Ingénierie des Systèmes d'Information
Vol. 28, No. 1, February, 2023, pp. 205-209
Journal homepage: http://iieta.org/journals/isi

A Machine Learning Framework for Automatic Fake News Detection in Indian Tamil News Channels

Sudhakar Murugesan*, K.P. Kaliyamurthie

Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai 600073, India

Corresponding Author Email: sudhakarmtech@gmail.com

https://doi.org/10.18280/isi.280123

Received: 14 November 2022
Accepted: 20 January 2023

Keywords:
Indian news, fake news, Naïve Bayes, Logistic Regression, LSTM, classifications

ABSTRACT

With the development of technology and social media, many people started using the Internet. Today, everyone is creating and sharing content on social media. Before they transfer it to others, no one checks the originality of the content; it is com manually identifies the news content as fake or real manually. Due to this challenge, many people are started sharing fake news purposely to destroy the community, and some political and business purposes are spreading quickly. News channels and online newspapers have challenges in identifying trustworthy news sources. In this research paper, we collected various news articles from Indian news are gathered and will perform preprocessing, feature extraction, classification and prediction are going to be done using Naïve Bayes, Logistic Regression and LSTM. The proposed approach has 35,550 trustworthy news, and 15,450 fake news and TF-IDF techniques are used for the feature extraction. These three proposed algorithms are going to compare and predict the results. The Long Short-Term Memory will detect the accuracy of fake news in the Indian news channel is 99.7%.

1. INTRODUCTION

Today, Artificial Intelligence made many changes in the world, especially in the Information technology domain, like automatic driving systems, robotic surgery, facial recognition, reducing human errors and 24x7 [1]. The Internet is available to everyone today; due to this reason, the quality and quantity of news also increase every day. Many people are consuming information from online resources, constantly changing human life. After Covid-19, everyone started using the Internet and social media mainly, the school and college students; these young people used social media platforms or mobile applications to create content and share forwarded messages without checking the originality of the content. This type of online news may increase during election time, and politicians will start sharing fake news. Therefore, we must develop a model to detect fake news so that people will get only reliable information [2]. Social media have allowed users to share news content freely and faster than traditional news sources. This misinformation is not a problem for a particular community but is a big problem worldwide. People can view thousands of search results every time they enter their search query in the search engine (Google, Bing, Yahoo); from this search, people will get some helpful information and sometimes, it will show them misleading information [3].

Due to the technology improvement, everyone spends most of his or her online, and we receive more information from the Internet; it is our duty to control and fight against fake news. Some people intend to distribute fake news against the government and business companies, harming the company's reputation and the government [4]. The government and many other organizations motivate researchers to develop a system to detect the distribution of fake news. The European Commission established a committee in 2018 to advise on policies to fight against online misinformation, and the committee suggested that more research and innovation technologies be developed. Figure 1 will show an overview of the fake news identification method.

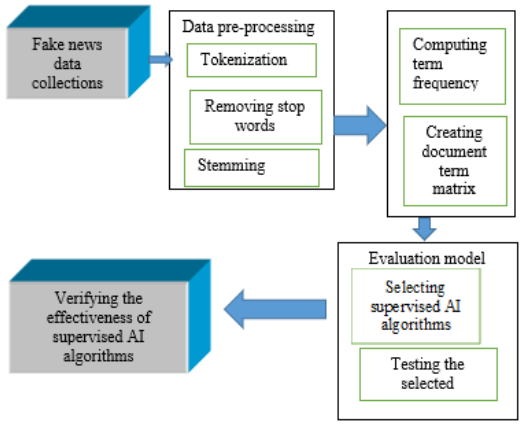

Figure 1. Overview of fake news identification methods

Today, 4.70 billion people are using social media around the world, according to a report from Kepios in July 2022. In India 460.00 million active users in India since last year, and millions of people have been joining. The following section is going to be discussed in this research paper. Section II will be the research problem; section III will be the proposed solution; section IV will be the related research work; section v will be the machine learning algorithms; section VI will be the datasets; section VII will be the results and finally, conclusions.

Tamil is spoken mainly in the Indian state of Tamil Nadu, Singapore, Malaysia, Mauritius, and Srilanka. The fake news is going to affect Tamil-speaking people worldwide. In India

2021-2022, more than 94 YouTube channel has been banned due to the spreading of fake news.

1.1 Research problem

Mainly this research is concerned with identifying a solution, and this solution is used to identify and filter out fake news. This one will help the news readers to get genuine information or news. This will help the readers, companies and research institutes involved in this issue.

There are several research studies have been proposed for automatic fake news detection, and the majority of research is based on languages like English, Spanish, Romanian, German, Italian and Hindi. This research work will be going to detect Tamil fake news.

1.2 Proposed solution

The tool is the proposed solution for fake news identification, and this tool is going to be used to detect and remove fake news from the sites. The proposed mechanisms are going to add web browsers.

2. LITERATURE REVIEW

Nowadays, everyone gets information from anywhere. It is the period of knowledge where a person can get to the events of various events worldwide in the solace of their own home. This will lead people to create a piece of fake news. Today, many people are using online media to refresh themselves with the news, and giving them reliable and charitable information is more critical. Here we will use specializes in existing solutions to predict and detect faux information, and many prevailing faux information detection techniques depend on characteristic extraction. The different classifiers used within side these solutions have been labelled accordingly.

Several natural language processing techniques are used to detect and automatically analyse political statements in Romanian news. The main text collected from the Factual and authentic or fake can be found in the different public reports [5]. There are several research approaches available to detect the fake news field, and today we will use them. Most researchers suggested machine learning algorithms for detecting fake news [6]. Some other authors also used the BERT algorithm model, and the algorithm's accuracy is 98.90% [7]. The Conventional Neural Network and RNN model will achieve accuracy for the detection of fake content at 82% [8].

2.1 Significance

As the internet and social media got wide, major problems emerged. Although social media acts as an excellent medium for spreading information at an exponential rate which is fluently accessible and available free of cost, at the same time, it also acts as an ideal place to produce and partake similar fake information. Today, with the increase in the number of individuals using the internet daily, exposing them to new information and stories regularly, phoney news at times becomes veritably deceptive and can spread exponentially. The implications of misinformation will be long lost, and it is challenging to correct them immediately. The individual will expose his or her thinking based on so many things, which may be intentional. When the news is exposed to be fake, they will establish logical-based lies. The malicious information will spread quickly, and it affects not only the individual, and it is also affecting the business and society [9].

As a human, it is challenging to detect counterfeit information, and social media users sometimes will not be aware of the many posts, tweets and articles that are solely created to shape consumers' beliefs. If the trusted person shares the counterfeit information, it is not easy to detect it. As a public, we absorb all the news and do not look at the authenticity of the news. Today 67% of youngsters depend on these platforms to know all the essential information. Some people will start believing that all the news is correct. Anyone who disagrees with his or her information will be tagged as biased, ending with the Confirmation-Bias. People are prepared to receive their perception news only, and they will not try to get proof against their perceptions [10].

The main aim of the fake news spreader is to confuse people, and they can't differentiate between the truth and false information [11]. Many of us try to consume the real news, but contributors to the false information are not necessary to be hsudhakan, and maybe they are social media bots, cyborgs and humans [12].

2.2 Datasets

The news articles will be classified into two types:(i) genuine and (ii) fake news articles, and in this research, in this research, we followed different strategies to identify fake news and reals news. Table I will show the various news sources for the annotated news articles, and from this source, we will collect only a lot of the report. After collecting the news, we will label the information as accurate or fake for that purpose, and we will follow some strategies. 1. The reliable news channel published the news. 2. The same news published by the other authentic news channels. 3. We can also confirm the relationship between the content and title of the news. If the news did not follow this strategy, we could ensure that the news was unauthentic.

Table 1. News channels

Name	URL	Origin
Dinamalar	https://www.dinamalar.com/	Tamil Nadu
Dailythanthi	https://www.dailythanthi.com/	Tamil Nadu
Dinamani	https://www.dinamani.com/	Tamil Nadu
One India	https://tamil.oneindia.com/	Tamil Nadu
Maalaimalar	https://www.maalaimalar.com/	Tamil Nadu
Dinakaran	https://www.dinakaran.com/	Tamil Nadu
Tamil Murasu	https://www.tamilmurasu.com.sg/	Singapore
Seithi	https://seithi.mediacorp.sg/	Singapore
Lankasri	https://lankasri.com/	Sri Lanka
Tamilwin	https://tamilwin.com/	Sri Lanka
Vanakam Malaysia	https://vanakkammalaysia.com.my/	Malaysia
Vanakam London	https://vanakkamlondon.com/	London

Table 2 will show us the split of the datasets and the dataset is split into three sessions; the first session is for training purposes, the second session is for validation, and the third session is for testing purposes. The datasets were equally

shared for fake news and accurate news and 59.48% of training, 20.26% of validation and 20.26% of testing.

Table 2. Distribution of datasets

Dataset types	Real news	Fake news	Total
Training Dataset	7732	7732	15464
Validation Dataset	3859	3859	7718
Test Dataset	3859	3859	7718
Total	15450	15450	30900

Table 3 will show us examples of fake news and real news datasets. The information is the same, but the news channel spreads different information around the country.

Table 3. Real & fake news dataset example

Label	Source	Information
Fake	https://tamil.oneindia.com/fact-check/bjp-sowdha-mani-spreads-fake-news-on-communist-citizenship-in-the-usa-479987.html?ref_medium=Desktop&ref_source=OI-TA&ref_campaign=Topic-Article	America will not provide visas for any communist party in India that will not allow them to enter their country, but the news is fake.
Real	https://www.tribuneindia.com/news/nation/us-bars-communists-but-new-norms-not-applicable-to-india-151211	It does not apply to communist parties in India. The targets are ruling communist parties in China, Cuba and North Korea.

Figures 2 & 3 will show us the distribution of words for the real news and fake news datasets.

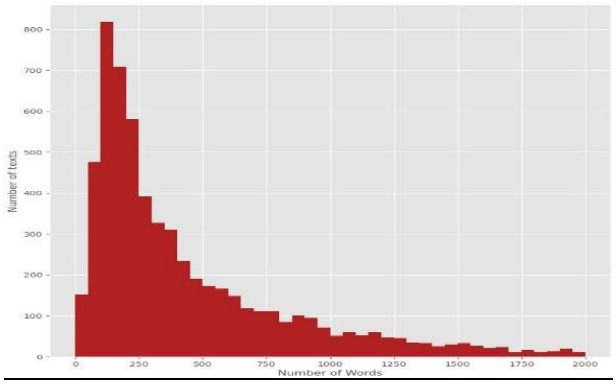

Figure 2. Distribution of words for fake news dataset

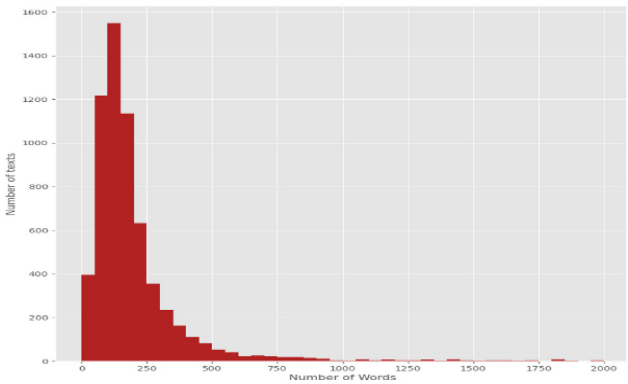

Figure 3. Distribution of words for real news dataset

3. CLASSICAL ALGORITHMS AND PROPOSED ARCHITECTURE

Machine learning algorithm classifiers are used for many purposes, and one of the purposes is fake news detection. First, the classifiers will be trained with a dataset, and then the classifiers will detect the phoney news automatically. We are going to discuss the various machine learning classifiers [13].

Figure 4 will show us the proposed architecture for the fake news detection mechanisms. In this architecture, we will use two classical algorithms of Naive Bayes, Logistic Regression and Support Vector Machine and one deep learning model of Long Short-Term Memory [14]. The software package used in this experiment is Cuda toolkit (11.8), Python (3.11.0), PIP Keras (2.10.0) and Tensor Flow (2.7).

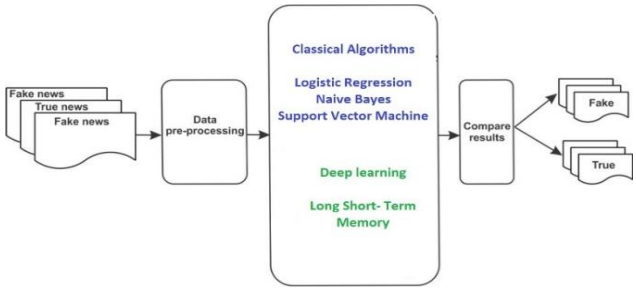

Figure 4. Proposed system

3.1 Naïve Bayes algorithm

It is one of the classification techniques. It is based on the Bayes theorem. This technique automatically predicts and assumes all the predictions are independent. It takes the fundamental probability knowledge, believes that all the inputs are separate, and signifies the output. If you want to construct a classification model, the Naive Bayes is the most straightforward approach, and the dataset was collected from Facebook news. It contains 3500 articles for training purposes and 1700 for testing [15]. The dataset is divided into training, testing, and validation. The training dataset is used to train the model; the validation dataset is used for global parameters. We will use the testing dataset for accuracy purposes. Here we used the steaming method to convert from words to root form, and the accuracy of this model is 74%. Data accuracy will be compared using the Support Vector Machine, Neural Networks and Naive Bayes classifiers [16]. The dataset was collected from various Twitter posts, and we used normalization techniques to clean the data before using it for training purposes. The accuracy for each classifier follows: 99.90% for SVM, 99.90% for NN and the Naive Bayes will provide 96.08%. Figure 5 will show us the workflow of the Naive Bayes Algorithm

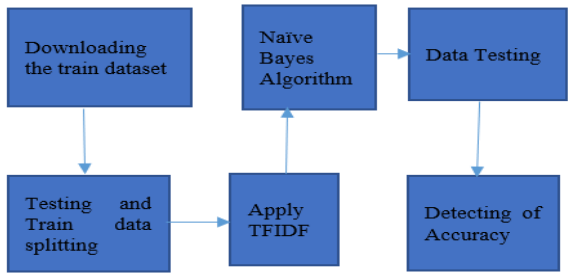

Figure 5. System architecture of NBA

3.2 Logistic regression

It is one of the classification algorithms used in machine learning to predict variables. Binary variables used here in the Logistic Regression and the binary code 1 indicate success, and 0 indicates failure. In the research liar, dataset and three algorithms (Naive Bayes, Logistic Regression, and Random Forest) were used for fake news detection [17]. Their research detected fake news from political information and utilised Logistic Regression's Naive Bayes classifier. These two algorithms used materials and methods to compare their performance and test the data. In their research, 44,000 datasets were used, and the accuracy for the Naive Bayes is 94.84, and Logistic Regression is 98.70 [18, 19].

3.3 Long short-term memory

The Long Short Term-Memory is based on the Recurrent Neural Network Architecture and it is capable of learning a mapping between the input and output patterns.

4. RESULT

Table 4 will show the result for the validation dataset and we used three classical algorithms and one deep learning model in this experiment. The F1 score for the Naive Bayes algorithm is 97.50%, the F1 score for the Support Vector Machine is 94.40%, and the F1 score for the Logistic Regression is 96.50%. The Naive Bayes will provide the highest F1 score in the classical algorithms. The deep learning model LSTM F1 score is 96.7%. The total number of the dataset for the validation and test is 7718. Figure 6 will show us the accuracy of each classifier.

Table 5 will show the result for the test dataset and we used three classical algorithms and one deep learning model in this experiment. The F1 score for the Naive Bayes algorithm is 97.50%, the F1 score for the Support Vector Machine is 94.70%, and the F1 score for the Logistic Regression is 96.50%. The Naive Bayes will provide the highest F1 score in the classical algorithms. The deep learning model LSTM F1 score is 99.7%. The total number of the dataset for the validation and test is 7718.

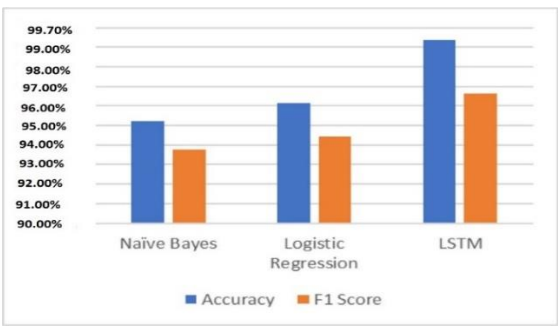

Figure 6. Accuracy for different classifiers

Table 4. Validation result

Type	Name	Acc	Pre	Rec	F1	TP	TN	FP	FN
Classical	SVM	0.944	0.904	0.988	0.944	3650	3657	354	57
	NB	0.976	0.957	0.994	0.975	3670	3773	251	24
	LR	0.965	0.954	0.976	0.965	3885	3590	225	18
Deep learning	LSTM	0.939	0.964	0.997	0.967	3673	3743	287	15

Table 5. Test result

Type	Name	Acc	Pre	Rec	F1	TP	TN	FP	FN
Classical	SVM	0.945	0.9120	0.985	0.947	3750	3557	354	57
	NB	0.975	0.964	0.986	0.975	3770	3673	251	24
	LR	0.965	0.954	0.976	0.965	3785	3690	225	18
Deep learning	LSTM	0.979	0.964	0.994	0.997	3773	3643	287	15

5. CONCLUSIONS

This research article proposes machine learning algorithms and systems to detect fake news in Indian online information and sources. Fake news is spreading daily in our lives, and we need an automatic detection system; for that, we are using machine learning models. Much research has been done on fake news detection, and this proposed model is the first model for detecting phoney news in Tamil news media. This proposed system will automatically detect faux news from online news media, and this system will help us to control phoney information. In this paper, we used machine learning classifier algorithms of Naive Bayes, Logistic Regression, Support Vector Machine, and the Convolutional Neural Network model of LSTM. The proposed model of LSTM will provide better accuracy.

REFERENCES

[1] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal,S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., Candlish, S.M., Radford, A., Sutskever, I., Amodei, D. (2020). Language models are few-shot learners. In Proceedings of the 33th Conference on Neural Information Processing Systems, pp. 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

[2] García, S.A., García, G.G., Prieto, M.S., Guerrero, A.J.M., Jiménez, C.R. (2020). The impact of term fake news on the scientific community. Scientific performance and mapping in web of science. Social Sciences, 9(5): 73. https://doi.org/10.3390/socsci9050073

[3] Tran, D.N., Nguyen, T.N., Khanh, P.C.P., Trana, D.T. (2022). An iot-based design using accelerometers in animal behavior recognition systems. IEEE Sensors Journal, 22: 17515-17528. https://doi.org/10.1109/JSEN.2021.3051194

[4] Yu, K.P., Lin, L., Alazab, M., Tan, L., Gu, B. (2021). Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system. In IEEE Transactions on Intelligent Transportation Systems, 22(7): 4337-4347. https://doi.org/10.1109/TITS.2020.3042504

[5] Rachana, B., Priyanka, T., Sahana, K.N., Supritha, T.R., Parameshachari, B., Sunitha, R. (2021). Detection of polycystic ovarian syndrome using follicle recognition technique. Global Transitions Proceedings, 2(2): 304-308. http://dx.doi.org/10.1016/j.gltp.2021.08.010

[6] Manzoor, S.I., Nikita, Singla, J. (2019). Fake news detection using machine learning approaches: A

systematic review. In Proceedings of the 3rd International Conference on Trends in Electronics and Informatics, pp. 230-234. http://dx.doi.org/10.1109/ICOEI.2019.8862770

[7] Kaliyar, R.K., Goswami, A., Narang, P. (2021). Fake BERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools and Applications, 80: 11765-11788. https://doi.org/10.1007/s11042-020-10183-2

[8] Ajao, O., Bhowmik, D., Zargari, S. (2018). Fake news identification on twitter with hybrid CNN and RNN models. In Proceedings of the 9th International Conference on Social Media and Society (SMSociety '18), pp. 226-230. https://doi.org/10.1145/3217804.3217917

[9] Buzea, M.C., Trausan-Matu, S., Rebedea, T. (2022). Automatic fake news detection for romanian online news. Information, 13(3): 151. https://doi.org/10.3390/info13030151

[10] Guo, Z.W., Shen, Y., Bashir, A.K., Imran, M., Kumar, N., Zhang, D., Yu, K.P. (2020). Robust spammer detection using collaborative neural network in Internet-of-Things applications. IEEE Internet of Things Journal, 8(12): 9549-9558. https://doi.org/10.1109/JIOT.2020.3003802

[11] Aslam, N., Khan, I.U., Alotaibi, F.S., Aldaej, L.A., Aldubaikil, A.K. (2021). Fake detect: A deep learning ensemble model for fake news detection. Complexity, https://doi.org/10.1155/2021/5557784

[12] Kumar, A., Singh, S., Kaur, G. (2019). Fake news detection of indian and united states election data using machine learning algorithm. International Journal of Innovative Technology and Exploring Engineering, pp. 1559-1563. http://dx.doi.org/10.35940/ijitee.K1829.0981119

[13] Vogel, I., Jiang, P. (2019). Fake news detection with the new german dataset "GermanFakeNC". International Conference on Theory and Practice of Digital Libraries, pp. 288-295. https://doi.org/10.1007/978-3-030-30760-8_25

[14] Suarez, P.J.O., Sagot, B., Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In Proceedings of the Workshop on Challenges in the Management of Large Corpora, pp. 9-16. https://doi.org/10.14618/ids-pub-9021

[15] Gilda, S. (2017). Notice of violation of IEEE publication principles: Evaluating machine learning algorithms for fake news detection. 2017 IEEE 15th Student Conference on Research and Development (SCOReD), pp. 110-115. http://dx.doi.org/10.1109/SCORED.2017.8305411

[16] Pardo, F.M.R., Giachanou, A., Ghanem, B., Rosso, P. (2020). Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on Twitter. CEUR Workshop Proceedings, 2696: 1-18. http://hdl.handle.net/10251/166528

[17] Sudhakar, M., Kaliyamurthie, K.P. (2022). Effective prediction of fake news using two machine learning algorithms. Measurements: Sensorrs, 24: 100495. http://dx.doi.org/10.1016/j.measen.2022.100495

[18] Tacchini, E., Ballarin, G., Vedova, M.L.D., Moret, S., de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. https://doi.org/10.48550/arXiv.1704.07506

[19] Sharma, U., Saran, S., Patil, S.M. (2020). Fake news detection using machine learning algorithms. International Journal of Creative Research Thoughts, 9(3): 1394-1402. http://dx.doi.org/10.17577/IJERTCONV9IS03104