Followers', 'No. Of Friends/Following', 'No. Of Tweets Posted', 'No. Of Retweets', 'Length of Username', 'Account Age', 'Verified Account', 'Length Of User Description', 'Total Tweets Liked By Account', 'Public List Mentions', 'URL Provided', 'Using Default Profile Image', 'Using Default Profile Theme'.

The XGBoost classifier was used. Its hyperparameter tuning was done using grid-search cross-validation. Accuracy, confusion matrix and ROC-AUC curve were used as metrics. The BERT-Base model has twelve encoder layers, seven hundred and sixty eight feed-forward networks and twelve attention heads. The batch size was taken to be 32. The maximum length of sequence at 152. The data preprocessing was done for the tweet texts using various natural language processing techniques.

Pytorch's implementation of transformers and ktrain was used. The training was carried out on Google Cloud using their Compute Engine by creating a virtual machine (VM) instance. This VM had virtual-CPUs, 30GB RAM and 1 NVIDIA Tesla T4 GPU. The training took around 10 hours to train. The dataset used had 27910 rows out of which 2791 samples were used as the test set, 2512 samples were used as the validation set and 22607 samples were used as the training set. Accuracy, confusion matrix and ROC-AUC curve were used as evaluation metrics. The scores of the model are compared with other baseline models. The learning rate was found by using the lr_find function by plotting the loss vs. the log learning rate. We selected the learning rate by taking a value in the middle of the falling loss i.e. 0.1e^-5.

We then used the autofit method that employs a triangular learning rate policy and trained 5 epochs.

## III. RESULTS AND DISCUSSION

The evaluation metrics of the classification of tweets based on user and tweet characteristics using XGBoost classifier are listed in Table I and its corresponding ROC curve is as shown in Fig. 1.

**Table I**. xgboost classification results

| | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| **False** | 0.78 | 0.81 | 0.80 | 2583 |
| **True** | 0.83 | 0.81 | 0.82 | 2999 |
| **Accuracy** | | | 0.81 | 5582 |
| **Macro avg** | 0.81 | 0.81 | 0.81 | 5582 |
| **Weighted avg** | 0.81 | 0.81 | 0.81 | 5582 |

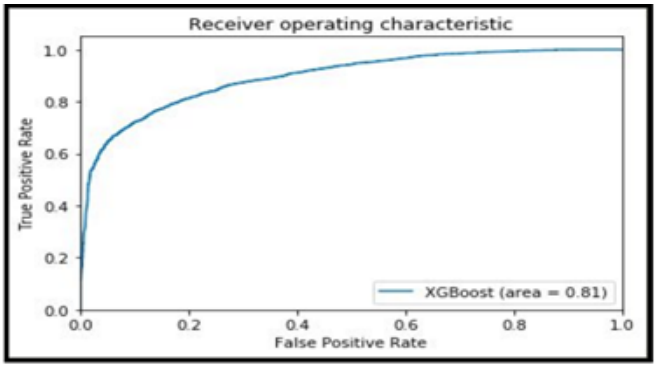


Fig. 1 : ROC curve for XGBoost Classifier

For the purpose of evaluation, we compare our proposed model with seven prominent models used by researchers in the field listed below and the comparative result of these models and our proposed fake news detection model using XGBoost classifier is outlined in Table II. It can be concluded that the proposed model is better than all the other models in terms of accuracy as well as F1 score.

BERT model for classification of tweets based on tweet text was trained for five epochs with a maximum learning rate of 0.1e-05. The model was trained on 22607 samples and validated on 2512 samples. The results obtained are listed in Table III.

TABLE II: COMPARISON OF PROPOSED XGBOOST MODEL WITH OTHER MODELS FOR CLASSIFICATION OF TWEETS BASED ON TWITTER USER CHARACTERISTICS

| **Method** | **Description** | **Accuracy** | **Real (F1)** | **Fake (F1)** |
|---|---|---|---|---|
| | | | | |
| DTC [18] | Decision Tree | 0.454 | 0.733 | 0.355 |
| SVM-RBF [19] | SVM (RBF Kernel) | 0.318 | 0.455 | 0.037 |
| SVM-TS [20] | SVM (Time Series) | 0.544 | 0.796 | 0.472 |
| DTR [8] | Decision Tree Ranking | 0.409 | 0.501 | 0.311 |
| GRU [21] | RNN | 0.646 | 0.792 | 0.574 |
| RFC [22] | Random Forest | 0.565 | 0.810 | 0.422 |
| PTK [23] | SVM (Propagation Tree Kernel) | 0.75 | 0.804 | 0.698 |
| Proposed | XGBoost | 0.81 | 0.82 | 0.8 |

Accuracy achieved by BERT is 98.53% and Matthews Correlation Coefficient (MCC) is 0.9706. BERT's performance is compared with other baseline models in Table IV. The baselines models considered are from work done by Ajao et. al. [24] The models compared are LSTM: vanilla model trained to detect fake tweets without preceding field knowledge of the subjects being deliberated, LSTMDrop: LSTM method with dropout regularization and LSTM-CNN: LSTM-CNN hybrid model. It can be clearly seen that the proposed BERT model for Classification of tweets based on tweet text far outperforms all other models.

TABLE III : BERT RESULTS

| | | Predicted | |
|---|---|---|---|
| | | True | False |
| Actual | True | 1331 | 26 |
| | False | 15 | 1419 |

TABLE IV: COMPARISON OF BERT AND OTHER MODELS

| Method | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| LSTM | 82.29 | 44.35 | 40.55 | 40.59 |
| LSTMDrop | 73.78 | 39.67 | 29.71 | 30.93 |
| LSTM-CNN | 80.38 | 43.94 | 39.53 | 39.70 |
| Our Approach (BERT) | 98.53 | 98.20 | 98.51 | 98.57 |

## IV. CONCLUSION

Through this study, we outlined a fake news detection model. We performed tweet classification based on user and tweet characteristics and achieved an accuracy of 81% using the XGBoost classifier. Further, we did tweet text classification using the BERT transformer which gave us an accuracy of 98%. The accuracies achieved by our models are superior to those attained by other baseline models. In the future, these models can also be extended for early detection of fake news taking the temporal features in consideration. These models can also be extended to news articles in the future. Future research work could also include enhancing the accuracy of a fake news detector model by making a hybrid of XGBoost and BERT classifiers using any ensembling technique such as bagging, boosting and stacking.

## V. ACKNOWLEDGMENT

We thank Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, Huan Liu from FakeNewsNet for making their data available, enabling our research.

## VI. REFERENCES

[1]. https://www.gartner.com/en/newsroom/press-releases/2017-10-03-gartner-reveals-top-predictions-for-it-organizations-and-users-in-2018-and-beyond

[2]. Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." Journal of Economic Perspectives, 31 (2): 211-36.

[3]. https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

[4]. Chen, Yimin & Conroy, Nadia & Rubin, Victoria. (2015). News in an Online World: The Need for an " Automatic Crap Detector ". The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015), Nov. 6?10.

[5]. Mi, Guyue & Gao, Yang & Tan, Ying. (2015). Apply Stacked Auto-Encoder to Spam Detection. 3-15. 10.1007/978-3-319-20472-7_1.

[6]. Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: three types of fakes. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST '15). American Society for Information Science, USA, Article 83, 1–4.

[7]. Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST '15). American Society for Information Science, USA, Article 82, 1–4.

[8]. Zhao, Zhe & Resnick, Paul & Mei, Qiaozhu. (2015). Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. 1395-1405. 10.1145/2736277.2741637.

[9]. Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In

Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press, 2972–2978.

[10]. Castellini, Jacopo & Poggioni, Valentina & Sorbi, Giulia. (2017). Fake Twitter followers detection by denoising autoencoder. 195-202. 10.1145/3106426.3106489.

[11]. Duong, Chi & Hung, Nguyen & Wang, Sen & Stantic, Bela. (2017). Provenance-Based Rumor Detection. 125-137. 10.1007/978-3-319-68155-9_10.

[12]. Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "FND: A Deep Learning Approach," SMU Data Science Review: Vol. 1: No. 3, Article 10. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/10.

[13]. Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and Fake News: Early Warning of Potential Misinformation Targets. ACM Trans. Web 13, 2, Article 10 (March 2019), 22 pages. DOI:https://doi.org/10.1145/3316809.

[14]. Liu, Yang & Wu, Yi-Fang. (2018). Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks.

[15]. Beautiful Data: The Stories behind Elegant Data Solutions, Toby Segaran and Jeff Hammerbacher, 2009, O'Reilly Media Inc.

[16]. Peter Norwig, How to Write a Spelling Corrector (norvig.com)

[17]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NIPS

[18]. Castillo, Carlos & Mendoza, Marcelo & Poblete, Barbara. (2011). Information credibility on Twitter. Proceedings of the 20th International Conference on World Wide Web. 675-684. 10.1145/1963405.1963500

[19]. Context and Dynamic Information for Studying Fake News on Social Media. ArXiv, abs/1809.01286.

[20]. Yang, Fan, Yang Liu, Xiaohui Yu and Min Yang. "Automatic detection of rumor on Sina Weibo." MDS '12 (2012).

[21]. Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Won. 2015. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15). Association for Computing Machinery, New York, NY, USA, 1751–1754. DOI:https://doi.org/10.1145/2806416.2806607.

[22]. Kwon, Sejeong & Cha, Meeyoung & Jung, Kyomin. (2017). Rumor Detection over Varying Time Windows. PLOS ONE. 12. e0168344. 10.1371/journal.pone.0168344.

[23]. Ma, Jing & Gao, Wei & Wong, Kam-Fai. (2017). Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. 10.18653/v1/P17-1066.

[24]. Context and Dynamic Information for Studying Fake News on Social Media. ArXiv, abs/1809.01286.

**Cite this article as :**

Srishti Sharma, Vaishali Kalra, "A Deep Learning Based Approach for Fake News Detection", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 8 Issue 3, pp. 388-394, May-June 2021. Available at doi : https://doi.org/10.32628/IJSRSET218366
Journal URL : https://ijsrset.com/IJSRSET218366