International Journal of Safety and Security Engineering
Vol. 12, No. 6, December, 2022, pp. 723-727
Journal homepage: http://iieta.org/journals/ijsse

Fake News Detection in the Medical Field Using Machine Learning Techniques

Sudhakar Murugesan*, Kaliyamurthie Pachamuthu

Bharath Institute of Higher Education and Research, Chennai 600073, India

Corresponding Author Email: sudhakarmtech@gmail.com

https://doi.org/10.18280/ijsse.120608

Received: 10 October 2022
Accepted: 17 December 2022

Keywords:
fake news, KNN, naïve bayes, SVM, BERT, decision tree

ABSTRACT

In today's world, fake news is the biggest and most challenging problem in the natural world environment. This type of fake news will create many problems in society, especially in the medical field. This research aims to detect automatic phoney news in the news. The following machine learning algorithm will help us to see phoney medical information based on dataset 1. KNN, 2. Naive Bayes, 3. Support Vector Machine, 4. BERT and 5. Decision tree, but the Decision tree will provide better accuracy. Findings: Performance measures such as accuracy, precision, recall, and f1-score showed 98.5% accuracy of our proposed Adaboost & Decision Tree algorithm. In this research work, we introduced and implemented the Proposed Ensembling (Adaboost & Decision tree) and achieved better accuracy. We collected and trained the dataset to identify misinformation in the medical area.

1. INTRODUCTION

Today's fake news is causing significant damage to the world because of the development of internet facilities, and many people have started using social media. People are not checking the originality of the message before sharing it with others. This types of fake news are a severe threat to society. It is difficult to see the phoney info manually because millions of information spread in work through social media, so we have to implement an automatic system to detect this type of artificial information. There are many reasons to create a piece of fake news; some will make it for fashion, motive, or sentiment and destroy the individuals. In Wuhan city (China) December 2019, Covid-19 first started. This virus started to spread all over China and gradually spread to almost all the world. Due to the spread of this virus, many people have been affected, and many people have died, particularly in the US, India, Italy, the United Kingdom and Spain, etc., because there is no medicine found to cure this virus [1]. World Health Organization announced the Covid-19 pandemic due to this massive spread and increase in death rate worldwide. Every day it is spreading from person to person. These led to many countries implementing lockdowns and social distancing. This lockdown caused many problems, but it helped nations to reduce the spread of Covid-19 [2].

In this medical field, any newly discovered disease will spread quickly, and reason is due to the uncertainties of scientific research around research. At the same time, people are desperate for information, and this gap is filled by misinformation; hence the situation gives a thriving platform to fake news on the Internet.

With the development of Information technology, almost everyone has started using the Internet and social media. During the Covid-19 lockdowns, all the schools, colleges, and universities moved to online teaching mode, so everyone had a smartphone. Millions of messages and posts on social media are flooded about Covid-19, but not all the messages are authentic; most are false. Some of them spread that vaccinations are killing people, and this type of message is causing problems for researchers and doctors and creating public panic [3]. This type of fake news will create more problems in the medical sector; sometimes, it will cause death. From January 2020 to April 2020, the International Fact-Checking Network researched the fake news spread on social media. Because of the fake news, some people are stopped eating non-vegetarian food because the virus is spread by animals and led to economic crises in some countries [4]

This type of misinformation will create more problems in the medical field. Removing false information and fake news immediately from the Internet or social media is complicated. We must develop a system to help the public identify whether this information is fake or real.

A tsunami of misinformation accompanies any pandemic or new disease. A study analyzing the main types, sources, and claims of COVID-19 reports that the fact-checkers increased by 900% from January to March 2020, highlighting the existence of misinformation during the early months of the pandemic [4]. During the pandemic era circulation of fake news is more harmful as its scope is broad, ranging from dangerous cures, antivaccination and false conspiracy theories to altering general public opinion.

Figure 1. The spread of misinformation (social media outlets)

Misinformation exists not only for new diseases but also for diseases like cancer and diabetes, and alternative cures are the main class of fake news about such illnesses. Shi observes a dramatic increase in online searches for cannabis as a cancer cure. The fake news stories about cancer treatments gained 4.26 million engagements compared to the original levels, which earned only 0.036 million engagements on social media [5] (Figure 1).

Here we used two methods, and the first method was to collect the dataset for Covid-19, Cancer, Ebola, Yellow fever, Malaria and HIV. This dataset is collected from various websites, forums, journals, and WHO official websites. The datasets named fake and accurate news and the same balanced datasets were used in both. The second method will train the dataset using machine learning algorithms [6].

2. LITERATURE REVIEW

Many people use social media to spread fake news faster because it will not take any money, unlike other mediums. Many research works are available regarding fake news. Many researchers use machine learning algorithms to detect fake news; some use deep learning algorithms that also notice fake news [7]. Some researchers focus on general unnatural news detection methods, and others on social media counterfeit news detection. Currently, many researchers are using a bot detection of fake news [8].

A two-step model was designed to detect fake news from social media Artificial intelligence algorithms [9]. Three types of datasets were used in this research, and unstructured datasets were used to obtain meaningful data. There are 23 results generated using supervised learning algorithms, and on another side, we used the Covid-19 pandemic biostatistical analysis using the KNN classifier. These researchers collect the data based on some information from social media and some news topics and summarize them. The KNN will predict the fake news accuracy of 80% [10].

Any new pandemic that comes to the misinformation will also follow it. The report shows that within three months (January 2020 to March 2020) Covid-19 related searches increased by 900 per cent. Sharing fake information during the pandemic is more harmful and changes public opinion. Misinformation is not only for the new disease; it is also available for long-time diseases such as cancer, malaria, HIV and diabetes [5]. Researched cancer and observed that it keeps increasing the search for cannabis. This cannabis is a cancer cure. More than 4.26 million people searched and gained about cancer treatment fake news. Still, factual information earned only 0.036 million on social media, and misinformation has a more significant influence than real news. This type of phoney information in the medical field will cause death, not like other fields.

Nowadays, health-related misinformation leads researchers to research the medical field misinformation [11]. He studied Thai healthcare and used a deep-learning approach for the dataset. Ciara & Cioca researched and succeeded in fake news management in healthcare and used the machine learning-based KNN-BSA approach, which produced an accuracy of 70 per cent.

The Marco L. Della Vedova research on fake news detection will produce an accuracy of 78.8 per cent. He collected the data from various social media platforms such as Messenger, Tweets and real-world applications. He implemented a content-based approach method and a social-based system. In this research, he got the accuracy of fake news is 81.7 %.

The Elyassami classifies news as fake or real using machine learning. An investigation of how voting strategy impacts ensemble learning models was conducted. The four performance measures used to evaluate the five classifiers were accuracy, F1-score, recall, and precision. We are encouraged by the results. It is possible to use these ensembles against fake news spreading since they outperform other classifiers when trained with random forest algorithms and gradient boosting algorithms. Galli et al. [12] present their complete framework for detecting fake news and describe their machine-learning-based solution. Using different datasets, we demonstrate our algorithms can see phoney news and produce high-accuracy results.

Researchers have attempted to address the problem of fake news using various techniques. Some research uses traditional machine learning models like Na¨ıve Bayes, Support Vector Machines [13], while others use deep learning models like CNN and BiLSTM [7]. The datasets used in most experiments are the standard fake news datasets like LIAR, BuzzFeed, ISOT, etc. While some studies focus on general phoney news detection, others focus on fake news on social media like Twitter or Facebook. For instance, Helmstetter and Paulheim [14] used weakly supervised learning techniques to detect fake news on Twitter automatically.

Ciora and Cioca [15] is another successful research on fake news detection in healthcare. Their research articles showed 70% accuracy achieved by using the KNN-BSA-based machine learning algorithms. The datasets used in the experiments are two publicly available datasets on COVID. Even though we must tackle the misinformation on COVID, we should not forget that the same problem exists for other diseases like cancer, Ebola, or ZIKA. Despite advancements in the studies of these diseases, there still exists a lot of misconception among people regarding their facts.

3. METHODOLOGY AND DATA DESCRIPTIONS

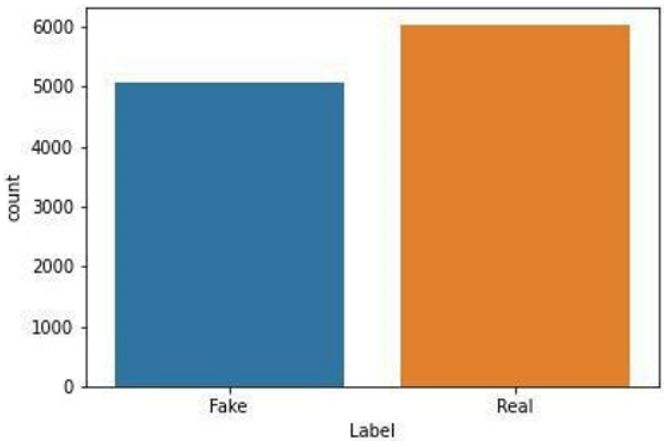

Figure 2. Class Distribution of the dataset

Our research will consider the most famous diseases like Covid-19, Malaria, Ebola, Cancer and HIV and Flu. We will collect the data based on the condition when the disease appeared, how many people were affected, which are countries affected, what are the symptoms and the number of people who died due to the disease. Also, we gathered information

from published research articles, web searches and Kaggle open source. Figure 2 shows the class distribution of the dataset. The dataset has a total of 11001 records, of which 6036 are accurate, and 5065 are fake.

Figure 3 will show the word count of each record and count the number of words in the text. The average count of word length is less than 20 in the description.

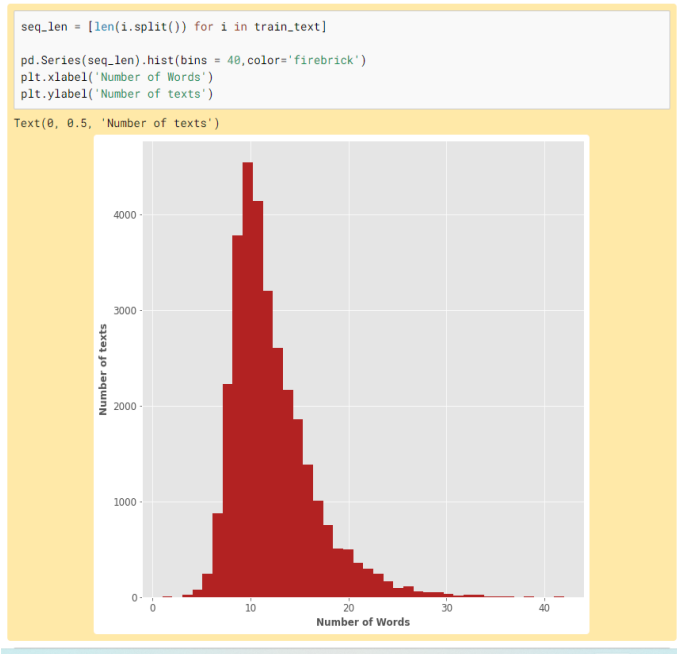

Figure 3. Sentence of length

Figure 4 will show the workflow of the news detection approach, and this classifier model will classify the news as fake or real.

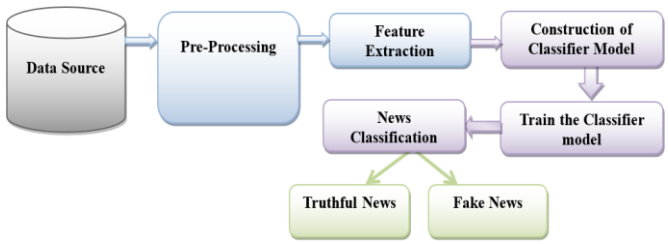

Figure 4. Workflow of fake news detection approach

Jacob Devlin introduced the BERT from Google in 2019, and it is a pre-trained model (Figure 5). This model will help us find the vectors that reflect their similarities in distance. Here we have used a token and tokenization algorithm; the word is based on the token, and the sub-word is based on the tokenization algorithm.

	precision	recall	f1-score	support
False	0.47	0.37	0.41	171
True	0.53	0.63	0.57	190
accuracy			0.51	361
macro avg	0.50	0.50	0.49	361
weighted avg	0.50	0.51	0.50	361

Figure 5. Sample classification report BERT

4. DATA ANALYSIS

We are going to perform the data analysis to highlight the key characteristics. There are three columns in the dataset; the first column indicates the title, the second column suggests text data, and the third column shows the text's label (Table 1).

Table 1. Attributes of datasets

Name of the data set	Total records	Real	Fake	Unique words	Average sentence length
MedHub	11101	6036	5065	30789	20
Diabetes Test	162	93	69	727	13
Covid-19 misinformation	13459	0	13459	23459	17

5. ALGORITHMS

Here we are going to discuss the algorithms that used in this research, and we are going to train and test each algorithm.

5.1 Naïve Bayes algorithm

Here we used the Bayes theorem to classify the news, and the Naïve Bayes classifier will work based on this theorem. The advantage of this theorem is that we can build very quickly, and it will work for large datasets. The disadvantage of this theorem is that it assumes all the variables are dependent. This algorithm will help us to find the nearest vectors by calculating the similarity between the two closest neighbours.

5.2 Logistic Regression algorithm

Logistic Regression is a supervised learning algorithm. It provides accurate results when new data is given to the trained model. It is a predictive analysis algorithm based on the concept of probability. The mathematical function of the sigmoid is used to map the predicted value to probabilities. The value of Logistic Regression must be between 0 and 1, which can be calculated using the equation: 1/(1+e^-value).

5.3 Decision Tree algorithm

	Text	Label
0	"Spraying chlorine or alcohol on the skin kill...	Fake
1	"Only older adults and young people are at risk"	Fake
2	"Children cannot get COVID-19"	Fake
3	"COVID-19 is just like the flu"	Fake
4	"Everyone with COVID-19 dies"	Fake

Figure 6. Dataset for training

A supervised learning algorithm is a Decision Tree. It may be used for Regression as well as classification. By learning fundamental choice rules from training data, a Choice Tree may be used to develop an innovative model that could be utilized to forecast the target variable's score (Figure 6). Below are the equations needed to do classification using a Decision Tree (1) Gini index defines the favour more significant probability (2) entropy is used to calculate the homogeneity of the sample (3) information gain is used to compare the samples before and after transformation:

5.4 Random Forest algorithm

The Random Forest method is a supervised learning approach to predict and classify data. A Random Forest is a meta-classifier that fits many Decision Tree classifiers to distinct sub-samples of the dataset and utilizes averaging to increase projected accuracy and control over-fitting. If bootstrap='True' is a default value, the subsamples are governed by the max sample size argument; otherwise, every tree will be generated using the entire dataset. It's a more advanced variation of the Decision Tree.

5.5 KNN algorithm

The KNN is a text classification algorithm and decision boundary-based classification algorithm. It is a straightforward algorithm to find the nearest neighbour by calculating the similarity between the closest vectors. Many researchers' used KNN for text classification on Facebook and Twitter. Here K is the value of the nearest neighbour value, and the KNN model achieved 85.5% of accuracy. The time complexity and data distribution is the drawback ok KNN when it comes to the text classification because it worked well up to 11000 records. The KNN has many disadvantages; one of the drawbacks is the high testing cost.

6. SYSTEM ARCHITECTURE

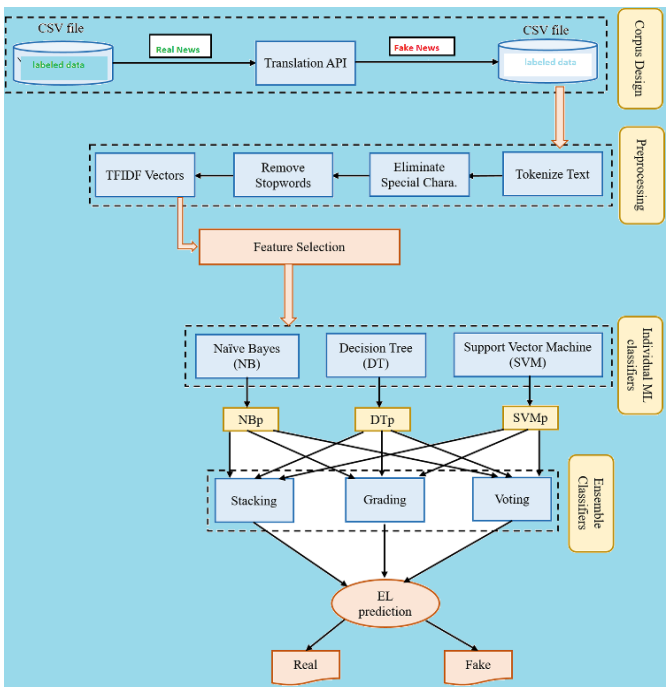

Figure 7. Architecture of automatically filtering news

In this architecture, we are going to train the data. In the processing field, it will remove the repeated words, common sentences and unique words (Figure 7). Then the algorithm will perform various tests and predict whether the news is fake or real.

7. RESULTS

Table 1 shows the experimental results of each algorithm. Three datasets are used here; two were downloaded from the Kaggle (DS1 & DS2); the third is DS1. This DS1 is the combination of DS2 and DS3. Table 2 shows the result of various machine learning algorithms. The proposed algorithm will have provided better results.

Table 2. Results

Result for each algorithm	Accuracy	Precision	Recall
Decision tree Adaboost	98.7	95.3	93.2
Logistic Regression	98.1	92.5	94.2
Support Vector Machine	95.3	94.5	93.2
K-Nearest Neighbours	92.5	85.5	86.5
Naïve Bayes	93.4	92.3	87.8

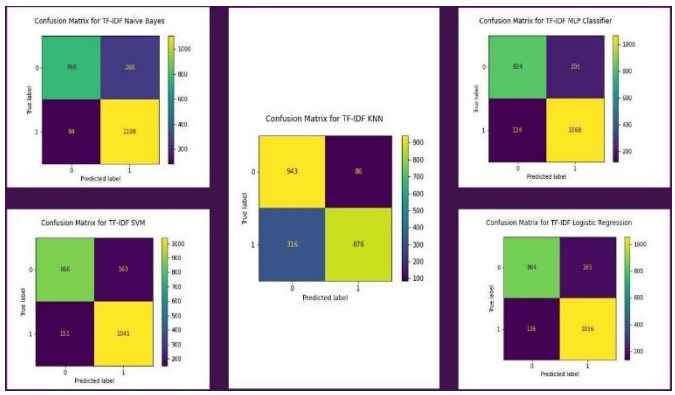

Figure 8. Confusion matrix

Figure 8 show the evaluation metrics and the confusion matrix for all models using TF-IDF vectorization. SVM, Logistic Regression, and Naive Bayes perform the best and are almost similar. Figure 9 will shows the accuracy value for the datasets.

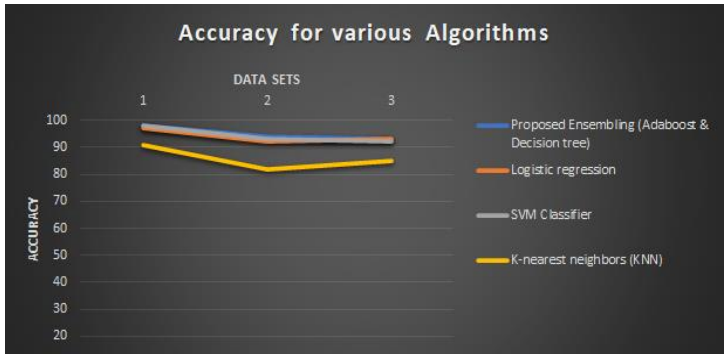

Figure 9. Accuracy value of datasets

All the models perform well and achieve more than 90% test accuracy except for KNN (Table 3). One of the desirable properties of the Fake news detection system is a high recall

rate for the Fake class; it should be able to identify fake records well compared to actual records.

Table 3. Value of Precision

Precision value result	Accuracy	Precision	Recall
Decision tree Adaboost	96.7	93.5	94.7
Support Vector Machine	95.2	75.3	76.2
Logistic Regression	98.2	74.3	93.4
K-Nearest Neighbours (KNN)	89.2	85.2	87.5
Naïve Bayes	92.5	78.6	79.5

8. CONCLUSIONS

Today's development of technology and the Internet leads to the spread of misinformation on social media. Many people will see this phoney information and share this fake news with others without verifying the truth of the content. This fake content will cause many problems for the medical industry. During the Covid-19, Ebola and HIV time, much misinformation spread among us, and most of the information was invalid. It caused many problems for the community. If you verify this type of fake, we can reduce the issues. Our proposed Adaboost & Decision tree algorithm will have provided better accuracy of fake news detection, and other algorithms also offered better results.

REFERENCES

[1] Livingston, E., Bucher, K. (2020). Coronavirus disease 2019 (COVID-19) in Italy. Jama, 323(14): 1335-1335. https://doi.org/10.1001/jama.2020.4344

[2] Pham, Q.V., Nguyen, D.C., Huynh-The, T., Hwang, W. J., Pathirana, P. N. (2020). Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: a survey on the state-of-the-arts. IEEE Access, 8: 130820. https://doi.org/10.1109/ACCESS.2020.3009328

[3] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological Science, 31(7): 770-780. https://doi.org/10.1177/0956797620939054

[4] Brennen, J.S., Simon, F.M., Howard, P.N., Nielsen, R.K. (2020). Types, sources, and claims of COVID-19 misinformation. Doctoral dissertation, University of Oxford.

[5] Shi, S., Brant, A.R., Sabolch, A., Pollom, E. (2019). False news of a cannabis cancer cure. Cureus, 11(1): e3918. https://doi.org/10.7759/cureus.3918

[6] Gadekallu, T., Soni, A., Sarkar, D., Kuruva, L. (2019). Application of sentiment analysis in movie reviews. In Sentiment Analysis and Knowledge Discovery in Contemporary Business, pp. 77-90. https://doi.org/10.4018/978-1-5225-4999-4.ch006

[7] Jiang, T., Li, J.P., Haq, A.U., Saboor, A. (2020). Fake news detection using deep recurrent neural networks. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, pp. 205-208. https://doi.org/10.1109/ICCWAMTIP51612.2020.9317325

[8] Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots?. arXiv preprint arXiv:2004.09531.

[9] Ozbay, F.A., Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540: 123174. https://doi.org/10.1016/j.physa.2019.123174

[10] Bandyopadhyay, S., Dutta, S. (2020). The analysis of fake news in social medias for four months during lockdown in COVID-19-a study: Biostatistical analysis of COVID-19. Preprints. https://doi.org/10.20944/preprints202006.0243.v1

[11] Payoungkhamdee, P., Porkaew, P., Sinthunyathum, A., Songphum, P., Kawidam, W., Loha-Udom, W., Boonkwan, P., Sutantayawalee, V. (2021). LimeSoda: Dataset for fake news detection in healthcare domain. In 2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), pp. 1-6. https://doi.org/10.1109/iSAI-NLP54397.2021.9678187

[12] Galli, A., Masciari, E., Moscato, V., Sperlí, G. (2022). A comprehensive Benchmark for fake news detection. Journal of Intelligent Information Systems, 59: 237-261. https://doi.org/10.1007/s10844-021-00646-9

[13] Jain, A., Shakya, A., Khatter, H., Gupta, A.K. (2019). A smart system for fake news detection using machine learning. In 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, pp. 1-4. https://doi.org/10.1109/ICICT46931.2019.8977659

[14] Helmstetter, S., Paulheim, H. (2018). Weakly supervised learning for fake news detection on Twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, pp. 274-277. https://doi.org/10.1109/ASONAM.2018.8508520

[15] Ciora, R.A., Cioca, A.L. (2021). Fake news management in healthcare. In 2021 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, pp. 1-4. https://doi.org/10.1109/EHB52898.2021.9657578