Fake News Detection using Machine Learning

Rajat Maloo[*]

August 2, 2022

Abstract

Due to biased and unbiased media, we encounter fake news from different sources like newspapers, television and social media (Facebook, Twitter, Instagram). However, we are not able to identify which news is real or fake. Some people or groups intentionally spread fake news for their personal benefit by manipulating people's opinions towards society. Fake news can be in any format like audio, videos, image or text. But in this paper, we are focusing on the news which is in the form of text. We used various natural language processing(NLP) techniques like removing stopwords, tokenization, stemming, lemmatization and Machine learning algorithms like Logistic regression, Support Vector Machine, Random Forest, Naive Bayes, K-Nearest Neighbor to build a model to identify which news is legit and fake. To analyze these machine learning algorithms we are using different metrics like accuracy, precision, recall and F1-Score. From this, we will also get to know which algorithm performs good and which performs worst. In this research, our goal is to perform binary classification to predict whether given news is real or fake.

1 INTRODUCTION

In this report, we will try to detect fake news using machine learning algorithms. But before diving deep into this we need to understand the importance of this detection. News is basically information about events, especially current events. There are various sources of this information like radio, newspaper, TV, Magazines, online sources, social media platforms and so on. The purpose of news can be to inform people, purpose can be interest in a particular subject or environment. News can help us to be happy, healthy, safe and smart. It can create a better

---
[*]Department of Mathematical Sciences, University of Essex, CO4 3SQ Colchester, United Kindom (e-mail: rm21385@essex.ac.uk).

understanding or a life-changing opportunity for some. Sometimes the news can be very critical for an individual. It can be related to job, organization, weather, health or politics. Earlier TV and newspaper used to be the primary source of information and it used to come after a lot of research, evidence and only dedicated people can capture and publish the news.

In today's world, we get news at our fingertips with the digital transformation. It takes seconds to spread the news globally. Definitely, with the advancement in technology, there are a lot of advantages like sharing some critical information with a large audience irrespective of physical distance become easier. The basic fundamental still remains the same principle like trust, honesty, accuracy, accountability and truth are very important.

The current technological advancement cannot be the only reason for fake news. It used to exist earlier as well like the myths or an article misinterpreted and shared. But still, there were fewer sources and most of them used to be reputed. Today's problem is that there is so much information and there are lot many sources that the chances of it being false have increased a lot. Now we need to understand what exactly fake news is, so fake or false news is wrong information generated with or without any bad intention. The impact of spreading this type of news varies to a very large extent sometimes it can be harmless and sometimes it can be quite dangerous. It's very easy to create fake news and most of the time they appear to be real. For example, an article is designed in such a way that it copies the font and pattern from trusted sources. This type of news can be hosted on a webpage that looks very similar to a reputed news agency. There are many researches that proved that this kind of representation easily convenience the majority of people and then they are biased to forward with some intensity.

Now we need to understand why fake or false news spreads. There are a few fundamentals reason for spreading false news.

Political reason: There are many reports which claim that certain news is created and spread to influence political and geopolitical matters. The simplest example can be creating biased opinions for a candidate during the election.

Financial Profit: Some websites create money based on the number of clicks or redirection they have. Advertisement companies pay this website to show their products on this website.

Cybercrime: Sometimes news articles are created to make people click on a given link which can do some malicious attack or phishing attack or force people to make payment transitions.

Hatred: Sometimes the news is created to spread false information about individuals, groups or organizations, etc.

Not always fake news is created with bad intentions. Sometimes it's created for joke or entertainment. By reading all the above points one might wonder if there are so many issues with the online content one should not refer to them but the

point here is there are a lot more advantages than disadvantages. In the below paragraph few of the advantages are highlighted.

- Most of the information is available free of cost.

- Online information is easy to create and update whereas the physical news-paper will reach to a person the next morning only.

- Unlike TV or radio, one can skip the content in which he or she is not interested.

- There are multiple options available so you can read health-related articles from one source and sports related from another.

- The online article can be much more creative by adding videos and interactive images.

After understanding about the fake news, below are the section which are covered in this report. Section 2 presents a literature survey and related research done by various authors regarding fake news. Section 3 gives information related to the dataset, NLP, data pre-processing and all the machine learning algorithms which are used to predict fake news. In Section 4 we will analyze the outcome of the machine learning algorithm which we used in the previous section and in the last section i.e Section 5 we will conclude the paper and summarizes the models which give the best accuracy.

The main objective of this paper 'Fake news detection using Machine learning is to identify whether the news is real or not. We explore the recognition of fake news using various models and predict the accuracy of different models. In this paper, we inspect which model will give high accuracy and assorted the news into real or fake.

2   LITERATURE SURVEY

Mykhailo Granik et. al. in their paper [6] present an easy method for fake news detection using a naive Bayes algorithm. This method was executed and tested on a dataset of Facebook news feeds. They were collected from Facebook pages and mainstream political news pages. They obtain an accuracy of around 74%. In this article, they also discuss how to improve their model accuracy.
Marco L. Della Vedova et. al. [5] submit a paper Machine learning fake news detection method. In his paper he combines news content and social context features, After applying machine learning algorithms they got an accuracy of 78.8%.To verify their method they use Facebook Messenger Chatbot in this the chatbot accepts

Facebook URLs as input and analyzes whether the news is real or fake. They got an accuracy of 81.7%.

J. Zhang, B. Dong and P. S. Yu [11] propose a paper known as **FakeDetector: Effective Fake News Detection with Deep Diffusive Neural Network.** This paper's goal is to analyze methodologies and algorithms for detecting fake news articles, creators and subjects from the internet. This paper acknowledges the challenges generated by the unknown aspect of fake news and various connections between news articles, creators and subjects. This paper introduces an automatic fake news detector model, known as FAKEDETECTOR. It builds a deep diffusive network model to learn the characterization of news articles, creators and subjects. They use a real-world fake news dataset to find the accuracy of the FAKEDETECTOR model.

In his paper **Shahbazi, Zeinab and Byun, Yung-Cheo** [8] proposed a combination of machine learning and blockchain techniques to detect fake news. They also use NLP for data preparation. In their article, they show the combination of blockchain and machine learning to provide solutions and build an architecture towards hared news online. They applied reinforcement learning to make a strong decision-making architecture and combine it with the blockchain framework.

3   METHODOLOGY

3.1   Dataset description

The initial step in this classification problem is dataset collection, then data pre-processing, implementation features selection, then splitting of the dataset into training and testing and lastly applying the algorithm. For the project, we used Fake News dataset from Kaggle. This dataset contains 20,800 rows and 5 columns. [1] Each record was categorized using the following attributes:

- id: unique id for a news article

- title: title of a news article

- author: author of the news article

- text: the text of the news article

- label: a label that indicates the article is real or fake.1 represents the news article is fake and 0 represent the news article is real.

From below figure we can say that fake new and real news contain approximately 10,000 articles each.

Figure 1: News articles in count with fake and real category
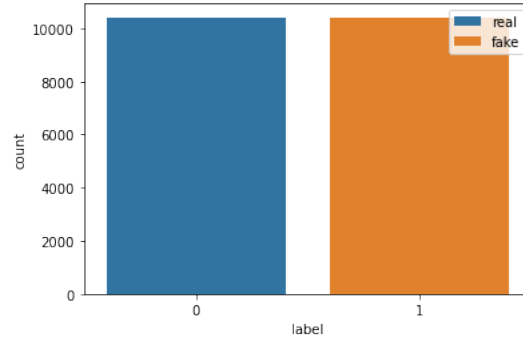
3.2 Data visualization

Figure 2 is the word cloud of reliable news and figure 3 is the word cloud of unreliable news. In Fig. 2, the frequently used words are Time, many, United, State, now, and Trump, which means that these words are equally imparted in the original news dataset. While in Fig. 3, the bolder and louder words are one, will, people, Trump, vote, and the U.S which means that these words equally contribute to the fake news dataset. The larger the term, the more often it appears in a dataset and the more important it is.

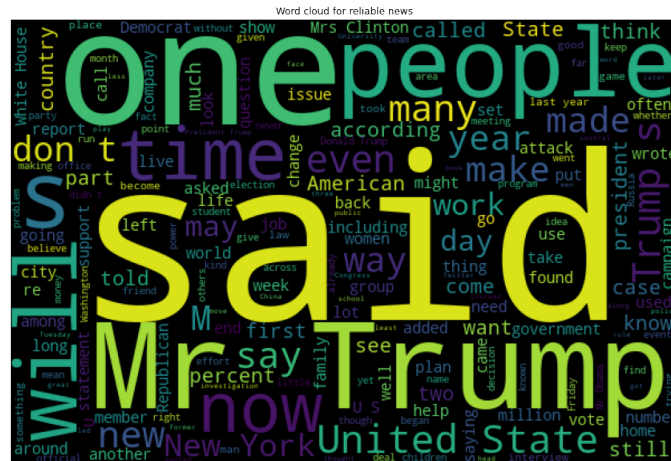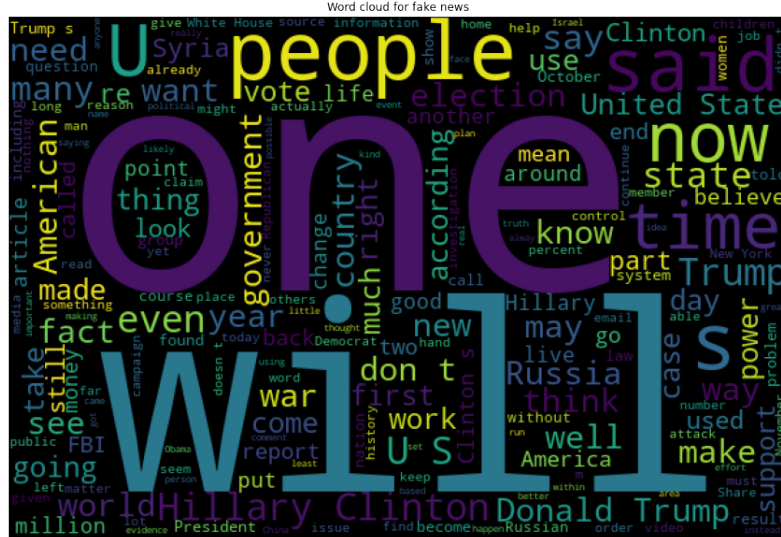Figure 2: Word cloud for real news.

5

Figure 3: Word cloud for fake news.

3.3 Data preprocessing

Before applying a machine learning algorithm to our dataset we have to do some basic data pre-processing.In the data pre-processing first, we will check if any null values are present in the dataset then we will sum up all null values in each category and replace all the null values with white space. Also, we are creating a new variable **content** in our dataset. In this variable, we saved the author's name and the title of the news article. In variable content, we are applying NLP techniques like removing special characters from the text because for analyzing text articles we are using only numbers and alphabets, converting all the words in lowercase, tokenization, removing stopwords from the article and applying stemming on words. After pre-processing the dataset now we have to convert text data into numerical form so that they can be used as an input for the training models. To train the different machine learning algorithms we have to split our dataset into two parts training and testing. We are splitting them into 80:20 ratios. 80 % for training the model and 20% for testing the model.

3.4 Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that gives the machine the ability to read, understand and derive meaning from human language. Every day humans interact with each other through the internet, social media transferring data to each other. This data is extremely useful in understanding human behavior and habits. NLP is related to the area of human-computer

interaction. [8] The data preparation process produces clean texts of news contents before being sent to a feature extraction process for converting to feature data vectors and saving the database. The basic preprocessing steps for NLP are segmentation, tokenization, removing stop words, stemming, speech tagging and named entity tagging. [4] Let's know some of them in detail below:

3.4.1 Segmentation

Segmentation is a process to break the paragraph into meaningful sentences. It is an initial and very important step to perform NLP.

3.4.2 Tokenization

After breaking down the article into sentences now we have to break down this sentence into words. These words are called tokens and the process is known as Tokenization. The case converter has been used to transform text data to lowercase or uppercase. In this paper, we have been converted into lowercase.

3.4.3 Stop-words removal

To make the learning faster we have to remove stop words from sentences. Stop words are words that do not add much value to the sentence, although they are often used to complete the sentence. These words are language-specific words, which do not convey information. There are around 400–450 stop-words in the English language. Some of the stop words are a, an, the, by, but, that, does, on, of, once, after, until, too, where, all, am, and, any, against, and so on.

3.4.4 Stemming

Stemming is the process in which different grammatical forms of a word like noun, verb, adjective, etc. have been converted into their root form. The objective of stemming is to obtain the basic forms of the words whose meanings are the same, but the word forms are different from each other. For example, the words, eating, eaten, eats can be converted to the word eat.

Some of the applications of Natural Language Processing are sentimental analysis, Chatbot, speech recognition, machine translation, spam filter, etc. In python, we use NLTK(Natural Language Toolkit) library for NLP and text analysis.

3.5 Machine Learning Algorithms

3.5.1 Logistic regression

Logistic regression is an example of a supervised machine learning algorithm. It is used to calculate or forecast the probability of a binary(0/1, yes/no, true/false) event occurring. In linear regression, the end result is continuous and can be any value. However, in the case of logistic regression, the predicted result is discrete and constricted to a finite number of values. Logistic regression is used to solve classification problems, and the most frequent use case is binary logistic regression, where the outcome is binary. [3] We use the logistic function or sigmoid function to calculate probability in logistic regression. The logistic function is a simple S-shaped curve used to convert data into a value between 0 and 1.

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)}}$$

- y : Output of logistic function.

- β_0 : y-intercept.

- β_1, β_2: Slope of dependent variable.

- X_1, X_2: Independent variable.

3.5.2 Support Vector Machine

Support Vector Machine is also supervised learning that can be used for both classification and regression problems. However, it is mostly used in classification problems. The algorithm creates a line or a hyperplane which split the data into classes. The dimension of the hyperplane depends upon the number of features. If the number of input features is three, then the hyperplane becomes a 2-D plane. [2]

3.5.3 Random Forest

Random Forest is a popular and commonly used machine learning algorithm that belongs to supervised learning. Random forest constructs multiple decision trees and joins them together to get a more accurate and stable prediction. Rather of depend on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, it predicts the end result. It uses bagging and feature randomness when constructing each tree to try to create a forest of trees whose prediction by the group is more accurate than that of any single tree. A random forest can classify a big dataset with high accuracy. [10]

3.5.4 Naive Bayes

Naive Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem and used for solving classification problems. Based on training data, Naive Bayes approximate the posterior probability P(A—B) of each class, A, of a given object, B. In Naive Bayes, the occurrence of one feature is independent of another feature. [9] It is mostly used in text classification that includes a high-dimensional dataset. [7]

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)}.$$

Where,

- P(A) : Probability of hypothesis before observing the evidence.

- P(B): Probability of Evidence.

- P(A—B) Probability of hypothesis A on the observed event B.

- P(B—A): Probability of the evidence given that the probability of a hypothesis is true.

3.5.5 K-Nearest Neighbors (KNN)

It is one of the easy-to-implement Machine Learning algorithms based on Supervised Learning. In K-NN we supply training data to the model and let it decide to which specific cluster a data point belongs. KNN model calculates the distance of a new data point to its neighbors, and the value of K estimates the majority of its neighbors' votes; if the value of K is 1, then the new data point is assigned to a class that has the nearest distance. The mathematical formula to calculate the distance between two points is as follows

$$d\,(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Where, p and q are two points.

3.6 Performance Metrics

To analyze the performance of algorithms for fake news detection, we used the most frequently used metrics in classification tasks i.e. confusion matrix A confusion matrix is a tabular form of a classification model performance on the test

dataset for which the true values are known. In confusion metrics, a number of correct and incorrect predictions are compiled with count values and broken down by each class. In Confusion metrics row represent actual labels and the column represent a predicted table. [9] It consists of four parameters: True Positive (TP): when predicted fake news pieces are actually classified as fake news. True Negative (TN): when predicted true news pieces are actually classified as true news. False Negative (FN): when predicted true news pieces are actually classified as fake news. False Positive (FP): when predicted fake news pieces are actually classified as true news.

Table 1: Confusion matrix

	Predicted True	Predicted False
Actual True	True positives(TP)	False negatives(FN)
Actual False	False positives(FP)	True negatives(TN)

3.6.1 Accuracy

With the help of confusion metrics we can calculate the accuracy of any model. It is define as ratio of number of correct predictions and total predictions. It measures the similarity between predicted fake news and real fake news. Below formula is used to calculate accuracy of any model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

In python we will use accuracy score function to calculate the accuracy of our model.

3.6.2 Precision

It is the ratio of correctly predicted true to a total number of predicted true values. In this case, It is the ratio of total fake news to all the fake news predicted. Below shows the formula to calculate the precision.

$$Precision = \frac{TP}{TP + FP}.$$

3.6.3 Recall

It is the ratio of a total number of positively classified to a total number of positive samples. In this project, it represents the number of fake news predicted as true out of the total number of true fake news.

$$Recall = \frac{TP}{TP + FN}.$$

3.6.4 F1-Score

If two models have low precision and high recall or vice versa, it is tough to compare these models. So, for this purpose, we can use the F1 score. It calculates the harmonic mean of recall and precision. The F1-score is maximum if the recall is equal to the precision. The good value of the F1 score would be 1 and the worst would be 0. It can be calculated using the below formula:

$$F1 = \frac{2 * precision * recall}{precision + recall}.$$

4 Analysis of results

The analysis results of the datasets using the five algorithms have been represented using the confusion matrix. The five algorithms used for the detection are as follows:
1 Logistic Regression.
2 Random Forest.
3 Support Vector Machine(SVM).
4 Navies Baised.
5 K-Nearest Neighbors (KNN).
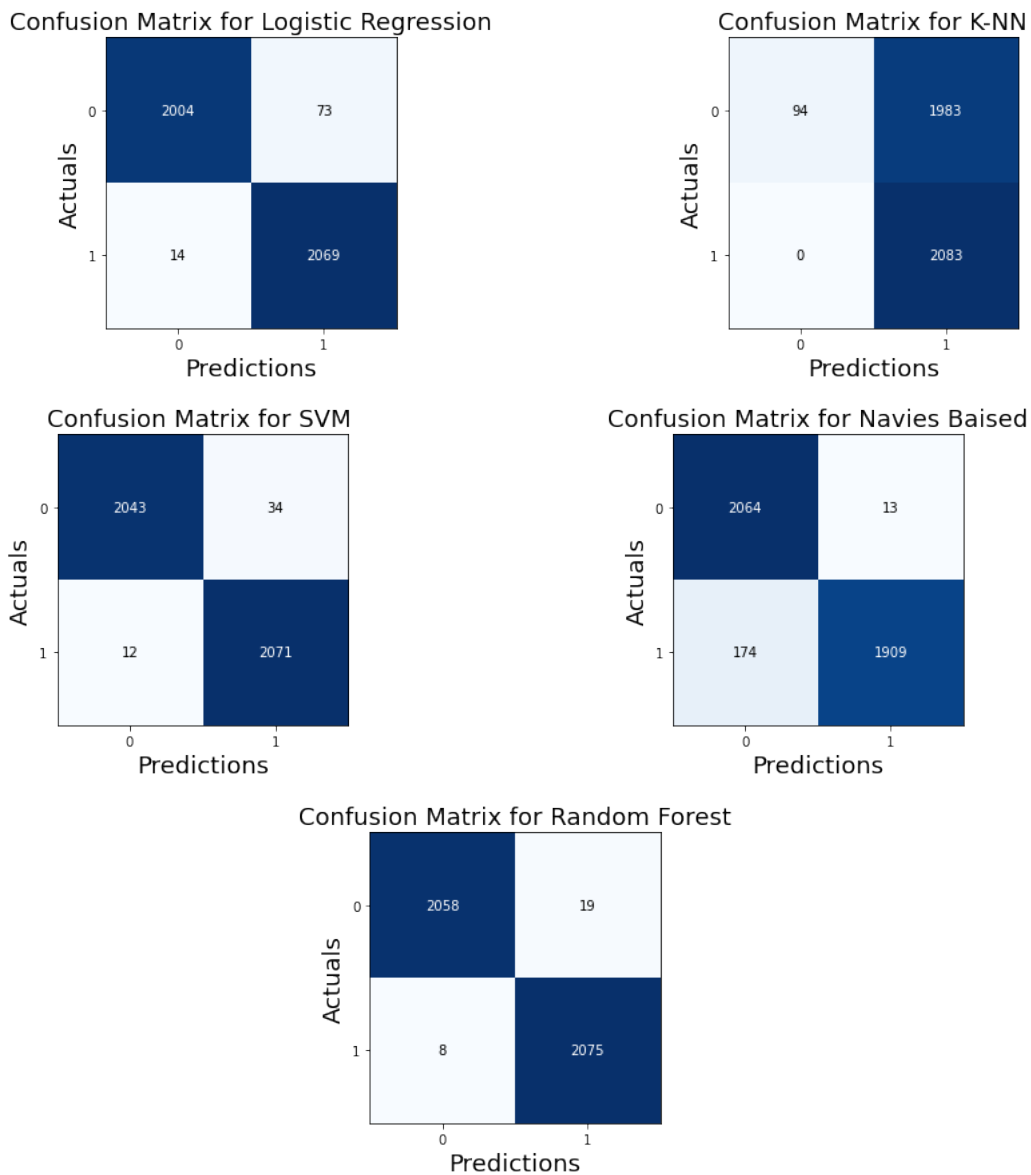   The Confusion Matrix for all the algorithms is shown below :

Figure 4: Confusion Matrix

From the above plot, we can find the accuracy of each model. Accuracy for logistic regression is 97.90 %, accuracy for the random forest is 99.35%, accuracy for support vector machine (SVM) is 98.89%, accuracy for Navies baised is 95.5% and accuracy for K-Nearest Neighbors (KNN) is 52.3%. After finding accuracy for each model we can say that all models got excellent accuracy except KNN. Random forest gives the highest accuracy which is 99.3%. The below tables show the classification report i.e. precision, recall, and F1 score:

Table 2: Classification Report for Logistic regression

Type of News	Precision	Recall	F1-score
FAKE	0.99	0.97	0.98
Real	0.96	0.99	0.98
Total	0.98	0.98	0.98

Table 3: Classification Report for Random Forest

Type of News	Precision	Recall	F1-score
FAKE	1.00	0.99	0.99
Real	0.99	1.00	0.99
Total	0.99	0.99	0.99

Table 4: Classification Report for SVM

Type of News	Precision	Recall	F1-score
FAKE	0.99	0.98	0.99
Real	0.98	0.99	0.99
Total	0.99	0.99	0.99

Table 5: Classification Report for Navies baised

Type of News	Precision	Recall	F1-score
FAKE	0.92	0.99	0.95
Real	0.99	0.92	0.96
Total	0.96	0.96	0.95

Table 6: Classification Report for KNN

Type of News	Precision	Recall	F1-score
FAKE	1.00	0.51	0.68
Real	0.05	1.00	0.09
Total	0.52	0.76	0.38

5 Conclusion

There are many ways to detect fake news which are NLP, machine learning algorithms and neural networks. Depending upon our requirement we choose the best suitable option to detect fake news. In this report, we are using natural language processing for preprocessing the dataset and different machine learning algorithms like Logistic regression, random forest, SVM, Navies Baised and KNN. After applying all these algorithms to our preprocessed dataset we got good accuracy for all the models except KNN. Also in this project, we concentrated on text news only but in the future, we can build an algorithm that helps us to detect audio or video news from different sources.

It's very important to identify and report fake news. Big tech companies like Facebook and Google are given the option to report fake news.

References

[1] https://www.kaggle.com/competitions/fake-news/data, February 2018.

[2] https://www.geeksforgeeks.org/support-vector-machine-algorithm/, January 2021.

[3] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake news detection using machine learning ensemble methods. *Complexity*, 2020, 2020.

[4] Tavishee Chauhan and Hemant Palivela. Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, 1(2):100051, 2021.

[5] Marco L. Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, and Luca de Alfaro. Automatic online fake news detection combining content and social signals. *FRUCT*, pages 272–279, 2018.

[6] Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. *IEEE Access*, pages 900–903, 2017.

[7] Phayung Meesad. Thai fake news detection based on information retrieval, natural language processing and machine learning. *SN Computer Science*, 2(6):1–17, 2021.

[8] Zeinab Shahbazi and Yung-Cheol Byun. Fake media detection based on natural language processing and blockchain approaches. *IEEE Access*, 9:128442–128453, 2021.

[9] Jasmine Shaikh and Rupali Patil. Fake news detection using machine learning. pages 1–5, 2020.

[10] Uma Sharma, Sidarth Saran, and Shankar M Patil. Fake news detection using machine learning algorithms. *International Journal of Creative Research Thoughts (IJCRT)*, 8(6):509–518, 2020.

[11] Jiawei Zhang, Bowen Dong, and Philip S. Yu. Fakedetector: Effective fake news detection with deep diffusive neural network. pages 1826–1829, 2020.