Ensembling Machine Learning Models for Urdu Fake News Detection

Hammad Akram[1], Khurram Shahzad[2]

[1]Department of Computer Science, National University of Computer and Emerging Sciences, Lahore, Pakistan
[2]Department of Data Science, University of the Punjab, New Campus, Lahore, Pakistan

Abstract

Fake news detection is recognized as a key natural language processing task. Recognizing the importance of the task, several attempts have been made for fake news detection in Western languages. However, fake news detection in Urdu has received little attention of researchers. A key reason to that is the scarcity of fake news datasets for Urdu. In order to promote research and development in the area, a track is dedicated to the second fake news detection in the Urdu language, UrduFake'21. This study has proposed to use ensembling machine learning models for Urdu fake news detection. The proposed approach employs a voting-based approach of the three most effective techniques to decide that the given news article is fake or real. For the evaluation of the proposed approach, experiments are performed using several classical machine learning techniques, three types of features, unigram, bigram and trigram and the released dataset. The results of the experiments revealed that the proposed approach is more effective than the individual techniques. According to the results released by the organizers our proposed approach achieved a macro average F1 and accuracy scores of 0.621 and 0.713, respectively, and it is ranked 4^{th} among the 19 submissions.

Keywords

Fake news, Urdu fake news, Machine learning, Classical machine learning

1. Introduction

Over the years, news and media have become an integral part of our life. According to Statista, the worldwide news, entertainment and media market has a market value of 2.1 trillion US dollars [1]. A news article that is intentionally and verifiably false is called fake news [2]. Several individuals, as well as organizations, purposefully publish fake news to support their purposes and interests. Typically, such news discredit individuals, organizations, communities and political parties undermine peace and stability in the society or to gain political mileage. The advent of social media has also played a prominent role in making such news viral, thus having a significant impact on the society. Therefore, it is desirable to detect fake news and control them from spreading.

Recognizing the importance of fake news, several studies have been conducted on fake news detection [3, 4]. In particular, since the 2016 presidential elections, fake news detection has

Forum for Information Retrieval Evaluation 2021, December 16–20, 2021, India

hammad.fast.nu@gmail.com (H. Akram); khurram@pucit.edu.pk (K. Shahzad)
https://www.linkedin.com/in/muhammad-hammad-akram-3a754265/ (H. Akram);
https://www.linkedin.com/in/mkhurramshahzad/ (K. Shahzad)
0000-0002-7392-1957 (H. Akram); 0000-0001-8433-6705 (K. Shahzad)

received considerable attention of researchers [5]. Consequently, resources have been developed for fake news detection NLP task for English and Chinese language, however little attention has been paid to fake news detection in South Asian languages. More specifically, fewer studies have been conducted for fake news detection and there is scarcity of the relevant resources for this task in the Urdu language despite having over 100 million speakers worldwide.

To the best of our knowledge, [6, 7] commenced fake news detection in Urdu. To promote research and developed in Urdu fake news detection, Center for Computing Research (CIC), Instituto Politécnico Nacional (IPN), Mexico, introduced a track at the 12^{th} Forum for Information Retrieval Evaluation 2020 (FIRE 2020)[8]. The organizers released a dataset of fake Urdu news containing a substantial amount of news articles, 42 teams participated in the competition and 9 teams submitted their final experimental results on the test dataset. As a final result, a BERT based approach achieved a macro F1 score of 0.900. This year, the second track is announced at the 13^{th} Forum for Information Retrieval Evaluation 2021 (FIRE 2021). The new track is advanced and more challenging composed to the preceding year, as the revised dataset has a larger number of news. This study has used ensembling machine learning models for Urdu fake news detection.

The rest of the paper is organized as follows. Section 2 provides an overview of the related work. Section 3 presents the specifications of the dataset used for the experimentation. The Section 4 presents our proposed ensembling approach and the machine learning techniques used for the experimentation. The results of the experiments are presented in Section 5. Finally, section 6 concludes the paper.

2. Related Work

It is observed that majority of the studies have been conducted on fake news detection in rich resource languages which includes, English, Chinese and Italian, whereas, little work has been done on low resource languages, such as Urdu [9]. Therefore, this section provides an overview of the notable existing studies conducted on Urdu fake news detection.

Table 1 provides a summary of the techniques proposed in literature for Urdu fake news detection. It can be observed from the table that the existing studies have used diverse techniques for fake news detection. For instance, a study from Urdufake'20 has used different variants of XGBoost and RoBERTa [10]. The results of the experiments showed that XGBoost outperformed the other techniques for Urdu fake new detection. Similarly, another study [11] used XLNet model with the AR pre-training method, whereas [12] used ensemble of machine learning techniques along with Multi-layer Dense Neural Network which achieved a very high F1 score.

3. Urdu Fake News Corpus

The UrduFake'21 track at the Forum for Information Retrieval Evaluation 2021 (FIRE2021) has provided a corpus of 1600 news articles. The key strength of the corpus is that it includes real and fake news from diverse domains. That is, it includes news articles from business, health, showbiz, sports and technology. Another notable observation is that the corpus is balanced

Table 1
Related Work Summary

Paper	Techniques
[9]	LR, RF, SVM, AdaBoost, CharCNN-Roberta, XLNet pre-trained model, Dense Neural Network, Bi-directional GRU model, ULMFiT model,
[10]	XGBoost with multiple features, RoBERTa
[11]	XLNet with AR pre-training
[12]	Ensemble approach, Multi-layer Dense Neural Network
[13]	RF, Bi-directional Gated Recurrent Unit, Multi-head self-attention based transformer
[14]	SVN, RR, BERT, MLP, AdaBoost Gradient boosting, Extra trees
[15]	HTC, TL model, Ensemble techniques

as it includes equal number of news articles from all the five domains, hence, providing equal opportunity for learning about fake news from all domains.

A summary specifications of the Urdu fake news corpus is presented in Table 2. It can be observed from the table that the dataset is composed of 950 real and 650 fake news. The presence of imbalance in the dataset may impede the effectiveness of supervised learning techniques. It can also be observed from the table that the training dataset is composed of 1300 news, whereas the testing dataset is composed of 300 new articles. Furthermore, the testing dataset is also imbalanced in favor of real news with a ratio of 2:1.

Table 2
Summary of the dataset

Item	Count
Total no of news	1600
Total no of real news	950
Total no of fake news	650
Total no of news in the training dataset	1300
Total no of real news in the training dataset	750
Total no of fake news in the training dataset	550
Total no of news in the testing dataset	300
Total no of real news in the testing dataset	200
Total no of fake news in the testing dataset	100

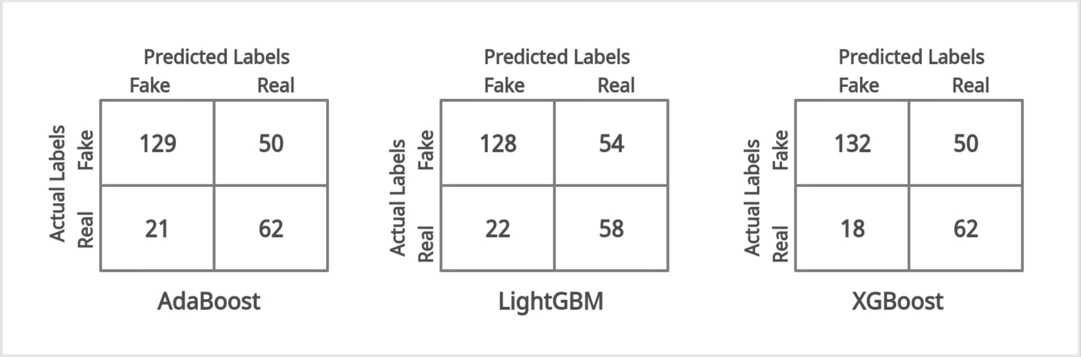
4. The Proposed Approach

The focus of this study is to use of classical machine learning techniques for Urdu fake news detection. We have employed a systematic approach for ensembling classical machine learning techniques using a voting scheme. Furthermore, it involved the use of the training dataset and the preliminary testing dataset released before the submission of the notebook to UrduFake'21 track. However, the results presented in Section 5 are generated using a unseen testing dataset released by the organizers of UrduFake'21 for the competition.

As a starting point of the approach, a comprehensive set of classical machine learning techniques were identified. The set of the techniques identified for this study are presented in Table 3. It can be observed from the table that we identified nine techniques for the initial experimentation. Subsequently, experiments were performed using the initially released dataset and precision, recall and F1 score were computed for each technique. Accordingly, the top three most effective techniques, AdaBoost, LightGBM and XGBoost, were identified for further processing. To develop a deeper understanding of the three top performing techniques, the confusion matrices for each technique were generated. The confusion matrices of the three techniques are presented in Figure 1.

Table 3
The classical techniques used for ensembling

Item	Details
Classical techniques	Decision Tree, Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbor, AdaBoost, LightGBM, XGBoost



AdaBoost

Actual Labels \ Predicted Labels	Fake	Real
Fake	129	50
Real	21	62

LightGBM

Actual Labels \ Predicted Labels	Fake	Real
Fake	128	54
Real	22	58

XGBoost

Actual Labels \ Predicted Labels	Fake	Real
Fake	132	50
Real	18	62

Figure 1: Confusion matrix of AdaBoost, LightGBM and XGBoost

The proposed ensemble technique employs a voting-based approach for predicting the final label of each news. An overview of the proposed approach is presented in Figure 2, whereas the details of the proposed approach are as follows:

- Convert the complete annotations into numeric form by replacing real news (R) with 0 and fake news (F) with 1. That is, convert the benchmark values, as well as the predictions, into binary numbers.
- Calculate the sum of all predicted values and store them separately. That is, for one news article, the sum of the results is 0 if all of the three top performing techniques predicted the news as a real, and 3 if all of the three techniques predicted the news as fake. Similarly, the value varies between 0 and 3 depending upon the predictions of the three techniques.

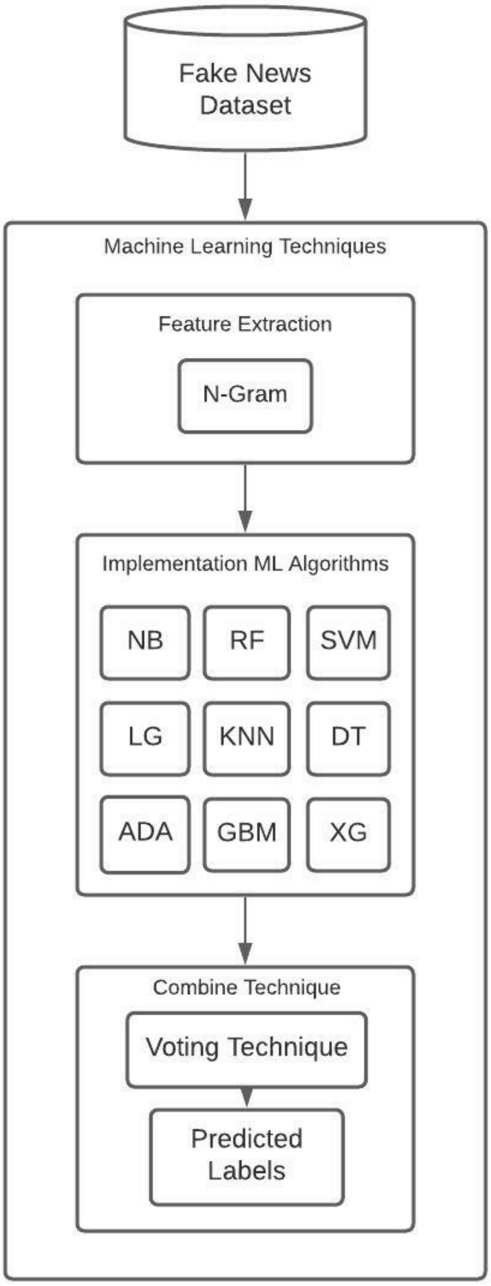


Figure 2: The proposed approach

- Produce the prediction labels by using the sum of the results that were stored in an earlier step. That is, if any of the 2 techniques predicted that the news is fake declare the news as fake (1), otherwise declare the news as real (0).

- As a final step, determine the final label of each news by using the voting scheme. That is, if a label is declared as 1 (fake), whereas the prediction of XGBoost is real (0) and at the same time sum of results is 0, declare the final label as 1 (fake). Similarly, if the label is 0 (real), the prediction of XGBoost is 1 and the sum of separately stored results is 2, declare the final label as 0 (real). Whereas, if both conditions are false, the prediction of the XGBoost should be considered as the final label.

5. Results

Experiments are performed using ten techniques, nine classical machine learning techniques and our proposed technique. For these techniques, unigram, bigram and trigram features are used. For the experiments the training and the testing dataset discussed in Section 3 is used. That is, 1300 news articles are used for training and 300 news articles are used for testing. The code used for the experiments can be downloaded from GitHub[1]. Subsequently, Precision, Recall and F1 scores are computed for the two classes. Also, macro average F1 and accuracy score are computed for all the techniques. The details of the results of unigram features are presented in Table 4 as the effectiveness of unigram features achieved a higher effectiveness score than bigram and trigram features.

Table 4
Results For machine Learning and Deep Learning Techniques

Technique	Fake Class			Real Class				
	P	R	F1	P	R	F1	F1 Macro	Accuracy
XGBoost	0.600	0.330	0.425	0.726	0.890	0.800	0.612	0.703
AdaBoost	0.697	0.300	0.419	0.727	0.935	0.818	0.618	0.723
LightGBM	0.583	0.350	0.437	0.729	0.875	0.794	0.616	0.700
Naive Bayes	0.490	0.260	0.339	0.700	0.865	0.774	0.556	0.663
Decision Tree	0.444	0.280	0.343	0.696	0.825	0.755	0.549	0.643
Random Forest	0.617	0.210	0.313	0.703	0.935	0.802	0.558	0.693
Logistic Regression	0.515	0.340	0.409	0.717	0.840	0.774	0.591	0.673
K-Nearest Neighbors	0.333	0.210	0.257	0.666	0.790	0.723	0.490	0.596
Support Vector Machine	0.453	0.440	0.446	0.724	0.735	0.729	0.588	0.636
Proposed approach	**0.634**	**0.330**	**0.434**	**0.729**	**0.905**	**0.808**	**0.621**	**0.713**

It can be observed from the table that three techniques achieved a higher accuracy score of greater 0.70. These are Random Forest, AdaBoost and our ensembling approach. It can also be observed from the table that the macro average F1 score of two techniques is greater than 0.60, whereas, the macro average F1 score of the third technique is less than 0.60. The two techniques that achieved a macro average F1 score greater than 0.60 are AdaBoost and our ensembling approach.

As the macro average score of these two techniques are exactly equal to 0.62, therefore a further comparison of these two techniques is performed in-terms of the F1 scores of each type

[1] https://github.com/socialmedialisteninglab/FakeNewsDetectionUrdu2021

of news. It can be observed from the table that the F1 score of fake news class for the ensembling approach is slightly higher than that of AdaBoost. Furthermore, the recall score achieved by the proposed approach for the fake news class is higher than that of AdaBoost. This represents that the proposed technique is slightly more effective than AdaBoost for identifying fake news.

6. Conclusion

The importance of fake news detection is well-established and a number of studies have been conducted on fake news detection for Western and Asian languages. However, Urdu fake news detection has received less attention despite the fact that the risk posed by fake news in comparable with the Western world. To that end, UrduFake'21 track at the Forum for Informational Retrieval Evaluation 2021 (FIRE2021) has taken a significant leap towards promoting fake news detection research in the Urdu language. In this study, an ensemble of classical machine learning models is proposed for fake news detection. The proposed approach relies on using a voting scheme between the three most effective techniques for the detection of Urdu fake news. Experiments are performed with the using the unseen testing dataset released by UrduFake'21, nine classical machine learning technique and the proposed approach. The results of the 19 teams released by the organizers ranked our proposed approach 4^{th} among the 19 submissions.

References

[1] A. Guttmann, Value of the global entertainment and media market 2011-2024, https://www.statista.com/statistics/237749/value-of-the-global-entertainment-and-media-market/, 2020. Accessed: 2021-10-09.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (2017) 22–36.

[3] N. A. Patel, R. Patel, A survey on fake review detection using machine learning techniques, in: 4^{th} International Conference on Computing Communication and Automation (ICCCA), IEEE, 2018, pp. 1–6.

[4] R. Varma, Y. Verma, P. Vijayvargiya, P. P. Churi, A systematic survey on deep learning and machine learning approaches of fake news detection in the pre-and post-COVID-19 pandemic, International Journal of Intelligent Computing and Cybernetics 14 (2010) 617–646.

[5] J. Schiffer, Media literacy in the EFL classroom, Ph.D. thesis, The University of Vienna, 2019.

[6] M. Amjad, G. Sidorov, A. Zhila, H. Gómez-Adorno, I. Voronkov, A. Gelbukh, "Bend the truth": Benchmark dataset for fake news detection in Urdu language and its evaluation, Journal of Intelligent & Fuzzy Systems 39 (2020) 2457–2469.

[7] M. Amjad, G. Sidorov, A. Zhila, Data augmentation using machine translation for fake news detection in the Urdu language, in: Proceedings of the 12^{th} Language Resources and Evaluation Conference, 2020, pp. 2537–2542.

[8] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, P. Rosso, UrduFake@FIRE2020: Shared track

on fake news identification in Urdu, in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 37–40.

[9] M. Amjad, G. Sidorov, A. Zhila, A. F. Gelbukh, P. Rosso, Overview of the shared task on fake news detection in Urdu at FIRE 2020., in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 434–446.

[10] N. Lina, S. Fua, S. Jianga, Fake news detection in the Urdu language using CharCNN-RoBERTa, in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 447–451.

[11] A. F. U. R. Khiljia, S. R. Laskara, P. Pakraya, S. Bandyopadhyaya, Urdu fake news detection using generalized autoregressors, in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 452–457.

[12] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@UrduFake-FIRE2020: Multi-layer dense neural network for fake news detection in Urdu news articles., in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 458–463.

[13] S. M. Reddy, C. Suman, S. Saha, P. Bhattacharyya, A gru-based fake news prediction system: Working notes for Urdufake-FIRE 2020., in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 464–468.

[14] N. N. A. Balaji, B. Bharathi, SSNCSE_NLP@Fake news detection in the Urdu language (UrduFake) 2020, in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 469–473.

[15] F. Balouchzahi, H. Shashirekha, Learning models for Urdu fake news detection., in: Proceedings of the Forum for Information Retrieval Evaluation, volume 2826, CEUR-WS, 2020, pp. 474–479.