International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Fake News Detection Using Machine Learning Ensemble Methods

MOHD SHOEB MOHD SAEED MOMIN

Department of BACHELOR OF VOCATIONAL IN ARTIFICIAL AND DATA SCIENCE ,
Anjuman I Islam's Abdul Razzaq Kalsekar Polytechnic, New Panvel , Maharashtra , India

ABSTRACT

The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality.

Automated classification of a text article as misinformation or disinformation is a challenging task. Even an expert in a particular domain has to explore multiple aspects before giving a verdict on the truthfulness of an article. In this work, we propose to use machine learning ensemble approach for automated classification of news articles.

Our study explores different textual properties that can be used to distinguish fake contents from real. By using those properties, we train a combination of different machine learning algorithms using various ensemble methods and evaluate their performance on 4 real world datasets.
Experimental evaluation confirms the superior performance of our proposed ensemble learner approach.

1. Introduction

)e advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. Besides other use cases, news outlets bene2tted from the widespread use of social media platforms by providing updated news in near real time to its subscribers.)e news media evolved from newspapers, tab-loids, and magazines to a digital form such as online news platforms, blogs, social media feeds, and other digital media formats [1]. It became easier for consumers to acquire the latest news at their 2ngertips. Facebook referrals account for 70% of tra=c to news websites [2].)ese social media platforms in their current state are extremely powerful and useful for their ability to allow users to discuss and share ideas and debate over issues such as democracy, education, and health. However, such platforms are also used with a negative perspective by certain entities commonly for monetary gain [3, 4] and in other cases for creating biased opinions, manipulating mindsets, and spreading satire or absurdity.)e phenomenon is commonly known as fake news.

)ere has been a rapid increase in the spread off ake news in the last decade, most prominently observed in the 2016 US elections [5]. Such proliferation of sharing articles online that do not conform to facts has led to many problems not just limited to politics but covering various other domains such as sports, health, and also science [3]. One such area a4ected by fake news is the 2nancial markets [6], where a rumor can have disastrous consequences and may bring the market to a halt.

Our ability to take a decision relies mostly on the type of information we consume; our world view is shaped on the basis of information we digest.)ere is increasing evidence that consumers have reacted absurdly to news that later proved to be fake [7, 8]. One recent case is the spread of novel corona virus, where fake reports spread over the Internet about the origin, nature, and behavior of the virus [9].)e situation worsened as more people read about the fake contents online. Identifying such news online is a daunting task.

2. Materials and Methods

In the following, we describe our proposed framework, followed by the description of algorithms, datasets, and performance evaluation metrics.

2.1. Proposed Framework. In our proposed framework, as illustrated in Figure 1, we are expanding on the current literature by introducing ensemble techniques with various linguistic feature sets to classify news articles from multiple domains as true or fake.)e ensemble techniques along with Linguistic Inquiry and Word Count (LIWC) feature set used in this research are the novelty of our proposed approach. are numerous reputed websites that post legiti-
mate news contents, and a few other websites such as PolitiFact and Snopes which are used for fact checking. In addition, there are open repositories which are maintained by researchers [11] to keep an up-to-date list of currently available datasets and hyperlinks to potential fact checking sites that may help in countering false news spread. However, we selected three datasets for our experiments which contain news from multiple domains (such as politics, entertainment, technology, and sports) and contain a mix of both truthful and fake articles.)e datasets are available online and are extracted from the World Wide Web.)e 2rst dataset is ISOT Fake News Dataset [23]; the second and third datasets are publicly available at Kaggle [24, 25]. A detailed description of the datasets is provided in Section 2.5.

)e corpus collected from the World Wide Web is preprocessed before being used as an input for training the models.)e articles' unwanted variables such as authors, date posted, URL, and category are 2ltered out. Articles with no body text or having less than 20 words in the article body are also removed. Multicolumn articles are transformed into single column articles for uniformity off ormat and structure.)ese operations are performed on all the datasets to achieve consistency off ormat and structure.

3. Results and Discussion

Table 2 summarizes the accuracy achieved by each algorithm on the four considered datasets. It is evident that the maximum accuracy achieved on DS1 (ISOT Fake News Dataset) is 99%, achieved by random forest algorithm and Perez-LSVM. Linear SVM, multilayer perceptron, bagging classi2ers, and boosting classi2ers achieved an accuracy of 98%.)e average accuracy attained by ensemble learners is 97.67% on DS1, whereas the corresponding average for individual learners is 95.25%.)e absolute di4erence between individual learners and ensemble learners is 2.42% which is not signi2cant. Benchmark algorithms Wang-CNN and Wang-Bi-LSTM performed poorer than all other algorithms. On DS2, bagging classi2er (decision trees) and boosting classi2er (XGBoost) are the best performing algorithms, achieving an accuracy of 94%. Interestingly, linear SVM, random forest, and Perez-LSVM performed poorly on DS2. Individual learners reported an accuracy of 47.75%, whereas ensemble learners' accuracy is 81.5%. A similar trend is observed for DS3, where individual learners' accuracy is 80% whereas ensemble learners' accuracy is 93.5%. However, unlike DS2, the best performing algorithm on DS3 is Perez-LSVM which achieved an accuracy of 96%. On DS4 (DS1, DS2, and DS3 combined), the best performing algorithm is random forest (91% accuracy). On average, individual learners achieved an accuracy of 85%, whereas ensemble learners achieved an accuracy of 88.16%.)e worst performing algorithm is Wang-Bi-LSTM which achieved an accuracy of 62%.

Table 2: Overall accuracy scor e for each dataset.

	DS1	DS2	DS3	DS4
Logistic regression (LR)	0.97	0.91	0.91	0.87
Linear SVM (LSVM)	0.98	0.37	0.53	0.86
Multilayer perceptron	0.98	0.35	0.94	0.9
K-nearest neighbors (KNN)	0.88	0.28	0.82	0.77
Ensemble learners				
Random forest (RF)	0.99	0.35	0.95	0.91
Voting classi2er (RF, LR, KNN)	0.97	0.88	0.94	0.88
Voting classi2er (LR, LSVM, CART)	0.96	0.86	0.92	0.85
Bagging classi2er (decision tree	0.98	0.94	0.94	0.9
Boosting classi2er (AdaBoost	0.98	0.92	0.92	0.86
Boosting classi2er (XGBoost)	0.94	0.98	0.94	0.89
Benchmark algorithms				
Perez-LSVM	0.99	0.79	0.96	0.9
Wang-CNN	0.87	0.66	0.58	0.73
Wang-Bi-LSTM	0.86	0.52	0.57	0.62

4. Conclusion

task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. In this research, we discussed the problem of classifying fake news articles using machine learning models and ensemble techniques.)e data we used in our work is collected from the World Wide Web and contains news articles from various domains to cover most of the news rather than specifically classifying political news.)e primary aim of the research is to identify patterns in text that differentiate fake articles from true news. We extracted different textual features from the articles using an LIWC tool and used the feature set as an input to the models.)e learning models were trained and parameter-tuned to obtain optimal accuracy. Some models have achieved comparatively higher accuracy than others. We used multiple performance metrics to compare the results for each algorithm.
ensemble learners have shown an overall better score on all performance metrics as compared to the individual learners.

Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.

Data Availability

Previously reported data were used to support this study and are available at https://www.kaggle.com/c/fake-news and https://www.kaggle.com/jruvika/fake-news-detection.

Conflicts of Interest

authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] A. Douglas, “News consumption and the new electronic media,” 7e International Journal of Press/Politics, vol. 11, no. 1, pp. 29–52, 2006.
[2] J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016.
[3] D. M. J. Lazer, M. A. Baum, Y. Benkler et al., “)e science of fake news,” Science, vol. 359, no. 6380, pp. 1094–1096, 2018.
[4] S. A. Garc´ıa, G. G. Garc´ ıa, M. S. Prieto, A. J. M. Guerrero, and C. R. Jim´ enez, “)e impact of term fake news on the scientific community scientific performance and mapping in web of science,” Social Sciences, vol. 9, no. 5, 2020.
[5] A. D. Holan, 2016 Lie of the Year: Fake News, Politifact, Washington, DC, USA, 2016.
[6] S. Kogan, T. J. Moskowitz, and M. Niessner, “Fake News: Evidence from Financial Markets,” 2019, https://ssrn.com/abstract=3237763.
[7] A. Robb, “Anatomy of a fake news scandal,” Rolling Stone, vol. 1301, pp. 28–33, 2017.
[8] J. Soll, “)e long and brutal history of fake news,” Politico Magazine, vol. 18, no. 12, 2016.
[9] J. Hua and R. Shaw, “Corona virus (covid-19) “infodemic” and emerging issues through a data lens: the case of China,” International Journal of Environmental Research and Public Health, vol. 17, no. 7, p. 2309, 2020.
[10] N. K. Conroy, V. L. Rubin, and Y. Chen, “Automatic de ception detection: methods for finding fake news,” Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.
[11] F. T. Asr and M. Taboada, “Misinfotext: a collection of news articles, with false and true labels,” 2019.
[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
[13] S. Vosoughi, D. Roy, and S. Aral, “)e spread of true and false news online,” Science, vol. 359, no. 6380, pp.1146–1151, 2018.
[14] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, 2017.
[15] V. L. Rubin, N. Conroy, Y. Chen, and S. Cornwell, “Fake news or truth? using satirical cues to detect potentially misleading news,” in Proceedings of the Second Workshop on Computa tional Approaches to Deception Detection, pp. 7–17, San Diego, CA, USA, 2016.
[16] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert),” Applied Sciences, vol. 9, no. 19, 2019.
[17] H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using n-gram analysis and machine learning tech niques,” in Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, pp. 127–138, Springer, Vancouver, Canada, 2017.
[18] W. Y. Wang, Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection, Association for Computa tional Linguistics, Stroudsburg, PA, USA, 2017.
19] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, “A simple buttough-to-beat baseline for the fakenewschallenge stance detection task,” 2017, https://arxiv.org/abs/1707.03264.
[20] N. Ruchansky, S. Seo, and Y. Liu, “Csi: a hybrid deep model for fake news detection,” in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 797–806, Singapore, 2017.
[21] V. P´ erez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” 2017, https://arxiv.org/abs/1708.07104.
[22] P. B¨uhlmann, “Bagging, boosting and ensemble methods,” in Handbook of Computational Statistics, pp. 985–1022, Springer, Berlin, Germany, 2012.
[23] H. Ahmed, I. Traore, and S. Saad, “Detecting opinion spams and fake news using text classification,” Security and Privacy, vol. 1, no. 1, 2018.
[24] Kaggle, Fake News, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/c/fake-news.
[25] Kaggle, FakeNewsDetection,Kaggle,SanFrancisco, CA,USA, 2018, https://www.kaggle.com/jruvika/fake-news-detection.