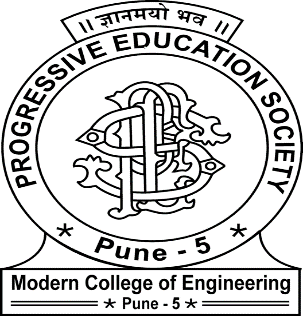
## Progressive Education Society’s MODERN COLLEGE OF ENGINEERING

Pune 411005.



## A PROJECT REPORT ON

FAKE NEWS DETECTION

## By

***SHRADDHA BHOITE (35009) POOJA JADHAV(35033) PRAJAKTA KANK(35040) SANKET KURLE(35051)***

Under the guidance of

## Prof. Yogita Fatangare

PES’s Modern College of Engineering,

## Progressive Education Society's Modern College of Engineering, Pune-05

**2021-22**

Certificate



This is to certify that, project entitled **Fake news detection**, submitted by Shraddha Bhoite , Pooja Jadhav, Prajakta Kank, Sanket Kurle is record of bonafide work carried out by them, under the supervision of Prof. Yogita Fatangre, in fulfillment of the requirement for the award of the T.E. of Bachelor of Engineering in **Information Technology,** Savitribai Phule Pune University.

Prof.Ms.Yogita Fatangare Prof. Mrs. S. D. Deshpande GUIDE HOD (IT)

Date: 02-06-2022

Place: Pune

# ACKNOWLEDGEMENT

Today on completion of this Project report, the persons need to thank the most who have helped me throughout the making of this Seminar report, and without whose help, it would not have seen the light of the day. Primarily, I am Thankful to Mrs. Yogita Fatangare for their constant motivation and support I truly appreciate and value their esteemed guidance and encouragement from the beginning to the end of this Project.

I submit my gratitude and sincere thanks to our HOD Dr. Mrs. S.D. Deshpande for their unwavering moral support and motivation during the entire course of the work.

I would also like to thank our Principal Dr. Mrs. K.R. Joshi who encouraged us and created a healthy environment for all of us to learn in best possible way.

Shraddha Bhoite

Pooja Jadhav

Prajakta Kank

Sanket Kurle

Semester-VI(B.E. in IT)

P.E.S’s Modern College Engineering, Pune

# ABSTRACT

A lot of things read online, especially in the social media feeds, may appear to be true, often it's not. Fake news is news, stories created to deliberately mislead or deceive readers. Usually, these stories are created to influence people's views, push a political agenda or cause confusion. Spreading fake news always delivers a bad and negative impact on society. This paper mainly focuses on machine learning classification techniques for detecting fake news. This paper focuses on using different applications of NLP (Natural Language Processing) techniques and various machine learning methods like logistic regression, naive Bayes classifier, Support Vector Machine (SVM). for differentiating the fake news.[1]

The objective of the paper is to detect whether the given news is true or fake.

Keywords: Fake news, Machine Learning, NLP, SVM, Logistic Regression, Naïve Bayes classifier**.**

**CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter No.** | | **Name of Chapter** | **Page**  **Number** |
| 1. | | Introduction  Introduction to Fake News Detection. Motivation behind seminar topic selection   * 1. Aim and Objective(s) of the work | 1 |
| 2. | | Literature Survey | 4 |
| 3. | | Motivation, purpose and scope and objective of | 9 |
| 4. | . | Details of design/technology/Analytical and/or  experimental work | 10 |
| 5. | . | Result | 17 |
| 6. | 2. | Conclusion | 20 |
| 7. | | Bibliography/References (in IEEE Format), | 21 |

# List of figures

|  |  |  |
| --- | --- | --- |
| **Figure Number** | **Figure Title** | **Page Number** |
|  |  |  |

**List of Table**

|  |  |  |
| --- | --- | --- |
| **Table Number** | **Table Title** | **Table Number** |
|  |  |  |

1. **INTRODUCTION 1.1Introduction to Fake News Detection**

* Fake news is a type of propaganda that consists of deliberate disinformation or hoxes

spread via online social media platforms like WhatsApp, Facebook, Instagram, and Twitter. Fake news is news, stories, or hoaxes created to deliberately mislead readers. Nowadays we read a lot of things online, especially in the social media feeds, which may appear to be true, often is not. Usually, these stories are created to influence people's views, push a political agenda and can often be a profitable business for online publishers. Online platforms are helpful for users because they can easily access news. But the problem is this allows cybercriminals to spread fake news through these platforms.

* The best example of fake news is that the pandemic situation occurring within the entire world. There is a variant of news articles till presently that are falsified and used merely to create confusion and disturbance inside the minds of individuals and to misguide their minds to believe that false news.[1]
* So, there is a need for machine learning classifiers that can detect fake news. By considering this project's ideology people can at least be able to check whether the news they have got in the front of their eyes is legit or not and the people will become more aware of the fake news circulation.
* There has been a rapid increase in the spread of fake news in the last decade, most prominently observed in the 2016 US elections [5]. Such proliferation of sharing articles online that do not conform to facts has led to many problems not just limited to politics but covering various other domains such as sports, health, and also science [3]. One such area affected by fake news is the financial markets [6], where a rumor can have disastrous consequences and may bring the market to a halt.

## Motivation behind project topic

The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to digital technology may be easily fooled. A similar problem is a fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough to acknowledge this problem takes on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news. Fake news really is a false alarm to any person as it always just misleads the readers, and the person always ends up being confused and not acting in the right way. Their daily life with their naked eyes.[2]

## Aim and Objective(s) of the work:

Aim of the Project:

* + This project aims to develop a model for detecting and classifying fake news using machine learning techniques.
  + The project will contribute to the start of a new revolution against one of the most prevalent hazards i.e. spread of fake news.
  + To create awareness is the main motive and makes citizens more responsible.[3]

The objective of the project:

* + The main objective is to detect fake news, which is a classic text classification problem with a straightforward proposition.
  + So, it is most important to build a model that can differentiate between “Real news” and “Fake news”.

.

# LITERATURE SURVEY

Table 2.1: Literature survey of Fake News Detection

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Paper Title | Author Name | Year of Publ icati on | Journ al/ Confe rence Name | Objective/ Purpose of Paper | Methodolog y/ Techniques Used | Analysis/ Result | Research Gap/ Limitation s of  Existing System | Futur e Work |
| 1 | Fake News Detecti on Using Machin e Learnin g approac hes: A systema tic Review | 1)Syed Ishfaq Manzo or  2) Dr Jimmy Singla 3) Nikita | 2019 | Resea rchGa te -  IEEE | The easy  access and exponential growth of the information available on social media networks has made it  intricate to distinguish between false and true information. Machine learning approaches are used in detection of fake and fabricated news. | Detection Methods 1)Linguistic basic Algorithm used   1. Naïve Bayes 2. Descision trees 3. SVM 4. Nueral Networks 5. Random Forest 6. XG Boost | Automatic fake news inference model named as FakeDetec torIt  is based on textual classificati on and builds a deep diffusive network model to  learn the representat ions  of new articles, authors and subjects simultaneo usly[4] | The impact and penetration of social media have dramaticall y changed reach of falsified informatio n.  The introductio n of smart gadgets and very low cost internet cost have added to its reach. In India even the remotest village has access to smart phones and internet services.  Although there are numerous advantages of these services  but it | In future a resear ch can be on the use of the unsup ervise d machi ne learni ng classi fiers for detect ion of fake news |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  | comes at a cost ,  the cost of rapid disseminati on of  falsified informatio n  along with substantiat e informatio  n |  |
| 2  . | A  Bench mark Study on Machin e Learnin g Method s for  Fake News Detecti on | 1. Junae d Younus Khan 2. Md. Tawkat Islam Khonda ker   3)  Anindy a Iqbal 4)  Sadia Afroz2 | 2019 | Resea rchGa te | The proliferation of fake news and its  propagation on social  media have  become a major concern due to its ability to create devastating impacts.  Different machine learning approaches have been attempted to detect it. | Traditional Machine Learning Models:   1. SV M, LR and Deci sion Tree Mod els 2. Naiv e Bay es Mod el 3. Naiv e Bay es Mod el   Neural Network- Based and Deep Learning Models:  1) CN N Mod el | we use 3 different datasets.  The performan ce of  LSTM-  based models greatly depends on the  length of the dataset as well as informatio n given in a news article.  LSTM-  based models have a higher probability to overcome overfitting  .  C-LSTM  have shown high  promise that | Most of the existing researches have been focused on classifying online news and social media posts.  Different methods have been proposed by different researches for deception detection | Futur e plan is to experi ment on a larger datase t to  find how the traditi onal model like Naive Bayes comp etes again st highl y comp utatio nal neural netwo rk- based model s to  detect |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | 1. LST M 2. Bi- LST M 3. HA N   Convolution al HAN | demands further attention on these models in fake news detection. Finally, perform a topic- based analysis that exposes the difficulty to correctly detect political, health and research related  deceptive news. |  | fake news |
| 3  . | Fake News Detecti on Using Machin e Learnin g | 1)Pragn esh Bugade 2)Pooja Sarode 3)Tanv ee Pimple | 2021 | IRJET | Fake news nowadays is an important aspect in the life of social media, and in the political world. Fake news really is a false alarm to any person as it always just misleads the readers,  and the  person always ends up being confused and not acting in the right way. Fake news  detection is done by using | Three Machine Learning Algorithms: 1)Logistic Regression 2)Decision Tree Classifier 3)Random Forest Classifier | Three machine learning algorithms that used  in this project; each algorithm has its own accuracy percentage when implement ed on the dataset. people can at least be able to check whether the news they have  got in the | some tools that have been developed to spot fake news that spreads through examining lexical selection that seems in headlines and different intense language structures | The Syste m certai nly needs more impro veme nts in the near future by using a flask. |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | machine learning. different datasets and models are used to detect fake news . |  | front of their eyes are legit or not and the people will become more aware of the fake news circulation  . |  |  |
| 4  . | Fake news detectio n using machin e learnin g | 1)Ms.C H.UM A DEVI 2)R.PR IYANK A 3)P.S.S UREN DRA 4)B.S.P RIYAN KA 5)CH.N  .D.L.NI KHILA  6)Habe ebullah Hussain i Syed | 2019 | Resea rchGa te | Fake news is news, stories or hoaxes created to deliberately mislead or deceive readers.  Usually, these stories are created to either influence people’s views, push a political agenda or cause confusion and can often be a profitable business for online publishers with the help of applications of NLP  (Natural Language Processing) techniques for discerning the 'fake news', which is the  misguiding | Three methods: They are Naïve Bayes, Logistic regression, and Support Vector Machine (SVM). | The task of automatic identificati on of fake news is achieved. two new fake news datasets are used. classificati on models that rely  on a  combinati on of  lexical, syntactic, and semantic informatio n, as well features representi ng text readability properties. Our best per- forming models achieved accuracies that are com- parable to  human | Deep Syntax analysis using Probabilist ic Context Free Grammars (PCFG)  have been shown to be particularl y valuable in combinatio n with n- gram methods.  These models do not consider the important qualities like word ordering and context. It is very possible that two articles that are  similar in their word | The future scope of the syste m is conne cting this metho dolog y to the intern et news which gives result s even for the test data that is not prese nt in the traini ng data sets. We can even  chang |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | news that is being published through unknown sources. |  | ability to spot fake content. It gives the appropriat e result for the test data that is present in the training datasets. | count will be completely different in their meaning. | e to  some other better classi fier to classi fy the data other than naïve bayes and logist ic regres  sion |

# MOTIVATION, PURPOSE AND SCOPE AND OBJECTIVE OF

## Scope of the project:

* This project will contribute to the start of a new revolution against one of the most prevalent hazards i.e., the spread of the Fake News. It will serve as root and branch eradication of the same[3].
* This project will help to create the next level of awareness and make the citizens more responsible. This project will help the people of a nation to take meaningful and informed decisions.

## Motivation behind the project:

* The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to digital technology may be easily fooled. A similar problem is a fraud that may happen due to spam or malicious emails and messages.
* So, it is compelling enough to acknowledge this problem takes on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news. Fake news really is a false alarm to any person as it always just misleads the readers, and the person always ends up being confused and not acting in the right way. Their daily life with their naked eyes.[3]

## The objective of the project:

* + The main objective is to detect fake news, which is a classic text classification problem with a straightforward proposition.
  + So, it is most important to build a model that can differentiate between “Real news” and “Fake news”.

# DETAILS OF DESIGN/TECHNOLOGY/ANALYTICAL AND/OR EXPERIMENTAL WORK

Fig of System Design

Training of the classifiers of machine learning is an important task. This plays an important role for the accuracy of results of these classifiers. A classifier must have to be trained in a proper way with proper data set. Different researchers have trained the machine learning classifiers to detect the fake news. The main problem that occurs while training these classifiers is that mostly the training data set in an imbalanced form. Researchers in have used the supervised machine learning classifiers for fake news detection. To train these classifiers they have used the three different models for feature extraction. Actually, these features are used to train the classifiers. These models are the TF-IDF Model, N-Gram Model, Bag of Words Model. These models extract the features from the training data set and then the classifier is trained through these features. Researchers has trained some machine learning classifiers to detect the fake news. For the training purpose, they have used a training data set. They have first removed the unnecessary words and the words are transformed to its single form. So that the training dataset that is given to these classifiers should only have the valuable data.[4]

|  |
| --- |
| **Dataset** |
| 𝖴 |
| Removing Stop Words or Stemming |
| 𝖴 |
| Test dataset & Training Dataset |
| 𝖴 |
| Training Classifier |
| 𝖴 |

Fig of methodology

Different Machine Learning Classification Algorithms that can help in detecting news is true or false:

* 1. **Support Vector Machine (SVM) :** Support vector Machine(SVM) is the classification method of supervised learning. There uses the hyperplane to splits two data class’s point with the maximum margin.
  2. **Naïve Bayes Classifier:** Naïve Bayes is a family of probabilistic algorithms that calculate the possibility that any given data point may fall into one or more of a group of categories (or not).

3. 

1. **Logistic Regression:** Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values).
2. **Decision Tree:** A decision tree is a supervised machine learning classification algorithm used to build models like the structure of a tree. It classifies data into finer and finer categories: from “tree trunk,” to “branches,” to “leaves.” It uses the if-then rule of mathematics to create sub-categories that fit into broader categories and allows for precise, organic categorization.
   * + Dataset Description

In Case of Social Media we have twiiter datasets of popular tweets with more than 10k retweet and comments which is -

<https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset>

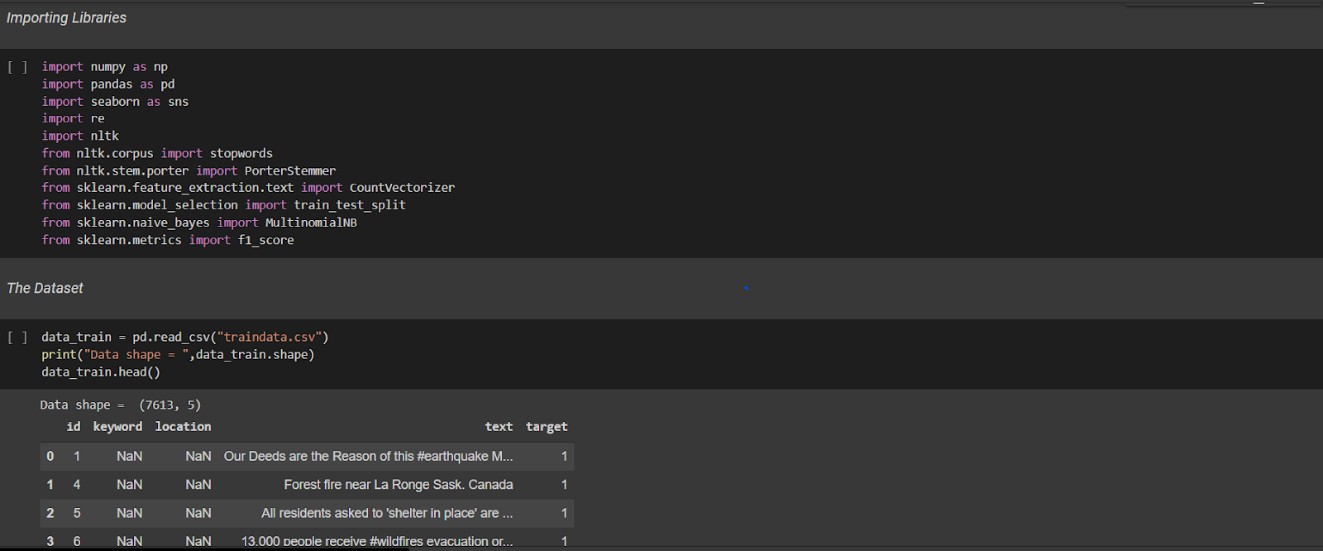
* + - Data Preprocessing from nltk.corpus import stopwords

**Stop Words:** A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages. You can find them in the nltk\_data directory.

from sklearn.feature\_extraction.text import CountVectorizer

**CountVectorizer** is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for using in further text analysis).

from sklearn.naive\_bayes import MultinomialNB



* + - Data Model Selection & Model Building

## Nltk corpus

**What is a corpus?**

A corpus can be defined as a collection of text documents. It can be thought as just a bunch of text files in a directory, often alongside many other directories of text files.

**How to create wordlist corpus?**

WordListCorpusReader class is one of the simplest CorpusReader classes. It

line

* WordListCorpusReader – It is one of the simplest CorpusReader classes.
* This class provides access to the files that contain list of words or one word per
* Wordlist file can be a CSV file or a txt file having one word in each line. In

our **wordlist file**

* + we have added : geeks

for geeks

welcomes

you to nlp

* + Two arguments to give
  + directory path containing the files
  + list of filenames

**Porter Stemmer**

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding ([NLU](https://www.techtarget.com/searchenterpriseai/definition/natural-language-understanding-NLU)) and natural language processing ([NLP](https://www.techtarget.com/searchbusinessanalytics/definition/natural-language-processing-NLP)).

Stemming is a part of linguistic studies in morphology and artificial intelligence

([AI](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence)) information retrieval and extraction. Stemming and AI knowledge extract meaningful information from vast sources like [big data](https://www.techtarget.com/searchdatamanagement/definition/big-data) or the Internet since additional forms of a word related to a subject may need to be searched to get the best results. Stemming is also a part of queries and Internet [search engines.](https://www.techtarget.com/whatis/definition/search-engine)

Recognizing, searching and retrieving more forms of words returns more results. When a form of a word is recognized it can make it possible to return search results that otherwise might have been missed. That additional information retrieved is why stemming is integral to search queries and information retrieval.

When a new word is found, it can present new research opportunities. Often, the best results can be attained by using the basic morphological form of the word: the lemma. To find the lemma, stemming is performed by an individual or an algorithm, which may be used by an AI system. Stemming uses a number of approaches to reduce a word to its base from whatever inflected form is encountered.

It can be simple to develop a stemming [algorithm](https://www.techtarget.com/whatis/definition/algorithm). Some simple algorithms will simply strip recognized prefixes and suffixes. However, these simple algorithms are prone to error. For example, an error can reduce words like *laziness* to *lazi* instead of *lazy*. Such algorithms may also have difficulty with terms whose inflectional forms don't perfectly mirror the lemma such as with *saw* and *see*.

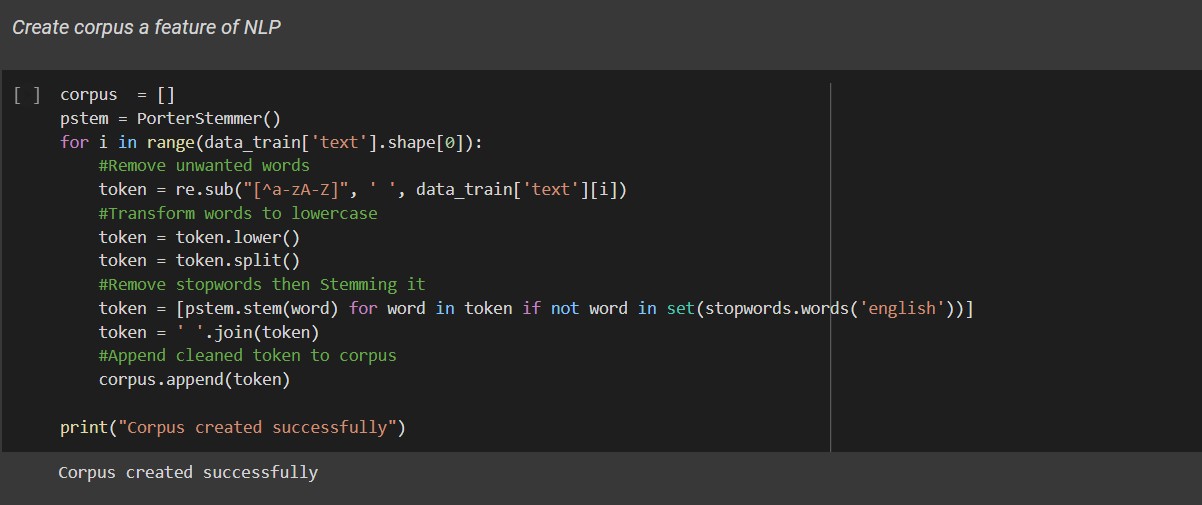
Examples of stemming algorithms include:

Lookups in tables of inflected forms of words. This approach requires all inflected forms be listed.

Suffix strippi . Algorithms recognize known suffixes on inflected words and remove them.

[Lemmatization](https://www.techtarget.com/searchenterpriseai/definition/lemmatization). This algorithm collects all inflected forms of a word in order to break them down to their root dictionary form or lemma. Words are broken down into a part of speech (the categories of word types) by way of the rules of grammar.

Stochastic models. This algorithm earns from tables of inflected forms of words. By understanding suffixes, and the rules by which they are applied, an algorithm can stem new words.

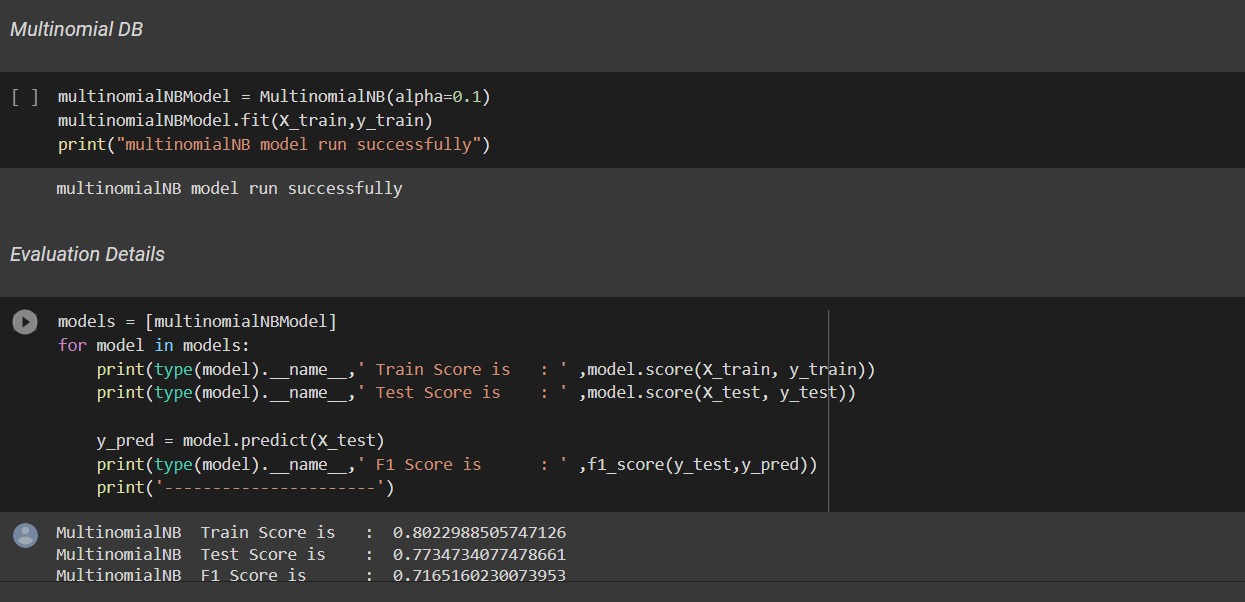


# Result

Naive Bayes Classifier Algorithm is a family of probabilistic algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair feature.

Bayes theorem calculates probability P(c|x) where c is the class of the possible outcomes and x is the given instance which has to be classified, representing some certain features. **P(c|x)=P(x|c)\*P(c)/P(x)**

Naive Bayes are mostly used in natural language processing (NLP) problems. Naive Bayes predict the tag of a text. They calculate the probability of each tag for a given text and then output the tag with the highest one.



**MultinomialNB Train Score: 80.22% MultinomialNB Test Score: 77.34% MultinomialNB Train Score: 71.65%**

# CONCLUSION

Hence, this project will make people more informed. It will contribute to starting a new revolution against Fake News. It will serve a root and branch eradication of the same and people will become more aware of the fake news circulation. Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.

# BIBLIOGRAPHY/REFERENCES (IN IEEE FORMAT)

1. S. I. Manzoor, J. Singla, and Nikita, “Fake news detection using machine learning approaches: A systematic review,” *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, no. June, pp. 230–234, 2019, doi: 10.1109/ICOEI.2019.8862770.
2. A. Nagaraja, K. N. Soumya, P. Naik, A. Sinha, and J. V. Rajendrakumar, “Fake news detection using machine learning methods,” *ACM Int. Conf. Proceeding Ser.*, pp. 185–192, 2021, doi: 10.1145/3460620.3460753.
3. S. Gupta and P. Meel, “Fake news detection using passive-aggressive classifier,” *Lect. Notes Networks Syst.*, vol. 145, no. March, pp. 155–164, 2021, doi: 10.1007/978-981-15-7345-3\_13.
4. A. A. A. A. Et al., “Detecting Fake News using Machine Learning: A Systematic Literature Review,” *Psychol. Educ. J.*, vol. 58, no. 1, pp. 1932–1939, 2021, doi: 10.17762/pae.v58i1.1046.