

Urdu Fake News Detection using TF-IDF Features and TextCNN

Muhammad Abdullah Ilyas¹, Khurram Shahzad²

¹*Department of Computer Science, University of the Punjab, Lahore, Pakistan*

²*Department of Data Science, University of the Punjab, New Campus, Lahore, Pakistan*

Abstract

Social media platforms are widely used to exchange ideas and share news with each other. Consequently, any idea, post, or new posted on social media is likely to become viral in a short span of time. However, if a fake news becomes viral it can have consequences for the society. Therefore, the detection and eradication of fake news is desired. Recognizing the need for fake news detection, several studies have been conducted for this task in English and Western languages. However, the research in Urdu fake news detection has merely commenced. To promote research and development in this area, the second time in two years, Forum for Information Retrieval Evaluation 2021 has dedicated a track for fake news detection task for Urdu, called UrduFake'21. In this study, we have used three deep learning techniques for Urdu fake news detection, including the state-of-the-art Text CNN with three types of word embeddings, as well as TF-IDF features. The results of the experiments have shown that TextCNN with TF-IDF is the most effective technique that achieved an accuracy of 0.716 and macro average F1 score of 0.663. Furthermore, according to the results released by the organizers of UrduFake'21 our approach is ranked 2nd among the 19 submissions.

Keywords

Fake news, Urdu fake news, News classification, Machine learning, Deep learning, Convolutional Neural Network (CNN), TextCNN

1. Introduction

Entertainment and media market is recognized as a 2.1 trillion US dollars market [1]. Furthermore, it is estimated that more than 50% of the adult population read news papers. Among these, over 2.5 billion read in-print and more than 600 million read news in the digital form [2]. Given the increased use of social media services, people around the world use these services as a source of news. According to the 2021 survey of Statista, more than 70% respondents from different countries, such as South Africa and Malaysia, accepted the use of social media as a source of news [3].

Fake news refers to the news articles that are intentionally and verifiably false [4]. Fake news is considered as a remarkable issue as the propagation of disinformation which may undermine the peace and stability of a society. Given the importance of the task, several studies have been

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

✉ abdullah.ilyas@pucit.edu.pk (M. A. Ilyas); khurram@pucit.edu.pk (K. Shahzad)

❶ <https://www.linkedin.com/in/m-abdullah-ilyas-2170744a> (M. A. Ilyas);

<https://www.linkedin.com/in/mkhurramshahzad/> (K. Shahzad)

❷ <https://orcid.org/0000-0003-1223-1060> (M. A. Ilyas); 0000-0001-8433-6705 (K. Shahzad)

 © 2021 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)

conducted for fake news detection, which have been categorized into four types of methods, style-based, propagation-based, knowledge based and source-based methods [5]. However, there is a scarcity of studies that focus on fake news detection in Urdu language despite the fact that there are millions of Urdu speakers across the globe [6]. In the quest for promoting research in this area, a track of the International Forum for Information Retrieval (FIRE'20) was dedicated to Urdu fake news detection [7]. This year again, a track of the FIRE'21 is dedicated to Urdu fake news detection <https://www.urdufake2021.cicling.org/>.

In this study, we have performed experiments using three deep learning techniques and multiple features. In particular, we have used Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and TextCNN with the classical features, as well as word embeddings. The classical features used in this study are: Term Frequency Inverse Document Frequency (TF-IDF), whereas, the three types of embeddings fed to the deep learning techniques are: Word2Vec, fastText and GloVe.

The rest of the paper is organized into six sections. Section 2 presents the related work about fake news detection in the Urdu language. Section 3 provides an overview of the dataset released as a part of UrduFake'21 task. An overview of the proposed approach is presented in Section 4, whereas the results of the experiments are presented in Section 5. Finally, conclusions are presented in Section 6.

2. Related Work

As discussed in the preceding section, numerous studies have been conducted for fake news detection in Western languages, whereas the contemporary Urdu language task has received little attention of researchers. The initial work on fake news detection in Urdu developed the first-ever dataset of Urdu fake news detection [8]. The researchers generated another datasets by automatically translating fake news from the English fake news into Urdu [4]. The developed dataset developed is completely balanced as it is composed of 200 fake news and 200 real news. The study used word and character n-gram with Support Vector Machine (SVM) for the development of an automatic system for fake news detection which achieved accuracy scores ranging from 0.83 to 0.87. However, a key limitation of the approach is that dataset is composed of a small number of news which may not be suitable for learning and prediction of deep learning techniques.

A notable study [9] presented a comparison of linguistic features of fake news with real news. Whereas, [10] contend that fake news detection in English news can be done using semantic, syntactic and lexical features. Besides English, studies have also been done for all the major Western languages, including Spanish, German and French. These studies have used lexical features which includes bag of words, n-grams and parts of speech, for fake news detection [11].

As deep learning has achieved groundbreaking results for several natural language processing tasks, some studies have proposed the use of Long Short Term Memory (LSTM) to distinguish between real and fake news [12] [13]. Similarly, another study [14] used Recurrent Neural Network (RNN) for fake news detection, whereas, [15] have used ensemble learning and label smoothing with CharCNN and RoBERTa model for the same task. In contrast, [16] proposed the use dense neural network for the classification of fake news detection on Urdu dataset which

Table 1
Dataset Distribution

Category	Business	Health	Showbiz	Sports	Technology	Total
Real	150	150	150	150	150	750
Fake	80	130	130	80	130	550

Table 2
Distribution of training, development and testing dataset

Class	Training	Development	Testing	Total
Real	600	150	200	950
Fake	438	112	100	650
Total	938	262	300	1600

achieved an F1 score of 0.801. Similarly, another study implemented autoregressor to achieve an accuracy of 0.840 and F1 Score of 0.837 [17]. However, during UrduFake'21, the highest F1 score of 0.902 was achieved by using Random Forest and TF-IDF features [18].

3. Urdu Fake News Dataset

This section provides an overview of the UrduFake'21 corpus released for the Urdu fake news detection task. As per the standard procedure, initially a training dataset and a sample testing dataset was released for developing and tuning techniques. The specifications of the dataset of UrduFake'21 are presented in Table 1. It can be observed from the table that the released dataset is composed of 1300 Urdu news articles, including 750 real and 550 fake news articles. Another notable observation is that the news articles are collected from five categories: showbiz, health, sports, business and technology, which shows the diversity of the dataset. The presence of news from diverse domains makes it a more challenging as these domains have different vocabulary which limits the learning capabilities of machine learning techniques. That is, the vocabulary used in the news from the healthcare domain substantially differs from that of the sports domain.

In the second phase of UrduFake'21, an unseen test dataset was released which is used for the final evaluation of the techniques. That is, the testing dataset released in the second phase included a corpus of news, whereas, the labels of the news articles, either real or fake, were not provided. The participants were asked to submit their code and predictions against the unseen dataset.

This study has divided the dataset into three parts, training, development and testing dataset. The training dataset and the sample testing dataset released in the first step is referred to as training and development dataset, whereas, the unseen dataset released in the second round is referred to as testing dataset. The detailed specifications of the dataset from the two rounds are presented in Table 2. It can be observed from the table that the dataset is composed of 1600 news articles including 950 real and 650 fake news articles.

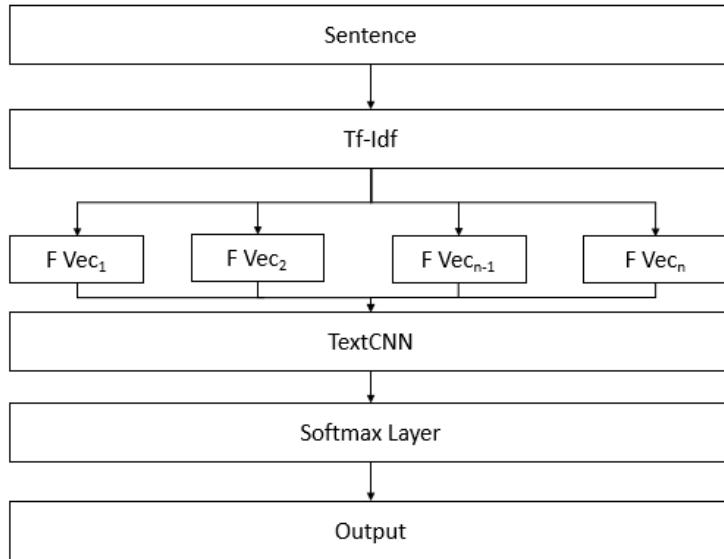


Figure 1: TextCNN with TF-IDF Model

4. Using TextCNN with TF-DF

This study has used Text Convolutional Neural Network (TextCNN) which builds upon Convolutional Neural Network (CNN) as it has achieved a remarkable accuracy on the key sentence classification task. The approach was originally proposed by [19] which applied different number of convolutional layers. In this study, we have applied three convolutional layers which uses 512 filters in each convolutional layer, wheresa the kernal size for three layers is set to 3, 4 and 5, respectively.

An overall architecture is presented in Figure 1. It can be observed from the figure that the first step extracts feature values from a news article. The proposed approach used Term Frequency - Inverse Document Frequency (TF-IDF) which computes the importance of a word in the news corpus. Once the feature values are computed and the convolutional layer with the kernel size 16 is used. Finally, the relu activation function is used in order to distinguish between fake and real news.

The choice of the TF-IDF stems from the experiments on the training and development dataset. In particular, for the choice of features, experiments were performed using Word2Vec, GloVe and fastText embeddings as features. The results of the experiments established that TF-IDF features are more effective than the three types of word embedding. A further examination of the embeddings and news vocabulary revealed the underlying reasons for the higher performance of TF-IDF over word embeddings. The first reason is that there are several words in the development dataset whose embeddings were not available due to the absence of these words from the training dataset. And the second reason is that there were some words that were more frequently used in one type of technique than the other technique.

5. Results

Experiments are performed using the TextCNN approach discussed in the preceding section. In addition to that, experiments are also performed using two other types of neural networks, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The choice of these techniques stems from the fact that techniques have achieved groundbreaking results for various NLP tasks. Recall from the preceding section, this study has used three types of word embeddings, Word2Vec, GloVe and fastText. The reason for choosing word embeddings is that the use of embeddings have achieved a very high F1 score for another NLP tasks in Urdu language [20].

For the experiments, the training dataset provided by UrduFake'21 organizers has been used. The initially released testing dataset has been used for development, whereas the finally released unseen dataset that is used for the competition has been used for testing and results generation. The code used for the experiments can be downloaded from GitHub¹ Accordingly, the generated results are presented in Table 3. That is, the table contains precision, recall and F1 score for the two classes, real news and fake news. Also, macro average F1 scores and Accuracy scores of the deep learning techniques are presented in the table.

Table 3

Precision, Recall F1 Score of Fake and Real Class

Sr	Technique	Feature	Fake class			Real class			Macro F1	Acc
			P	R	F1	P	R	F1		
1	CNN	Word2Vec	0.460	0.350	0.397	0.709	0.795	0.75	0.573	0.646
2	CNN	GloVe	0.428	0.210	0.281	0.685	0.860	0.762	0.522	0.643
3	CNN	fastText	0.362	0.370	0.366	0.681	0.675	0.678	0.522	0.573
4	RNN	Word2Vec	0.337	0.560	0.421	0.671	0.450	0.538	0.479	0.486
5	RNN	GloVe	0.238	0.05	0.082	0.659	0.920	0.768	0.425	0.630
6	RNN	fastText	0.355	0.270	0.306	0.674	0.755	0.712	0.509	0.593
7	CNN	TF-IDF	0.514	0.540	0.526	0.764	0.745	0.754	0.640	0.676
8	TextCNN	TF-IDF	0.592	0.480	0.530	0.762	0.835	0.797	0.663	0.716

It can be observed from the table that the proposed TextCNN achieved the highest accuracy score of 0.716 and macro average F1 score of 0.663 with TF-IDF features. Furthermore, it can be observed from the table that the F1 score achieved the proposed techniques for the fake news class is 0.530 and 0.797 for the real news class. These results represent that the proposed approach is more effective for the identification of real news than for the fake news. One possible reason for the difference in the F1 scores of the two classes stems from the smaller number of fake news in the training dataset.

6. Conclusion

A plethora of studies have been conducted for fake news detection in English and other Western languages. However, fake news detection in Urdu language is yet to receive the attention of

¹https://github.com/mabdullahilyas934/URDU_FAKE_2021

researchers. To promote research and development in Urdu fake news detection, a track of FIRE 2021, named UrduFake'21, is dedicated to this remarkable task. Given that the state-of-the-art deep learning techniques has achieved groundbreaking results for various NLP tasks. Therefore, this study has focused on the use of several deep learning techniques for fake news detection in Urdu. The techniques used in this study are, Convolutional Neural Networks and Recurrent Neural Networks. These techniques are fed with classical features, TF-IDF, as well as word embeddings, Word2Vec, GloVe and fastText embeddings. The results show TextCNN achieved a macro F1 score of 0.663 and an accuracy of 0.716 using TF-IDF features. According to the ranking released by the organizers, our is ranked 2nd in among the 19 submissions.

References

- [1] A. Guttmann, Value of the global entertainment and media market 2011-2024, 2021. URL: <https://www.statista.com/statistics/237749/value-of-the-global-entertainment-and-media-market/>, last accessed 4 October 2021.
- [2] D. Ponsford, More people read newspapers worldwide than use web, 2021. URL: <http://www.ifabc.org/news/More-People-Read-Newspapers-Worldwide-Than-Use-Web>, last accessed 4 October 2021.
- [3] A. Watson, Social media as a news source worldwide 2021, 2021. URL: <https://www.statista.com/statistics/718019/social-media-news-source/>, last accessed 4 October 2021.
- [4] M. Amjad, G. Sidorov, A. Zhila, Data augmentation using machine translation for fake news detection in the Urdu language, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 2537–2542.
- [5] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR) 53 (2020) 1–40.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing, 2017, pp. 2931–2937.
- [10] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, ACL, 2017, p. 3391–3401.
- [11] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake

- news in a new corpus for the Spanish language, *Journal of Intelligent & Fuzzy Systems* 36 (2019) 4869–4876.
- [12] A. Giachanou, P. Rosso, F. Crestani, Leveraging emotional signals for credibility detection, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2019, pp. 877–880.
 - [13] B. Ghanem, P. Rosso, F. Rangel, An emotional analysis of false information in social media and news articles, *ACM Transactions on Internet Technology (TOIT)* 20 (2020) 1–18.
 - [14] Y. Liu, Y.-F. B. Wu, Early detection of fake news on social media through propagation path classification with Recurrent and Convolutional Networks, in: Proceedings of the Thirty-second AAAI conference on Artificial Intelligence, 2018, pp. 354–361.
 - [15] 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.
 - [16] 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.
 - [17] 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.
 - [18] 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.
 - [19] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners' guide to) Convolutional Neural Networks for sentence classification, in: Proceedings of the 8th International Joint Conference on Natural Language Processing, ACL, 2017, pp. 253–263.
 - [20] S. Kanwal, K. Malik, K. Shahzad, F. Aslam, Z. Nawaz, Urdu named entity recognition: Corpus generation and deep learning applications, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19 (2019) 1–13.