2023-03

Fake News Detection Using Machine Learning Techniques

Sultana, Achhiya

Independent University, Bangladesh

Draft Version

Fake News Detection Using Machine Learning Techniques

Achhiya Sultana
Department of Computer Science and Engineering
Independent University, Bangladesh
1821707@iub.edu.bd

Mahmudul Islam
Department of Computer Science and Engineering
Independent University, Bangladesh
mahmud@iub.edu.bd

Mahady Hasan
Department of Computer Science and Engineering
Independent University, Bangladesh
mahady@iub.edu.bd

Farruk Ahmed
Department of Computer Science and Engineering
Independent University, Bangladesh
farruk@iub.edu.bd

***Abstract*—A lot of information is spread by people in the social media to update their status and share crucial news with others. But the majority of these platforms don't promptly validate the individuals or their posts and people aren't able to identify the fake news manually. Therefore, there is a need for an automated system capable of detecting fake news. This research has proposed to build a model using four machine learning algorithms. The dataset employed in the experiment is a composite of two datasets containing almost equal amounts of true and fake news articles on politics. The preprocessing stages begin with cleaning the data by removing punctuation, tokenization, special characters, white spaces, redundant word elimination, numerals, and English letters followed by stemming and stop with data discretization. Then, we analyzed the collected data and 80% of the data has been used to train each model initially. After that, the four manifested classification algorithms are applied. For identifying fake news from news articles, methods like Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting Classifier were used. The trained classifiers' accuracy has been evaluated using the remaining 20% of the data. The results show that the decision tree model produces the best accuracy of 99.60% and gradient boosting of 99.55%. Besides, the random forest shows 99.10% along with the logistic regression 98.99%. Moreover, we have explored the best model to achieve the highest precision, recall, F1-score based on the confusion matrix's outcome.**

***Index Terms*—Social Media, Fake News Detection, Machine Learning, Classifier, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting.**

I. Introduction

Our world is changing so quickly; without a question, the digital world has many benefits, but it also has drawbacks. One of them is fake news (FN). Day by day, we are relying upon social media sites for instance Facebook, LinkedIn, Twitter, Instagram, and others. Individuals seek information from these online platforms, yet some evil people distribute wrong information, driving them insane. To fool the public, phony news is created in a verifiable and factitious untrue manner [1], [2]. Fake propaganda is disseminated in order to destroy the status of a person or an entity. This could be misinformation directed at a political party or an organization. Artificial intelligence's machine learning component assists in creating systems that can learn and carry out various tasks [3]. Different machine learning methods are available, including supervised, unsupervised, and reinforcement learning. A data set named as the train data set must be used to train the algorithms first. After training, the algorithms might be utilized to perform many tasks. Machine learning algorithms are frequently used to make predictions or find something that is concealed.

Though, people might benefit from online platforms because they have easy access to news. However, the issue is that it allows cybercriminals to spread phony news through the platforms. People peruse the information and begin to believe everything before it has been verified. Researchers discovered that the volume of false information is increasing four times over [4]. That is why it is necessary to recognize phony news. In our research, we have used Artificial Intelligence(AI) models to identify false information. Machine learning techniques are used to enable computers to learn from given information in order to do particular tasks. We have developed a supervised strategy for this research. Each Machine Learning model needs a suitable dataset to produce an exact result. The dataset must be collected and preprocessed before doing analysis. Then, We were going to start by training the machine using the Machine Learning (ML) model. Finally, we tested some of the dataset's data.

In the beginning, 80% of the data is used to train every model. The accuracy of the trained classifiers is assessed using the rest 20% of the data. Since, the algorithms are trained. As a result, the machine learning algorithms automatically detect bogus news. The aim of this research motivated us to investigate the following research question:

RQ: How effectively can various machine learning models handle the challenges of detecting fake news?

The prime contribution of this paper as follows:

- The best accuracy is established for all applied algorithms.

- The best precision, recall, and f1-score are established for all algorithms.

This study's remaining sections are organized as follows. Related works are covered in Section 2. The method has been described in Section 3. Section 4 of our research presents the experimental findings. Finally, section 5 has drawn conclusions and future work.

II. Related Works

When online news and interpersonal contextual factors are combined, a unique machine learning (ML) false information identification strategy that exceeds other approaches in the journal and boosts accuracy to 78.8% [5] was initially suggested by Marco L. Della Vedova et al. Secondly, researchers used the technique on a Facebook Messenger chatbot and tested it on a practical situation; researchers were capable of identifying false propaganda with just an efficiency level 81.7%. Researchers first discussed the datasets they utilized for the experiment and showed the evidenced technique, they employed and the way they suggested combining it into a progressive strategy that was previously proposed in the literature. Their objective was to categorize a newspaper article as neither authentic nor bogus. The finalized dataset comprises 15,500 messages across 32 pages (14 intrigue pages, 18 scientific pages), with higher than 2,300,00 views from more than 900,000 individuals. 6,577 (42.4%) posts are non-hoaxes, whereas 8,923 (57.6%) are hoaxes.

Maxson Fernandes and Arvinder Pal Singh Bali [6] demonstrated ML and NLP approaches in 2019. The estimation was performed on three standard datasets using a fresh collection of characteristics taken from the contents and headlines. Furthermore, the performance of seven ML models about precision and F1 values was evaluated. With such an efficiency of 88%, gradient boosting typically performed better than algorithms.

Zineb Ferhat Hamida and Ahlem Drif [7] introduced a combination of Convolutional Neural Network and LSTM Recurrent Neural Network structure, utilizing both LSTM's long-distance dependencies and CNN's poorly graded region of interest, in 2019. The dataset utilized was the contents information of hoaxes, and its volume was 20,761. The highest efficiency in CNN-LSTM is 0.725 when compared to the CNN and Support Vector Machine(SVM) standards.

SVM [8], Naive Bayes (NB) [9], LR [10], K-Nearest Neighbour (KNN) [11], RF [12] as well as DT [13] are some other machine-learning techniques that have recently been applied to identify false propaganda. Several techniques have really been effective at categorizing false news according to a variety of criteria. Numerous neural network techniques, including Recurrent Neural Networks (RNN) based models [13]–[15] for user propensity, Long Short-Term Memory (LSTM) with Linguistic Inquiry and Word Count (LIWC) features [16], and CNN-based models [17], [18] with feature points, were used to identify false information because feature extraction takes a lot of time.

III. Methodology

In this study, we discussed our proposed technique to detect fake news. Initially, we have collected datasets from several sources. Then, we have done the pre-processing steps and analyzed the data, trained the model, applied the algorithms. After that, we have tested the model to get the outcome which is mentioned in Fig.1. Every stage of our system is shortly stated in the following sections.

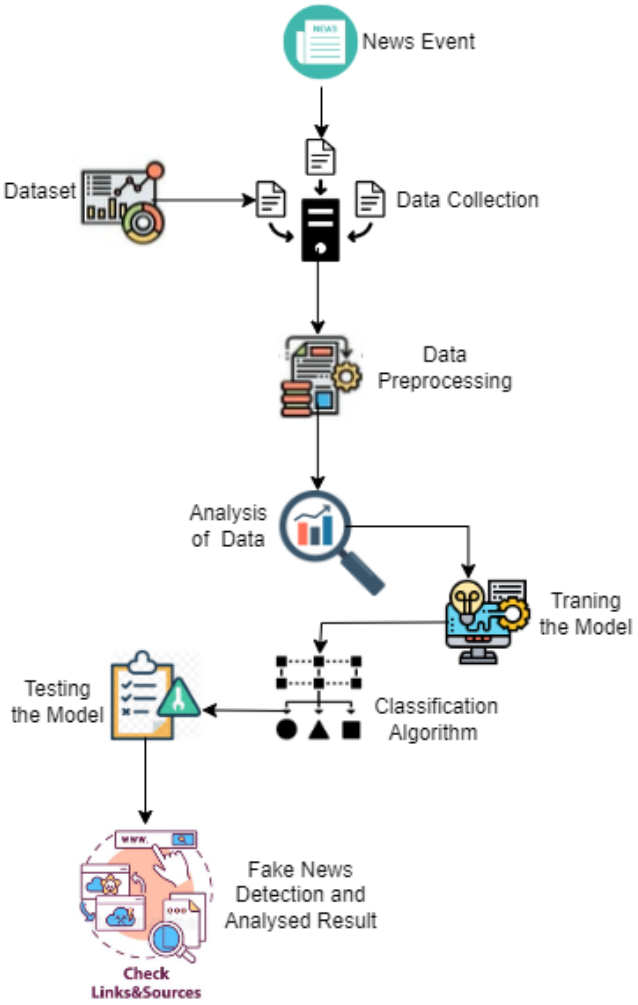

Fig. 1. Methodology for fake news detection

A. Data Collection and Preprocessing

In our research, We collected the required data from Kaggal website. The majority of social media data is unstructured, including errors, slang, and improper grammar, among other things [15]. Moreover, the data could be text (semi-structured, structured, unstructured), audio, video, photos, and so on. The data must be cleaned and it may be used to achieve better insights before modeling. Basic preprocessing was performed on news for this purpose. Data preprocessing steps are shown in Fig.2.

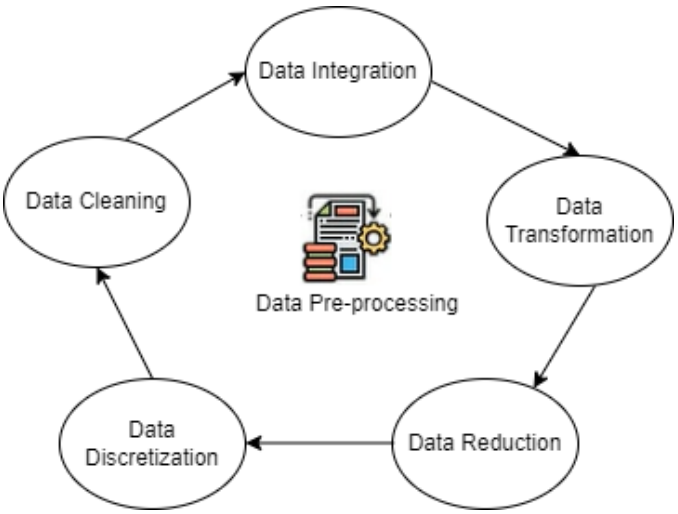

Fig. 2. Steps for Data Pre-processing

1) Data Cleaning: We can get the data in either a structured or an unstructured format while collecting it. Unstructured data does not have a proper framework while the structured data has a clearly established pattern. We also have a semi-structured format that sits between the two structures and it is better than the unstructured format. To highlight, the data must be cleaned before training the model. The data cleaning (or preprocessing) process typically includes the following steps:

- Remove punctuation: Punctuation can give a sentence grammatical context that assists in our perception. Nevertheless, it brings no value to our vectorizer, which only enumerates the words and ignores context. So, we exclude all special characters. Using an example: What are you playing now? - What are you playing now
- Eliminate redundant words: Redundant words are frequent words that almost always exist in texts. We eliminate them because they don't give much information about the data. For instance, gold or silver is acceptable to him - gold, silver, acceptable.
- Tokenization: Tokenization divides the text into smaller pieces. For example - words or phrases. It gives previously unstructured form. A good example is Plata o Plomo - 'Plata','o','Plomo'.
- Stemming: Stemming boosts decrease a term to its stem form. Treating terms that are related similarly often makes sense. It eliminates suffixes such as "ment", "able", and "ism", "less" using a straightforward rule-based method. Although the number of words decreases, the actual words are frequently overlooked. Such as, Amusement - Amuse.

2) Data Integration: A method of integrating data from several sources is called data integration. The objective is to give users a comprehensive understanding of the data. It might be better understood as a method for integrating information from several resources. It's regarded as one of the most important processes in the preparation of data. It employs a number of strategies, including data virtualization, streaming data integration, and data replication.

3) Data Transformation: Data transformation, broadly speaking, is the method of transforming the original data into the form or pattern that is more suitable for model construction and data discovery. It is a crucial stage in feature engineering that makes insight discovery easier. For joining two datasets (True data and Fake data) and removing column names from our dataset, we used data transformation.

4) Data Reduction: The technique of scrimping on functionalities in a computation that uses a lot of resources without missing crucial information is referred to as a feature reduction or dimension reduction. Having fewer features means having fewer variables, which makes the computer's work simpler and faster.

5) Data Discretization: The process of obtaining persistent data and transforming it into distinct containers is known as data discretization. The direct viability of the information is another characteristic of discretization. It turns out that building a model with discrete data is easier and faster than trying to build one with consistent data. Discretization can help us achieve some degree of harmony between the two in this situation.

B. Statistical Information

Every piece of data in our dataset is collected from Kaggal.com. There are six columns in total. The quantity of rows in the dataset is 23481 for false news and 21417 for real news. The dataset is kept as a CSV file.

C. Split method

In our research, we have used a split strategy to create classification and regression models. The splitting method of train-test datasets is an approach for evaluating a machine learning algorithm's performance. We used this technique to deal with clustering or relapse concerns. The system includes splitting up a dataset into two subsets. One subset has 80% of the data which are used to train the model. And the remaining set contains 20% which are used to test the model.

D. Algorithms

The following sections describe a quick explanation of each methods which are used to develop the model.

1) Logistic Regression: A logistic regression (LR) model is used to classify text based on a large feature set with a binary output because it provides the easy equation to categorize problems into binary or many classes and may classify difficulties into true/false or true article/fake article states [19]. While several parameters are tested before obtaining the maximum accuracy from the LR model, we did hyperparameter tuning to obtain the best outcome for each particular dataset. The following equations are a mathematical definition of the logistic regression hypothesis function [20]:

$$h_\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{1}$$

In logistic regression, the output is converted to a probability value using a sigmoid function, and the goal is to minimize the cost function to obtain the best probability. As shown in the cost function calculation:

$$\text{Cost } (h_\theta(x), y) = \begin{cases} \log\left(h_\theta(x)\right), & y = 1 \\ -\log\left(1 - h_\theta(x)\right), & y = 0 \end{cases} \quad (2)$$

2) Decision Tree Classification: The particular learning category includes the decision tree algorithm. They can be used to address problems with relapse and characterization. In this case, to determine if the article is bogus or real, we employed the decision tree (fig.3).

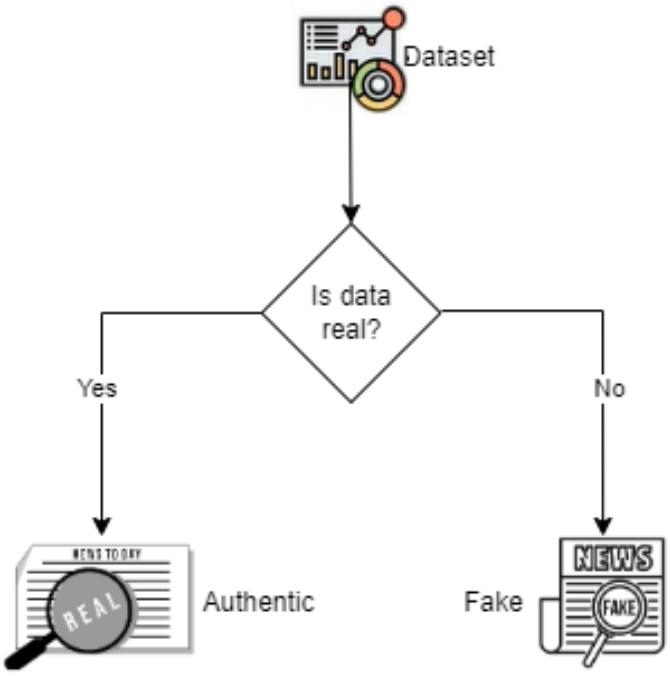

Fig. 3. Decision Tree for fake news identification

Finding the feature of the root node in every step of the Decision Tree presents a significant problem. The procedure in discussion is called feature selection. There are two widely used methods for selecting features:

- Information Gain
- Gini Index

Information Gain: Tv is the subset of T with A = v, Values (A) is the set of all possible values of A, and if T is a set of instances and A is an attribute, then

$$\text{Gain}(S, A) = \text{ Entropy}(S) - \sum_{veValues} (A)\frac{|S_v|}{|S|} \cdot \text{Entropy}\left(S_v\right) \quad (3)$$

Gini Index: The Gini Coefficient (Index) is a metric that used to determine how frequently a randomly selected piece would be recognized wrongly. It implies that an attribute with a lower Index value ought to be chosen. Sklearn supports the "Gini" criterion for the Gini Index and by definition, it uses the "Gini" value. The formula for calculating the Gini Index is provided below.

$$\text{GiniIndex } = 1 - \sum_j p_j^2 \quad (4)$$

3) Gradient Boosting Classifier: Gradient boosting combines various machine learning models, mostly decision trees, each of which makes a prediction. If we arrange all the decision tree models in a straight line, we can claim that each model will attempt to lessen the flaws of the one before it. The architecture of gradient boosting classifier is presented in fig.4.

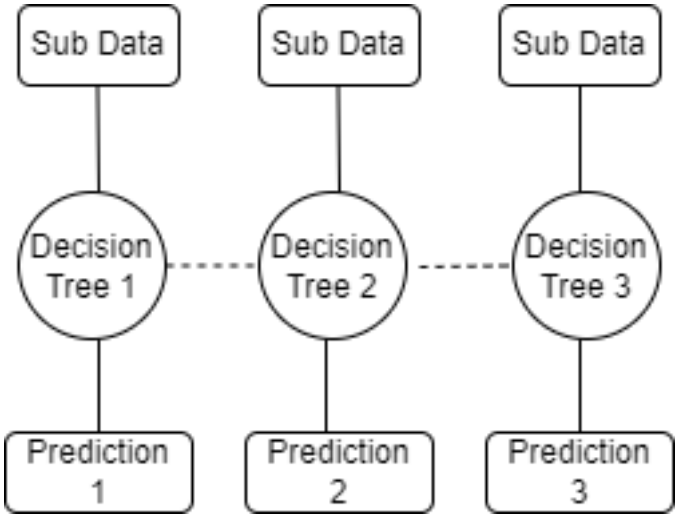

Fig. 4. Architecture of Gradient Boosting classifier

As we can see from the architecture, each decision tree makes a prediction, but do they all use the same dataset? Because they are using the same dataset, they can provide identical findings, negating the value of having several forecasts. The dataset, we're using, is one, but it's broken up into smaller subdatasets, each of which contains the same number of data points as the original dataset. Each subdataset is put into a decision tree model, and because each subdata is unique, we get the different outcomes.

4) Random Forest Classification: The core learning models of Random Forest are multiple decision trees. By selecting rows and attributes from the dataset at random, we produce sample datasets for each model. This part is referred to as Bootstrap.

The Random Forest regression technique must be approached similarly to other machine learning techniques. Create a specific query or set of facts, then ask the source to provide the needed information. Check that the data is in an accessible format, or convert it if necessary. List any visible abnormalities and missing data that may be needed to obtain the desired data.

Build a machine learning model. Decide on the baseline model you wish to reach. Train the machine learning model

using the data. Give the model some context with test data, and then compare the test data and the model's projected data performance metrics. If it falls short of what you were hoping for, you can try updating your model to reflect this, dating your data, or utilizing another data modeling technique. At this point, you interpret the information you have learned and make the appropriate reports.

Higher accuracy and over-fitting are prevented by the larger amount of trees present in the forest. The workflow of the Random Forest algorithm is illustrated in figure 5.

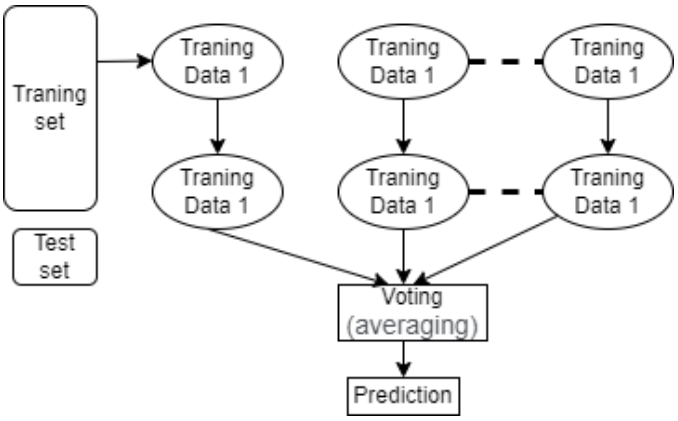

Fig. 5. Workflow of the Random Forest algorithm

E. Limitation

Obviously, human interaction is required at some point in any classification analysis. Although 100 percent accuracy may not be attainable, identifying the characteristics of false news would represent a positive development. While the results of the analysis suggest a model, some external features, such as the news source, author, starting location, and time stamp, were not taken into account and may have an impact on the model's output.

IV. Experimental Results

This section gives a full overview of a classification model's effectiveness as well as the types of errors it generates.

A. Results

The machine provides almost an authentic outcome. Usually, no machine can generate a 100% accuracy. But our prepared model gives a compatible outcome for all algorithms. We train our model using the split method. We have used different methods of machine learning algorithms to get the genuine result. And finally, we get 99.60% accuracy using the Decision Tree model.

On the test set, the accuracy of the Logistic Regression classifier was 98.99%. We have calculated the confusion matrix.

n = 8980	Predicted=Yes	Predicted=No
Actual Yes	TP(4239)	FN(42)
Actual No	FP(49)	TN(4650)

Table 1: Confusion Matrix for LR Model

The accuracy of the Decision Tree model was 99.60%. We have also computed the confusion matrix.

n = 8980	Predicted=Yes	Predicted=No
Actual Yes	TP(4256)	FN(25)
Actual No	FP(11)	TN(4688)

Table 2: Confusion Matrix for DT Model

The accuracy of the Random Forest model was 99.10%. Table 3 depicts the confusion matrix for Random Forest.

n = 8980	Predicted=Yes	Predicted=No
Actual Yes	TP(4253)	FN(28)
Actual No	FP(56)	TN(4643)

Table 3: Confusion Matrix for RF Model

The accuracy of the Gradient Boosting classifier was 99.55%. Table 4 represents confusion matrix for Gradient Boosting model.

n = 8980	Predicted=Yes	Predicted=No
Actual Yes	TP(4271)	FN(10)
Actual No	FP(26)	TN(4673)

Table 4: Confusion Matrix for GB Model

We can also compute recall, precision, and F1-score from the confusion matrix. The quantity of class positives that are actually class positives is referred to as precision. Recall measures how many accurate class predictions were made using the datasets' whole collection of successful examples. F1-score demonstrates a balance between recall and precision.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F-Measure} = \frac{2 * \text{ Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{7}$$

TP represents True Positives while FP represents False Positives. Similarly, TN represents True Negatives and FN represents False Negatives. Table 5 shows the models' performance measurement in details.

Model	Accuracy	Precision	Recall	F1-Score
LR	98.99%	98.857%	99.02%	98.94%
DT	99.60%	99.74%	99.416%	99.57%
GB	99.55%	99.39%	99.76%	99.57%
RF	99.10%	98.70%	99.34%	99.01%

Table 5: Performance for our Models

The chart which consists of several parameters of different methods is shown in Fig.6 below:

The accuracy for several models in our research is shown in Table 6 below.

Fig. 6. Accuracy, recall, precision, f1-score of different algorithms

Model	Accuracy
Logistic Regression	98.99%
Decision Tree Classification	99.60%
Gradient Boosting Classifier	99.55%
Random Forest Classification	99.10%

Table 6: Accuracy of models

The Decision Tree Classification achieves 99.60% accuracy which is the highest one. A comparison of this study with existing studies is provided in Table 7 below.

Studies	Algorithms			
	LR	DT	GB	RF
Marco L. Della et al.[5]	✓	-	-	-
Maxson and Arvinder Pal[6]	✓	✓	-	✓
Zineb and Ahlem Drif [7]	-	-	-	-
Proposed System	✓	✓	✓	✓

Table 7: Comparative Analysis of Fake News Detection Systems

V. Conclusion and Future Work

Social media has been used to distribute false news, which has a detrimental effect on both individual users and society. In this study, we investigated the issue of bogus news by analyzing previous research. We explained the fundamental ideas and guidelines of fake news. We examined current machine learning methods for detecting false news including model building. The Decision Tree helps the suggested model achieve its most notable precision. The precision score is 99.74% which is the highest one.

The dataset and its decomposition will be expanded in the next stage of our study to enhance the model's performance. Only a few machine learning models are used in our work. There are some suggestions for identifying false news - Create a large dataset of fake news; Identify its patterns; Learn which websites are frequently utilized to produce false news; Improve the model; Strive for greater accuracy.

We employed four different machine learning methods for this project, but there are more (including these four) those can be used, such as support vector machines (SVM) [8], naive bayes (NB) [9], and k-nearest neighbours (K-NN) [12] . Moreover, In the future, We would like to employ the LSTM (Long-Short Term Memory), Deep Learning method as well as NLP(Natural Language Processing) in the next work.

References

[1] Zhou, X. and Zafarani, R., 2018. Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315, 2.

[2] Zhou, X., Jain, A., Phoha, V.V. and Zafarani, R., 2020. Fake news early detection: A theory-driven model. Digital Threats: Research and Practice, 1(2), pp.1-25.

[3] Donepudi, P.K., 2019. Automation and machine learning in transforming the financial industry. Asian Business Review, 9(3), pp.129-138.

[4] Zhou, X. and Zafarani, R., 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53(5), pp.1-40.

[5] Della Vedova, M.L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M. and De Alfaro, L., 2018, May. Automatic online fake news detection combining content and social signals. In 2018 22nd conference of open innovations association (FRUCT) (pp. 272-279). IEEE.

[6] Bali, A.P.S., Fernandes, M., Choubey, S. and Goel, M., 2019. Comparative performance of machine learning algorithms for fake news detection. In Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3 (pp. 420-430). Springer Singapore.

[7] Drif, A., Hamida, Z.F. and Giordano, S., 2019. Fake news detection method based on text-features. France, International Academy, Research, and Industry Association (IARIA), pp.27-32.

[8] Zhang, H., Fan, Z., Zheng, J. and Liu, Q., 2012. An improving deception detection method in computer-mediated communication. Journal of Networks, 7(11), p.1811.

[9] Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M. and Whittaker, S., 2017. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. arXiv preprint arXiv:1709.05295.

[10] Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S. and De Alfaro, L., 2017. arXiv preprint arXiv:1704.07506.

[11] Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F. and Flammini, A., 2015. Computational fact checking from knowledge networks. PloS one, 10(6), p.e0128193.

[12] Mugdha, S.B.S., Ferdous, S.M. and Fahmin, A., 2020, December. Evaluating machine learning algorithms for bengali fake news detection. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[13] Zhang, J., Cui, L., Fu, Y. and Gouza, F.B., 2018. Fake news detection with deep diffusive network model. arXiv preprint arXiv:1805.08751.

[14] Ma, J., Gao, W. and Wong, K.F., 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

[15] Ruchansky, N., Seo, S. and Liu, Y., 2017, November. 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).

[16] Rashkin, H., Choi, E., Jang, J.Y., Volkova, S. and Choi, Y., 2017, September. 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 (pp. 2931-2937).

[17] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z. and TI-CNN, P.S.Y., 2018. Convolutional neural networks for fake news detection. arXiv preprint arXiv:1806.00749, 2(6).

[18] Liu, Y. and Wu, Y.F., 2018, April. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[19] Ranjan, A., 2018. Fake news detection using machine learning (Doctoral dissertation).

[20] Mitchell, T.M., 2006. The discipline of machine learning (Vol. 9). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.