A Project Report

on

Fake News Detection Using Machine Leaning

Submitted in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

GALGOTIAS UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Dr. Ajay Shanker Singh
Assistant Professor

Submitted By

S.no	Name	AdmissionNo	EnrollmentNo
1.	Rishab Bharti	18SCSE1010347	18021011579
2.	Kunal Singh Teotia	18SCSE1010516	18021011744

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

December, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **"Fake News Detection Using Machine Learning"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of July, 2021 to December and 2021, under the supervision of Dr.Ajay Shanker Singh, Assistant Professor, Department of Computer Science and Engineering, Galgotias University, Greater Noida

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

Kunal Singh Teotia, 18SCSE1010516
Rishab Bharti, 18SCSE1010347

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr.Ajay Shanker Singh
Assistant Professor

CERTIFICATE

The Final Project Viva-Voice examination of Student 1 Name: Rishab Bharti, Admission No: 18SCSE1010347, Student 2 Name: Kunal Singh Teotia, Admission No: 18SCSE1010516 has been held on _____ and his/her work is recommended for the award of Bachelor of Technology in Computer Science and Engineering.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021
Place: Greater Noida

Acknowledgement

We would like to express our deep and sincere gratitude to our reviewers for positive response towards our project. We are extremely grateful to our project's guide, Dr. Ajay Shanker Singh Sir for providing vital guidance throughout the project. The team members helped each other throughout the project and project has been completed successfully with expected outputs.

Rishab Bharti (18021011579)

Kunal Singh Teotia (18021011744)

Abstract

With the popularity of mobile technology and social media growing, information is readily available. Mobile App and social media platforms have overturned traditional media in the distribution of news. Alongside the increment in the utilization of online media stages like Facebook, Twitter, and so forth news spread quickly among a large number of clients with an extremely limited ability to focus time. Machine learning and Knowledge-based approach and approach are the two techniques utilized for investigating the truthiness of the content. Public and private assessments on a wide assortment of subjects are communicated and spread persistently through various online media. Most methodologies are utilized, for example, regulated AI. The spread of phony news has extensive results like the making of one-sided feelings to influencing political race results to support certain applicants. Additionally, spammers utilize engaging news features to produce income utilizing notices through click baits. In this paper, we intend to perform a parallel grouping of different news stories accessible online with the help of thoughts identifying with Artificial Intelligence, Natural Language Processing, and Machine Learning. The result of the project determines the fake news detection for social networks using machine learning and also checks the authenticity of the publishing news website.

In our modern era where the internet is everywhere, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to winning election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits.

In this project, we aim to perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news. This project could be practically used by any media company to automatically predict whether the circulating news is fake or not.

Although many attempts have been made to solve the problem of fake news, any significant success is yet to be seen. With huge amounts of data collected from social media websites like Facebook, Twitter, etc., the best models improve every day. With the use of deep neural networks, the future work in this field seems a lot more promising. The limitations that come packaged with this problem is that, the data is erratic and this means that any type of prediction model can have anomalies and can make mistakes. For future improvements, concepts like POS tagging, word2vec and topic modelling can be utilized. These will give the model a lot more depth in terms of feature extraction and fine-tuned classification.

Contents

List of Figures

CHAPTER 1

INTRODUCTION

With the advancement of technology, digital news is more widely exposed to users globally and contributes to the increment of spreading hoaxes and disinformation online. Fake news can be found through popular platforms such as social media and the Internet. There have been multiple solutions and efforts in the detection of fake news where it even works with artificial intelligence tools. However, fake news intends to convince the reader to believe false information which deems these articles difficult to perceive. The rate of producing digital news is large and quick, running daily at every second, thus it is challenging for machine learning to effectively detect fake news.

Internet is one of the important inventions and a large number of persons are its users. These persons use this for different purposes. There are different social media platforms that are accessible to these users. Any user can make a post or spread the news through these online platforms. These platforms do not verify the users or their posts. So some of the users try to spread fake news through these platforms. These fake news can be a propaganda against an individual, society, organization or political party. A human being is unable to detect all these fake news. So there is a need for machine learning classifiers that can detect these fake news automatically.

World is changing rapidly. No doubt we have a number of advantages of this digital world but it also has its disadvantages as well. There are different issues in this digital world. One of them is fake news. Someone can easily spread a fake news. Fake news is spread to harm the reputation of a person or an organization. It can be a propaganda against someone that can be a political party or an organization. There are different online platforms where the person can spread the fake news. This includes the Facebook, Twitter etc. Machine learning is the part of artificial intelligence that helps in making the systems that can learn and perform different actions (Donepudi, 2019). A variety of machine learning algorithms are available that include the supervised, unsupervised, reinforcement machine learning algorithms. The algorithms first have to be trained with a data set called train data set. After the training, these algorithms can be used to perform different tasks. Machine learning is using in different sectors to perform different tasks. Most of the time machine learning algorithms are used for prediction purpose or to detect something that is hidden. Online platforms are helpful

for the users because they can easily access a news. But the problem is this gives the opportunity to the cyber criminals to spread a fake news through these platforms. This news can be proved harmful to a person or society. Readers read the news and start believing it without its verification. Detecting the fake news is a big challenge because it is not an easy task (Shu et al., 2017). If the fake news is not detected early then the people can spread it to others and all the people will start believing it. Individuals, organizations or political parties can be effected through the fake news. People opinions and their decisions are affected by the fake news in the US election of 2016 (Dewey, 2016). Different researchers are working for the detection of fake news. The use of Machine learning is proving helpful in this regard. Researchers are using different algorithms to detect the false news. Researchers in (Wang, 2017) said that fake news detection is big challenge. They have used the machine learning for detecting fake news. Researchers of (Zhou et al., 2019) found that the fake news are increasing with the passage of time. That is why there is a need to detect fake news. The algorithms of machine learning are trained to fulfill this purpose. Machine learning algorithms will detect the fake news automatically once they have trained.

Implications of fake news detection:

In the discourse of not being able to detect fake news, the world would no longer hold value in truth. Fake news paves the way for deceiving others and promoting ideologies. These people who produce the wrong information benefit by earning money with the number of interactions on their publications. Spreading disinformation holds various intentions, in particular, to gain favor in political elections, for business and products, done out of spite or revenge. Humans can be gullible and fake news is challenging to differentiate from the normal news. Most are easily influenced especially by the sharing of friends and family due to relations and trust. We tend to base our emotions from the news, which makes accepting not difficult when it is relevant and stance from our own beliefs. Therefore, we become satisfied with what we want to hear and fall into these traps.

1.1 MACHINE LEARNING

Arthur Samuel in 1959, an American pioneer, termed Machine Learning that computer gaming and computing, and stated, “It gives computers the power to be told without being explicitly programmed”.

In 1997, Tom Mitchell gave a “well-posed” mathematical and relational definition that; a computer virus is alleged to be told from

- Experience E
- Task T
- Performance measure P

If its performance on task T, as measured by P, improves with experience E.

Classification of Machine Learning

Machine learning implementations are classified into three major categories, counting on the nature of the training “signal” or “response” available to a learning system which are as follows:-

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning: When an algorithm learns from example, data and associated target responses, which will incorporates numeric values or string labels, such as classes or tags; to later predict the proper response when posed with new examples comes under the category of supervised learning. This approach is indeed almost like human learning under the supervision of a lecturer .The teacher provides good examples for the scholar to memorize, and the student then derives general rules from these specific examples.

Unsupervised learning: Whereas when an algorithm learns from plain examples with none associated response, leaving to the algorithm to determine the information patterns on its own. This kind of algorithm tends to restructure the information into something else, like new features that will represent a class or a replacement series of un-correlated values. They are quite useful in providing humans with insights into the meaning of knowledge and new useful inputs to supervised machine learning algorithms.

As a form of learning, it resembles the methods humans use to work out that certain objects or events are from the identical class, like by observing the degree of similarity between objects. Some recommendation systems that you simply find on the net within the sort of marketing automation are supported this kind of learning.

Reinforcement learning: After you present the algorithm with examples that lack labels, as in unsupervised learning. However, you'll accompany an example with positive or feedback consistent with the answer the algorithm proposes comes under the category of Reinforcement learning, which is connected to applications that the algorithm must make decisions (so the product is prescriptive, not just descriptive, as in unsupervised learning), and the decisions bear consequences. Within the human world, it is rather like learning by trial and error.

Errors facilitate your learn because they need a penalty added (cost, loss of your time, regret, pain, and so on), teaching you that a particular course of action is a smaller amount likely to succeed than others. A remarkable example of reinforcement learning occurs. When computers learn to play video games by themselves.

In this case, an application presents the algorithm with samples of specific situations, like having the gamer stuck in a very maze while avoiding an enemy. The application lets the algorithm know the result of actions it takes, and learning occurs while trying to avoid what it discovers to be dangerous and to pursue survival. You will have a glance at how the corporate Google DeepMind has created a reinforcement-learning program that plays old Atari's videogames. When watching the video, notice how the program is initially clumsy and unskilled but steadily improves with training until it becomes a champion.

Semi-supervised learning: Where an incomplete training signal is given: a training set with some (often many) of the target outputs missing. There is a special case of this principle referred to as Transduction where the whole set of problem instances is understood at learning time, except that a part of the targets are missing.

1.2 EXISTING SYSTEM

There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature.

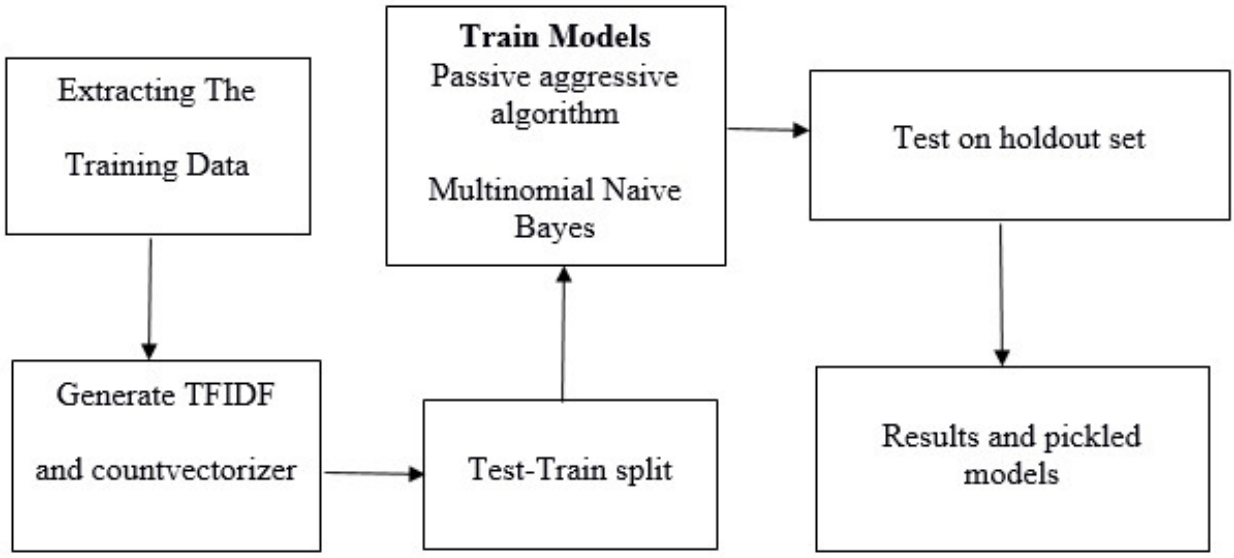
Conroy, Rubin, and Chen outlines several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech (POS) tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars (PCFG) have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora.

Feng and Hirst implemented a semantic analysis looking at 'object:descriptor' pairs for contradictions with the text on top of Feng's initial deep syntax model for additional improvement. Rubin, Lukoianova and Tatiana analyze rhetorical structure using a vector space model with similar success. Ciampaglia et al. employ language pattern similarity networks requiring a pre-existing knowledge base.

Top Five Unreliable News Sources		Top Five Reliable News Sources	
Before It's News	2066	Reuters	3898
Zero Hedge	149	BBC	830
Raw Story	90	USA Today	824
Washington Examiner	79	Washington Post	820
Infowars	67	CNN	595

1.3 PROPOSED SYSTEM

The model is build based on the count vectorizer or a tfidf matrix (i.e) word tallies relatives to how often they are used in other artices in your dataset) can help . Since this problem is a kind of text classification, Implementing a Naive Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizer vs tfidf vectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for countvectorizer or tfidf-vectorizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text dataset.



1.4 SYSTEM REQUIREMENTS

Hardware Requirements

The hardware requirements include the requirements specification of the physical computer resources for a system to work efficiently. The hardware requirements serve as the basis for a contract and for the implementation of the system

- Processor - Any Processor above 500 MHz
- Ram - 4 GB
- Hard Disk - 4 GB
- Input device - Standard Keyboard and Mouse.
- Output device - VGA and High Resolution Monitor.

Software Requirements

- Operating System: Windows 7 or higher
- Programming : Python 3.6 and related libraries
- Software Description: Python

Python is an interpreted high-level programming language for all-purpose programming. Created by Guido Van Rossum and first released in 1991, created by Guido Van Rossum. It provides constructs that alter clear programming on each tiny and huge scales. Python is a dynamic type system and automatic memory management. It can support multiple programming parameters, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many OS.

System Analysis

System analysis is a method of enquiring and understanding an existing problem system, defining requirements and analysing the best possible solution is the best way to increase the efficiency of the existing system. It is also considered to be a great problem solving methodology that will break the existing system into different component modules and evaluate each module to determine the efficiency of the system individually, by doing module wise analysis we will be able to individually increase the speed of each module.

Documentation plays a very important role in analysis of any system as it provides the target for design and development. ince it is impossible to stop the spread of fake news at once for all, we need to consider the ways in which we can prevent them. Detection and prevention of fake news exists as long as people tend to believe the fake news or the rumors. Fake news issue is being addressed by most of them within the previous number of years, especially when it was an outbreak among the people of US when the general election took place in 2016. It is hard to figure out the fake news among the bundle of real news since the fake news are generated as a shadow of real news.

However, fake news can be prevented from spreading only by showing keen knowledge toward real news. The foremost form of spreading this fake news or rumors are social media websites such as Google Plus, Facebook, Twitter and many others. Most of the people just have a habit of forwarding anything which they receive over the internet, they don’t even verify whether it is fake or real.

1.5 FUNCTIONAL REQUIREMENTS

A function is a set of all inputs, structure and output. A functional requirement is a function that defines the function of the concerned system. It includes specific functionality such as technical detail and data manipulation etc. documentation of the operation and the activities that a system should perform.

It contains the information about the implementation phase and the functional requirement. **Data Collection-** The dataset used I this project is UCI, which is known as UCI SMS span and the dataset has 5575 SMS classified into spam and ham.

Data Pre-processing- Pre-processing is done before training and testing the data so that the data scientist finds it easy to understand the dataset and feed it into the Doc2Vec model. By doing this, we are able to obtain a result that is more precise.

Dataset Splitting- The data should to split into three sub division that are Training, testing and validation.

Model Training- After the pre-processing done by the data scientist, which he has processed into train and test the data is ready for model training. This process starts with feeding the algorithm with the training data, an algorithm will process the data and will display an output through which we can be able to determine the accuracy of the project.

Model evaluation and testing- The goal of this process is to develop a simple model that would be able to target a value fast and with high accuracy rate. This can be achieved by model tuning.

1.6 NON-FUNCTIONAL REQUIREMENTS

While the functional requirement specifies the behavior of the system, non-functional requirements are concerned with how the system is supposed to be. It contains the details of how to implement these non-functional requirements. Documentation plays a very important role in analysis of any system as it provides the target for design and development. Since it is impossible to stop the spread of fake news at once for all, we need to consider the ways in which we can prevent them. Detection and prevention of fake news exists as long as people tend to believe the fake news or the rumors.

It also tells the details about the criteria that are used to evaluate the operation of the system, instead of a specific behavior. In other words they are also called as quality attributes and non-behavioral requirements.

Since it is impossible to stop the spread of fake news at once for all, we need to consider the ways in which we can prevent them. Detection and prevention of fake news exists as long as people tend to believe the fake news or the rumors. Fake news issue is being addressed by most of them within the previous number of years, especially when it was an outbreak among the people of US when the general election took place in 2016.

It is hard to figure out the fake news among the bundle of real news since the fake news are generated as a shadow of real news. However, fake news can be prevented from spreading only by showing keen knowledge toward real news. The foremost form of spreading this fake news or rumors are social media websites such as Google Plus, Facebook, Twitter and many others.

Most of the people just have a habit of forwarding anything which they receive over the internet, they don’t even verify whether it is fake or real.

The following is the details of non-functional requirements.

- Response time: A function is a set of all inputs, structure and output. A functional requirement is a function that defines the function of the concerned system. It includes specific functionality such as technical detail and data manipulation etc. documentation

of the operation and the activities that a system should perform. In a view to acquire a standard accuracy rate for our data, we applied a Naive Bayes classifier.

- Maintainability: After the pre-processing done by the data scientist, which he has processed into train and test the data is ready for model training. This process starts with feeding the algorithm with the training data, an algorithm will process the data and will display an output through which we can be able to determine the accuracy of the project.
- Usability: After the pre-processing done by the data scientist, which he has processed into train and test the data is ready for model training. This process starts with feeding the algorithm with the training data, an algorithm will process the data and will display an output through which we can be able to determine the accuracy of the project.
- Availability: It also tells the details about the criteria that are used to evaluate the operation of the system, instead of a specific behavior. In other words they are also called as quality attributes and non-behavioral requirements. Since it is impossible to stop the spread of fake news at once for all, we need to consider the ways in which we can prevent them. Detection and prevention of fake news exists as long as people tend to believe the fake news or the rumors.
- Stability: It is hard to figure out the fake news among the bundle of real news since the fake news are generated as a shadow of real news. However, fake news can be prevented from spreading only by showing keen knowledge toward real news. The foremost form of spreading this fake news or rumors are social media websites such as Google Plus, Facebook, Twitter and many others.

CHAPTER 2

LITERATURE REVIEW

1. Mykhailo Granik et. al. in their paper shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it is fake news.

2. Himank Gupta et. al. gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected 400,000 tweets from HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non- spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-of-Words model. 4. They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately18%.

3. Marco L. Della Vedova et. al. first proposed a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing its accuracy up to 78.8%. Second, they implemented their method within a Facebook Messenger Chabot and validate it with a real-world application, obtaining a fake news detection accuracy of 81.7%. Their goal was to classify a news item as reliable or fake; they first described the datasets they used for their test, then presented the content-based approach they implemented and the method they proposed to combine it with a social-based approach available in the literature. The resulting dataset is composed of 15,500 posts, coming from 32 pages (14 conspiracy pages, 18 scientific pages), with more than 2, 300, 00 likes by 900,000+ users. 8,923 (57.6%) posts are hoaxes

and 6,577 (42.4%) are non-hoaxes.

4. Cody Buntain et. al. develops a method for automating fake news detection on Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets: CREDBANK, a crowd sourced dataset of accuracy assessments for events in Twitter, and PHEME, a dataset of potential rumors in Twitter and journalistic assessments of their accuracies. They apply this method to Twitter content sourced from BuzzFeed's fake news dataset. A feature analysis identifies features that are most predictive for crowd sourced and journalistic accuracy assessments, results of which are consistent with prior work. They rely on identifying highly retweeted threads of conversation and use the features of these threads to classify stories, limiting this work's applicability only to the set of popular tweets. Since the majority of tweets are rarely retweeted, this method therefore is only usable on a minority of Twitter conversation threads. In his paper, Shivam B. Parikh et. al. aims to present an insight of characterization of news story in the modern diaspora combined with the differential content types of news story and its impact on readers. Subsequently, we dive into existing fake news detection approaches that are heavily based on text-based analysis, and also describe popular fake news datasets. We conclude the paper by identifying 4 key open research challenges that can guide future research. It is a theoretical Approach which gives Illustrations of fake news detection by analyzing the psychological factors.

5. Paper Name: - Fake News Detection on Social Media: A Data Mining Perspective. Author: - Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang and Huan Liu. In this paper to detect fake news on social media, a data mining perspective is presented that includes the characterization of fake news in psychology and social theories. This article looks at two main factors responsible for the widespread acceptance of fake messages by the user which is naive realism and confirmatory bias. It proposes a general two-phase data mining framework that includes 1) feature extraction and 2) modeling, analyzing data sets, and confusion matrix for detecting fake news.

6. Paper Name: - Media Rich Fake News Detection: A Survey. Author: - Shivam B. Parikh and Pradeep K. Atrey. Social networking sites read news mainly in three ways: The (multilingual) text is analyzed with the help of computational linguistics, which semantically and systematically focuses on the creation of the text. Since most publications are in the form of text, a lot of work has been done on analyzing them. Multimedia: Several forms of media are integrated into a single post. This can include audio, video, images, and graphics. This is very attractive and attracts the viewer's attention without worrying about the text. Hyperlinks allow the author of the post to refer to various sources and thus gain the trust of viewers. In practice, references are made to other social media websites, and screenshots are inserted.

7. Paper Name: - Fake News Detection using Naive Bayes classifier. Author: - Mykhailo Granik and Volodymyr Mesyura. This article describes a simple method of fake news detection based on one of the artificial intelligence algorithms called the Naive Bayes classifier. The goal of the research is to examine how this particular method works for the particular problem with a manually labeled (fake or real) dataset and to support the idea of using machine learning to detect fake news. The difference between this article and articles on similar topics is that this article is extensively based on a Naive Bayes classifier which is used for the classification of fake news and real news; In addition, the developed system was tested on a relatively new data set, which provided the opportunity to evaluate its performance against the most recent data.

CHAPTER 3
DATA FLOW AND ARCHITECTURE DIAGRAMS

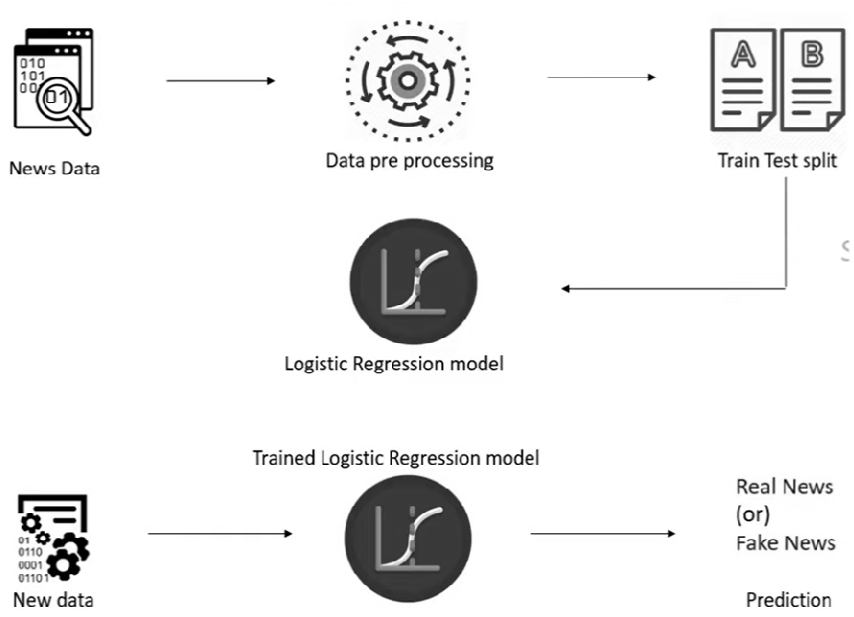


Fig. 1 – Data Flow Model

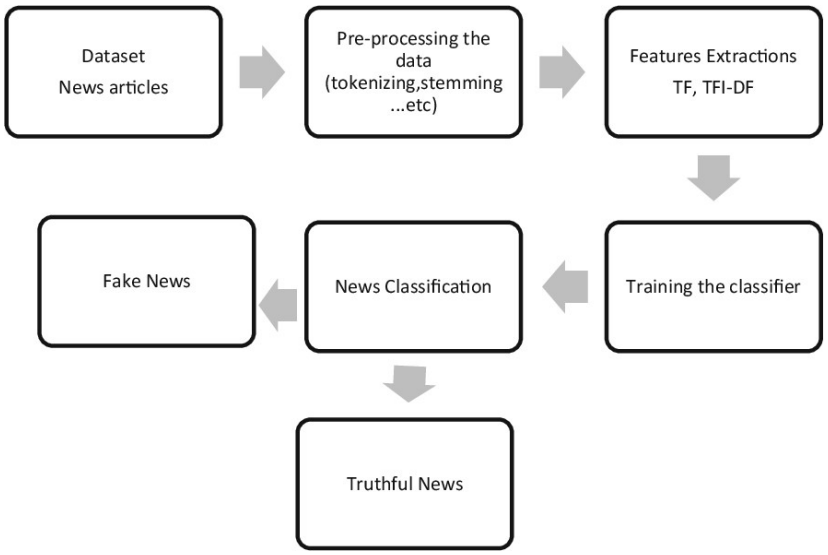


Fig. 2 – Architecture Diagram

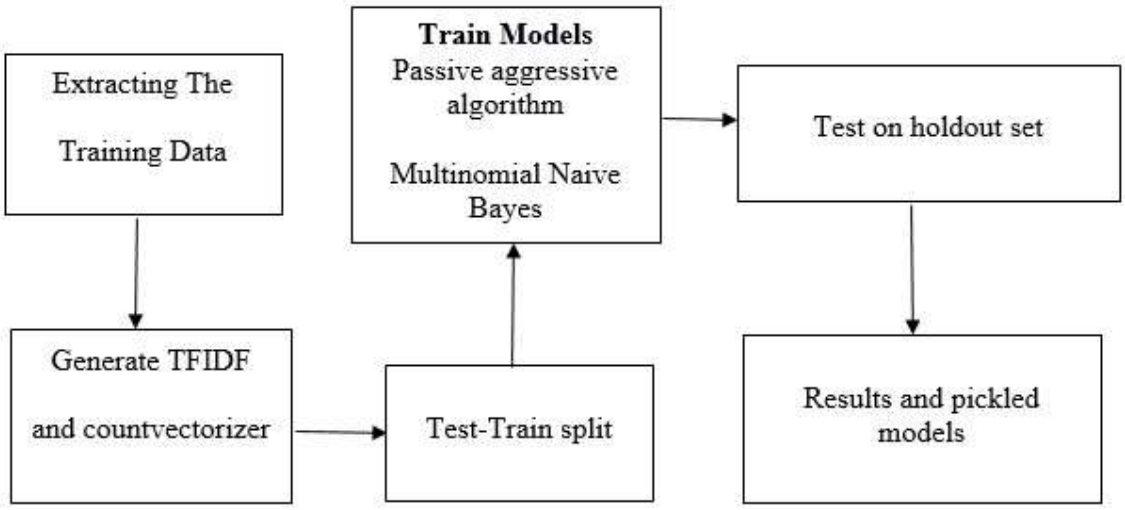


Fig. 3 – Use Case Diagram

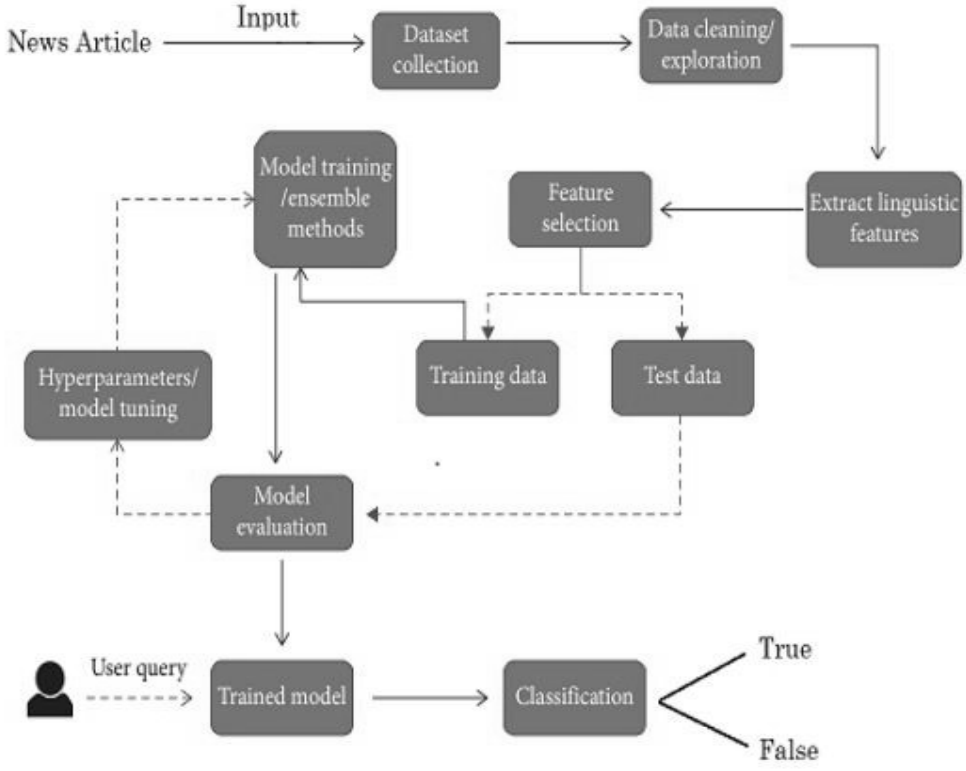


Fig. 4 – Flow of module in fake news detection system

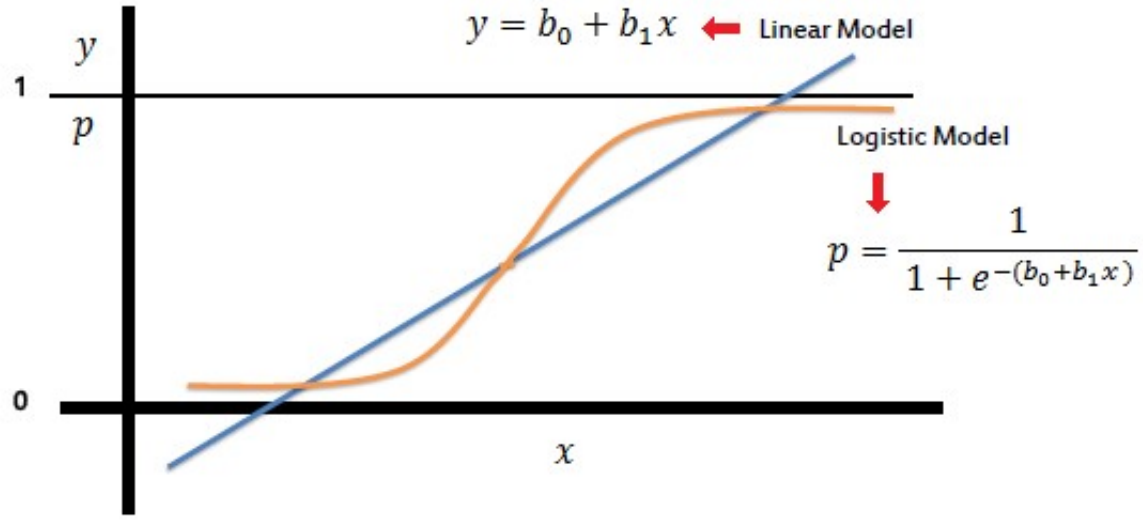
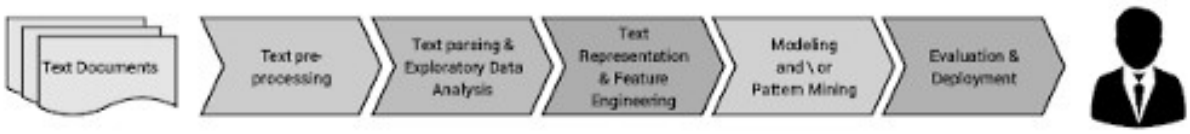


Fig. 5 – Logistic Regression

CHAPTER 4
MODULES DESCRIPTION



1. **Natural Language Processing:**

   Machine learning data only works with numerical features so we have to convert text data into numerical columns. So we have to preprocess the text and that is called natural language processing.

   In-text preprocess we are cleaning our text by steaming, lemmatization, remove stopwords, remove special symbols and numbers, etc. After cleaning the data we have to feed this text data into a vectorizer which will convert this text data into numerical features.

2. **Dataset:**

   I downloaded these datasets from Kaggle. There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news. Both datasets have a label column in which 1 for fake news and 0 for true news. We combined both datasets using pandas built-in function.

```
True_news.head()
```

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	0
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	0
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	0

```
Fake_news.head()
```

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1

In this Dataset there are no missing values otherwise we have to remove that information or we have to impute some value.

Our final dataset is balanced because both categories have the approximate same no. of examples.

3. Cleaning Data:

We can't use text data directly because it has some unusable words and special symbols and many more things. If we used it directly without cleaning then it is very hard for the ML algorithm to detect patterns in that text and sometimes it will also generate an error. So that we have to always first clean text data. In this project, we are making one function 'cleaning_data' which cleans the data.

Examples :

Stay, Stays, Staying, Stayed — -> Stay

House, Houses, Housing — -> House

Stop words: words that occur too frequently and not considered informative

Examples :

{'the', 'a', 'an', 'and', 'but', 'for', 'on', 'in', 'at' …}

4. **Split the data:**

   Splitting the data is the most essential step in machine learning. We train our model on the trainset and test our data on the testing set. We split our data in train and test using the train_test_split function from Scikit learn.

   We split our 80% data for the training set and the remaining 20% data for the testing set.

5. **Tfidf Vectorizer:**

   Tfidf-Vectorizer : (Term Frequency * Inverse Document Frequency)

   a. Term Frequency: The number of times a word appears in a document divided by the total number of words in the document. Every document has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

   b. Inverse Document Frequency: The log of the number of documents divided by the number of documents that contain the word *w*. Inverse data frequency determines the weight of rare words across all documents in the corpus.

$$idf(w) = log(\frac{N}{df_t})$$

   c. Finally Tfidf vectorizer:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

This vectorizer is already predefined in Scikit Learn Library so we can import by:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

Now fit this vectorizer on our training dataset and transform its values on the training and testing dataset with respect to the vectorizer.

```
vec_train_data = vectorizer.fit_transform(train_data)
```

```
vec_train_data = vec_train_data.toarray()
```

```
vec_test_data = vectorizer.transform(test_data).toarray()
```

After vectorizing the data it will return the sparse matrix so that for machine learning algorithms we have to convert it into arrays. toarray function will do that work for us.

6. **Multinomial Naive Bayes Classifier:**

   Naive Bayes: The Naive Bayes Classifier technique is based on the Bayesian theorem and is particularly suited when then high dimensional data.

   Multinomial Naive Bayes :

It is used for classification when is in discrete form. It is very useful in text processing. Each text will be converted to a vector of word count. It cannot deal with negative numbers.

It is predefined in Scikit Learn Library. So we can import that class in our project then we create an object of Multinomial Naive Bayes Class.

1. Fit the classifier on our vectorized train data
2. When the classifier fitted successfully on the training set then we can use the predict method to predict the result on the test set.

```
from sklearn.naive_bayes import MultinomialNB
```

```
clf = MultinomialNB()
```

```
clf.fit(training_data, train_label)
y_pred  = clf.predict(testing_data)
```

7. Classification Metrics:

To check how well our model we use some metrics to find the accuracy of our model. There are many types of classification metrics available in Scikit learn

1. Confusion Matrix
2. Accuracy Score
3. Precision
4. Recall
5. F1-Score

Confusion matrix: Basically this metrics how many results are correctly predicted and how many results are not correctly predicted

Accuracy Score: It is the number of correct prediction over the total no. of predictions.

8. **Save the model:**

   After the good performance on data, we can save our model so that next time we can use it directly. ‘joblib’ and ‘pickle’ library used to save the machine learning model. From the Following step, you can save and load your model.

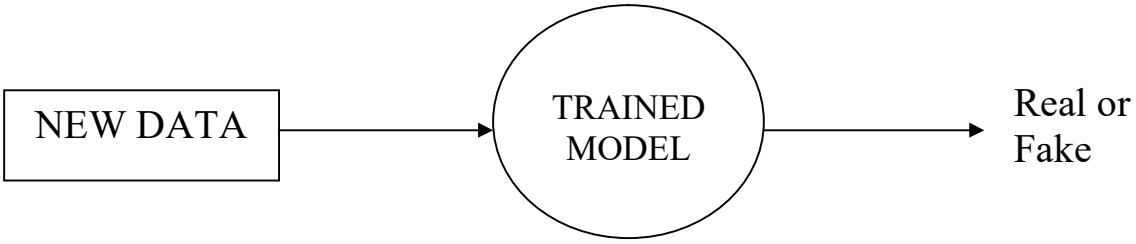
Save the Model

```
import joblib
```

```
joblib.dump(clf , 'model.pkl')
```

```
['model.pkl']
```

```
model = joblib.load('model.pkl')
```



ALGORITHMS USED

1. **Naïve Bayes Classifier:**

   This classification technique is based on Bayes theorem, which assumes that the presence of a particular feature in a class is independent of the presence of any other feature. It provides way for calculating the posterior probability.

$$P(x) = \frac{P(c) * P(c)}{P(x)}$$

   P(c|x)= posterior probability of class given predictor P(c)= prior probability of class P(x|c)= likelihood (probability of predictor given class) P(x) = prior probability of predictor.

2. **Logistic Regression:**

   It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected). Mathematically, the log odds of the outcome are modelled as a linear combination of the predictor variables.

   Odds = p/(1-p) = probability of event occurrence / probability of not event occurrence
   ln(odds) = ln(p/(1-p))
   logit(p)=ln(p/(1-p))= b0+b1X1+b2X2+b3X3....+bkXk

   Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters

of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from *logistic unit*, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

In a binary logistic regression model, the dependent variable has two levels (categorical). Outputs with more than two values are modeled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression (for example the proportional odds ordinal logistic model). The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

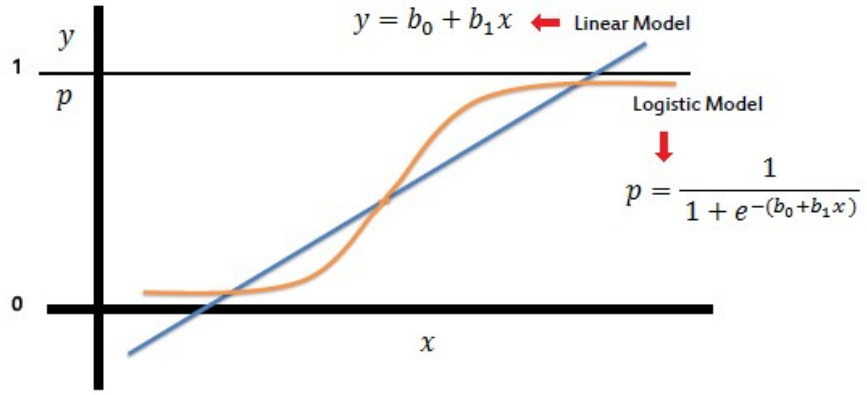


Fig. – Logistic Regression

CODE:

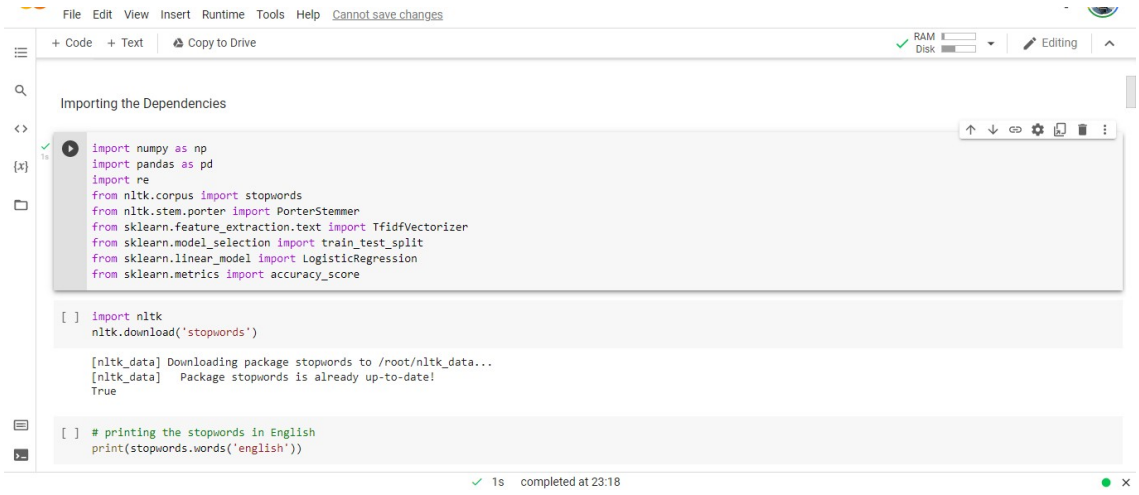


Fig 2. Importing the dependencies

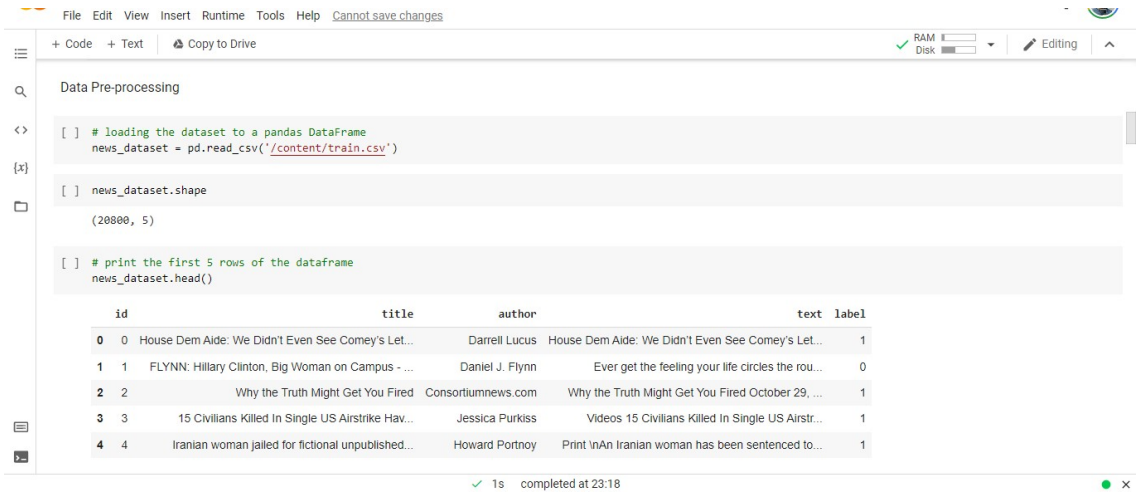


Fig 3. Data Pre processing

Fig 4. Stemming

Fig 5. Training Logistic Regression

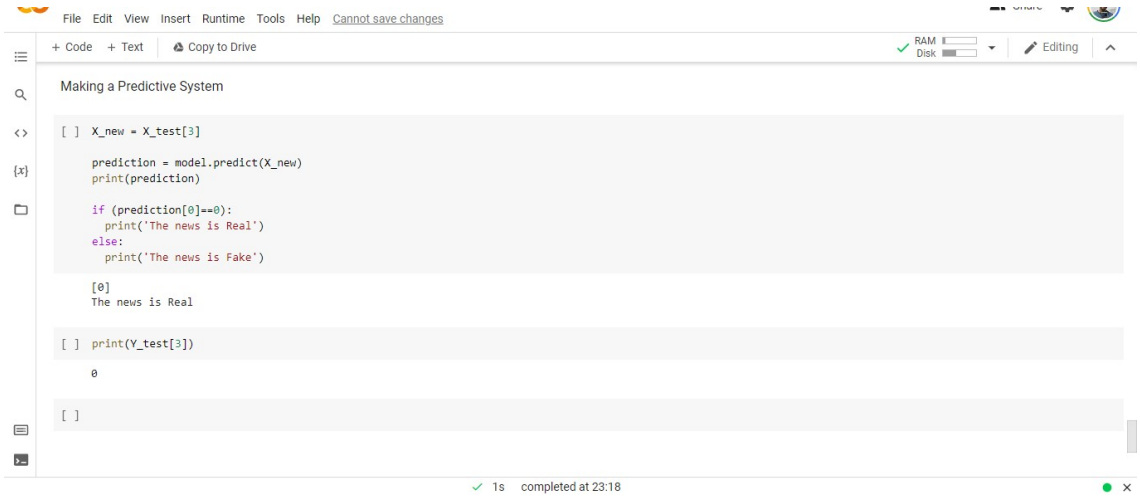


Fig 6. Making a predictive system

ADVANTAGES

Fake News Detection system will help in controlling the spread of fake news over social media. This way, we can help the people to make more informed decisions, and they are not made to think about what others are trying to manipulate to believe. A Fake News Detection system will reduce the burden to check the authenticity of the news manually and saves lots of time.

DISADVANTAGES

The accuracy of detecting fake news will not be 100%. Therefore some articles may be predicted as false.

FUTURE SCOPE

Fake news is categorized as any kind of cooked-up story with an intention to deceive or to mislead. In this paper we are trying to present the solution for fake news detection task by using Machine Learning techniques. Many events have resulted to a rise in the prominence and spread of phony news. The widespread impacts of the massive onset of fake news can be seen, humans are conflicting if not outright poor detectors of fake news. With this, endeavours are being made to automate the task of fake news detection. The most mainstream of such actions include blacklisting of sources and authors that are unreliable. Even though these tools are useful, but in order to produce a progressive complete end to end solution, we are required to represent for tougher cases where reliable sources and authors are responsible for releasing fake news. Here, the purpose of this project was to build a model that help us to recognize the language patterns that can be used to classify fake and real news with the help of ML (machine learning) techniques. The outcomes of this project shows the capability of ML to be fruitful in this task. We have tried to build a model that helps in catching many intuitive indications of real and fake news as well as in the visualization of the classification decision. Now-a-days fake news is such a big problem that it is affecting our society as well as our facts and opinions. The problem that needs to be solved can be solved using AI and Machine learning techniques.

CHAPTER 5
RESULTS AND DISCUSSION

Web is one of the incredible wellsprings of data for its clients (Donepudi, 2020). There are diverse online media stages that incorporates Facebook or Twitter that assists individuals with interfacing with others. Distinctive sort of information are likewise shared on these stages. Individuals these days really like to get to the report from these stages in light of the fact that these are not difficult to utilize and simple to get to stages. One more benefit to individuals is that these stages give choices of remarks, responds and so on These benefits draw in individuals to utilize these stages (Donepudi et al., 2020b). Be that as it may, as like their benefits, these stages are additionally utilized as the best source by the digital hoodlums. These people can get out the phony word through these stages. There is additionally an element of sharing the post or news on these stages and this element additionally demonstrates accommodating for getting out such phony word.
Individuals begin trusting in such news just as offers the news with different people groups. Scientists in (Zubiaga et al., 2018) said that it is hard to control the bogus word from getting out on these online media stages.

Anybody can be enlisted on these stages and can begin getting out the word. An individual can make a page as a wellspring of information and can get out the phony word. These stages don't confirm the individual regardless of whether he is truly trustworthy distributer. Thusly, anybody can get out the word against an individual or an association. These phony news can likewise hurt a general public or an ideological group. The report shows that it is not difficult to change individuals assessments by getting out fake word (Levin, 2017). Subsequently, there is a requirement for distinguishing these phony word from getting out so the standing of an individual, political party or an association can be saved.

Why machine learning is required to detect fake news?

Increasing use of internet has made it easy to spread the false news. Different social media platforms can be used to spread fake news to a number of persons. With the share option of these platforms, the news spread in a fast way. Fake news just not only affects an individual but it can also affect an organization or business So controlling the fake news is mandatory. A person can know the news is fake only when he knows the complete story of that topic. It

is a difficult task because most of the people do not know about the complete story and they just start believing in the fake news without any verification. The question arises here how to control fake news because a person cannot control the fake news. The answer is machine learning. Machine learning can help in detecting the fake news. Through the use of machine learning these fake news can be detected easily and automatically. Once someone will post the fake news, machine learning algorithms will check the contents of the post and will detect it as a fake news. Different researchers are trying to find the best machine learning classifier to detect the fake news. Accuracy of the classifier must be considered because if it failed in detecting the fake news then it can be harmful to different persons. The accuracy of the classifier depends on the training of this classifier. A model that is trained in a good way can give more accuracy. There are different machine learning classifiers are available that can be used for detecting the fake news that will be answered in the next question.

How machine learning classifiers are trained for detecting fake news?

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 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 have 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 classifier should only have the valuable data.

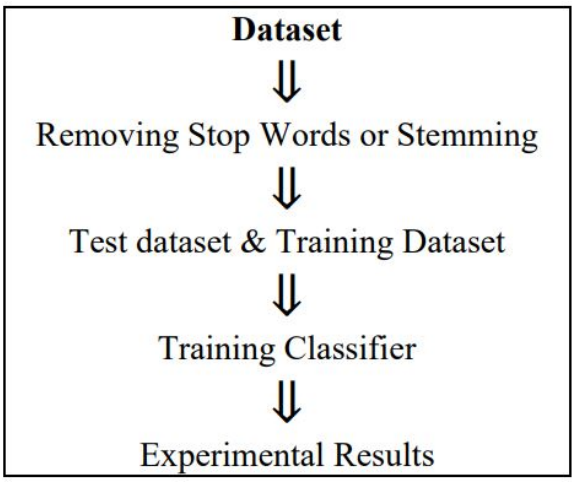


Fig. - Shows the steps that are used while training a classifier.

CHAPTER 6
CONCLUSION

Due to increasing use of internet, it is now easy to spread fake news. A huge number of persons are regularly connected with internet and social media platforms. There is no any restriction while posting any news on these platforms. So some of the people takes the advantage of these platforms and start spreading fake news against the individuals or organizations. This can destroy the repute of an individual or can affect a business. Through fake news, the opinions of the people can also be changed for a political party. There is a need for a way to detect these fake news. Machine learning classifiers are using for different purposes and these can also be used for detecting the fake news. The classifiers are first trained with a data set called training data set. After that, these classifiers can automatically detect fake news.

The data we used in our work is collected from the Kaggle.com and contains news articles from various domains to cover most of the news rather than specifically classifying political news. The learning models were trained and parameter-tuned to obtain optimal accuracy.
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. Machine learning techniques can be employed to identify the key sources involved in spread of fake news.

REFRENCES

1. M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2019 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2019, pp. 900-903.

2. H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2019 11th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2019.

3. M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2019 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2019, pp. 272- 279.

4. C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2019 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2019.

5. Jadhav, S. S., & Thepade, S. D. (2019). Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier, Applied Artificial Intelligence, 33(12), 1058-1068, https://doi.org/10.1080/08839514.2019.1661579

6. Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet–A deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32-44. https://doi.org/10.1016/j.cogsys.2019.12.005

7. Kaur, S., Kumar, P. & Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. Soft Computing, 24(12), 9049–9069. https://doi.org/10.1007/s00500-019-04436-y

8. Kesarwani, A., Chauhan, S. S., & Nair, A. R. (2020). Fake News Detection on Social Media using K-Nearest Neighbor Classifier. 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), Las Vegas, NV, USA, pp.1-4, https://doi.org/10.1109/ICACCE49060.2020.9154997

9. Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., & Afroz, S. (2019). A benchmark study on machine learning methods for fake news detection. Computation and Language. https://arxiv.org/abs/1905.04749

10. Rahman, M. M., Chowdhury, M. R. H. K., Islam, M. A., Tohfa, M. U., Kader, M. A. L., Ahmed, A. A. A., & Donepudi, P. K. (2020). Relationship between Socio-Demographic Characteristics and Job Satisfaction: Evidence from Private Bank

Employees. American Journal of Trade and Policy, 7(2), 65-72. https://doi.org/10.18034/ajtp.v7i2.492

11. Reis, J. C. S., Correia, A., Murai, F., Veloso A., & Benevenuto, F. (2019). Supervised Learning for Fake News Detection. IEEE Intelligent Systems, 34(2), 76-81, https://doi.org/10.1109/MIS.2019.2899143

12. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2020). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

13. Abdullah-All-Tanvir, Mahir, E. M., Akhter S., & Huq, M. R. (2019). Detecting Fake News using Machine Learning and Deep Learning Algorithms. 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp.1-5, https://doi.org/10.1109/ICSCC.2019.8843612

14. Ahmed, H., Traore, I., & Saad, S. (2019). Detection of online fake news using n-gram analysis and machine learning techniques. Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 127–138, Springer, Vancouver, Canada, 2019. https://doi.org/10.1007/978-3-319-69155-8_9

15. Ahmed, H., Traoré, I., & Saad, S. (2020). Detecting opinion spams and fake news using text classification. Secur. Priv., 1(1), 1-15. https://doi.org/10.1002/spy2.9

16. Al Asaad, B., & Erascu, M. (2019). A Tool for Fake News Detection. 2019 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, Romania, 2019, pp.379-386. https://doi.org/10.1109/SYNASC.2018.00064

17. https://www.irjet.net/archives/V7/i5/IRJET-V7I51102