RESEARCH ARTICLE

Effect of Corpora on Classification of Fake News using Naive Bayes Classifier

Farzana Islam Adiba[1], Tahmina Islam[1], M Shamim Kaiser[2], Mufti Mahmud[3], Muhammad Arifur Rahman[4*]

[1]*Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh*
[2]*Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh*
[3]*Department of Computing & Technology, Nottingham Trent University, Nottingham, UK*
[4]*Department of Physics, Jahangirnagar University, Dhaka, Bangladesh*

Abstract

At the present world, one of the main sources of the news is an online platform like different websites and social media i.e. Facebook, Twitter, LinkedIn, YouTube, Instagram and so on. However, due to the lack of proper knowledge or deliberate activity of some cunning people, fake news is spreading more than ever. People in general are struggling to filter which news to trust and which one to discard. Even the sly people take advantage of the situation by spreading false news and misleading the people. Natural Language Processing (NLP), one of the major branches of Machine Learning, the wealth of research is remarkable. However, new challenges underpinning this development. Here in this work, Naive Bayes Classifier, a Bayesian approach of Machine Learning algorithm has applied to identify the fake news. We showed, besides the algorithms, how the wealth of corpora can assist to improve the performance. The dataset collected from an open-source, has been used to classify whether the news is authenticated or not. Initially, we achieved classification accuracy about 87% which is higher than previously reported accuracy and then 92% by the same Naive Bayes Algorithm with enriched corpora.

Key Words: *Fake news; Social media; Machine learning; Natural language processing; Naive bayes*

***Corresponding Author**: Muhammad Arifur Rahman, Associate Professor, Department of Physics, Jahangirnagar University, Dhaka, Bangladesh, Tel: +88 017 2642 8888; E-mail: arif@juniv.edu*

***Received Date**: August 10, 2020, **Accepted Date**: September 26, 2020, **Published Date**: October 30, 2020*

***Citation**: Adiba FI, Islam T, Kaiser MS, et al. Effect of Corpora on Classification of Fake News using Naive Bayes Classifier. Int J Auto AI Mach Learn. 2020;1(1):80-92.*

1. Introduction

In this era of the Internet of Things (IoT), one of the main sources of the news is online platform like different websites and social media. We are depending significantly on this online news while we are on the go and busy with other daily usual works. "Fake news" which is not a technical issue in the media, is deliberately human activity and also it isn't new in our lives. However, the combination of truth and falsehood that creates chaos in the society. There is so much information online that it is becoming impossible to decipher the true from the false. A piece of false news very quickly spread from society to society and from country to country through online news. Some misleading news which has a loss of credibility through social media. As the fraud people use flashy headlines so that people get easily attracted and click on it. Thus, this leads to the problem of fake news [1]. The source of false news can be related to financial or political. News of deception that covers the truth, so that there is a lack of truth. In the present context, the world is battling with Corona virus disease 2019 (COVID-19) and fake news is spreading online about at an alarming rate. The most popular social media sites such as Google Plus, Twitter, Facebook, Instagram etc. which are the biggest sources of spreading fake news. In the Figure 1 the knowledge graph is used to explain how any hoax or false news expands from one country to another via social media.

Any kind of incorrect information or fake rumors can be defined as False news. The news or information which spread through social media users but is not verified or justified from any reliable source can be defined as false rumors [2]. Here rumors are the form of statements that are exchanged from person-to-person and used as a weapon against any person, company, or even a country to create chaos or harass anyone, deliberately. Nowadays rumors are spread through social media from one user to another. Thus, any fake news is distributed widely. For example, in Figure 1 a false news that created a stir in social media has been explained. There was an image that got viral in social media that the president of Russia Vladimir Putin left lions in the street so that people do not come out during the corona virus. But in reality, the picture was clicked from Johannesburg, South Africa in 2016. Another example, in Bangladesh there is a rumor spread during corona virus that *Centella Asiatica* (locally known as- Thankuni pata) can be curative for the Corona virus. Though it is scientifically reported that the herbal extract of *Centella Asiatica* is good for kidney and liver disease, improve brain health, Asthma treatment, skincare, etc. But there is no verification that *Centella Asiatica* has any healing capability or relation to resisting COVID-19. So, it is false news that distracted common people easily.

Nowadays, the field of discourse, text research technologies is based on Natural Language Processing. 'Human-like language processing' reveals that NLP is considered a discipline within Artificial Intelligence (AI) [3]. NLP lies in different sectors such as text processing, sentence analyzing, linguistic analysis, information extraction, artificial intelligence, robotics, speech recognition, etc. The use of NLP can retrieve spoofing news from various online sectors.

The objective of our paper is to demonstrate the effect of the size of corpora while classifying real and false news using NLP (Natural Language Processing). To investigate better performance, we shall use a Bayesian approach of natural language processing Naive Bayes classifier to accurately classify the REAL and/or FAKE news. For the best performance of our knowledge, we consider a proposal to identify how the corpus size affects the classification result of true and false news from

the train and test data collected from the Kaggle dataset. We are discussed about the existing dataset. Our goal is to collect all data and identify the best accuracy from the text corpus. As a text corpus can be defined as the collection of texts in the language and corpora is the plural form of the corpus [4]. For tokenization text corpus is used to achieve tokens from the corpus.

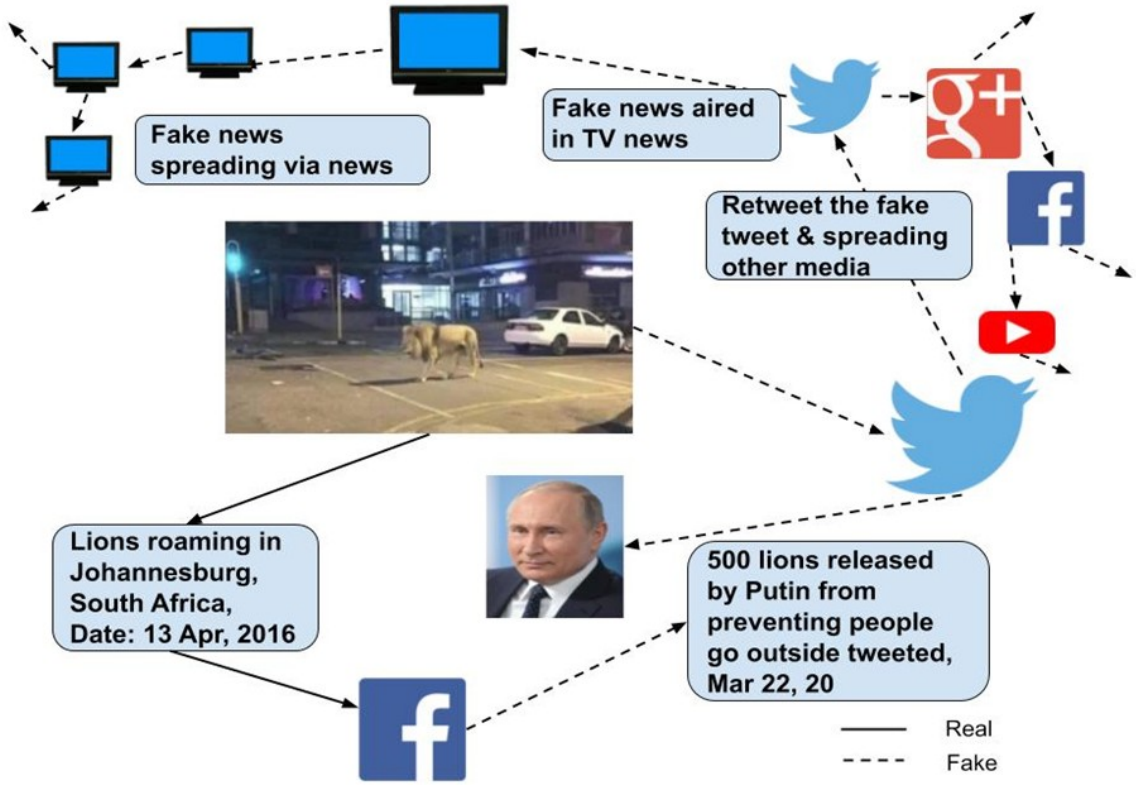

Figure 1: *Knowledge graph showing how fake news spread via media as this fake news was spread through a tweet, and then broadcast in news channel in different countries.*

To improve our performance, we compared our results with other reported papers and get maximum accuracy using TF-IDF (TF represents 'Term Frequency' and IDF represents 'Inverse Document Frequency' for automated text analysis) and Counter Vector.

Detecting fake news has become a challenging topic nowadays. This is because normal users aren't aware of fake news; unknowingly they engage in this news and spread it through social networks. The paper is organized in a different section as follows. In section 2 we describe the literature review. In section 3 our goal is to describe the format of collecting data, use the NLTK (Natural Language Toolkit) toolkit, describe the mathematical model of the Naive Bayes classifier, etc. In section 4 we discuss the dataset which is collected from the open-source platform. In Section 5 the algorithm is applied to compare to both datasets to show the results.

2. Literature Review

As initially Fake news is published through social media to give a huge sharing among many users. Terry Taylor et.al. developed a fake news detector with the help of Text blob, Natural Language, and SciPy Toolkits [5]. They took the document and each paragraph of the document

was counted and tokenized. A-score algorithm was used to label quotes as to either real or fake based on the results of the machine learning classification. The tool identified 96% of the quotes. For understanding the overall result, they showed a performance table where they labeled the results as True Positive, False Positive, True Negative, False Negative and F-score.

Sherry Girgis et al. [6] proposed a classifier to predict fake news using RNN (Recurrent Neural Network) technique models as vanilla, GRU (Gated Recurrent Unit) and LSTMs (long short-term memories). They used LIAR dataset which includes 12,836 short statements labeled for truthfulness, subject, context, venue, speaker, state, party, and prior history. For preparing the dataset they split each sentence and erase unnecessary words. Finally, they established three experiments as Vanilla, GRU and LSTM and compared the results according to their accuracy. The more improved result was shown for GRU than Vanilla and LSTM.

Fake news is published through social media to spread social disturbance and earn money through clickbait. Monther Aldwairi et al. [7] proposed a logistic regression to identify fake posts. The proposed tool would define fake news on the basis of some characteristics. The syntactical structure analyzed to separate the words containing misleading effects on the users. The outstretched uses of punctuation and exclamatory marks will be identified as repellent and non-clickable news. Weka is a collection of machine learning algorithms for data mining tasks [8]. Weka classifiers were used such as Logistic, NaiveBayes, BayesNet etc and showed a comparison among the classifiers. When the content of the social web page is filled with fake news then it can also hide the main information in a different sector. They implemented with Weka machine learning and, then rank these data according to the algorithms. Using numerous classifiers to choose the best data from the dataset. The classifiers are compared based on Precision, Recall, F-Measure and ROC (Receiver Operating Characteristic).

Social media is the most popular platform for acquiring news, so false or spurious news can be spread spontaneously. Ning Xin Nyow et al. [9] categorized fake news as Fake news on traditional media and Fake news on social media. A comparison among various existing fake news datasets is demonstrated. FakeNewsNet was selected as it includes multi-dimension information from news content from political and entertainment sources. FakeNewsNet repository containing 4 attributes as id, URL, title, tweet ids which can be obtained from tweet properties. The results were analyzed from both the news aspects and tweet-aspects.

S. I. Manzoor et al. [10] analyzed readers' psychological factors that persuade naive users. The authors explained with a Facebook post which includes an image to believe on that news. Three major forms of data were discussed as text, multimedia (images, video, etc.), and hyperlinks containing links of various sites. Various types of fake news were also summarized as Visual-based, User-based, Knowledge-based, etc.

Online Fake news can be identified in machine learning such as supervised learning, unsupervised learning, etc. In Shuo Yang et al. [11] who are explained their paper using an unsupervised learning framework that exploits a probabilistic graphical model to model the truth of the news. Gibbs sampling algorithm is used to solve the problem in their paper. Using these algorithms to detect verified and unverified users using truth estimation. In verified users who are already reliable but, unverified users must have a piece of false news. So, LIAR and BuzzFeed two public datasets issued to detect the unverified users who are scattered the fake news. They categorized

five methods to identify accuracy, precision, recall, F1-score. Two datasets are compared to find the outcome of the best method.

That researcher considered identifying fake news on top positions on social media. Kelly Stahl [1] used to detect false information using Linguistic Cue (text-based communication) and Network Analysis (content-based) approaches. Three methods Naive Bayes Classifier, Support Vector Machines, and Semantic Analysis are used to detect false news on social media. Firstly, they demerging both the Naive Bayes classifier and support vector machine are more relevant to detect the fake news. But they found that Naive Bayes classifier which has some drawback as textual processing. So, semantic analysis is also merged to solve the Naive Bayes problem.

Online social media are saturated with so much news that can be real or fake. Vivek Singh et al. [12] who tried to detect automatically unverified news from the "Kaggle Fake News" dataset. This paper used LIWC (Linguistic Analysis and Word Count) package to obtain linguistic features and also normalized the data.

This paper considers the logistic regression, support vector machine, random forest, decision tree, k-NN (k-Nearest Neighbors) classifier, etc. and focused on the test set [12]. Finally, they found the accuracy by using the SVM (support vector machine).

Nowadays, people are most interested in the internet. So, much enormous news is also disclosed on the internet. Samir Bajaj [13] who used a deep learning process and pre-trained 300-dimensional Glove embeddings. The model is used such as logistic regression, Two-layer Feedforward Neural Network, Recurrent Neural Network (RNN), Long Short-Term Memories (LSTM), Gated Recurrent Units, Bidirectional RNN with LSTMs, Convolution Neural Network with Max Pooling, etc. After classifying these models, the result comes with the best precision, recall, and F1-measure.

Other authors have noted fake news as different perspectives. From the social network, misleading information can affect people's normal life. Mykhailo Granik et al. proposed a simple approach to detect fake news using the Naive Bayes Classifier [14]. They tried to find similarities between spam messages and fake news. As Naive Bayes is a simple probabilistic classifier, so they were used to detect spam messages using text classification. The dataset they used was collected from BuzzFeed News. The news articles used were collected from Facebook API (Application Programming Interface). For creating a mathematical model Bayes theorem was applied.

3. Methodology

Collecting data: Collecting real-time data is the first step for developing this system. So, data must be collected very carefully. We have selected open-source data and apply an algorithm on it. Two open-source datasets on "Fake and Real news" is selected.

Preparing NLP environment: There are different NLP tools are available. But among them, Natural Language Toolkit (NLTK) is the most popular tools recently used. Natural Language Toolkit (NLTK) is a library for Natural Language Processing (NLP) which is written in the python programming language. We used NLTK Toolkit which can support tokenization, stemming, lemmatization etc. various functionalities.

Data Processing: Fake news datasets should be prepared for the model evaluation and analyzing the corpora effect for different datasets. For data processing, some steps should be followed as Tokenization, Stemming, TfIdfVectorization, CountVectorization [15]. Tokenization is the technique of chopping a character sequence into pieces or tokens. So, the paragraph will be chopped into sentences, then sentences will be chopped into words with this process. Stemming is the process of adding affix with any root word. Porter stemmer is better for handling root words [15]. For Data Cleaning stop words will be removed (Figure 2).

Feature Extraction: For word processing Bag of Words is a common technique to transform the dataset into fixed-length vectors. This technique is generally applied to determine the occurrence of the words in the entire dataset. For feature extraction generally, two vectors are used.

- CountVectorizer
- TfIdfVectorizer

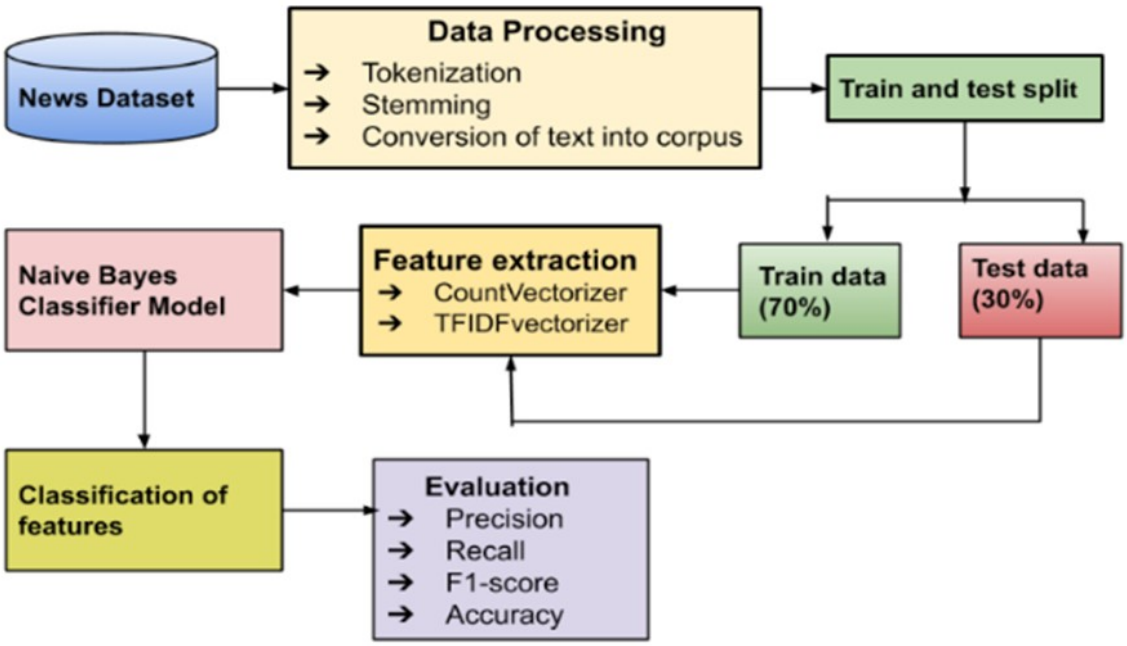

Figure 2: *System design for the development of Naive Bayes classifier showing the effect of corpora for detecting fake news.*

CountVectorizer extracts all the unique words and creates a list of them. Then it counts how many times they occurred in the entire text corpus. It will be used for presenting the frequency of each word in the entire text. Another feature extraction technique we will be used TfIdfVectorizer. Here TF represents 'Term Frequency' and IDF represents 'Inverse Document Frequency'. This technique is used to evaluate the frequency of a word in each sentence. TF technique is used to evaluate the frequency of a word in each sentence. But it may sometimes count less important word many times. It will decrease the efficiency of the model. For solving this problem, IDF is used. It can analyze less and higher relevant words. Thus, the average weight is increased using this vectorizer. The equation 1 for TF-IDF is given below [16] –

$$W(d,t)=TF(d,t)\times\log\frac{N}{df(t)} \quad (1)$$

Here N is used to present the number of documents, and df(t) is used to present the number of documents comprising the term *t* in the documents or corpus. After generating the text corpus, the vectors are generated for n number of sequences called n-gram. In this work, the range is (1,3), which is from unigram to tri-gram. Max_feature is another parameter which works with the top max_features according to the term frequency which has been created in the overall corpus. Here for Dataset-1 max_features=5000 across the corpus for 6256 unique values and for Dataset-2 max_features=15000 across the corpus for 20387 values.

After preparing text then, the dataset will be prepared for training and testing. We have used 70% data for training and 30% for testing.

Mathematical Model for Naive Bayes Classifier: Bayes Theorem is one of the most common classifiers used in the text classification. If *X* is a test instance with *d* different features having values (x_1, ..., x_d). The Bayes rule is derived as following [17] described in equation 2:

$$P\left(Y\left(T\right)=i\middle|x_1.....x_d\right)=P\left(Y\left(T\right)=i\right).\frac{P\left(x_1.....x_d|Y\left(T\right)=i\right)}{P\left(x_1.....x_d\right)} \quad (2)$$

For fake data classification we can calculate an expression for P (label | features). P (label | features) is equal to the probability that an input has a particular label and the specified set of features, divided by the probability that it has the specified set of features [15] described in equation 3:

$$P\left(label\middle|features\right)=\frac{P\left(features,label\right)}{P\left(features\right)} \quad (3)$$

Evaluation Metrics: For evaluation of the performance confusion matrix is used. It is suitable for visualization of the performance of an algorithm. From the confusion matrix, some metrics can be derived. Some common measures are Recall, Precision, F1-score and accuracy. Here Precision issued to the number of true positives already have found and else already described. Recall describes the accuracy among true positives and false negatives. Accuracy describes the number of true samples which have been covered from the total samples. One of the most important metric F1-score issued to display the overall scenario about the performance of the classification [18]. The equations for the metrics are describing below [16] in equation 4, 5, 6 and 7:

$$Recall=\frac{TP}{TP+FN} \quad (4)$$

$$Precision=\frac{TP}{TP+FP} \quad (5)$$

$$F1-score=\frac{2\times Recall\times Precision}{Recall+Precision} \quad (6)$$

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

In the equation 4, 5 and 7, TP is for True Positive, FN is for False Negative, TN is for True Negative and FP for False Positive. The equation 6 uses the value achieved from equation 4 and 5. The confusion matrix terms can be denominated as follows:

TPR (True Positive) or Sensitivity: This is the occurrence of the true positive value when both actual and predicted results are correct simultaneously. It can also be defined as sensitivity.

TNR (True Negative Rate) or Specificity: This is the occurrence of the true negative value when both actual and predicted results are incorrect simultaneously.

FP (False Positive): If the actual results are incorrect and the predicted results are correct then it is the occurrence of false-negative values.

FN (False Negative): If the actual results are correct and the predicted results are incorrect then it is the occurrence of false-negative values.

4. Data Set

In this work, a dataset is collected from Kaggle which is the largest community of data scientist [19]. So, we have collected a dataset named "REAL and FAKE news dataset" [20] which is the dataset-1. Figure 3 show a sample of our used corpora (dataset). This dataset has a shape of 7796 rows having 4 columns or attributes. It has four attributes which are index, title, text and label. The title index consists of 6256 unique values and in the text index, there are 6060 unique values. Here label column is for the decision whether the news is false or true which represents using 'REAL' and 'FAKE'. This column consists of about 50% Fake and 50% Real news.

5316	Newly Approved GM Potatoes Have Potential to Silence Human Genes	Late last week, the US Department of Agriculture (USDA) approved two new strains of genetically engi...	FAKE
1524	Accord reached after Sanders sues the DNC over suspended access to critical voter list	The presidential campaign of Sen. Bernie Sanders of Vermont filed a lawsuit against the Democratic N...	REAL
2974	Why the latest Patriot Act reform won't be enough to	Recent debates over US government spying have focused on one	REAL

Figure 3: *A sample of our used corpora (dataset). The corpora contain four columns. Column 1represent by an integer shows the ID, Column 2 shows the heading of the news, Column 3 contains the full news (showed partially here) and Column 4 represents the label FAKE or REAL.*

Another dataset will be used, which is also collected from Kaggle [21]. This dataset consists of 20387 unique values in five columns and it is dataset-2. It has five attributes which are id, title, author, text and label. The label is for the Fake and real news, where '0' is for fake and '1' is for real data.

5. Results and Analysis

A confusion matrix is used to describe the classification summary in a table. It can be called as the summary of the predicted and actual data. Finally, a confusion matrix is generated by and analyzing confusion matrix the True Positive rate (TPR), True Negative rate (TNR), False Positive rate (FPR) False Negative Rate (FNR) are specified.

For the dataset-1 both TfIdfVectorizer and CountVectorizer confusion matrix is generated which shows the accuracy about 0.874 and 0.872 respectively. But for the second dataset, the output for the two features is 0.920 and 0.919 respectively. From the accuracy gained from the dataset, we can observe the result is improved for the larger dataset.

In Figure 4 and Figure 5 a confusion matrix of the fake and real news has been shown for TfIdfVectorizer for the two datasets respectively. For the explanation of the confusion matrix Table 1 is created. The Table 2 is showing the classification outputs. In Figure 6 and Figure 7 shows the comparison for our dataset with other papers in the perspective to Accuracy of the performance.

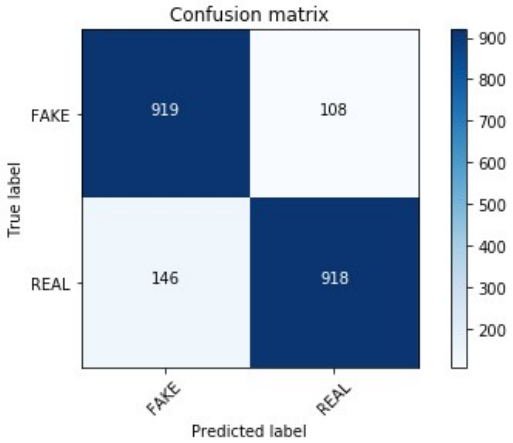

Figure 4: *Confusion Matrix for TfIdfVectorizer for dataset-1 with 7796 rows having 4 columns or attributes. Here, x−axis represents the Predicted level and the y−axis represents the True level*

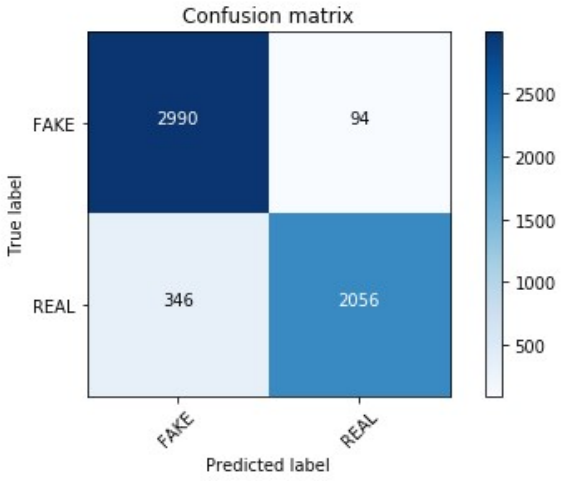

Figure 5: *Confusion Matrix for TfIdfVectorizer for dataset-2 having 20387 unique values with training and test dataset. Here, x−axis represents the Predicted level and the y−axis represents the True level.*

Table 1:
Table shows the Naive Bayes classifier performance for various evaluation metrics for the features TfIdfVectorizer and CountVectorizer for two datasets.

Classification Types	Dataset-1		Dataset-2	
	TfIdfVector	CountVector	TfIdfVector	CountVector
True Positive	0.895	0.866	0.969	0.948
True Negative	0.86	0.877	0.856	0.880
False Positive	0.142	0.127	0.112	0.093
False Negative	0.101	0.129	0.039	0.066

Table 2:
Table shows classification metrics for the features TfIdfVectorizer and CountVectorizer for two different datasets for the Fake and Real data.

Evaluation Metrics	Dataset-1				Dataset-2			
	TfIdfVector		CountVector		TfIdfVector		CountVector	
Types	Fake	Real	Fake	Real	Fake	Real	Fake	Real
Precision	0.86	0.89	0.87	0.87	0.90	0.96	0.91	0.93
Recall	0.89	0.86	0.87	0.87	0.97	0.86	0.95	0.88
F1-Score	0.87	0.87	0.87	0.87	0.93	0.90	0.93	0.90
Accuracy	0.874		0.872		0.920		0.919	

In this paper, Mykhailo Granik et al. [14] proposed some method how to improve the output of the classification. Since they have used 2000 data, which is one of the shortfalls of their work. They also suggested removing and using stemming for identifying similar words efficiently. So, we have used the stemming method and removed stop words. Here the total accuracy of our dataset is about 87% and 92%. With comparison to the paper of Mykhailo Granik et al. which is much better. So, we used a larger dataset compared to them and applied stemming method for better output. Another work by Terry Traylor et al. [5] represented the Supervised Learning estimator which shows the precision result for fake news about 63.3%. Our work showed better precision for fake news compared to them for both Dataset-1(87% and 86%) and Dataset-2(91% and 90%) in case of Tftdf and count feature vectors. They also used 30% of the data for testing purpose. The overall accuracy of their classification is 69%. But our overall classification accuracy is about 87% for dataset-1 and for dataset-2 accuracy is 92%.

Pratiwi et al. [22] applied Naive Bayes for 70% training set and 30% testing set. They got accuracy about 78.6% which is less than ours. They found precision value for fake news about 67.1%, but we achieved precision value for fake news is about 86% and for Dataset-2 the fake news precision is about 91%. Figure 6 shows the comparison among the accuracy of different papers.

To calculate the relevant result is the best part to find the highest accuracy using the TF-IDF and Counter Vector from the dataset. The input of the Chile Earthquake Dataset 2010 [23] was taken to achieve the highest accuracy. According to Mahir et al. [23], their performance of accuracy for count vector and TF-IDF vector was 84.56% and 89.06%. But in our working for Dataset-2 performance of TF-IDF and counter vector is 92% and 91.9% which is better result comparison of their performance because of having 70% data for training and 30% for testing. Where Mahir et al. used the dataset into 60% train data and 20% test data including 20% validation LSTM, RNN model [23].

As can be seen in other works that Poddar et al. [24] comparison the accuracy with different classifiers where Naive Bayes classifiers enumerate the CountVectorizer as 86.3% and TF-IDF as 85.4%. We get a better result for CountVectorizer as 87.2% from dataset- 1 and best performance are obtained from dataset-2 as 92%. In Figure 7 the difference in the accuracy for TF-IDF and Count Vector is displayed where the x−axis shows different paper authors name and the y−axis is for the accuracy perspective to the vectors.

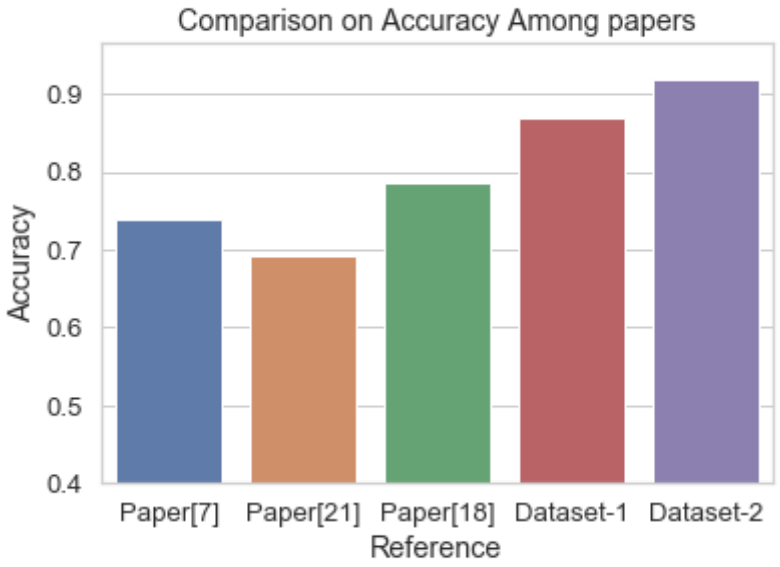

Figure 6: *Comparison of overall Accuracy among different papers where x−axis expresses the references Granik and Mesyura, 2017 [14], Traylor et al. 2019 [5], Pratiwi et al. 2017 [22] and our used Dataset-1 and Dataset-2.y−axis represents the accuracy.*

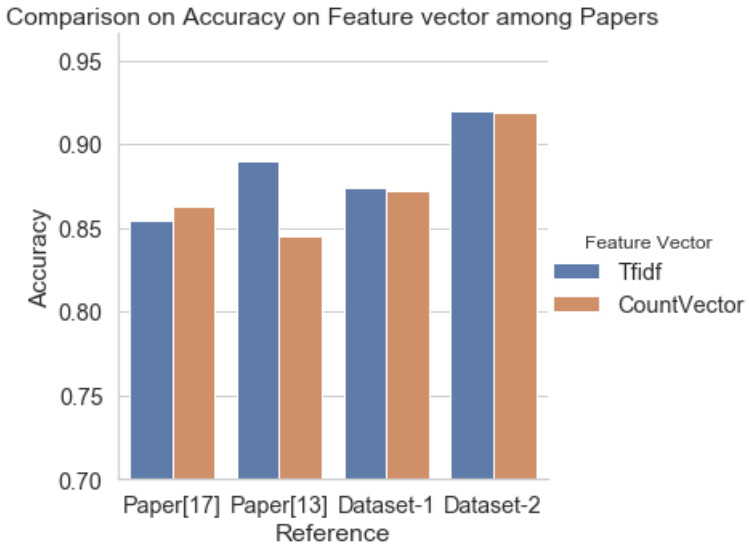

Figure 7: *Comparison of accuracy subject to Features as TF-IDF and CountVectorizer among various papers where x-axis shows references Poddar et al. 2019 [24], Mahir et al. 2019 [23] and for our two datasets and y-axis shows the accuracy.*

6. Conclusion

Naive Bayes is one of the prominent Machine Learning Algorithms to solve text classification problem. This algorithm is suitable when the dimensionality of the inputs is high and still the method is relatively simple. For the classification of the FAKE news from the REAL news we used this algorithm and showed it can improve the accuracy significantly for identifying FAKE news with the increased corpora. Initially, we acquired 87% accuracy with moderate sized corpora. As we later found, if we use enriched corpora, it can reach up to 92% of accuracy which is higher than any reported paper so far. From the observation of the two different dataset it is perceptible that by increasing the number of data had a significant impact in case of accuracy.

Conflict of Interest: The authors declare that there is no conflict of interest.

References

1 K. Stahl. Fake news detection in social media. California State University Stanislaus, 2018.

2 https://en.wikipedia.org/wiki/Fake_news

3 Liddy ED. Natural language processing. In Encyclopedia of Library and Information Science. (2ndedn), 2001.

4 Hoey M, Mahlberg M, Stubbs M, et al. Text, discourse and corpora: theory and analysis. (1stedn), Corpus and Discourse. Bloomsbury Academic. 2007.

5 Traylor T, Straub J, Gurmeet, et al. Classifying fake news articles using natural language process-ing to identify in-article attribution as a supervised learning estimator. 2019 IEEE 13th International Conference on Semantic Computing (ICSC), Newport Beach. 2019.

6 Girgis S, Amer E, Gadallah M. Deep learning algorithms for detecting fake news in online text. 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt. 2018.

7 Aldwairi M, Alwahedi A. Detecting fake news in social media networks. Procedia Comput Sci. 2018;141:215–22.

8 https://www.cs.auckland.ac.nz/courses/tutorials/

9 Nyow NX, Chua HN. Detecting fake news with tweets' properties. 2019 IEEE Conference on Application, Information and Network Security (AINS), Pulau Pinang, Malaysia. 2019.

10 Manzoor SI, Singla J, Nikita. Fake news detection using machine learning approaches: A systematic review. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India. 2019.

11 Yang S, Shu K, Wang S, et al. Unsupervised fake news detection on social media: A generative approach. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33:5644–51.

12 Singh V, Dasgupta R, Sonagra D, et al. Automated fake news detection using linguistic analysis and machine learning. In International Conference on Social Computing, Behavioral-Cultural Modeling, &Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), 2017.

13 Bajaj S. The pope has a new baby! Fake news detection using deep learning. 2017.

14 Granik M, Mesyura V. Fake news detection using naive bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, Ukraine. 2017.

15 Bird S, Klein E, Loper E. Natural language processing with python: analyzing text with the natural language toolkit. (1stedn), O'Reilly Media, 2009.

16 Elaziz M, Al-qaness M, Ewees A, et al. Recent advances in NLP: the case of arabic language. studies in computational intelligence. Springer, Cham. 2019.
17 Aggarwal CC. Data classification: algorithms and applications. (1stedn), Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor&Francis, 2014.
18 Hardeniya N, Perkins J, Chopra D, et al. Natural Language Processing: Python and NLTK. Packt Publishing, 2016.
19 https://www.kaggle.com/datasets
20 https://www.kaggle.com/nopdev/real-and-fake-news-dataset
21 https://www.kaggle.com/kagglepankaj/fake-news-dataset
22 Pratiwi YR, Asmara RA, Rahutomo F. Study of hoax news detection using naive bayes classifier in indonesian language. 2017 11th International Conference on Information Communication Technology and System (ICTS), Surabaya, Indonesia. 2017.
23 Mahir EM, Akhter S, Huq MR, et al. Detecting fake news using machine learning and deep learning algorithms. 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia. 2019.
24 Poddar K, Umadevi KS, Amali GB. Comparison of various machine learning models for accurate detection of fake news. 2019 Innovations in Power and Advanced Computing Technologies (i-PACT). Vellore, India. 2019.