Fake News Detection in Urdu Language using BERT

Snehaan Bhawal[1], Pradeep Kumar Roy[2]

[1]*Kalinga Institute of Industrial Technology, Odisha, India*
[2]*Indian Institute of Information Technology Surat, Gujarat, india*

Abstract

With the increase in popularity of social media, we can see an increase in the amount of Fake News in circulation, leading to misleading public opinion. Thus a system of Fake News detection is necessary to avoid such consequences. Most of such existing Fake News detection systems work with resource-rich languages like English and Spanish, but very few systems can work with low resource languages like Urdu. The current study focuses on detecting Fake News in the Urdu language using Machine and Deep learning techniques. The 'UrduFake' data is used in this research, provided to us as a shared task of FIRE-2021. The experimental outcomes of various models showed that the Transfer learning models performed better than the Machine learning models and achieved a weighted average F1-score of 0.87 and 0.61 on the validation and test dataset.

Keywords

Fake News Detection, Urdu, Deep Learning

1. Introduction

There has been a steady rise in internet traffic throughout the world. Connectivity between people has increased with the popularity of social media [1]. Such media houses have now become the principal source of information for the general public. Due to the unrestricted nature of such media, there is little to no oversight in the articles being posted. Although it promotes freedom of speech, it can also be misused to spread Fake News [2]. Most of such platforms do not verify the articles and promote them according to popularity, leading to the faster spread of such unverified articles.

Rubin et al.[3] categorized such deceptive news into three broad groups: i) Serious Fabrication, ii) Hoaxes and iii) Satire. There have been many cases where these kinds of Fake News was intentionally spread via social media platforms to mislead the general public [4][5]. This can be used to target people by discrediting them or creating a situation of political unrest, and undermining society's stability. Such articles are usually based on polarizing topics [6] and garner massive popularity on social media, which in turn helps to promote the same to a wider audience. Thus there is an urgent need to detect and stop such volatile articles at an earlier stage of circulation to prevent further spread by assessing the credibility of the said article and determining it to be trustworthy or not.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

mailtosnehaan@gmail.com (S. Bhawal); pkroynitp@gmail.com (P. K. Roy)
0000-0002-1072-5326 (S. Bhawal); 0000-0001-5513-2834 (P. K. Roy)

However, most of the research regarding the detection of fake news has been done in resource-rich languages like English, and Spanish [7]. Despite Urdu having more than 100 million speakers, it has seen very little development in such detection systems due to the absence of properly labelled data and very few resources for NLP tasks. The event organizers [8, 9] provided a benchmark data set for Fake News detection in Urdu [10]. The current study utilizes this data to implement and compare different Machine and Deep Learning models for Fake News Detection in the Urdu language.

The rest of the article is summarized as follows: Section 2 discusses related work, while the task description and data set distribution is explained in Section 3. Section 4 provides the preprocessing steps taken, followed by the explanation of the proposed methodology in Section 5. The experiment results are discussed in Section 6. Section 7 conclude this research with limitations and future scope.

2. Literature Review

Automating fake news detection has been a challenging task for a long time, particularly for low resource languages. Researchers are creating their own data sets [11] [12] due to the presence of insufficient benchmarked datasets. Zhou and Zafarani [13] introduced four techniques for fake news detection based on (i) knowledge, (ii) content, (iii) propagation and (iv) source of origin.

Rubin et al. [14] developed a model using content-based approach by picking up on the satirical cues present in a the news articles and implementing a SVM based algorithm with five features- (i) Absurdity, (ii) Humor, (iii) Grammar, (iv) Negative affect, and (v) Punctuation. They tested their combinations on 360 different news articles and were able to detect satirical or potentially misleading news with a F1 score of 0.87. Another study by [15] follows the propagation-based approach by exploring the social context during news propagation on social media by looking into the relationship between publishers, articles, and users.

Regarding Fake News detection in Urdu, the number of research works that has been conducted is very less. To the best of our knowledge, the data set [10] provided by the organizers serves as the single proper available data for the required task. For works relating to Fake News Detection in the Urdu language, we can refer to the works done in the previous iteration of the shared task of FIRE. The study reported by [16] topped the leader board. They used an ensemble model of a RoBERTa and a CNN model with word and character embedding, respectively.

3. Task and Data description

Nowadays, social networking platforms are one of the primary sources of information used to spread Fake news. Mostly, the existing system is built with a non-Urdu language dataset. Hence, the news written in Urdu may not be detected by the system. The current study implements and shows a comparison of different Machine and Deep Learning models in Fake News Detection in the Urdu Language for the UrduFake-2021 task[1]. Table 1 shows the category-wise distribution

[1] https://www.urdufake2021.cicling.org/home

Table 1
Category wise Article Distribution Training data

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

Table 2
Label Distribution in the given data

Data Set	Real	Fake	Total
Train	600	438	1038
Validation	150	112	262
Test	200	100	300

of the articles present in the Training Data set from Six different domains, namely, Business, Health, Showbiz, Sports and Technology. Table 2 provides the distribution of Real and Fake classes in the Train and Test data.

4. Data Preprocessing

The dataset[2] provided to us by the organizers is already processed as discussed by the authors of [10]. Additionally, we have removed any numerals, URLS, email ids and all website links. The punctuations were replaced with spaces and extra spaces were removed in each article.

5. Methodology

In the study, three different approach were used :

i Conventional Machine learning models.
ii Neural Network models
iii Transfer learning models

5.1. Conventional Machine Learning based models

Under Conventional ML-based models, we have explored the use of 1-5 gram word TF-IDF features. The features were first extracted and then provided to the different Machine Learning classifiers, namely, Logistic Regression(LR), Naive Bayes (NB), Random Forest (RF), XGBoost (XGB) and Support Vector Machine (SVM). The detailed results of the classifier models are shown in Table 3 of the Results section.

[2] https://www.urdufake2021.cicling.org/dataset

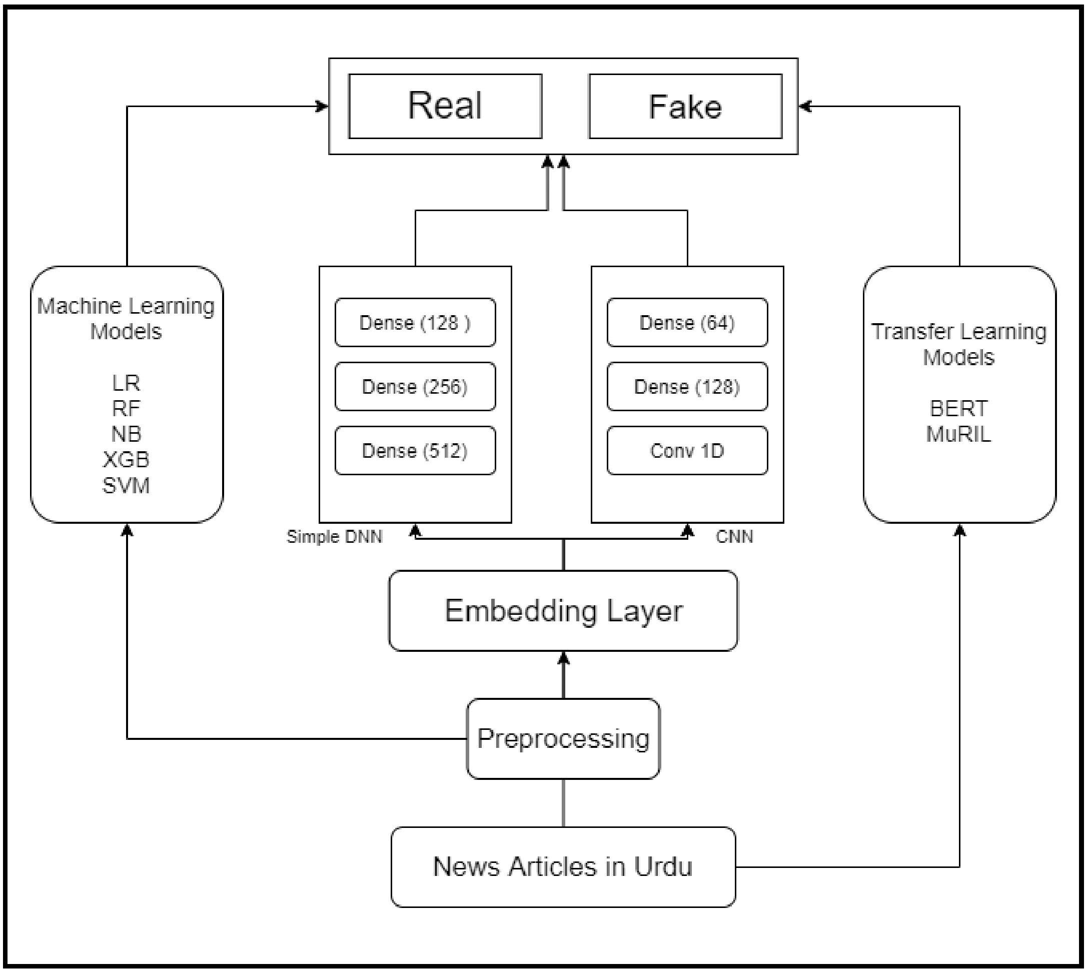

Figure 1: Framework used to predict Fake News

5.2. Neural Network based models

In the Neural Network based models, we have reused the previously extracted 1-5 gram TF-IDF features and used them as the input to a simple Deep Neural Network (DNN). The DNN consists of three fully connected layers consisting of 512, 256, 128 layers, followed by a single output neuron. Only a single neuron was chosen as the output because of the binary nature of classification required in the problem. The ReLU activation function is used in the hidden layers and the sigmoid activation function is used at the output layer. Adam and binary cross-entropy were chosen to be the optimizer and loss function for all Neural Network-based models.

This was followed by a Convolutional Neural Network (CNN) based approach. The CNN model consisted of one Conv1D layer followed by a Global Max Pooling layer and a dropout layer. This was then connected to a sequential network consisting of two hidden layers comprising of 128 and 64 neurons, respectively. As the input, we used an embedding layer of 100 dimensions with input length set to 512, resulting in an input layer of dimension (512, 100). The Convolutional

layer was made of 64 filters, with kernel size being 3.
As the final Neural Network based model, a Bidirectional Long Short-Term Memory model (Bi-LSTM) was chosen. It consists of 256 memory units followed by a Global Max Pooling and Batch Normalization layer. An embedding layer of 50 dimensions was taken as the input layer, with the padding length being fixed at 512, followed by dense layers of 20 and 10 neurons in the first and second layers, respectively. The output layer was the same as the other models, with a single neuron with a sigmoid activation function.
After successive hyperparameter tuning, we found out that the best results were achieved for the Neural network models by setting the max sequence length to 512, further increase led to a decrease in F1 scores and an increase in training time. The learning rate was set to 0.00001, and the optimizer was Adam. The codes for the current study can be found in the GitHub repository[3].

5.3. Transfer Learning based models

We have implemented BERT (Bidirectional Encoder Representations from Transformers) models to work with the transfer learning capabilities. For these models, no further preprocessing was done. The limitation of such BERT-based models is that they cannot accommodate all the tokens in each article as the maximum sequence length for such models is 512. Still, this issue was ignored as we saw that increasing the sequence length in Neural Network models led to diminishing returns.
Two different variants of BERT models were studied.

i BERT (multilingual)
ii MuRIL

The BERT [17] multilingual model was trained on 102 languages with masked language modelling. Here, the pooled output from the pre-trained model was fed to a dropout layer and finally to the output neuron.

The last model that we used is MuRIL [18] (Multilingual Representations for Indian Languages). This is a BERT model trained on a large corpus of 17 Indian languages, including Urdu, collected from Wikipedia and the Dakshina dataset [19]. This model is also trained with the translated and transliterated data and the monolingual corpus. Which gives it an advantage in processing code mixed languages.

6. Results

This section presents the experimental results of all the models mentioned in Section 5. These results were obtained on the validation data with the model being trained with training sample shown in Table 2 and presented in precision, recall, and weighted F1-score. A particular model is best if it reports the best-weighted average of precision, recall, and F1 score among all other models. The value in bold represents the highest value achieved for a particular data set.

[3] https://github.com/Sbhawal/NEWUrduFake-FIRE-2021-CODES.git

Table 3
Results of Conventional Machine Learning Models

Model	Class	Precision	Recall	F1-score
LR	Real	0.72	0.93	0.81
	Fake	0.84	0.52	0.64
	Weighted Avg	**0.77**	**0.75**	**0.74**
RF	Real	0.64	0.84	0.73
	Fake	0.64	0.38	0.47
	Weighted Avg	0.64	0.64	0.62
NB	Real	0.69	0.92	0.72
	Fake	0.81	0.46	0.58
	Weighted Avg	0.74	0.72	0.70
XGB	Real	0.70	0.89	0.78
	Fake	0.76	0.49	0.60
	Weighted Avg	0.73	0.72	0.70
SVM	Real	0.58	1.00	0.74
	Fake	1.00	0.04	0.09
	Weighted Avg	0.76	0.59	0.46

Table 4
Results of Neural Network based models

Model	Class	Precision	Recall	F1-score
DNN	Real	0.72	0.97	0.82
	Fake	0.92	0.49	0.64
	Weighted Avg	**0.80**	**0.76**	**0.75**
DNN+ Emb	Real	0.73	0.91	0.81
	Fake	0.82	0.55	0.66
	Weighted Avg	0.77	0.76	0.75
CNN	Real	0.74	0.85	0.79
	Fake	0.75	0.61	0.67
	Weighted Avg	0.74	0.74	0.74
Bi-LSTM	Real	0.72	0.83	0.77
	Fake	0.71	0.57	0.63
	Weighted Avg	0.72	0.72	0.71

By observing the outcomes of the experimented tradition ML-based model shown in Table 3, it is found that the LR classifier performed the best with precision, recall and F1-score of 0.77, 0.75 and 0.74, respectively.

The outcomes of the LR model is almost similar to that of the simple Deep Neural Network model which achieved precision, recall and F1 score of 0.80, 0.76 and 0.75 respectively as Shown in Table 4. But, in general, the performance of the traditional ML models shown in Table 3 are low as compared to the Neural Network models. These comparative outcomes confirmed that the neural network-based models are better choices for developing an automated Urdu fake news detection system.

Finally, we have experimented with the Transfer Learning based models- BERT and MuRIL. The outcomes of the models are shown in Table 5. The MuRIL model performed the best with

Table 5
Results of Transfer Learning based models

Model	Class	Precision	Recall	F1-score
BERT	Real	0.87	0.89	0.88
	Fake	0.84	0.82	0.83
	Weighted Avg	0.86	0.86	0.86
MuRIL	Real	0.86	0.92	0.89
	Fake	0.88	0.80	0.84
	Weighted Avg	**0.87**	**0.87**	**0.87**

weighted precision, recall and F1-score values of 0.87, 0.87, and 0.87, respectively, beating the multilingual BERT model, which achieved an F1 score of 0.86 on the validation data.

7. Conclusion

Fake news on the social media platforms are a big issue at the current date. This research suggested a Transfer learning based framework for Urdu fake news detection. Many traditional ML models and NN based models have been experimented to achieve the best prediction accuracy. We found, the MuRIL- a transfer learning model, outperform the traditional Machine Learning and other NN based models in the Fake News Detection task. The transfer learning based MuRIL model achieved accuracy and macro F1 score of 0.743 and 0.610 respectively on the test dataset. The developed model used an Urdu dataset for training. Hence, the fake news posted in other languages may not be detected by it. Due to the use of BERT based models, we have limited the sequence length to 512 which can be improved by using an ensemble of DNN and BERT models which will be explored in the future.

References

[1] P. K. Roy, S. Chahar, Fake profile detection on social networking websites: A comprehensive review, IEEE Transactions on Artificial Intelligence 1 (2020) 271–285. doi:10.1109/TAI.2021.3064901.

[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] V. L. Rubin, Y. Chen, N. K. Conroy, Deception detection for news: three types of fakes, Proceedings of the Association for Information Science and Technology 52 (2015) 1–4.

[4] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of economic perspectives 31 (2017) 211–36.

[5] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, Nature communications 9 (2018) 1–9.

[6] 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.

[7] 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, European Language Resources Association, Marseille, France, 2020, pp. 2537–2542. URL: https://aclanthology.org/2020.lrec-1.309.

[8] 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: FIRE (Working Notes), 2020, pp. 434–446.

[9] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, P. Rosso, Urdufake@ fire2020: Shared track on fake news identification in urdu, in: Forum for Information Retrieval Evaluation, 2020, pp. 37–40.

[10] M. Amjad, G. Sidorov, A. Zhila, H. Gomez-Adorno, I. Voronkov, A. Gelbukh, Bend the truth: A benchmark dataset for fake news detection in urdu and its evaluation, Journal of Intelligent & Fuzzy Systems 39 (2020) 2457–2469. doi:10.3233/JIFS-179905.

[11] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, arXiv preprint arXiv:1708.07104 (2017).

[12] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).

[13] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR) 53 (2020) 1–40.

[14] V. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the Second Workshop on Computational Approaches to Deception Detection, Association for Computational Linguistics, San Diego, California, 2016, pp. 7–17. URL: https://aclanthology.org/W16-0802. doi:10.18653/v1/W16-0802.

[15] K. Shu, S. Wang, H. Liu, Beyond news contents: The role of social context for fake news detection, in: Proceedings of the twelfth ACM international conference on web search and data mining, 2019, pp. 312–320.

[16] N. Lina, S. Fua, S. Jianga, Fake news detection in the urdu language using charcnn-roberta, Health 100 (2020) 100.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[18] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. Talukdar, Muril: Multilingual representations for indian languages, 2021. arXiv:2103.10730.

[19] B. Roark, L. Wolf-Sonkin, C. Kirov, S. J. Mielke, C. Johny, I. Demirşahin, K. Hall, Processing South Asian languages written in the Latin script: the Dakshina dataset, in: Proceedings of The 12th Language Resources and Evaluation Conference (LREC), 2020, pp. 2413–2423. URL: https://www.aclweb.org/anthology/2020.lrec-1.294.