**Edelweiss Applied Science and Technology**
*ISSN: 2576-8484*
*Vol. 8, No. 6, 6132-6146*
*2024*
*Publisher: Learning Gate*
*DOI: 10.55214/25768484.v8i6.3346*


# Bangladeshi False message detection using the Tree and characterized by procrastination classifiers

Rasidul Haque[1], Md. Shafiul Alam Chowdhury[2], Md. Farukuzzaman Khan[3], Mohammed Sowket Ali[4], Md. Amanat Ullah[5*], Md. Abdul Mannan[6]
[1]Stanley College, WA, Australia; rasidul.haque@sche.edu.au (R.H.).
[2]Department of Computer Science and Engineering, Uttara University, Dhaka,Bangladesh; hafiul.cse@uttarauniversity.edu.bd (M.S.A.C.).
[3]Department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh; mfkhan@cse.iu.ac.bd (M.F.K.).
[4]Department of Information and Communication Technology (ICT), Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh; sowket@baust.edu.bd (M.S.A.).
[5,6]Department of Mathematics, Uttara University, Dhaka, Bangladesh; amanat.cse@uttarauniversity.edu.bd (M.A.U.) mannan.iu31@uttarauniversity.edu.bd (M.A.M.).

**Abstract:** False message is much moreexotericdue to the rapid use of social media. The ability to detect False message is recognized among the riskiest types of manipulation, as it is formed with the malicious purpose of misleading consumers. Many researchers have already suggested several False message detection systems based on social context and diffusion. In this paper, we have presented a system that can expose False message related to the subject of Bangladeshi digital message content. Moreover, machine learning classifiers named Extra tree and Procrastination have been selected. For feature extractions, we have used Term-Frequency-Inverse Document Frequency (TF-IDF) and countvectorizer simultaneously, and countvectorizer separately. After doing two types of experiments, the Extra tree classifier gives the best result for the first experiment and the Procrastination classifier gives the average result for the first experiment,where both classifiers give the average result for the second experiment.

***Keywords:*** *GNB, LR, MLP, RF, SVM, TF-IDF matrix, VEC.*

## 1. Introduction

False message is a paradox, as it ignores the reality of legitimate message and earns public interest. Nowadays, people are more engaged with using social media and the internet. Thus, dependents much on online content, social networking platforms are the primary stage for creating False message. [1]. Digital content seems to have a major influence on the emotions and decisions of users. Now, many people are more engaged with buying things on e-commerce sites and social media selling groups. The online review system is one of the efficient ways to buy the desired product. Some detractors are now relating to the recent iteration of falsehoods as a result of the spreading dissemination of misrepresenting information. False message has increased attention in recent years.

Misconceptions and False message stories are being used to instill fear, promote racist ideas, and incite violence against common people. Even misinformation has serious constitutional implications. It is one of the challenging matters to stop the spreading of False message. Sometimes, False message and articles mislead and give false information, hide proper information, and cause misunderstanding among general people. Although spreading false message is not a new thing, being able to detect it is thought to be a very challenging task. False message can be classified into several types, including stories, news, post articles, and so on. People are mostly puzzled by False message and are mostly influenced by false information. False message about well-known organizations is sometimes spreading and causing harm to their reputation. Some days ago, one piece of message was published that the Sonali bank will give


* Correspondence: amanat.cse@uttarauniversity.edu.bd

1600/- taka per customer in Bkash which is a rumor [2]. People become excited after publishing that message when they hear that message is nothing but False messag. It impacts their minds. In another message article published some days ago about Barack Obama, the former US President Barack Obama prepares a 'dust cloud of nonsense' [3]. It hasabruptly turned into a conversation topic in the political industry. Day by day, people become more dependent on social media and get attracted by False message.Numerous researchers in identifying fake and real message have done and invented different technologies based on English message datasets. Using the Bangladeshi dataset, we investigate the use of various classifications for the False message discovery system in our experiment. In this experiment, we examine two machine learning algorithms with TF-IDF and count vectorize. Besides, also determine which features, in the case of the Bangladeshi dataset, are most predictive of identifying False messag. We compared six machine learning classifiers, including an extra tree classifier and a Procrastination classifier.

## 2. Previous Study

As False message has become a global issue, it has also become a challenge for all sectors of society. Granik et.al [13] described a recognized Naive Bayes method for detecting False message, which is known as a well-known technology. There are many similarities between False message and spam messages detection like having a lot of grammatical errors, influencing reader opinion, and using a similarly limited set of words. As a result, the same method is used for False message detection. This experiment results in 74% accuracy for the particular dataset. For experimental results Naive, they mainly used BuzzFeed News, which contains Facebook posts. Only a few of them are used, which are mainly linked with the Facebook post with articles and level of text. The Data section is mainly a combination of true and false news. On the other hand, no factual quantities are ignored. After cleaning the datasets, they got 1771 message articles. This experiment shows similar results for both true and false news, but the False message range shows a little worse result compared to true message identification due to the skewness of the dataset. Their experimental results show that the precision for the given classifier is 0.71. On the other hand, recall for the same classifier is 0.14. Yerlekar et.al[12] discussed a unique methodology. At first, they cleaned the dataset by modifying information that is incorrect, incomplete, duplicated, and improperly formatted. Data is split into two categories testing and training stages, and it is split into a 30-70% range. They have built a flask application on the cloud. After experimenting with the dataset, they found 80% accuracy.Mr. R. Praveen Kumar et al. [14] presented a falsemessage detection model. In their proposed method, two feature extraction named count vectorize and a TF-IDF matrix were used. They also used MNB and SVM as classifiers. Benchmark Dataset for False message Detection" is used as a dataset. They have divided the dataset into three categories such as training, test, and validation. Count vectorization has been used for feature extension. After that, use GridSearchCV and a 5-hold cross-validation set. After examination, the SVM classifier is giving 56% accuracy and naive Bayes is giving 78% accuracy.M. G. Hussain et al. [15] proposed a method for Bangladeshi False message detection. For feature extension, they have used Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorize. Their experiment shows that SVM has performed better than MNB. For experimental purposes, the open-source dataset was used to classify False message. They have made a dataset from ProthomAlo1, ProthomAlu2, BalerKontho3, Motikontho4. Dataset is made with 2500 message articles for experimental purposes. They have removed special characters, digits, English characters, and emotions from the dataset. Datasets are divided into two segments, 30% are used in tests and 70% are used in training. Their experiment shows that SVM including linear kernels shows a 96.64% accuracy and where MNB shows 93.32% accuracy.An uncommon technique is proposed by Deokate[16] for False message detection. Different types of datasets are used. There is a disparity among their datasets. BuzzFeed's dataset captures stories shared on Facebook. On the other hand, CREDBANK and PHEME datasets are Twitter-based. CREDBANK was mainly a large-scale dataset of Twitter and rumors of Twitter. Different features are used for better performance. At first, the dataset is preprocessed. Three types of features named structural, User, and content must be selected. Their experiment shows that the Mean absolute error gives an accuracy of 0.0116%, the Root means squared error gives an accuracy of

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

0.1075%, and Relative absolute error of 21.3112 %, compared to this algorithm better accuracy result obtained by Root relative squared error accuracy of 21.4997 %. Comparative analysis is done by F. Islam [17]. They examined three classifiers for identifying False message and showed which technology works better for False message detection. They mainly use well-known classifiers for Bangladeshi False message. Furthermore, they had made their dataset. Dataset is made with a sparse matrix that contains nonzero values. They have a real and fake dataset together. For training and testing the dataset, it is divided into two parts: 75% and 25% ratio. Naïve Bayes does not give better output. On the other hand, the Logistic Regression classifier performs better in each experiment. The experiment is done with both title and body. In both cases, it gives an accuracy of 81%. Random forest gives better output with 83.5% accuracy for the title, 82.9% accuracy for body text and when combined we have got 85% accuracy.Another comparative analysis is done in [18] for Bangladeshi False message detection. They mainly focus on Gaussian Naïve Bayes classifiers. They have made their dataset and tested it with six different classifiers. Two features named TF-IDF including Extra Tree Classifier are used. They have tested their dataset with multiple classification methods named SVM, LR, MLP, RF, VEC, and GNB. Findings of the experimental result are Gaussian Naïve Bayes gives better output of 54.15% accuracy and after using an extra tree classifier it gives 85.52% accuracy and without feature, it gives 45% accuracy.

## 3. Methodology

In this case, we extracted features and then ran them through classifiers to ensure accuracy. In our dataset, we used various classifiers to detect False message. The process is explained in detail in the following section.

In this research, two classifiers were used to distinguish between False message from the Bangladeshi news. The classification model's structure flow diagram is depicted in Figure 1, and the remaining stages are explained below. The strategy of the exploration activity has been discussed in this section. Each examination action requires a distinct strategy for dealing with it. Provide an itemized discussion of how to use a model, along with a brief description of each component of the technique.

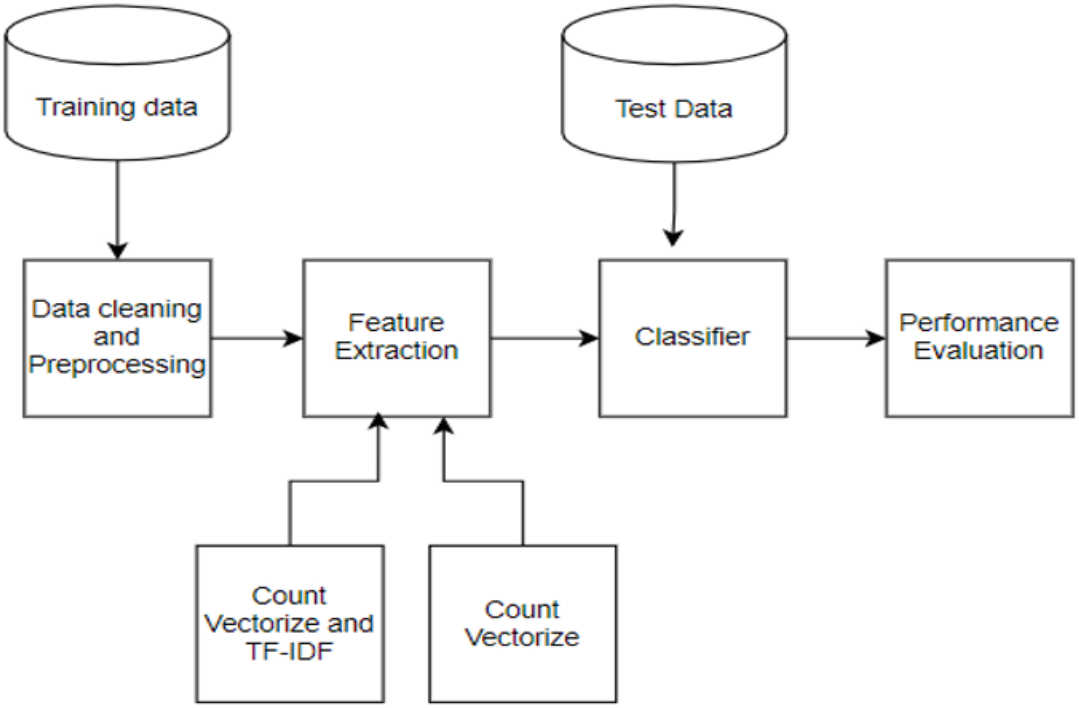


**Figure 1.**
System flow diagram.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

*3.1. Data Collection*

We properly gathered the information for the investigation. All information comes from online media and message websites like channel Dhaka. News, earki.com, and motikontho.wordpress.com. It is somewhat unpredictable to gather information from online sites for security issues. We mainly collected the dataset from [4]. A total of 8501 data points are collected from online sources. For experiment purposes, around 1299 false data and 7202 true data have been used. The image below depicts the amount of False message and real message used in this study.

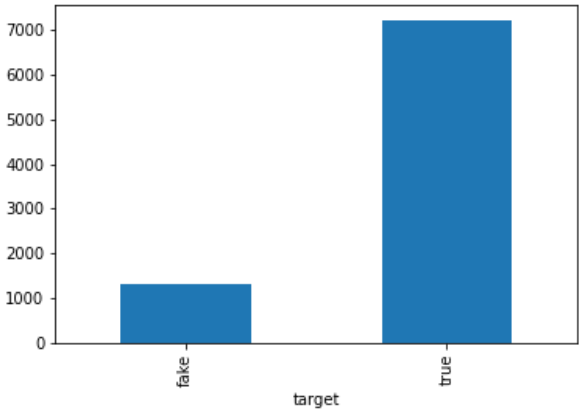


**Figure 2.**
Fake Vs real message dataset.

*3.2. Data Preprocessing*

To implement codes, the dataset must be preprocessed, and a correct dataset must be used. Natural Language Processing (NLP) is a subcategory of Data science that mainly relates to textual data. However, before the data can be used for analysis or prediction, it must first be processed. The purpose of preparing a text content dataset for designing the model is known as text preprocessing. For NLP experimental purposes, it is the initial stage. Some examples of preprocessing steps are as follows Figure 3 depicts the key steps of the preprocessing period. It is essential to preprocess the input text dataset before injecting it into the specific classifier. A raw dataset may carry unnecessary symbols or other aspects that are even though irrelevant to our classification. We also removed special characters such as @, #! And so on from our dataset. These components could either work on improving or deteriorate the accuracy of classifications. Our dataset is complete after we remove numerical values, stop words, punctuation marks, and special symbols for improving the performance.

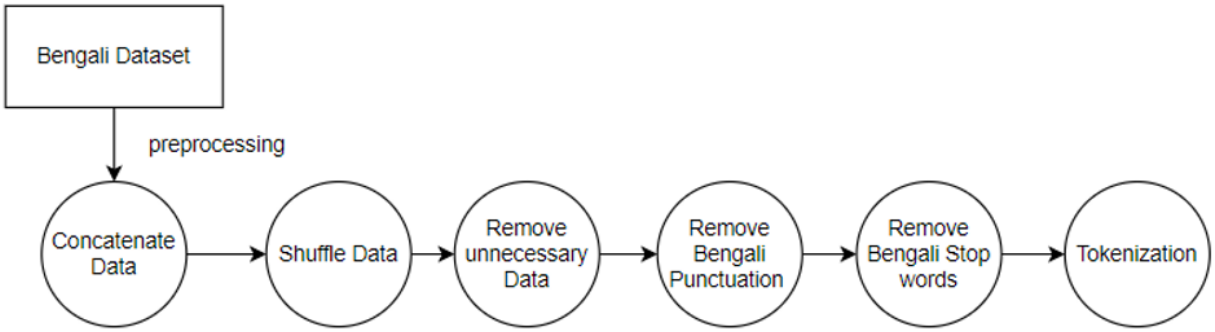


**Figure 3.**
Data preprocessing.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

*3.2.1. Concatenating and Shuffling the Dataset*

First, Concatenate true and false data. The order in which the observations are fed into the model influences internal decisions. Before each training iteration, the training data should be shuffled at random. Even if the algorithm is impenetrable. It is regarded as best practice. It will prevent biases.

| | articleID | domain | date | category | headline | content | label | target | source | relation |
|---|---|---|---|---|---|---|---|---|---|---|
| 5068 | NaN | bd24live.com | 2018-09-19 17:54:49 | National | ভালো নেই আফজাল শরীফ, সহায়তা চাইলেন প্রধানমন্ত্রীর | জাতীয় চলচ্চিত্র পুরস্কার প্রাপ্ত কমেডি অভিনেতা... | 1.0 | true | জাতীয় চলচ্চিত্র পুরস্কার প্রাপ্ত কমেডি অভিনেতা | Related |
| 6460 | NaN | banglatribune.com | 2018-10-05 20:14:55 | National | কুষ্টিয়ায় জমির বিরোধের সংঘর্ষে নিহত ১ | কুষ্টিয়ার দৌলতপুরে জমি নিয়ে বিরোধের সংঘর্ষে নি... | 1.0 | true | দৌলতপুর থানার ভারপ্রাপ্ত কর্মকর্তা (ওসি) | Related |
| 2676 | NaN | dailyjanakantha.com | 2018-09-21 17:49:04 | Miscellaneous | ২১ সেপ্টেম্বর বিশ্ব আলঝেইমারস দিবস। | অনলাইন রিপোর্টার ।। যে রোগে মানুষ সব ভুলে যায় ... | 1.0 | true | Reporter | Related |
| 5821 | NaN | bd-pratidin.com | 2018-09-20 16:41:24 | National | পুলিশের জন্য ভারত থেকে ২০টি প্রশিক্ষণপ্রাপ্ত য... | বেনাপোল স্থলবন্দর দিয়ে আজ বৃহস্পতিবার সকালে পু... | 1.0 | true | বেনাপোল পোর্ট থানার ভারপ্রাপ্ত কর্তকর্তা | Related |
| 6305 | NaN | bangla.bdnews24.com | 2018-09-20 03:25:27 | National | ডিআইজি মিজানকে ফেরু দুলকে তলব | আগামী ৩০ সেপ্টেম্বর সকাল ১০টায় দুলক প্রধান কার... | 1.0 | true | দুলকের জনসংযোগ কর্মকর্তা | Related |
| 546 | 550.0 | motikontho.wordpress.com | 2012-04-29T17:55:36+00:00 | Miscellaneous | ডেসটিনিকে সস্কার দিলেন না মালাই লামা | দৈনিক ... | বিশেষ মতিবেদকআলোচিত এম এল এম কম্পানী ডেসটিনির ... | 0.0 | fake | NaN | NaN |
| 6693 | NaN | jagonews24.com | 2018-09-20 16:36:31 | Crime | কটূক্তি: থানেদার বিরুদ্ধে গ্রেফতারি পরোয়ানার ... | বঙ্গবন্ধু শেখ মুজিবুর রহমান ও ক্ষমতাসীন আওয়াম... | 1.0 | true | Reporter | Related |
| 461 | NaN | bd24live.com | 2018-09-20 08:15:10 | National | কষ্টসাধ্য জয় নিয়ে যা বললেন রোহিত শর্মা | হংকংয়ের বিপক্ষে এশিয়া কাপের প্রথম ম্যাচে ভারত ... | 1.0 | true | ভারত অধিনায়ক | Related |
| 8212 | NaN | jagonews24.com | 2018-09-26 16:59:18 | National | ট্রাকসহ ভেঙে পড়ল বেইলি সেতু | টাঙ্গাইল-আরিচা মহাসড়কের নাগরপুর উপজেলার বরাপু... | 1.0 | true | নাগরপুর থানা পুলিশের ভারপ্রাপ্ত কর্মকর্তা | Related |
| 4646 | NaN | channelionline.com | 2018-09-21 18:37:06 | National | চট্টগ্রামের সঙ্গে ঢাকা ও সিলেটের রেল যোগাযোগ বন্ধ | ব্রাহ্মণবাড়িয়া প্রতিনিধি ব্রাহ্মণবাড়িয়ার কসবা... | 1.0 | true | আখাউড়া রেলওয়ে থানা পুলিশের ভারপ্রাপ্ত কর্মকর্তা | Related |

**Figure 4.**
Concatenation and shuffling of data.

*3.2.2. Removing Unnecessary Dataset*

Some of the dataset's unnecessary columns, such as 'date' and 'label,' are not used during modeling, so we remove them. After removing all unnecessary data, the dataset will look like Figure 5.

| | articleID | category | headline | content | target |
|---|---|---|---|---|---|
| 2069 | NaN | Sports | স্মৃতির শহর ভ্যালেন্সিয়ায় আজ খেলবেন রোনালদো | ক্রীড়া ডেস্ক : ক্রিস্টিয়ানো রোনালদো ২০০৯ সাল থ... | true |
| 506 | 510.0 | Miscellaneous | ঈদের দিন বন্ধুর বাসায় সেমাইয়ের প্লেট চেটেপুটে ... | ঈদের দিন দাওয়াতে গিয়ে বন্ধুর বাসায় সেমাই খাওয়া... | fake |
| 1089 | NaN | Entertainment | যেভাবে রোবট 'চিট্টি' হলেন রজনীকান্ত (ভিডিও) | ভারতের সবচেয়ে বড় তারকা রজনীকান্তের ছবি 'রোবট' ... | true |
| 6967 | NaN | National | নাজিরপুরে নির্বাচনি গণসংযোগ করলেন শ ম রেজাউল ... | আগামী সংসদ নির্বাচনে পিরোজপুর-১ (পিরোজপুর সদর-... | true |
| 2137 | NaN | Editorial | মজাদার স্নো-বল কাস্টার্ড | ঘরোয়া উৎসবে একটি জনপ্রিয় খাবার হচ্ছে স্নো-বল ক... | true |
| 4869 | NaN | National | চট্টগ্রামে ইয়াবাসহ রোহিঙ্গা গ্রেপ্তার | গ্রেপ্তার তিনজন হলেন- মো. রফিক (২৭) ও রিয়াজুল... | true |
| 1072 | 1078.0 | International | পুরো জীবন টয়লেটে কাটিয়েছেন যে মানুষটি ! | প্রেমরাজ দাস, ৪০ বছরের এই ব্যক্তি তার পুরো জীব... | fake |
| 6465 | NaN | National | নরসিংদীতে দুই ভাইকে প্রকাশ্যে কুপিয়ে হত্যা | নরসিংদীতে মাদক ব্যবসাকে কেন্দ্র করে দুই ভাইকে ... | true |
| 563 | NaN | National | আসন্ন নির্বাচনের হুমকি 'সাইবার ক্রাইম' | ঢাকা: আসন্ন একাদশ জাতীয় সংসদ নির্বাচনে 'সাইবার... | true |
| 6396 | NaN | National | খানাখন্দে ভরা শরীয়তপুর-নওডোবা সড়ক | নির্মাণ কাজ শেষে পদ্মা সেতুর সংযোগ সড়কের সুবিধ... | true |

**Figure 5.**
Removing unnecessary data.

*3.2.3. Removing Punctuations*

Text cleaning, also known as text pre-processing, is a vital factor when working with text. Real-world human-writable text data contains a wide range of misspelled words, short words, special symbols (!"#$%&'() *+, -. /: <=>? @ [\] ^_` {|} ~।০৪), emojis, and other noise text. Before feeding this type of noisy text data to the machine learning model, it is needed to clean it. Various methods must be used to clean the data. This tutorial will show you how to deal with a special symbol or punctuation mark in text data. If your word embedding model does not support punctuation, it is perfectly fine to remove it from the text. From related work, such as plagiarism detection or author recognition, it is necessary to remove punctuations [6]. The diagram below depicts the visualization of a dataset after punctuation has been removed.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

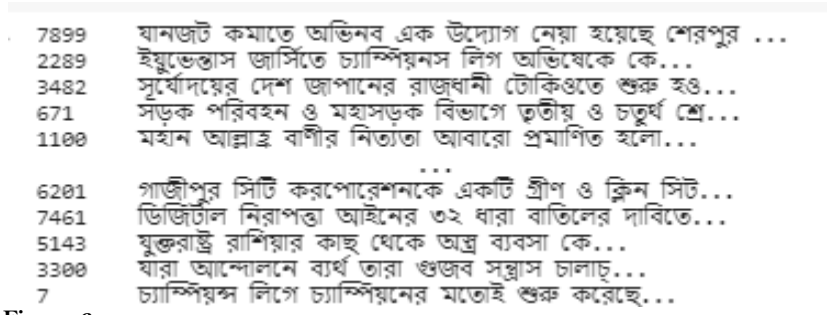


**Figure 6.**
After removing unnecessary data.

*3.2.4. Removing Stop Words*

Stop words are infinitesimal words in a language that can be used to categorize text features. They are mainly responsible for creating noise in text data. All those are words that are recurrently used in text to help connect ideas or sentences pattern [8]. It accomplishes having a similar purpose as Tokenization, as it transmits text data into a more processable format. As a result, once we examine our data, we will: be able to clear out the noise and emphasize the words. However, before delving into the data, stop words are easily eliminated by removing words from a pre-defined list. It's worth noting that there's no such thing as a universal list of stop words. Stop words removal in this case deducts the common Bangladeshi language prepositions such as 'অতএব', 'অথচ', 'অথবা', 'অনুযায়ী', 'অনেক', 'অনেকে', 'অনেকেই', 'অন্তত', 'অন্য', 'অবধি', 'অবশ্য', 'অর্থাত', 'আই', 'আগামী', 'আগে', 'আগেই', 'আছে', 'আজ', 'আদ্যভাগে', 'আপনার', 'আপনি', 'আবার', 'আমরা', 'আমাকে', 'আমাদের', 'আমার', 'আমি', 'আর', 'তেমন', 'তো', 'তোমার', 'থাকবে', 'থাকবেন', 'থাকা', 'থাকায়', 'থাকে', 'থাকেন', 'থেকে', 'থেকেই', 'থেকেও', 'দিকে', 'দিতে', 'দিন', 'দিয়ে', 'দিয়েছে', 'দিয়েছেন', 'দিলেন', 'দু', 'দুই', 'দুটি', 'দুটো', 'দেওয়া', 'দেওয়ার', 'দেওয়া', 'দেখতে', 'দেখা', 'দেখে', 'দেন', 'দেয়', 'দ্বারা', 'ধরা', 'ধরে', 'ধামার', 'নতুন', 'নয়', 'না', 'নাই', 'নাকি', 'নাগাদ', 'নানা', 'নিজে', 'নিজেই', 'নিজেদের', 'নিজের', 'নিতে', 'নিয়ে', 'নিয়ে', 'নেই', 'নেওয়া' and so on in Bangladeshi language. The stop words in the text data have been removed for data preprocessing, and the features have been extracted successfully [5]. The diagram below depicts the visualization of a dataset after stop words have been removed.

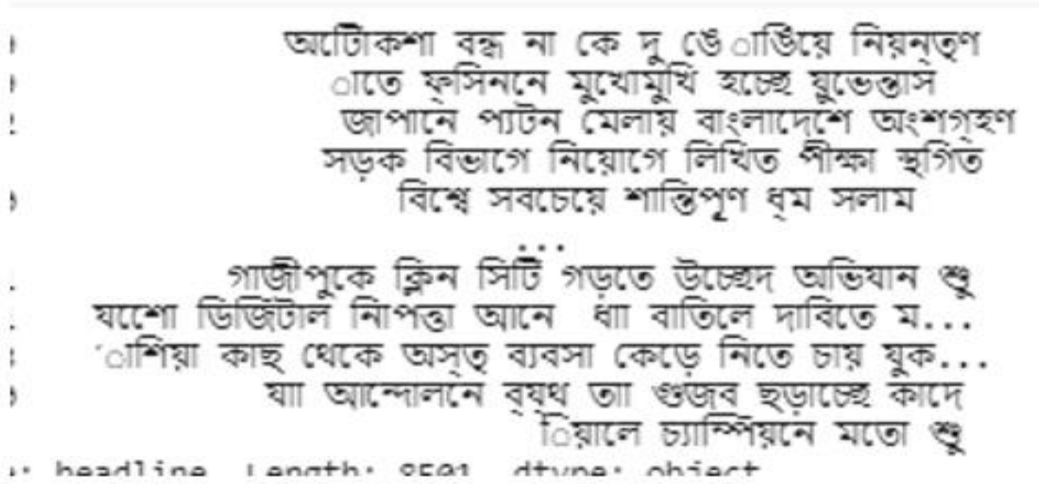


**Figure 7.**
After removing stop words.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

*3.2.5. Removing Digits*

Removing numbers such as "০১২৩৪৫৬৮৯" from the Bangladeshi Dataset When doing text clustering or getting key phrases, we usually remove Bangladeshi numbers because numbers don't give much importance to getting the main words. Diagram 8 depicts the representation of a dataset that has had all digits removed.


**Figure 8.**
After removing digits.

*3.2.6. Tokenization*

Tokenization is divided into words after removing unnecessary data using the remaining sentences, sack words that we use daily [7]. It divides a block of text into individual words based on a delimiter. Each of these smaller units is referred to as a token. Machine learning algorithms use these discrete elements as input. Tokenization is one of the most common tasks when working with text data.

First, identify the words that make up a string of characters before processing a natural language. As a result, tokenization is the most fundamental step in advance. This is significant because analyzing the words in the text allows the meaning of the text to be easily deduced. The dataset's output is shown below.

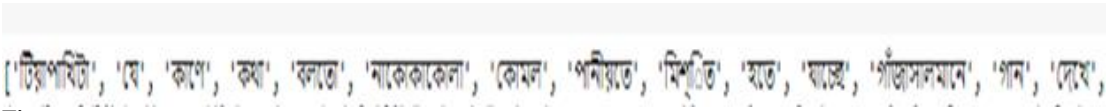


**Figure 9.**
Dataset after tokenizing.

## 4. Feature Extraction

The methodology below can be needed to facilitate machine learning methods in the system of innovation from Bangladeshi message headlines.

*4.1. Tf-IDF*

The Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical weighting performance measurement that is widely used in information extraction in natural language processing. It is an analytical criterion that has been used to evaluate the period in a dataset [8]. The emphasis of a term spreads in percentage to how many times it occurs in the document. This is compensated by occurrence of the words in the dataset [8][9]. Term frequency (TF) describes the number of words that appear in a document. Terms are words or phrases in the context of the natural language. Some fields such as text analytics, Deep Learning, and data capture render the use of term frequency. All terms are considered equal when evaluating TF. However, it is common knowledge that certain words,

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

such as 'গোলাম', 'আজমের', 'নেয়', 'সহজে', 'হাল', 'ছাড়তাম' appear frequently but have little meaning. However, TF-IDF is a multiplication of Tf and IDF.

TF-IDF score for term i in document j =TF (i, j)* IDF(i)

*4.2. Count Vectorize*

The data set must first be segmented to eliminate particular words before it can be used for predictive analysis. These words must first be transmitted as integers or floating-point values before they could be used as input data in machine learning algorithms. This process is referred to as vectorization. The data set must first be segmented to eliminate particular words before it can be used for predictive analysis. These words must first be transmitted as integers or floating-point values before they could be used as input data in machine learning algorithms. This process is referred to as vectorization. The Count Vectorize provides an easy way to tokenize a variety of content archives and create a jargon of known words, as well as encode new reports [19].

The Bangladeshi stop words, punctuation, and digits were first removed from all the message headlines using scikitlearn'sCountVectorizer, and then they were tokenized using spaces and punctuation marks as delimiters. After tokenizing all the headlines, a sparse matrix with all the message headlines as rows and the tokens as columns was restored. The CountVectorizer not only changes the value of a set of text datasets and also generates a terminology of recognized distinct words, but it also converts legal documents with that terminology [20].

**Table 1.**
Count vectorize for data.

| গোলাম | আজমের | নেয় | সহজে | হাল | ছাড়তাম |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 |

## 5. Classifier

*5.1. Extra Tree Classifier*

Extra Trees are recognized as Extremely Randomized Trees, as their performance is comparable to Random Forest. Throughout the architecture of the trees, some problems occur. Each build is constructed using all the data obtainable, especially in the training dataset. The best split for formulating the root node or any other node is determined by monitoring through a subset of randomly approved characteristics of conformation sqrt (number of features). The split for every single feature is randomly generated. The decision stump has a maximum depth of one.

Because of its explicit signification, simple characteristics, as well as ease of transformation to "if-then" principles, the extra-tree classifier was selected for this experiment. Because of its intellectual capacity to randomize statistical inputs, the extra-tree technique is preferred. This concept plays a vital role when dealing with issues that have a high proportion of datasets. This type of situation frequently leads to increased accuracy [10].

*5.2. Passive Aggressive Classifier*

When correct predictions are made, it remains passive, but when incorrect predictions are made, it reacts aggressively. Let's look at how to use Python to create an aggressive passive classifier. First, import the Python libraries required for this task. Procrastination algorithms are a type of large-scale learning algorithm [11]. They are similar to Perceptrons in the sense that they do not require a learning rate. They perform including a regularization parameter, unlike the Perceptron. To model and validate data using a classification algorithm, a procrastination classifier must be loaded.

## 6. Experimental Result

*6.1. Experimental Setup*

We tried out some different settings here. During the preprocessing stage, we used the Bangladeshi Natural Language Processing Toolkit (BNLP) to dispel punctuation, stop words, digits, and tokenization. We used Countvectorizer and TF-IDF together and Countvectorizer separately as feature

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

extraction methodologies, and calculated them with scikit-learn. In this study, we used an extra tree and an active-passive classifier for classification. We divided the dataset into two sections: 80 percent for training and 20 percent for testing. The research works in this paper were all operated on Google Colab, including some tools such as the Scikit-learn library and Python 3.5.

Using the Bangladeshi dataset, we tested several classification algorithms named the Extra Tree classifier and Procrastination classifier. Compare them to other well-known classifiers named Linear Vector Machine (SVM), Logistic Regression (LR), Multinomial Naive Bayes (MNB), Random Forest Classifier (RF), and Decision Tree.

### *6.2. Experimental Result 1*

**Table 2.**
Classification report after using TF-IDF and Count vectorize feature extraction.

| Model | Accuracy |
|---|---|
| Decision tree | 88.36% |
| Logistic regression | 89.48% |
| Multinomial naive bayes | 90.06% |
| Passive aggressive | 90.30% |
| Linear SVM | 90.89% |
| Random forest | 91.48% |
| Model | Accuracy |
| Extra tree classifier | 92.18% |

### *6.3. Experimental Result 2*

**Table 3.**
Classification report after using Count vectorize feature extraction.

| Model | Accuracy |
|---|---|
| Logistic regression | 88.65% |
| Passive aggressive | 88.71% |
| Decision tree | 89.12% |
| Extra tree classifier | 89.28% |
| Linear SVM | 90.59% |
| Multinomial Naïve Bayes | 90.71% |
| Random forest | 90.95% |

## 7. Visualization

The classification strongly involves the use of statistical visualization. Here's a word cloud visual representation of the data set based on real and False message. Word Cloud is a traditional technique that endeavors to design words that represent the word frequency that is used to check the information at a glance. We draw a word cloud by attributing colors from each segment after acquiring the similarity measures for both categories, as shown in Figure 14.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

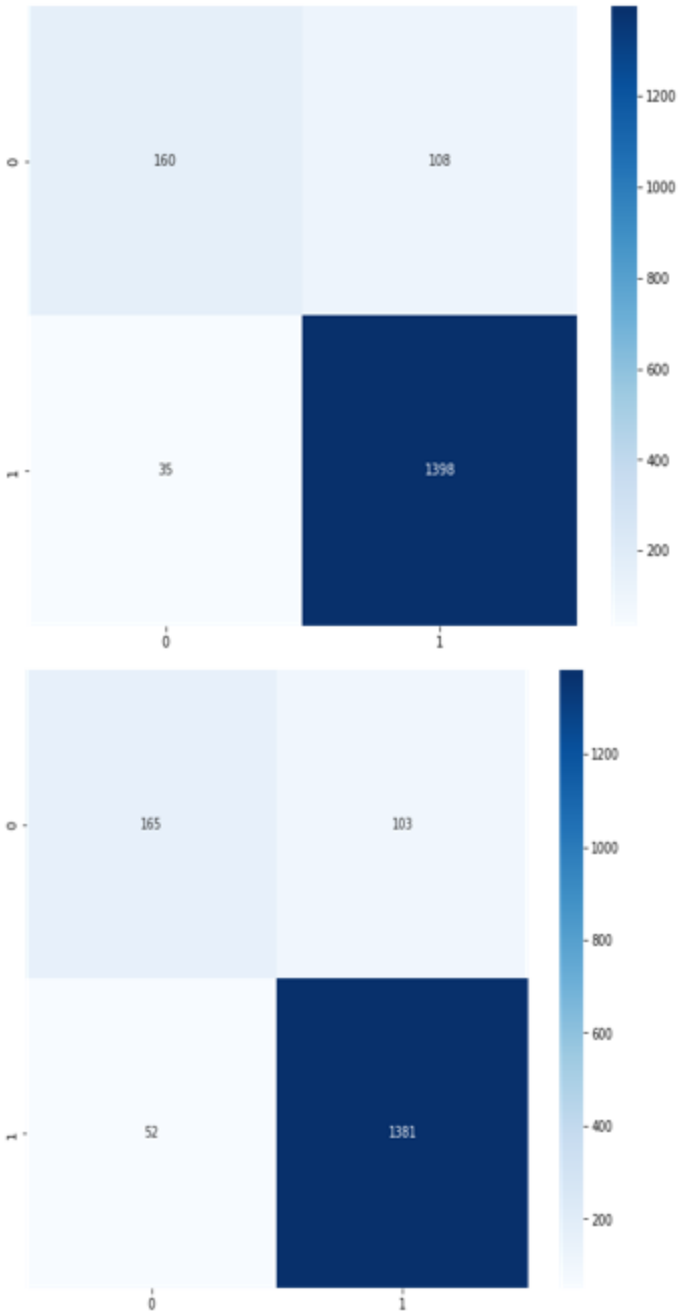


**Figure 10.**
Confusion-matrix for extra tree algorithm.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| fake | 0.82 | 0.60 | 0.69 | 268 |
| true | 0.93 | 0.98 | 0.95 | 1433 |
| accuracy | | | 0.92 | 1701 |
| macro avg | 0.87 | 0.79 | 0.82 | 1701 |
| weighted avg | 0.91 | 0.92 | 0.91 | 1701 |

Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| fake | 0.76 | 0.62 | 0.68 | 268 |
| true | 0.93 | 0.96 | 0.95 | 1433 |
| accuracy | | | 0.91 | 1701 |
| macro avg | 0.85 | 0.79 | 0.81 | 1701 |
| weighted avg | 0.90 | 0.91 | 0.90 | 1701 |

**Figure 11.**
Classification report for extra tree algorithm.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

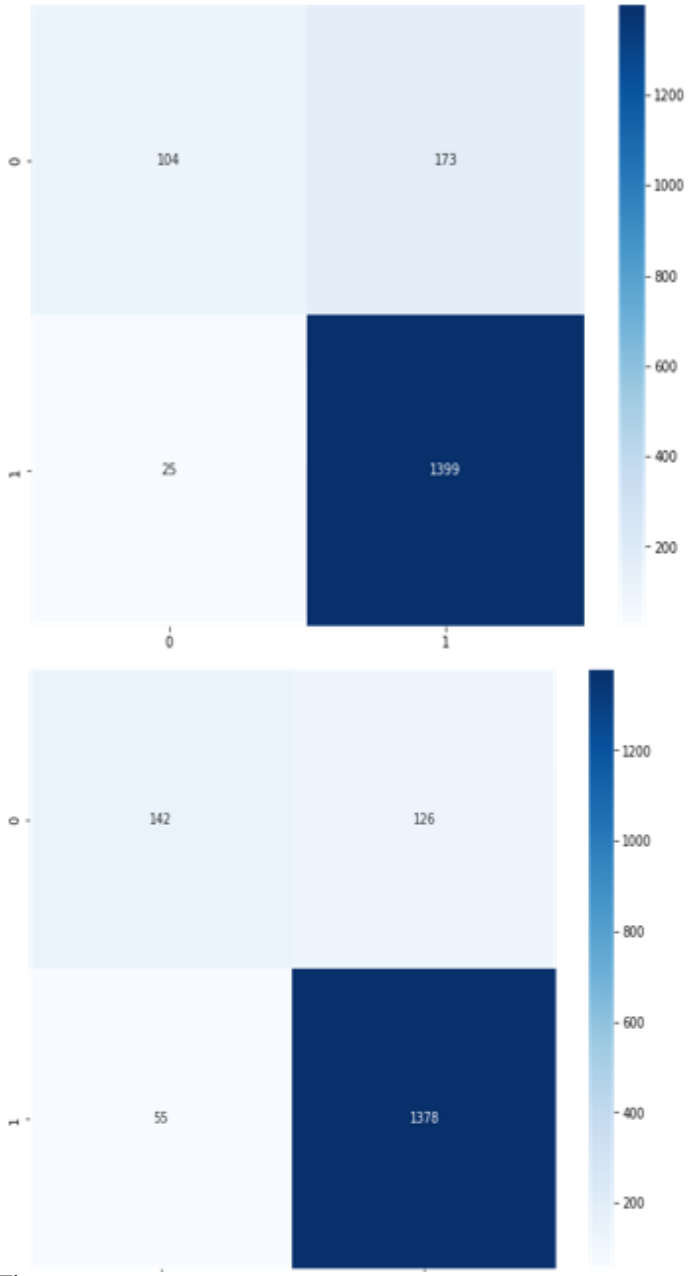


**Figure 12.**
Confusion-matrix for procrastination classifier.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| fake | 0.72 | 0.57 | 0.64 | 268 |
| true | 0.92 | 0.96 | 0.94 | 1433 |
| accuracy | | | 0.90 | 1701 |
| macro avg | 0.82 | 0.77 | 0.79 | 1701 |
| weighted avg | 0.89 | 0.90 | 0.89 | 1701 |

Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| fake | 0.72 | 0.57 | 0.64 | 268 |
| true | 0.92 | 0.96 | 0.94 | 1433 |
| accuracy | | | 0.90 | 1701 |
| macro avg | 0.82 | 0.77 | 0.79 | 1701 |
| weighted avg | 0.89 | 0.90 | 0.89 | 1701 |

**Figure 13.**
Classification report for procrastination algorithms.


**Figure 14.**
Word cloud in the dataset.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

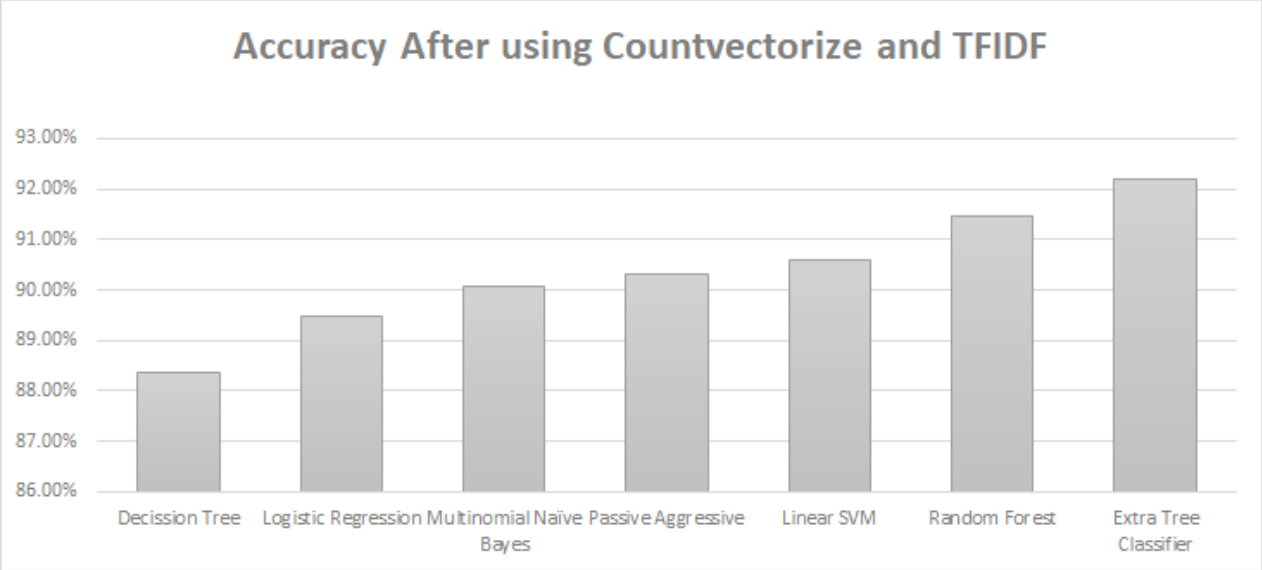


**Figure 15.**
Accuracy values after using Count vectorize and TF IDF feature extraction.

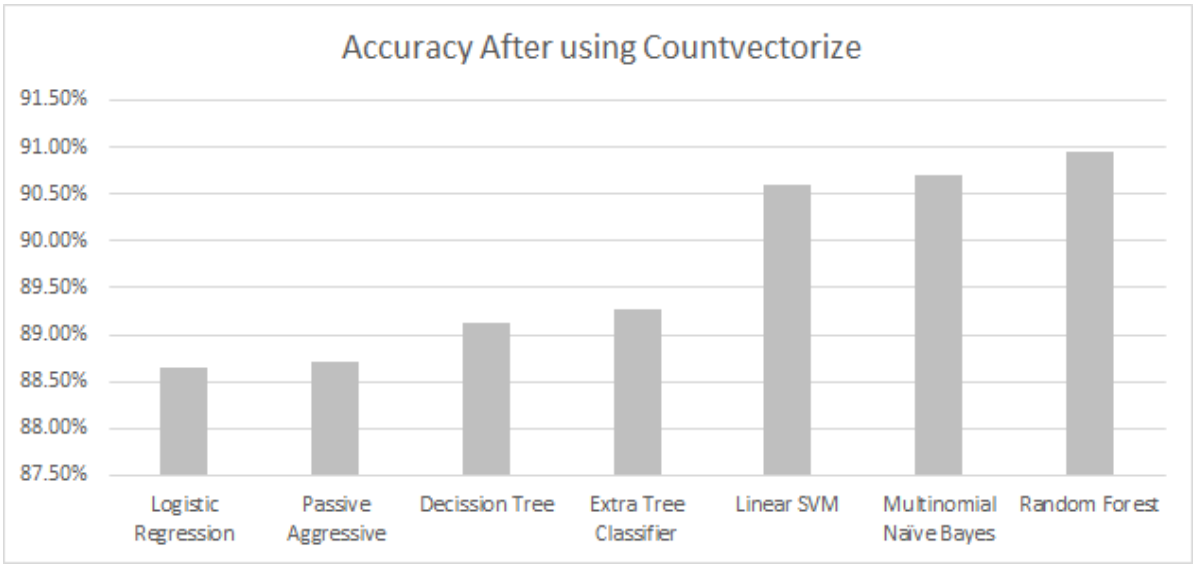


**Figure 16.**
Accuracy values after using count vectorize and TF IDF feature extraction.

## 8. Analysis

From Figure 14 and Figure 15 it is clear that the extra tree classifier gives the supreme accuracy of 92.18% and the decision tree classifier gives the lowest accuracy about 88.36% after using Count vectorize and TF-IDF. On the other hand, the random forest classifier performed better after using the Count vectorize classifier. After using the count vectorize classifier, some classifiers such as the decision tree and Random Forest give better results. However, procrastination classifiers give 93.30% accuracy for the count vectorize and TF-IDF classifier and 88.71% accuracy after using the count vectorize classifier separately.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 8, No. 6: 6132-6146, 2024*
*DOI: 10.55214/25768484.v8i6.3346*

*8.1. Conclusion and Future Work*

In this work, we have performed an innovative experiment on Bangladeshi False message identification in the substance of Bangladesh and other counties. Although internet-based False message is rapidly spreading in South Asia, no research has been conducted in this area until now; however, we hope that our work will open new ways for future research in this area. In the future, we will aim to establish a more accurate and precise classifier model than what we have so far. We plan to use neural networks and deep learning models as well. We believe to build our broad data set of Bangladeshi messages in the coming decades, as well as a more powerful framework for the feature extraction process to obtain characteristics more specifically. Furthermore, we will use different feature extractions for the experiment.

## Copyright:



## References

[1] Guo, Bin & Ding, Yasan & Yao, Lina & Liang, Yunji& Yu, Zhiwen. (2020). The Future of False Information Detection on social media: New Perspectives and Trends. ACM Computing Surveys. 53. 10.1145/3393880.

[2] https://www.tbsnews.net/thoughts/busting-top-3-fake-news-week-173236

[3] https://en.prothomalo.com/opinion/Who-shares-fake-news

[4] https://www.kaggle.com/cryptexcode/banfakenews

[5] Deokate, Suchitra. (2019). False message Detection using Support Vector Machine learning Algorithm. International Journal for Research in Applied Science and Engineering Technology. 7. 438-444. 10.22214/ijraset.2019.7067.

[6] Chong, Miranda & Special, Lucia & Mitkov, Ruslan. (2010). Using Natural Language Processing for Automatic Detection of Plagiarism.

[7] Mugdha, Shafayat & Binte Mohammed, Marian & Salsabil, Lubaba & Anika, Afra & Marma, Piya & Hossain, Zahid &Shatabda, Swakkhar. (2021). A Gaussian Naive Bayesian Classifier for False message Detection in Being Procrastination 81-33-4367-2_28.

[8] Ahmed, Hadeer & Traore, Issa & Saad, Sherif. (2017). Detection of Online False message Using N-Gram Analysis and Machine Learning Techniques. 127-138. 10.1007/978-3-319-69155-8_9.

[9] Mugdha, Shafayat & Binte Mohammed, Marian & Salsabil, Lubaba & Anika, Afra & Marma, Piya & Hossain, Zahid &Shatabda, Swakkhar. (2021). A Gaussian Naive Bayesian Classifier for False message Detection in Bangladeshi. 10.1007/978-981-33-4367-2_28.

[10] Sharaff, Aakanksha & Gupta, Harshil. (2019). Extra-Tree Classifier with Metaheuristics Approach for Email Classification. 10.1007/978-981-13-6861-5_17.

[11] B.Suganthi, K.Manohari "FAKEDETECTOR: Effective False message Detection with Passive Aggressive Algorithm", 2021, Science, Technology and Development

[12] Yerlekar, Prof. (2021). False message Detection using Machine Learning Approach Multinomial Naive Bayes Classifier. International Journal for Research in Applied Science and Engineering Technology. 9. 1304-1308. 10.22214/ijraset.2021.34509.

[13] Granik, Mykhailo, and Volodymyr Mesyura. "False message detection using naive Bayes classifier." *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*. IEEE, 2017.

[14] Mr.R.Praveen Kumar, P.Dineshwaran, Ravikumar R, "False message Detection Using Machine Learning", International Research Journal of Engineering and Technology (IRJET), 2019, pp.486-488.

[15] M. G. Hussain, M. Rashidul Hasan, M. Rahman, J. Protim and S. Al Hasan, "Detection of Bangla False message using MNB and SVM Classifier," 2020 International Conference on Computing, Electronics & Communications Engineering (accessDhaka20, pp. 81-85, doi: 10.1109/iCCECE49321.2020.9231167.

[16] Deokate, Suchitra. (2019). False message Detection using Support Vector Machine learning Algorithm. International Journal for Research in Applied Science and Engineering Technology. 7. 438-444. 10.22214/ijraset.2019.7067.

[17] F. Islam et al., "Bangladeshi False message Detection," 2020 IEEE 10th International Conference on Intelligent Systems (IS), 2020, pp. 281-287, doi: 10.1109/IS48319.2020.9199931.

[18] Mugdha, Shafayat & Binte Mohammed, Marian & Salsabil, Lubaba & Anika, Afra & Marma, Piya & Hossain, Zahid &Shatabda, Swakkhar. (2021). A Gaussian Naive Bayesian Classifier for False message Detection in Bangladeshi. 10.1007/978-981-33-4367-2_28.

[19] N. Smitha and R. Bharath, "Performance Comparison of Machine Learning Classifiers for False message Detection," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 696-700, doi: 10.1109/ICIRCA48905.2020.9183072.

[20] Vijayaraghavan, Sairamvinay, et al. "False message detection with different models." arXiv preprint arXiv:2003.04978 (2020).