

# **Fake News Identification**

Project report submitted in partial fulfillment of the requirement for the degree of  
Bachelor of Technology

In

**Computer Science and Engineering/Information Technology**

By

Anmol Rana (161359)

Under the supervision of

Dr. Hemraj Saini

To



Department of Computer Science & Engineering and Information Technology  
**Jaypee University of Information Technology Waknaghat, Solan-173234,  
Himachal Pradesh**

## Certificate

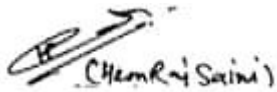
I hereby declare that the work presented in this report entitled “Fake News Identification” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** submitted in the Department of Computer Science and Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2019 to December 2019 under the supervision of **Dr. Hemraj Saini, Associate Professor, Computer Science/ IT.**

The Matter embodied in the report has not been submitted for the award of any other degree or diploma.

Anmol Rana  
Digitally signed by  
Anmol Rana  
Date: 2020.07.14  
22:20:20+05'30'

Anmol Rana (161359)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.



(Hemraj Saini)

Dr. Hemraj Saini  
Associate Professor  
Computer Science and Engineering/ Information Technology  
Dated:

## **Acknowledgement**

Working for the “Fake News Identification” project was interesting. I got to learn about different Natural Language Processing and Machine Learning techniques,

I am grateful to my project supervisor Dr. Hemraj Saini, Faculty member of Computer Science and engineering/ Information Technology Department for his guidance and advice on this project. The project would not have been a success without his enormous help and worthy experience.

I am also thankful to my parents for their cooperation and encouragement. I would also like to thank my friends who have helped me with their valuable suggestion which has been helpful in various phases of completion of project.

Although, this report has been prepared with utmost care and deep rooted interest. Even then we accept respondent and imperfection.

Anmol Rana (161359)

## **List of Abbreviations**

- ML - Machine Learning
- AI - Artificial Intelligence
- NLP - Natural Language Processing
- PAC - Passive Aggressive Classifier
- NB - Naive Bayes
- CFG - Context free grammar
- Tf-idf - Term frequency - inverse document frequency
- csv - comma-separated values
- POS - Part-of-Speech
- CNN - Convolution Neural Network
- GDU – Graphic Diffusive unit
- RNN – Recurrent Neural Network

## List of Figures

- Fig. 1. Increase in no. of fake news in Presidential Campaign months (U.S.)
- Fig. 2. Three essential components for fake news
- Fig. 3. Most Important Source of 2016 Election News
- Fig. 4. % of U.S. adults who get news on a social networking sites
- Fig. 5. Share of Visits to US News Websites by Source
- Fig. 6. Facebook engagement for top 20 political news
- Fig. 7. Increase in number of debunked fake news (between January and April)
- Fig. 8. Types of misinformation
- Fig. 9. Categories of misinformation and their prevalence following initial cases in India
- Fig. 10. Sample of tags for misinformation in three 10-day periods
- Fig. 11. Distribution of the categories of misinformation based on the novelty of propagation
- Fig. 12. Misinformation spread by widely followed members of the mainstream media
- Fig. 13. Screenshots of misinformation spread by mainstream news
- Fig. 14. Fake cover page of a magazine
- Fig. 15. % of users who believe that content they see is fake on social media
- Fig. 16. Relationships of Articles, Creators and Subjects.
- Fig. 17. Fake news on social media: from characterization to detection
- Fig. 18. Generalized scheme of the algorithm
- Fig. 19. Result of their model
- Fig. 20. Language differences in fake and legitimate content
- Fig. 21. Algorithm for feature extraction
- Fig. 22. Comparing the results they got from their models
- Fig. 23. General Approach
- Fig. 24. Training set used
- Fig. 25. Testing set used
- Fig. 26. Dependent variable associated with X\_train
- Fig. 27. Dependent variable associated with X\_test
- Fig. 28. Penalized Regression Coefficients
- Fig. 29. Frontend of the project
- Fig. 30. csv file where the data is stored
- Fig. 31. Relationships of Articles, Creators and Subjects

## **List of Tables**

Table 1. Difference between Bernoulli and Multinomial Naïve Bayes

Table 2. Comparing performances of different models used

# Table of Content

Acknowledgement .....	ii
List of Abbreviations .....	ii
List of Figures .....	iv
List of Tables .....	ii
Table of Content .....	iii
Abstract .....	iii
<b>Chapter 1</b>	
Introduction .....	1
Problem Statement .....	15
Objectives .....	16
Methodology.....	16
Organization .....	18
<b>Chapter 2</b>	
Literature Reviess.....	20
<b>Chapter 3</b>	
System Development.....	30
<b>Chapter 4</b>	
Performance Analysis .....	41
<b>Chapter 5</b>	
Conclusion.....	58
Future Scope.....	60
<b>Refrences .....</b>	<b>63</b>

## Abstract

Fake news is fairly an old phenomenon but today while the whole world is suffering from a pandemic it still constitutes of a huge part of our lives from news of bioweapon conspiracy to relation of the coronavirus to a technology like 5G. Though the concept is old but it gained friction only after the U.S.A. Presidential elections of 2016 as it can be seen in *Fig. 1* where the amount of fake news fed to the public skyrocketed in the months of November essentially before the date of elections. After this particular instance the use of fake news is severely on rise around the world to gain political advantage over the opposition parties to win elections. This practice is in direct contrast of the right of information of the people of any democratic country in the world and is also degrading the fourth pillar of democracy which is the media whose main purpose is to provide correct information to the people. But the biggest difficulty is how to define a fake news because it sometimes can be just used as satire, sometimes be a made up news and sometimes be just a propaganda put up by the government. The consumption of the news through online platforms is growing by the day and it is quite difficult to identify a reliable source so this makes it quite important to use the power of computer to deal with this problem which is in its own right a pandemic which is essential to be dealt with.

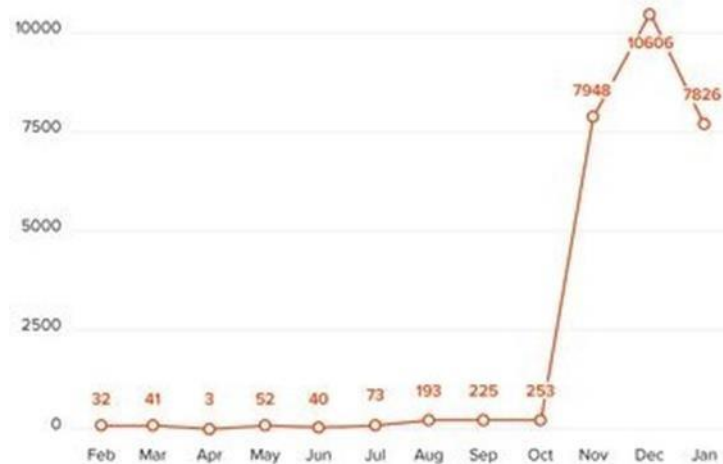


Fig. 1. Increase in no. of fake news in Presidential Campaign months (U.S.)



# Chapter 1: Introduction

## Introduction:

News is basically some kind of information which is backed up by some facts and data about a current event not a self-made intentional and sometimes unintentional story. This self-made story of any sorts to gain political power or financial prowess or sometimes just a satire is known as a fake news. Though the above statement is quite unclear but it provides a basic understanding of what it really means.

Identifying fake news is gaining a lot of attention in the community and many people are working towards it like many news channels and websites are running fact checking, there are also individual fact checking firms which are helping social media companies with fact checking and the data science community is also gaining interest in the field and trying to use their expertise to tackle this problem. As we noticed in the *fig. 1* the amount of fake news circulated was the highest in the month of November of 2016 since then that graph has seized to come down to the previous levels. At its height the top twenty fake news in 2016 were almost shared and liked on Facebook for at least 8 million times in the USA. This triggered the circumstances for the widespread use of the word fake news all around the world.

The main reason in this exponential rise in the fake news is the increase in the no. of users of the internet and due to that the increased number of users on the social media platforms. This is a problem for the whole world but it is bigger problem for developing and under developed countries where though the number of internet users are increasing majority of them are uneducated and are unable to understand the difference between what is wrong and what is right. Due to this increase in the usage of internet people are also producing fake news at a faster rate because first it is easier to publish and second it is way cheaper than other platforms. In a survey, it was found that almost 45 percent of the world has access to internet and only in India it is almost 31 percent i.e. approximately 300 million people.

Social media provides one of the most favorable platforms for the spread of fake news almost similar to that of fire. Some of the main reasons are the capability of internet to almost end the distance between the people and provide them with a very easy way to share, like and also motivate other people to join and participate on various different discussions on these topics

which are essentially fake.

There are seven different types of fake news which are: where the news body is different from the title which is also known as clickbait, sometimes correct news content is spread with incorrect context, correct information is intentionally made incorrect, sometimes it's just for satirical purpose i.e. no negative intentions, using the information totally in wrong sense to change perspective of a person on an issue, sometimes the original content is copied from an authentic news source and content that is totally false and made up from the ground.

Just like a fire need essentially three things to burn: O<sub>2</sub>, heat and something that burns i.e. fuel. Likewise, for the success of fake news it depends on three things and in absence of any of the three it is not able to hit its aspired target.

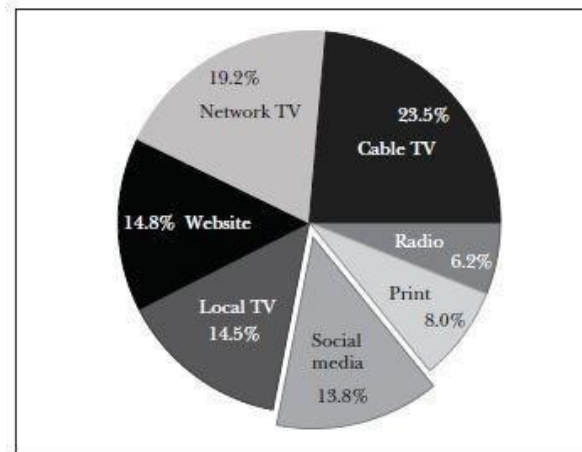


Fig. 2. Three essential components for fake news

The first essential part is tools i.e. the people which is quite easy to find, they can be either paid or can be the followers of someone and services are also easily available online. The second part is the social media which in this age of internet is readily available to most of the population and it is fairly easy to put up a fake news or some propaganda on these platforms where they can be easily circulated. The last part is the motivation behind the spreading of the fake news i.e. “why?” it was done at first place.

After the USA presidential elections were conducted there were many survey conducted on the influence of social media on people and amount of news people consume on social media platforms. Some similar kind of survey were conducted by the Pew Research Center and one of the questions asked in the survey was “What was the main source of news about the 2016

USA Presidential elections?” and we can see the result in *fig. 3* where almost 14 percent of people in the United States of America got their news from social media platforms like Reddit, Instagram, Twitter, Facebook etc. and almost 15 percent got them from other online services like news blogs and other websites which adds up to 30 percent of the US population getting their news from some kind of online platform where the amount of fake news is highest than the other traditional sources of news like television, print media or the radio.

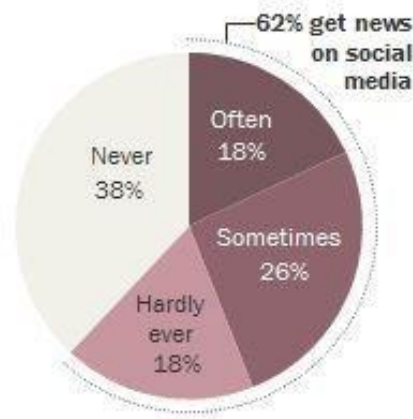


**Fig. 3. Most Important Source of 2016 Election News**

One of the other survey questions was “How often do people get their news from any social media platform?” And the result can be seen in *fig. 4* where 62 percent of people get their daily news from any social media platform which is 6 in every 10 Americans get their news on social media and another figure here is that 18 percent of people do this often i.e. they most often get their daily news from online social media platforms. But, the main problem starts after this where the people who believe what they see is the truth often share these fake news articles with their friends and families and since the people receiving tend to trust the person sharing them often start believing that the news they shared is the truth. This above mentioned situation is also called echo chamber effect where the effect of the fake news shared on social media is often intensified due to the people sharing and amount of engagement i.e. likes, comments etc. the news gets due to which people tend to believe that the news shared with them is correct and the source of news is also correct. But, sometimes human psychology also plays a role in believing that news is correct because most often the fake news is related to something that often confirms our apprehension, since as a species we are always looking out for potential danger for our survival. There was a research done where in a room of people everyone was smiling and only one person was frowning and one person had to choose the

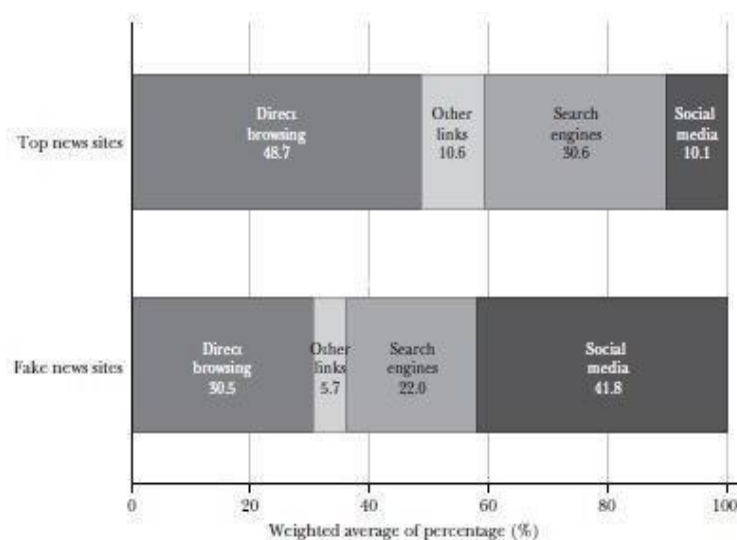
person who was frowning out of all the people and most often they were able to choose the correct option which shows how humans are able to sense impending danger to their survival, this experiment is called face in the crowd experiment.

*% of U.S. adults who get news on a social networking site ...*



**Fig. 4. % of U.S. adults who get news on a social networking sites**

In fig. 5, it is shown from which platform most often the two different news websites i.e. fake and real are accessed from, in this figure 690 real USA news websites and 65 websites with fake news were used as the dataset and it easily be seen that the fake news websites were accessed most often by any social media platform and the real news websites were directly accessed by the person on a browser.



**Fig. 5. Share of Visits to US News Websites by Source**

Another, very important figure can be seen below from where it can be deduced that how there is a sudden increase in the engagement, which is basically any reaction, comment etc. to a post on Facebook, of the top news related to elections in the year 2016 which were conducted in November. Nearly, 9 million people either liked, commented or shared any piece of fake news which could have also come in contact with other people. These kind of strategies are intensively used now a days in many countries taking inspiration from the USA 2016 presidential elections to come in power. Even in India due to intensive use of social media especially WhatsApp in the 2019 lok Sabha elations they were famously called “India’s first WhatsApp elections”. People can easily be pursued using social media and it is more frequent in countries like India, Brazil where vast amount of population is not properly educated but developed countries like United Kingdom and Australia were not immune to this and similar trend were seen in these countries as well during the time of elections. And countries like China are using fake news as a way to lie to their citizens and showing themselves in good light to brainwash them in believing something which is not true about their government. This is emerging as really big problem around the world as fake news is used as a weapon against the citizens of different country which is in direct contrast of their one of the most important right i.e. the right to information and people are also getting very confused what and what not to believe. According to a survey fake news has left about 64 percent of Americans disoriented about the most simple of the facts and 60 percent are not able to differentiate between which news is fake news and which is a real news.

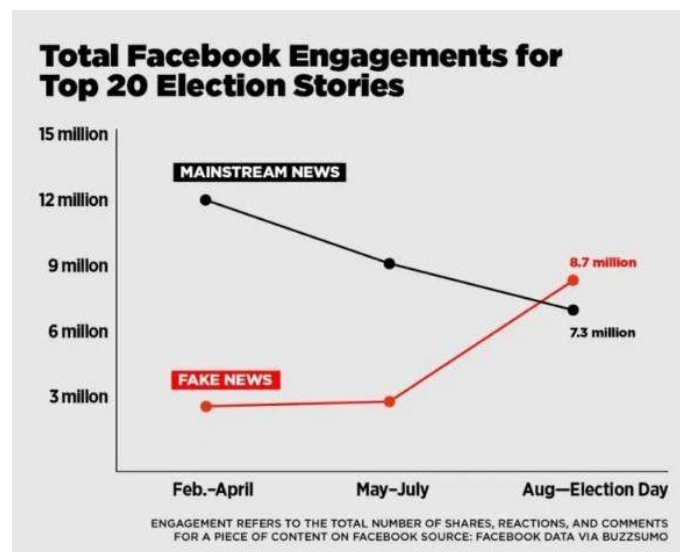


Fig. 6. Facebook engagement for top 20 political news

## Fake news and Coronavirus

Fig. 7 shows a study conducted by the University of Michigan between January and April 2020 which shows the number of exposed fake news which were in circulation in India and it can be seen sudden increase in the months of March where the Covid-19 started showing presence in India. Especially after the announcement of Janta Curfew by the Prime Minister of India the rise in exposed fake news just exploded rising from just two on the twentieth of January to fifty three in the last week of March. But this figure only show the news which were exposed but doesn't show all the fake news in circulation which was not found out by any of the firms which are working on debunking fake news from social media platforms especially WhatsApp and Facebook and which have affected lives of so many people.

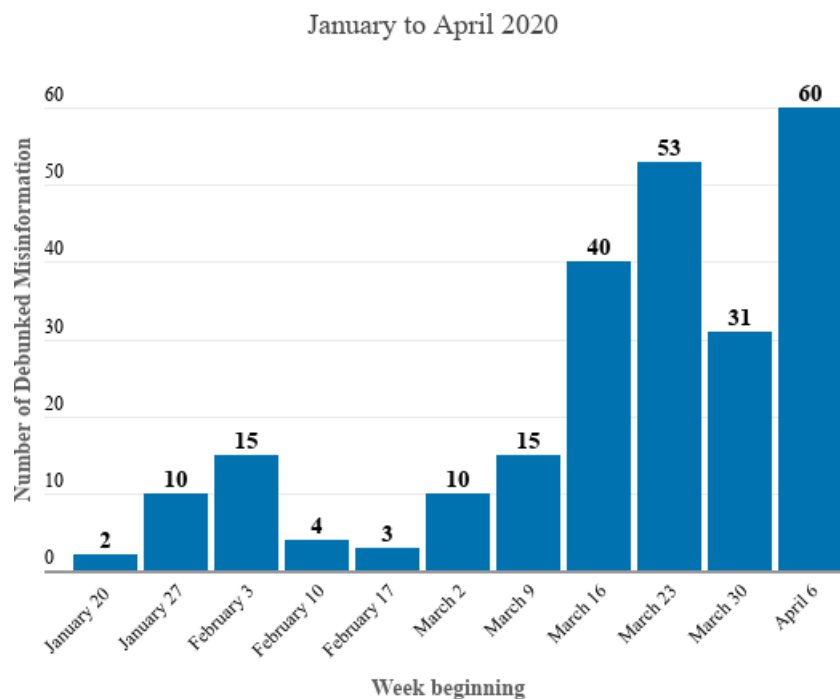


Fig. 7. Increase in number of debunked fake news (between January and April)

Though, the whole world is suffering from a pandemic like Covid-19 the amount of fake news in circulation has just become a part of our life from news of bioweapon conspiracy surrounding the pandemic to even drawing relation between a disease to a technology like 5G and even spreading fake news relating to a person

The fake news spreading in the community through different social media platforms was divided into seven different categories related to the Covid-19 which were related to a referencing a certain religion, related to preventions and vaccine related fake news, related to

environment, related to casualty and the number of people suffering, related to shortages of products to increase the sales i.e. panic buying, related to the any announcements made by the government and also referring different political parties, police etc. and the last was faking of numbers related to Covid-19 like number of cases in different areas or number of death in either The whole country or relating to some particular area.

It can be seen in *fig.8* that largest number of fake news were related to either the government or related to any culture. Numbers related to finding of a cure, remedy that can be used, also the preventions relating to Covid-19 and news related to number of deaths were also distributed online in a significant number.

Types of Misinformation		
Category	Instances	Definition
Culture	62	Messages with cultural references such as to a religious / ethnic / social group or a popular culture reference
Cure, Prevention & Treatment	37	Messages suggesting remedies (alternative or mainstream), preventive measures, and vaccines-related misinformation
Nature & the Environment	16	Messages that have references to animals and the environment.
Casualty	36	Messages relating to deaths, illness of people in the pandemic, including graphic images of suffering (not including doctored statistics)
Business and economy	15	Messages relating to scams, panic-buying and target businesses with fake positive cases.
Government	54	Messages have government announcements and advisories or refer to police, judiciary, political parties.
Doctored statistics	23	Messages that have exaggerated numbers of positive cases or death counts and fake advisories.

**Fig. 8. Types of misinformation**

Fig. 9 show different types of fake news distributed in the first few months of the year 2020 to first week of April where the number of Covid-19 cases were increasing and incident relating to a certain culture background took place. This data is till the twelfth of the month of April. It can be seen that till the starting of the month of March there were no significant of fake news is circulation relating to the Covid-19 but as soon the number starting to grow a little in our country i.e. the starting of March to when the janta curfew was announced it can be sent that fake news relating to government grew drastically which included doctored information relating to government orders and it can be seen that in the last week of march when the incident in Nizamuddin markaz, Delhi took place there was a shear rise in numbers relation to culture from 15 in the previous week to 33 in the first week of April.

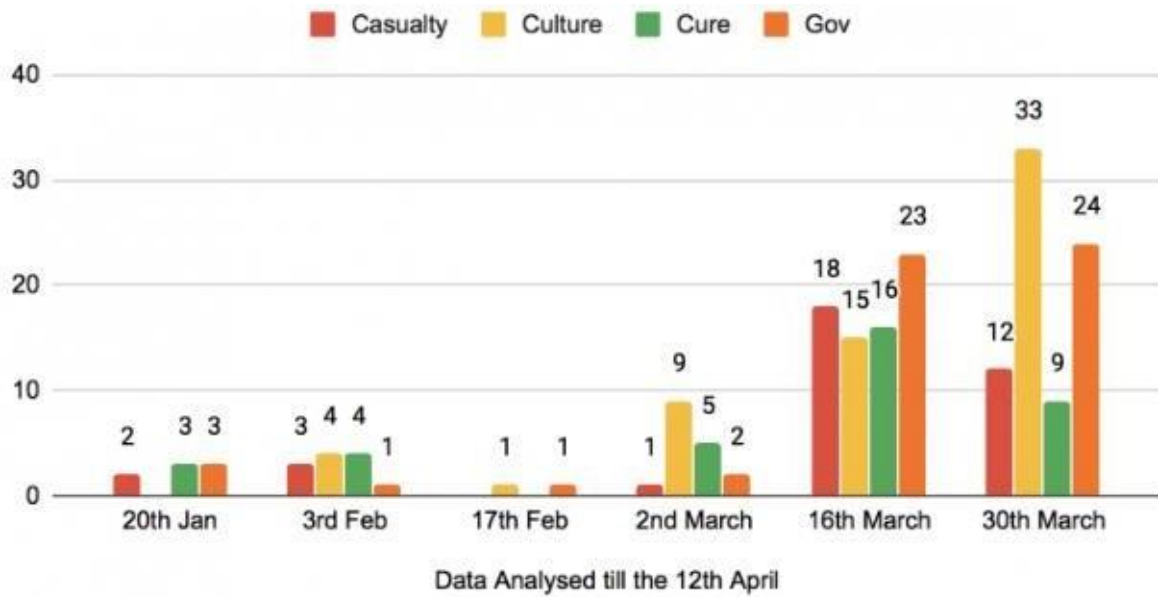


Fig. 9. Categories of misinformation and their prevalence following initial cases in India

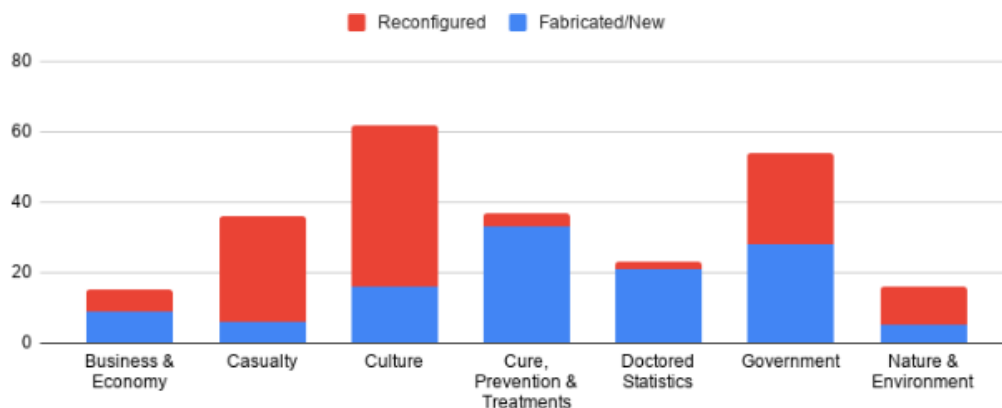
Below figure is showing different word clouds which expand through a whole month and it can be seen that how the usage of words to explain different time periods changed with changing circumstances. Between the first period i.e. between 14 and 23 March when the janta curfew and the lockdown were announced most significant words include lockdown, police, testing, janta-curfew, airport etc. But as we move toward the next period i.e. between 24 March and 2 April at the starting of this period the lockdown had just started and at the end Nizamuddin markaz, Delhi incident happened some of the most significant words were Doctor, hospital, Muslim, Religion etc.. And the last period where the Nizamuddin markaz, Delhi incident had caught significant amount of attention the words most used were Muslim, religion, spitting, and police etc. This type of data shows how the wrong information is spread across the country according to the events taking place around all of us to spread fear, hatred and sometime just for a person personal, political or financial gain.



Fig. 10. Sample of tags for misinformation in three 10-day periods



The below figure shows how many of the doctored news were created from either ground up i.e. totally new or how many were changed from some original content which was already present. It is no brainer that most of the new news made were related to either prevention and treatment or some statistics that were made up. Most of the other were mostly changed from some other content which was already present or it was right in the middle like for government related fake news there are approximately equal number of case on both the side.



**Fig. 11. Distribution of the categories of misinformation based on the novelty of propagation**  
 One of the most important role in spreading misinformation is played by the public figure who have influence over large number of people and people often tend to believe what they say. Not always they want to spread misinformation sometimes it can be unintentional i.e. they got a fact wrong or got a wrong news but most often it is intentional. Examples in *fig. 12*. Where it can be seen the tweets they have done have quite a number of retweets and also likes which means more people have been exposed to some kind of fake news.

Another significant role is played by the media either through print media, digital media or online news articles. Also, the misinformation through the means of media travels quite fast due to the number of people watching the news. Due to the lockdown the viewership might only have had increased so the influence of media over the citizen of country is quite big that is why the media is often known as the forth pillar of the democracy. But, nowadays the media houses not quite often state the facts and data which is coming around rather they want to sensationalism in their news to either increase the viewership or trying to influence people views toward certain event staking place around us. So, only the citizens must learn to differentiate between right and wrong or they should often try to verify the news that are being spread by the media.

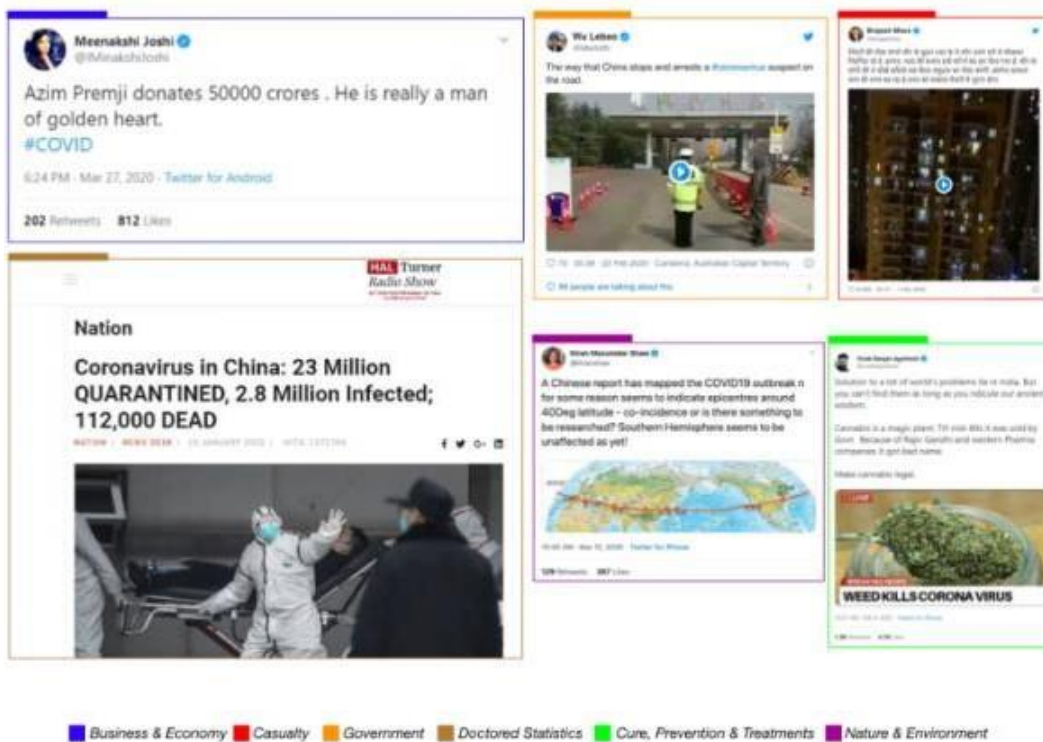


Image of misinformation spread by public figure

Fig.12. Misinformation spread by widely followed members of the mainstream media



Image of misinformation spread by news source

Fig. 13. Screenshots of misinformation spread by mainstream news

## History

Sensational things always sold well, in the early rise of the newspapers fake news was used to increase the flow of their newspapers. In the year 1835, a famous US based newspaper used a hoax in which it claimed that the moon has life to increase the circulation of their newspaper and made them one of the top selling newspaper. This shows that fake news has been used almost in our whole history to gain upper hand somewhere.

There are several other examples where misinformation was used like during Second World War to produce anti-Jew feeling in the mind of Germans. During the Lisbon earthquake of eighteenth century fake news was used by the Catholic Church to induce fake explanation for the earthquake to induce religious sentiments in citizen's minds. In the nineteenth century fake news was used to produce negative sentiments against the African-American by publishing about different crimes they didn't even commit.

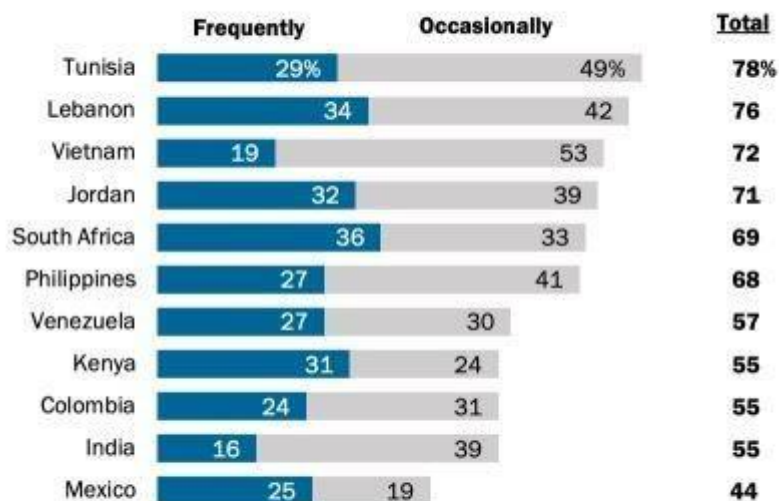


Fig. 14. Fake cover page of a magazine

But the present day misinformation is quite different from the misinformation produced in the past because in the past it was more about the sensational them to increase the number of newspaper selling but now though the sensationalism is present but the news is more often used to spread propaganda and fake news to induce in the mind of people the news which is not true about certain events. News is being used as a weapon against the people and they a constantly being fed wrong information.

## Fake news and India

*% of social media platform and messaging app users who \_\_\_ see articles or other content when they use social media that seems obviously false or untrue*



Note: Social media and messaging app users include those who said they use one or more of the seven specific online platforms measured in this survey.  
Source: Mobile Technology and Its Social Impact Survey 2018.

Fig. 15. % of users who believe that content they see is fake on social media

The above figure shows a very significant finding that in India 55 percent people think that the news article they see online or on social media are fake. This figure is quite huge if we consider this generally, it shows how big a menace fake news has become.

One of the biggest problem India is facing right now is dealing with the fake news in circulation. Fake news unfurl through internet based life inside the nation has become a noteworthy drawback, with its capability prompting horde viciousness, similar to the situation where at least twenty people were slaughtered in 2018 because of bogus data being coursed via web-based networking media.

Rasmus Kleis Nielsen, chief at Reuters Institute for the Study of Journalism, believes that "the issues of disinformation in a general public like India may be more advanced and more testing than they are in the West". The harm caused because of phony news via web-based networking media has expanded because of increment in number of web clients from 137 million of every 2012 to more than 600 million out of 2019. In India there are 230 million WhatsApp clients,

and thus one of the primary stages on which counterfeit news is spread. One of the principle issues is of beneficiaries thinking anything sent to them over web-based social networking because of absence of mindfulness. Different activities and practices have been begun and embraced to control the spread and effect of phony news.

Fake news was very pervasive during the 2019 Indian general political decision. Misleading data was predominant at all degrees of society all through the development to the political decision. The races were called by some as "India's first WhatsApp election", with WhatsApp being utilized by numerous individuals as an apparatus of purposeful publicity. As VICE and AltNews expresses, "parties have weaponized the stages" and "deception was weaponized" individually.

False and tricky news identified with Kashmir is broadly frequent. There have been different examples of pictures from the Syrian and the Iraqi common wars being made look like from the Kashmir struggle with the aim of energizing distress.

### **Dealing with fake news in India**

- Internet shutdowns are regularly utilized by the legislature as an approach to control internet based life bits of gossip from spreading.
- Ideas like connecting Aadhaar to online life accounts has been prescribed to the Supreme Court of India by the Attorney General.
- In India, Facebook has collaborated with certainty checking sites like BoomLive.
- Following more than thirty killings connected to gossipy tidbits spread over WhatsApp, WhatsApp acquainted various measures with control the spread of deception which included restricting the quantity of individuals a message could be sent to just as presenting a tip-line among different estimates, for example, suspending records and sending quit it letters.
- Fact-checking in India has become a business, rejecting the formation of truth checking sites, for example, Alt News, BOOM, Factly and SMHoaxSlayer.

- Indian service of data and broadcasting intends to set up a FACT checking module to counter the dissemination of phony news by ceaseless observing of online news sources and openly noticeable web based life posts. Module will take a shot at the four standards of Find, Assess, Create and Target (FACT).

### **Influence of Internet and Social Media**

Prior to the internet, it totally was much more costly to disperse information, developing trust took years, and there have been a great deal of simpler meanings of what realized news and media, making guideline or self-guideline easier. But the expansion of online networking has reduced a few of the limits that kept imagine news from spreading in majority rule governments. Over all it's permitted anybody to frame and plug information, especially individuals who have confirm generally skilled at "gaming" how interpersonal organizations work. Facebook and Twitter permitted people to trade data on a far bigger scope than at any other time, though distribution stages like WordPress, Wix.com permitted anybody to make a powerful site effortlessly. So, the boundaries to "making fake news" have been fixed.

## 1.2 Problem Statement

Every coin has two faces i.e. either heads or tails and the news on platforms like social media fall right under this category because though it has its boon but it is also full of bane. On one side it is really easy for people to access news on social media platforms because of the increase in number of internet subscribers, it is a really cost effective method for putting up news and it is really easy for a news to spread due to which people prefer reading news on these platforms but the other face of coin is though it makes it really easy to spread real news but it also make the spreading of misinformation i.e. the fake news really a lot more easy which often contains intended fake knowledge to distort people mind. This intensive use of fake news on social media platforms has really bad effect on people and the whole society. Due to these reasons the solution for an automatic detection of fake news has been getting a lot of attention from different communities especially the data science community and it is turning out to become a new area of research. But, the process of detecting fake news online and on social media has its own unique problems and poses different challenges in front of the researcher due to which the solutions earlier used to find fake news on the old platforms are deemed unusable and incapable to solve the problems because it most often uses context for detection of fake news which is quite tiresome process for huge number of fake news on the social media, this is the main reason for developing new systems to detect fake news.

Misinformation written in any news article often has an intention behind it to delude the person who is consuming the news article to change their opinion on some topic or make them trust some wrong information, which makes the problem more complex and difficult to detect on the basis of what is written in the news article. Other problems which are faced is the different number of subjects, different type of method in writing, the platforms on which the news is shared and the last is different number of diverse languages used to modify the true news article. Most often the misinformation spread across the social media platform is related to event happening, or any recent event that has taken place, or the news which are difficult to prove if it is fake or not due to lack of present facts and proof backing it up. Therefore it is quite important to develop a computerized and automatic system which can assist any human being who is using social media for getting their daily news in differentiating which post is fake and which is real so that they are not negatively affected by any misinformation which is being circulated on any online platform.

Therefore an automatic computerized system which is developed for finding out fake news just by looking at the linguistic and lexical of the news articles would be a great achievement since it will help all the citizens who are using social media for their daily news. It will be able to tell a fake news and notify the user when they are reading the article before either they start believing it to be true and further share it with their friends.

### **1.3 Objectives**

The objective of this project is to be able to find and differentiate between what is real news and what is fake news. It is based on different concepts of natural language processing and machine learning. The dataset we are working on includes news article, the head line and they are classified into whether they are fake news or real news so that we can train an algorithm and be able to detect real and fake news.

This system created will be able to detect fake news. It will be efficiently able to guide the user who use social media for their news in differentiating between what is real and what is fake which will help them make correct opinion and not make a negative opinion about something that is made up.

The system which is developed in the process of this project will be logical and will provide the flexibility i.e. the user can access it on different platforms like smart phones, PCs. The system will be easy to use and will be able to provide correct opinions about a news.

### **1.4 Methodology**

Natural language processing is the more popular technique for problems which consist of dealing with linguistics and provides us with great way to deal with the data i.e. to pre-process it. This technique is really helpful in the cases of text classification, sentiment analysis etc. Therefore we are using this technique hand in hands with different machine learning algorithms for the detection of fake news.



## Feature Extraction

When detecting fake news on traditional media platforms its main reliance is on the context in the news article and its contexts but in news articles circulating on the social media platforms, Meta data i.e. some kind of extra information related to the fake news is used to help the system detect misinformation. These different characteristics are being used to help explain the Meta or extra info about the news article. Some of these characteristics are:

- **Title:** This is a short line which summarizes the news content and is used to attract people. But sometimes it is not at all related to the content in the present in the news body, the term for this is ‘Clickbait’.
- **News body:** This is the area where the main news content is present i.e. where the news is explained in detail. It consists of some crucial allegations i.e. individually in highlight and it often mold the frame in which the writer is expressing.
- **Author:** It is the writer of the news article sometimes this attribute also consists of the publisher

On the basis of these characteristics, the features can be represented in many ways to draw out discerning features of fake articles. Most of the times the news article material we are working with is linguistic.

## Model Construction:

The problem that we are dealing with of fake news detection is a text-classification problem which basically means that something can be distributed between different classes example is our news articles can be classified between either fake or real, movie reviews can be classified between either being positive or being negative. When we write anything most often we use similar type of lexicon, linguistics and often possess some kind of pattern which is unique to us, so we can use this approach in which we find the most used words in a news article to find some kind of motif. One of the earliest steps is to classify the data on the basis of either fake or real. Then using this labelling we try to find what lexicons are appearing the most i.e. which words have the highest frequency. Then one of the most important step is to provide this data to the machine learning algorithms. In the end this gives us a classifier which is basically a

piece of code, which can be used to find the accuracy of the classifier by feeding the remaining labelled data to this classifier.

### **Dataset:**

There are many different ways to gather online news article example news firm's websites, social media platforms and different search-engines. But, the biggest problem we face is verifying the news articles i.e. the accuracy of news. Therefore, for this purpose specialized person with that particular domain knowledge who will be able to precisely analyze the news with the help of proper evidences, facts and various different reports. The most common ways to gather news articles which are already labelled are: Websites for particular task of fact verifying, People who are working in this field i.e. journalists and platforms like Mturk i.e. crowd sourcing. For this particular project a dataset of thirteen thousand online news article with title, the news and labelling as fake or real was used which was taken from Kaggle which helped in training our algorithm and later finding accuracy of the classifier.

## **1.5 Organization**

This project report has in total five chapters which are used to explain every small aspect of this project.

**Chapter 1:** This chapter gives us the formal introduction of the project. In this chapter the readers are introduced to various different terminologies used in the project and we also discuss the problem and the motivation which pushed us forward to take up this project. Including this we also discuss what the objective of the project are is and what methodology was used while executing the project.

**Chapter 2:** This chapter consists of various researches which were conducted in the recent past related to our project. Here, we are emphasizing more on the methodology that was used by the papers. Along with this, the outcome of their respective projects were also studied.

**Chapter 3:** In this chapter, we will go through various stages of our development and will learn about the design and algorithm implementation. Here we will also develop the model and

try to represent it from various aspects like analytical, computational, experimental, mathematical and statistical.

**Chapter 4:** In this chapter, we will go through the performance analysis of our project.

**Chapter 5:** This is our last chapter, here we will discuss the outcome of our project and also analyze our results. Along with this, we will also discuss future scope of the project and any upgrades that we can implement in the coming future. Also we will discuss some applications where the system can be helpful.

## Chapter 2: Literature Survey

In the below section we discuss various research taking place on the topic of Fake News Identification:

### Paper 1: FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network

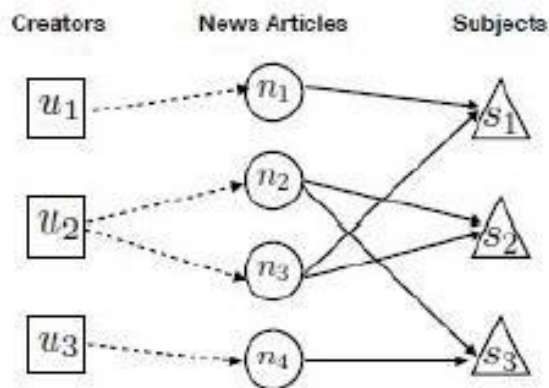


Fig. 16. Relationships of Articles, Creators and Subjects.

To study fake news on different social network platforms. They wanted to study the relationship between different types of data which includes both the content posted, the subject of that particular article and the person who had wrote the article or the post and try to establish a relationship between these different information which are data, subject and the profile, through this approach they want to find out the fake news from these online social media platforms. They gave a score for the news which was real and lower scores for the news they found were fake to differentiate between the credibility of the news on the platform. According to the problem they were facing to detect the news which was fake out of a dataset they gave a problem statement:

If we are given some data i.e. a news the algorithm to find the fake news wants to give a particular score to that particular data according to its credibility and in situ to score the news or post produces which will be its credibility. Therefore for the best results i.e. for the algorithm to learn different types of data must be provided which consists of both the news or data , the

person who wrote that particular data i.e. the profile, the subject of that data and the relation between these two set of information's as it is depicted in the figure above.

### **Dataset Analysis:**

The dataset which they used for this particular research included bunch of tweets which were posted by a non-profit center at its official twitter handle and also some article on their website which consists of data which is already checked for its credibility. After this they provided with some stats which were derived from the dataset itself, after this the data was thoroughly inspected for these different information's which are the subject, author and the article. What they found was the subject is highly correlated with the article and also the author which means they will get somewhat equal credibility drawing a relation between this data. For a single person they are able to write many number of news articles but one news article has only one person who has written it. A single body can be related to many different subjects, and each subject can also have many different bodies.

The news article body of the misinformation was able to reveal many significant findings. Since, many articles had similar author it was found that these articles and the author had used similar set of words which means these same words can be found in many different bodies, subject and used by authors on frequent basis which can be traced to fake news. Since these authors are given a credibility score and since they are using similar language in there text they can be easily found i.e. if a person is writing fake news over and over again they can be singled out in many articles. The system they made used GDU for drawing relations among the author, subject and the news body so that the credibility between all these actors can be related to each other. This can be seen in the fig. above where a single writer is related to one or more news articles and an article can be related to different number of subjects and one subject is related to many articles and one article is related to not more than one writer. If U2 writes some fake news he/she will be using a similar kind of language in all the articles therefore if he/she writes another article according to the language used we can trace back to the writer using the credibility of the article because as we know the credibility of similar content is similar to each other and the credibility of a writer will also be similar to how credible the article he/she has written. To draw these kind of relations on the basis of credibility of similar things they have used gated diffusive unit.

## Paper 2: Fake News Detection on Social Media: A Data Mining Perspective

The way they started this research was by defining what fake news really is and also finding if it has more than one meaning. They then differentiated between the definitions and found the definitions which are often confused as fake news but are not. They then explained what were the several different ways the fake news are spread either on traditional media platforms or the newer platforms like social media.

**Problem Definition:** Given the social news engagements among 'n' users for news article 'a', the task of fake news detection is to predict whether the news article 'a' is a fake news piece or not i.e.  $F(a) = 1$ ; if a is a piece of fake news,

0; otherwise.

where F is the classifier we want which can classify news into either fake or real.

**Dataset:** They were able to get news article from different sources like news websites, social media platforms like Facebook, Reddit, Twitter etc. and also different search engines like Bing, Google etc. But it is a difficult task to classify every piece of news article manually, therefore help of professional fake news tracker was taken because of the expertise and they can analyze the news article better than any person since they can find a context and use extra evidence to support the claims.

Fake news detection techniques used in traditional media platforms are not most often applicable on their huge dataset because they often use context to support their findings. But when working with a huge dataset it is often quite difficult to check a single piece of news for its authenticity therefore extra information about the news is used to check whether the news is fake or not. This Meta info often used in their findings was:

- **Source:** It consists of either the company which published the news i.e. on some platform or the author who wrote the news.
- **Body:** It consists of the whole news article which consists of the meaning and detail regarding the news article.
- **Headline:** It's the heading for the news body which is used to gain attention and can be used as a clickbait.
- **Media:** If any media i.e. photo or video is there.

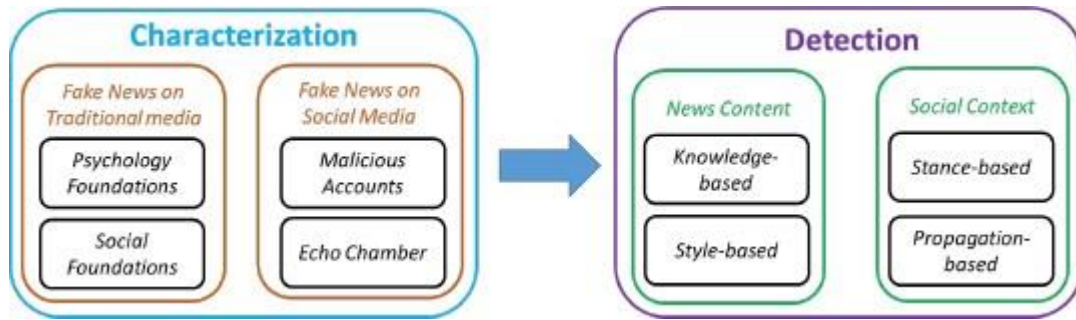


Fig. 17. Fake news on social media: from characterization to detection

The figure above shows that the fake news present on the traditional type of media and on the social media platforms both require different ways of detection. In the case of traditional it is easier because they need to only check out the context and the evidence but the news article on social media need different methods to detect fake news and stop them in the future to be even put up on the platform. Some methods are explained below:

**Stance Based:** This method used the user’s frame of reference from the applicable post subject to find out whether the news article is fake or real. There are two ways in which the news article post can be shown implicit stance or explicit stance. Implicit are taken out from the post. Explicit are the straightforward pint of view or sentiments such as either liking a post on Facebook, up vote on Reddit etc. This method is used to find out by itself that whether the person using thinks about the post positively or is having opposite thoughts that of the post.

**Propagation Based:** This method to find out whether the news is fake or not used the relations between the different news articles to predict whether the news is credible or not. This method assumes that the reliability of the news article is directly proportional to the reliability of the post on the social media post.

**Post Based:** This method of fake news detections uses the user’s belief or sentiments towards a news article through the post on the social media platform such as a strong belief towards something or a strong response to something. Therefore it easy to deduce that this method helps us find likely misinformation through the response of the used on any social media post. It main focal point is to pinpoint some info that is essential to find the credibility of any news article from various social media posts.

### Paper 3: Fake News Detection Using Naive Bayes Classifier

They have used the simplest approach which is most commonly used in the problems relation text classification i.e. detecting fake news using the Naïve Bayes classifier. They applied their point of view as a software and used a data set from news post in Facebook to test their approach. They were able to get accuracy of seventy four percent on the Facebook news post dataset which is a fair result if we also consider the fact that they are using the simplest approach of Naive Bayes Classifier. But this approach is not that bad to use because naïve bayes is often used for problem which consists of text classification or sentimental analysis. Naïve Bayes not just consists of one algorithm but some other algorithms like Gaussian, multinomial etc.

**Dataset:** the dataset used for machine learning and testing the naïve bayes classifier was collected by news platform Buzz Feed. It consists of news articles from different Facebook posts. They were gathered from Facebook pages of three biggest news platforms which are ABC, CNN, and Politico and first classified the data then first used the data to make the system learn and then tested the classified on already classified data to find out the result using the accuracy.

#### Generalized plan they followed:

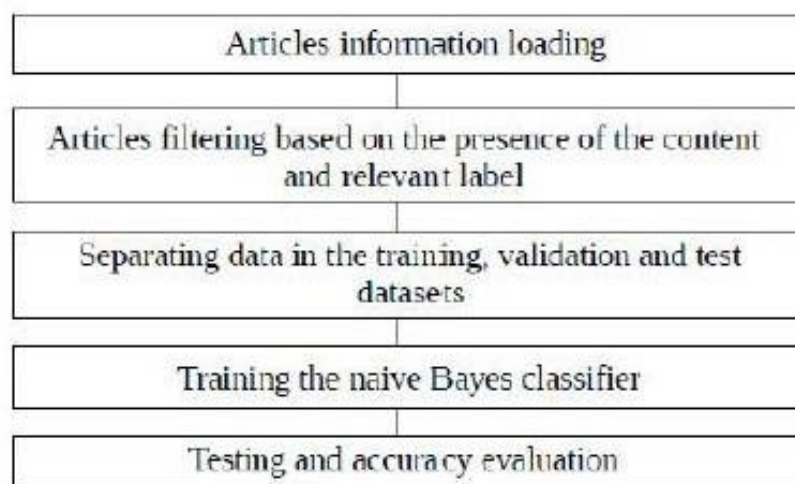


Fig. 18. Generalized scheme of the algorithm



Results they received using this dataset and approach were:

- Precision: 0.71
- Recall: 0.13

The reason recall is so low is due to the lopsided data which was used for learning and afterwards testing.

News article type	Total number of news in test dataset	Number of correctly classified news	Classification accuracy
True	881	666	75.59%
Fake	46	33	71.73%
Total	927	699	75.40%

Fig. 19. Result of their model

#### Ways to improve the result:

- The amount of data which is being used for testing should be increased. In machine learning increasing the dataset often helps in increasing the accuracy or the performance of the algorithm.
- Using longer single data i.e. the length of the news body used by them was quite short therefore using longer news article will help improving the accuracy since on Facebook sometimes the news articles are just a sneak peek.
- When the data is being preprocessed the stop words must be removed from the text body since computer does not understand the meaning of the sentence and the stop word deteriorate the performance.
- During the preprocessing, process like stemming can be used since explained in above point computer is not able to understand the meaning of sentence, and this is used to bring the words back to its stem like backing to back since back is the stem.
- Unigrams and bigram should be used because sometimes two words are used together to give a specific meaning.

#### **Paper 4: Automatic Detection of Fake News**

Just like in the paper 3 they also faced a similar problem in which they has to find features in the data to solve the problem since the traditional approach was not applicable on this type of data which was extracted from social media platforms therefore they used lexical properties of the text to build a classifier for the detection of misinformation. The lexical characteristics they used were:

**N-gram:** Since, some words present a true meaning when they are used together with each other like ‘Thank You’ is a bigram and presents a more viable meaning when used together, therefore they found out all the uni and bi grams from their dataset which was represented in bag-of-words form and later converted into tfidf values.

**Punctuation:** They used the count of words and software which helped them inquire linguistics to make a feature set which consisted of different punctuations. Punctuation feature set included characters like dashes, commas, question marks, exclamation, and period etc.

**Psycho-linguistic characteristics:** This essentially draws correlation between psychology and the characteristics relating to the linguistics used. They used a software to find out the words which often comes in this category. The software they used is based on a dictionary that represents psycho-linguistic features (negative sentiments), summaries, and also had the parts of speech (nouns, adverbs etc.).

**Syntax:** Using context free grammar rules they took out a feature set.

**Dataset:** Ehen they were trying to build a data set for the fake news detection, they used suggestions of many different writer on what should be included in the features:

- Must have equal number of fake as well as real news articles.
- Must contain news articles which are in textual form
- Must be evincible
- Must have similar length i.e. of the news article
- Must have a homogenous style of writing
- Must have news articles from a similar period of time

- Must be available to the citizens
- Must take into consideration the difference in linguistics
- Must have similar context

For the dataset they used six different topics of news. They collected news in two different phases, they started by first collecting real news about the six different topics. The news was collected from news firms which are well known among the citizens like CNN, New York Times, USA Today, Fox News, ABC News, Bloomberg etc. Then they started collecting fake news, they used crowd-based sourcing based platform MTurk which is an Amazon based service which is often used for the construction of different datasets for many different tasks. When they had the dataset containing both types of news they were able to use a linear support vector machine and k-fold cross validation, with all the factors like recall, accuracy etc. mean over k=5 iterations. In the below figure we can see that what types of words were most often used in real news and what type of words were used in the fake news. The most often used words in fake news were ‘Relative’, ‘Leisure’, ‘Prep’, ‘Time’ etc. and the most common words in real news were ‘Cogproc’, ‘Verb’, ‘Aux Verb’ etc.

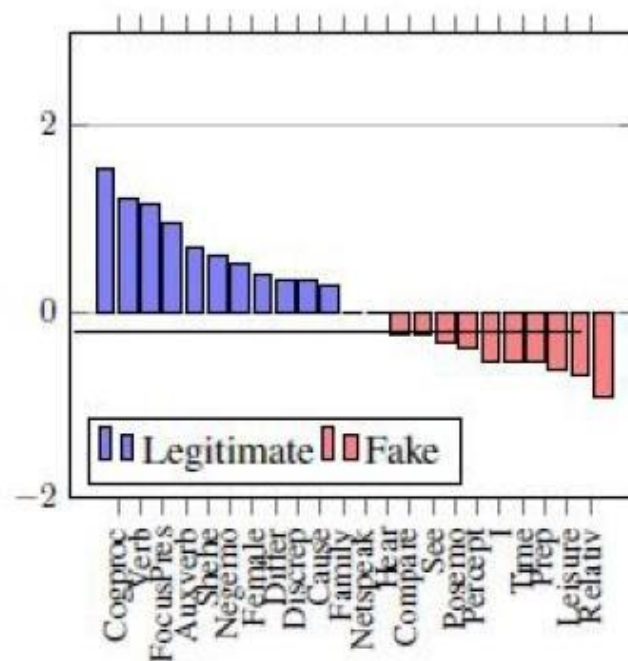


Fig. 20. Language differences in fake and legitimate content

## Paper 5: Detecting Fake News in Social Media Networks

```
1: Open URL file
2: for each title
3:   title starts with number? 1 → output file
4:   title contains ? and/or ! marks? 1 → output file
5:   all words are capital in title? 1 → output file
6:   users left the website after visiting? 1 → output file
7:   contents have no words from title? 1 → output file
8:   title contains keywords? NoKeywords → output file
9: end for
```

Fig. 21. Algorithm for feature extraction

They also started by first defining the problem i.e. what they are dealing with then they tried to find a good clickbait data related to online news articles. They then found all the properties of the dataset and the algorithm used can be seen in the figure above. Then the data was made for the Waikato Environment for Knowledge Analysis. But it is very difficult to find such data set at one place therefore they relied on platforms where finding clickbait news was easier like social media platforms example Facebook, Reddit etc.

After collecting all the uniform resource locators of the pages, a program was used to find out the characteristics from the urls the algorithm used can be seen above and a python code was used to do the same. The feature they found out from the articles were: heading which had numeric values at the starting, and also headings with ‘!’ OR ‘?’ in them, the main words used in English, whether if the user who opened the page left it immediately etc.

They used four different machine learning algorithms to work on the data they gathered which gave them some results. Though Waikato Environment for Knowledge Analysis comes with huge amount of different algorithms but they used the ones given below:

- Naïve Bayed Algorithm
- Random Tree
- BayesNet
- Logistic

The below table depicts the result which they got from the four algorithms. They are differentiated on the basis of 4 different parameters. The highest precision is that of Logistic, highest recall i.e. the sensitivity is of random tree and logistic, similar set of result for f measure and the receiver operatic characteristics area was of Naïve Bayes and bayesnet.

Classifier	Precision	Recall	F-Measure	ROC
Bayes Net	94.4%	97.3%	97.2%	100%
Logistic	99.4%	99.3%	99.3%	99.5%
RandomTree	99.3%	99.3%	99.3%	97.3%
Naive Bayes	98.7%	98.7%	98.6%	100%

Fig. 22. Comparing the results they got from their models

## Chapter 3: System Development

### 3.1 Analysis/Development/Algorithm/ Design

- **Objective:** To be able to detect fake news from dataset of fake and real news.
- **Analysis:** Using NLP and machine learning algorithms, the news articles have to be differentiated between either being fake or real. This project tries to give a solution to this circulation of fake news on social media platforms and also try to find out how real a news article is.
- **Design:** The method in which we are trying to find a solution for this problem is on the basis of the pre-processing of the data, then the formation of a proper dataset, creating classifiers using this dataset, testing these classifiers on the basis of testing dataset and using these classifiers to try to predict what type of news it is. There are in total four fields in the dataset, first field the title is used to find origin of the news which is accompanied with news body which we use for the classification of the news. The next step is that we intermix the data randomly and then the dataset is split up into two different unequal sub sets which are training and testing sets. Training set is used for the algorithm i.e. it is used to train the models which are constructed example bernoulliNB and multinomialNB. The other set is used for testing the classifier i.e. to check the performance of the classifier on basis of different parameters. The data set was divided in two categories training set which was 75 percent and testing set which was 25 percent of the entire dataset. From this training set we used the bag of words concept, this concept does not bother with how the sentence is structured but it only used the frequency of words in the text body.
- **Proposed Approach:** The figure below shows the approach we have followed for this project. Each of the section in the diagram is touched upon further.

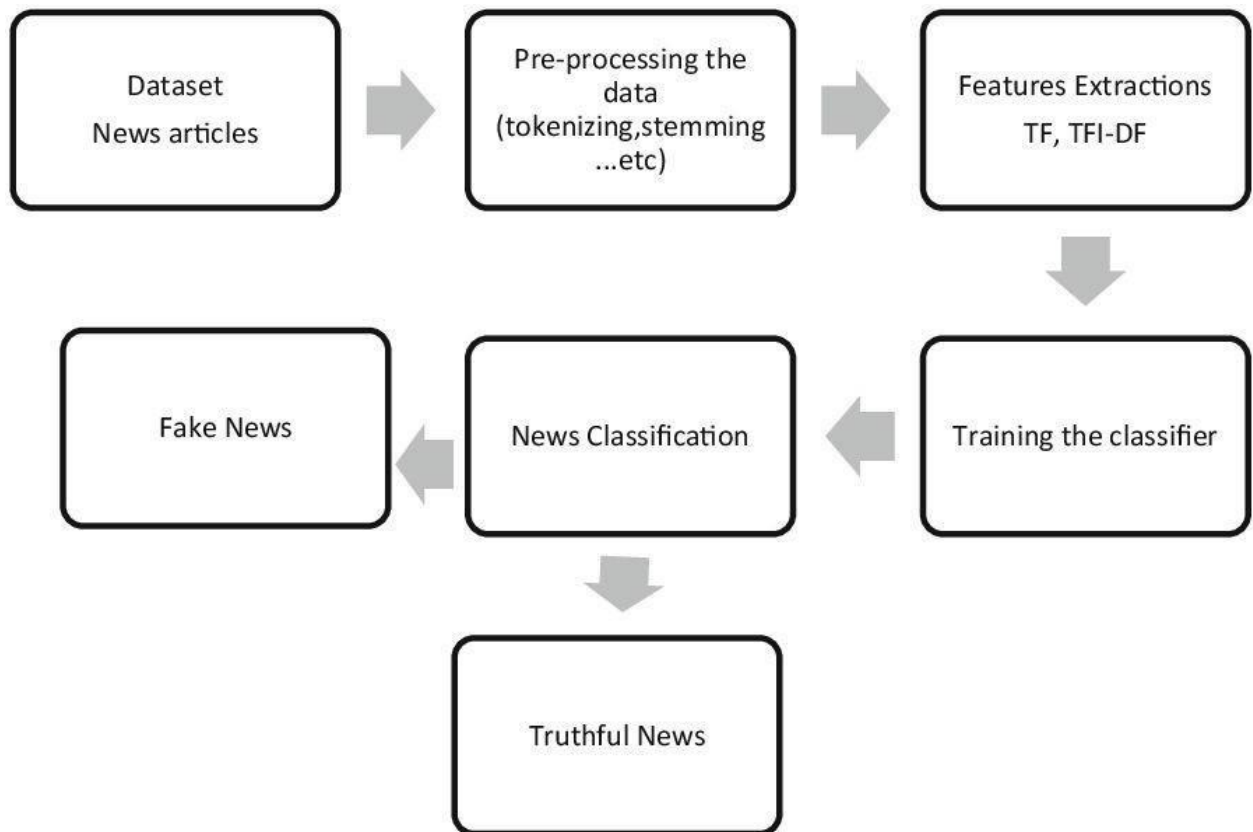


Fig. 23. General Approach

**Data Preprocessing:** When we get the dataset which consists of the news articles, the data is raw and the computer is not able to understand the text format. Data pre-processing is a process of cleaning and converting this incomprehensible and raw data into a format which is easily understandable to the machine i.e. numbers. The data set we got is a real world dataset of news articles and it can be not consistent, complete and more likely to have many mistakes like empty cells etc. So, the process of data pre-processing is great way to solve this problems. Different steps followed during data pre-processing are:

1. Importing of the libraries used
2. Getting the dataset and using pre-processing for empty and unqualified data.
3. Then this clean data is split into two sets of testing data and training data.
4. The last step is to extract the features.

## Step 1: Importing the required libraries

The required libraries for data pre-processing are imported. Python has many different libraries which are in use but the most often and commonly used are:

- Numpy
- Pandas

**Numpy:** It is one of the most basic and useful python library for calculations related to data science. It is used so that we can make use of arrays in python programming language. Though we have an array like data structure in python i.e. lists, but they are about fifty times slower to process. The reason for this far less speed is that the lists is not stored continuously in memory locations therefore it can be sometimes difficult to retrieve but numpy arrays are stored uninterrupted in the memory location. It also provides us with n-dimensional arrays for scientific purposes and possess many different functions in its library to work upon these n-dimensional arrays.

**Pandas:** It is most often used python libraries for manipulation of huge amount of data especially in data structures like tables which are known as dataframes. It also provide us with a very good tool to analyze the data which is present in tabular and n-dimensional form. It can also help in basic cleaning of the data with the help of many functions provided by pandas library like the empty cells, the cells which have null value etc. The tabular data often consists of excel spread sheets and different data bases. It can help us deal with many different types of extensions like comma separated values, xls, JavaScript object notation etc.

## Step 2: Importing the data set

Pandas is used to import the dataset which is '.csv' type which means comma separated values.

```
import pandas as pd
df=pd.read_csv(r'C:\Users\raanaa\python\data.csv')
```

We have first given alias to pandas which is pd, then we are using the function read\_csv which takes the path to the data as the input but in the form of a regular expression. We used a csv type of file because it is often very fast to access due to its very less weight. We can check our data imported using a function called 'head' i.e. df.head() which provides us with the whole table but we can also pass an integer like five to get only the first 5 columns of the table.



```
df.head()
```

	Unnamed: 0		title	text	label
0	8476		You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...		Google Pinterest Digg LinkedIn Reddit Stumbleu...	FAKE
2	3608		Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...		— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875		The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

### Step 3: Splitting the data:

When dealing with problem related to ML we often divide our data into two different sets of the original dataset:

- Training Set
- Testing Set

This process of splitting is quite common since we have labelled data in our possession we use some part of it to train our models and the rest of it for testing the classifiers we got after running the algorithm. The training set is always larger than the testing set because the more data we use to train our models more accurate they become. Most often the ratio that are used to split data our 70:30 or 80:20 which basically means eighty percent of data represents training set and twenty percent represents the testing set. Before splitting, the labels i.e. either fake or real are stored in another variable, and both this variable in our case 'Y' and remaining data which consists of news headlines and body.

```
X_train,X_test,Y_train,Y_test=train_test_split(df['text'],Y,test_size=0.25,random_state=53)
```

This piece of code was used to split data into four different categories:

- X\_train & X\_test: consists of the news articles components like news body and headlines.
- Y\_train & Y\_test: consists of the corresponding labels for the news articles.

X and y train are used to train the models and then for x\_test the results are predicted then the results are compared with Y\_test, to find how accurate our model is.

**X\_train:** represents the training set. Since, the data is shuffled the numbers seen on the left side are in that order. It consists of seventy five percent of the data set and consists of total around forty eight hundred news articles.

```
X_train
6073 Divisions among Democrats are deepening over S...
288 In a Tuesday editorial, the paper's opinion ed...
3565 \n- My Name is Fate (@Destini41) October 29, 2...
4964 Washington (CNN) The future of health care in ...
3852 6 Great Halloween Costume Ideas For Duos Poste...
...
662 -Debby Borza stood before a wall of photos of ...
3261 Presumptive Republican nominee Donald Trump ha...
5883 December's job growth numbers are in, and they...
2933 In a wide-ranging discussion, Trump also said ...
797 Top officials of the Cruz campaign are convinc...
Name: text, Length: 4751, dtype: object
```

Fig. 24. Training set used

**X\_test:** is the part of the dataset which is used to test our model. It represents the other part of the data set which is 25 percent of the whole dataset. It consists of around fifteen hundred news articles.

```
X_test
4221 Donald Trump threatened to sue the New York Ti...
1685 Planned Parenthood: Abortion pill usage now ri...
3348 In a last dash, final "hail mary" attempt to e...
2633 Washington (CNN) Donald Trump and Ben Carson n...
975 The Obama administration announced Friday it w...
...
5386 Shame to waste Corbyn on a snap election, says...
2106 14, 2016 How To Plan Farmer's Calendar All Y...
5036 "One should not insist on nailing [Trump] into...
991 Toward the end of our meeting with President O...
6186 Obama Looking For Justice Who Will 'Interpret'...
Name: text, Length: 1584, dtype: object
```

Fig. 25. Testing set used

**Y\_train:** represents the set of labels which are dependent on the part of dataset which is represented by x\_train. It can be seen easily from the fig since the first number in both x and y train are 6073 which means the order is maintained.

```
Y_test
4221    REAL
1685    FAKE
3348    REAL
2633    REAL
975     REAL
...
5386    FAKE
2106    FAKE
5036    REAL
991     REAL
6186    REAL
Name: label, Length: 1584, dtype: object
```

Fig. 26. Dependent variable associated with X\_train

**Y\_test:** It consists of the data which is used for checking the accuracy of the model is dependent on x\_train both have the first news article number 4221.

```
Y_test
4221    REAL
1685    FAKE
3348    REAL
2633    REAL
975     REAL
...
5386    FAKE
2106    FAKE
5036    REAL
991     REAL
6186    REAL
Name: label, Length: 1584, dtype: object
```

Fig. 27. Dependent variable associated with X\_train

## Step 4: Extracting features

The computer is not able to understand what the text means it only understand the language of numbers therefore we need a method to convert the news articles into some viable information. Most algorithms related to machine learning takes input in the form of vector. Therefore we need to convert text files and documents into vectors of numbers using a process called vectorization. One of the most often used model for dealing with the text files for ML is Bag of words model.

The model uses the approach in which it does not considers any type of Meta data about the text but its only focus is on the frequency of the words in a text file. This should be possible by doling out each word a one of a kind number. At that point any report we see can be encoded as a fixed-length vector with the length of the jargon of known words. The incentive in each position in the vector could be loaded up with a check or recurrence of each word in the encoded document. This is the bag of words model, where we are just worried about encoding plans that speak to what words are available or how much they are available in encoded archives with no data about request.

- **CountVectorizer or word counts:** The CountVectorizer, as we probably are aware gives an approach to tokenize an assortment of content reports and fabricate a jargon of known words, yet in addition to it encode new records utilizing that vocabulary. An encoded vector is returned with a length of the whole jargon and a whole number mean the occasions each word showed up in the document. Because these vectors will contain a great deal of zeros, we call them inadequate.
- **Tf-idf vectorizer:** Utilizing CountVectorizer is a decent beginning stage, yet is extremely fundamental. Words like "the" will seem ordinarily and their huge tallies are not be extremely valuable in the encoded vectors. An elective is to compute word frequencies, and by a wide margin the most well-known technique is called TF-IDF. This is an abbreviation than means "Term Frequency – Inverse Document" Frequency which are the parts of the subsequent scores doled out to each word.

**Term Frequency:** This calculates number of times a given word appears within a document i.e. the frequency of word in a news article

**Inverse Document Frequency:** This expels words that seem to appear a lot more often in the dataset

## **Machine Learning Models**

There are different distinctive algorithms that are used for classification, which can be utilized to group a news as "misinformation" or "genuine". In this project, we have utilized multinomial, Bernoulli Naive Bayes and PAC learning models. Every one of these methodologies can be exclusively utilized to distinguish news.

### **Naïve Bayes Algorithm**

Naive Bayes classifiers is one of the most basic classifier which works on simple probability based on applying Bayes hypothesis. It tries to guess probability relations for each class, for example, the likelihood that given record or information point has a place with a specific class. It expect that all the highlights are disconnected to one another so the nearness and nonappearance of an element doesn't influence the nearness and nonappearance of another element. It's anything but a solitary calculation for preparing such classifiers, be that as it may, a group of calculations dependent on a typical standard of the occasions a specific occasion has happened. It is a mainstream strategy for content order which is the issue of judging records as having a place with one class or the other with word frequencies as the highlights. With suitable pre-processing, it is practical in this area with further developed strategies including bolster vector machines bringing about improved precision.

### **Mathematical implementation of Naïve Bayes:**

As Bayes Theorem chips away at likelihood based on some conditions, which is the likelihood that an occasion will occur, given that a specific occasion has as of now happened. Utilizing this idea we can ascertain the likelihood of any occasion dependent on the probability of another occasion. The following is the recipe for ascertaining the conditional likelihood.

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Where,

- P(H) is the probability of hypothesis H being true. This is known as the prior probability.
- P(E) is the probability of the evidence (regardless of the hypothesis).
- P(E|H) is the probability of the evidence given that hypothesis is true.
- P(H|E) is the probability of the hypothesis given that the evidence is there.

The idea we use to characterize news with misinformation is that fake news stories on social media regularly utilize a similar arrangement of words while genuine news will have a specific arrangement of words. It is very detectable that couple of sets of words have a higher recurrence of showing up in fake news than in evident news and a specific arrangement of words can be found in high recurrence in obvious news. Obviously, it is difficult to guarantee that the article is fake as a result of the way that a few words show up in it, henceforth it is very improbable to make a framework entirely precise yet the presence of words in bigger amount influence the likelihood of this reality.

The question of ascertaining the likelihood of a condition of finding a particular word in fake news stories and in genuine news stories can be settled with the end goal that in a given preparing set, which contains bunches of news stories, marked as evident or misinformation, one can characterize the likelihood of finding a particular word in any fake news story as the proportion of fake news stories, which contain this word to the all-out number of fake news stories. The likelihood of finding a particular word in evident news stories can be characterized comparatively.

The equation for computing the probability of a condition of a certainty, that news story is a misinformation given that it contains some particular word looks as following:

$$P(F|W) = \frac{P(W|F) * P(F)}{P(W|F) * P(F) + P(W|T) * P(T)}$$

Where,

- P(F|W) is the probability that a news article is fake given that word W appears in it.
- P (W|F) is the conditional probability that word W will appear in the texts of fake news.
- P (F) is the probability that the news article will be a fake news article.
- P(W|T) is the conditional probability that the word W will appear in the texts of true news.
- P(T) is the probability that the text will be a true article.

### Different Types of Naïve Bayes Algorithms:

**Multinomial Naïve Bayes:** In multinomial naive bayes a text document is represented as feature vector whose value is equal to the count of that particular word in the document.

**Bernoulli Naïve Bayes:** In Bernoulli naive bayes a text document is represented as feature vector whose value is equal to either 0 or 1 based on the presence of word in document. It remains 1 even if the count is greater than 1.

	Bernoulli Naïve Bayes	Multinomial Naïve Bayes
Multiple Occurrences of words	Ignored	Considered
Document representation	Binary Vector i.e. 0 or 1	Integer Vector
Document Length	It is best for short documents	It can be used with long documents
Behavior with stop words	Since, stop words are a part of each text body, so the probability of them is 1.	Relative frequency are taken into account so the probability of stop words is 0.05.

Table 1. Difference between Bernoulli and Multinomial Naïve Bayes

## Chapter 4: Performance Analysis

The model was prepared for various varieties of the models for differing timeframes on a system having setup as follows:

- RAM: 8GB DDR3
- Core: Intel Core i5-6200 @ 2.3Ghz
- GPU: NVidia GeForce 940M
- OS: Windows 10 64 bit
- Language: Python
- Editor: JupyterLab

### Dataset:

The data-set used to test the effectiveness of the classifier is created by Amazon, containing 6335 news stories labelled as genuine or fake. It has 6335 rows and 4 columns. The 4 segments comprise of index number, title, content and label. Data-set incorporates business, science, technology, entertainment and well-being news classifications. The credibility of this data-set lies in the way that it was checked by columnists and afterward named as "Genuine" or "Fake". Title includes the insignificant data required to comprehend the news story like the heading of the paper which portrays the substance inside. Content involves an itemized depiction of the news story installed with idiosyncrasies like area, subtleties, individuals included and their experience and so forth. Mark is fundamentally a label which tells whether the news stories are "Fake" or "Genuine".

### Code:

To import data we first need to import some essential libraries from python's huge number of different libraries. 'os' is very important library to help python program to interact with the system os. 'Pandas' let us manipulate the imported dataset to our needs and the last library helps us split the dataset.

```
import os
import pandas as pd
from sklearn.model_selection import train_test_split
```



```
[26]: os.listdir(r"C:\Users\raanaa\python")
df=pd.read_csv(r'C:\Users\raanaa\python\hello.csv')
df
```

As we can see, the piece of code above provided us with the dataset which we want to use in our problem. The first function ‘listdir’ is a part of os library which is used to get the list of the files present on the path provided to the function in the form of a regular expression. Since, we are dealing with a csv file we use pandas function read\_csv to read the file and open it. It is stored in df variable.

```
[26]:
```

	Unnamed: 0		title	text	label
0	8476		You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...		Google Pinterest Digg LinkedIn Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy		U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...		— Kaydee King (@Kaydeeking) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters		It's primary day in New York and front-runners...	REAL
...	...				...
6330	4490	State Department says it can't find emails fro...		The State Department told the Republican Natio...	REAL
6331	8062	The 'P' in PBS Should Stand for 'Plutocratic' ...		The 'P' in PBS Should Stand for 'Plutocratic' ...	FAKE
6332	8622	Anti-Trump Protesters Are Tools of the Oligarc...		Anti-Trump Protesters Are Tools of the Oligar...	FAKE
6333	4021	In Ethiopia, Obama seeks progress on peace, se...		ADDIS ABABA, Ethiopia —President Obama convene...	REAL
6334	4330	Jeb Bush Is Suddenly Attacking Trump. Here's W...		Jeb Bush Is Suddenly Attacking Trump. Here's W...	REAL

The ‘df’ variable has our data stored in a form of a dataframe as we can see in the output above that we have our data with all our titles of the dataset which are the index, title, text and the label.

```
[6]: df.shape
```

```
[6]: (6335, 4)
```

The function above used i.e. shape is a part of pandas library and is used to get the shape i.e. the dimensions of the dataframe we have imported and the result we got is 6335 \* 5 which is correct since we are using that amount of news article to train our models.

```

Y=df.label

```

```

Y
0    FAKE
1    FAKE
2    REAL
3    FAKE
4    REAL
...
6330  REAL
6331  FAKE
6332  FAKE
6333  REAL
6334  REAL
Name: label, Length: 6335, dtype: object

```

```

df.drop('label',axis=1)

```

	Unnamed: 0		title	text
0	8476		You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...		Google Pinterest Digg LinkedIn Reddit Stumbleu...
2	3608	Kerry to go to Paris in gesture of sympathy		U.S. Secretary of State John F. Kerry said Mon...
3	10142	Bernie supporters on Twitter erupt in anger ag...		— Kaydee King (@KaydeeKing) November 9, 2016 T...
4	875	The Battle of New York: Why This Primary Matters		It's primary day in New York and front-runners...
...	...	...		...
6330	4490	State Department says it can't find emails fro...		The State Department told the Republican Natio...
6331	8062	The 'P' in PBS Should Stand for 'Plutocratic' ...		The 'P' in PBS Should Stand for 'Plutocratic' ...
6332	8622	Anti-Trump Protesters Are Tools of the Oligarc...		Anti-Trump Protesters Are Tools of the Oligar...
6333	4021	In Ethiopia, Obama seeks progress on peace, se...		ADDIS ABABA, Ethiopia —President Obama convene...
6334	4330	Jeb Bush Is Suddenly Attacking Trump. Here's W...		Jeb Bush Is Suddenly Attacking Trump. Here's W...

6335 rows x 3 columns

We can access the different headings of a dataframe using basically the variable allotted for dataframe and the name of the heading we want to select as we have used above `df.label`. Basically we are storing the labels of the news article in a news variable 'Y'. Then, in the next step we used drop function which is used to drop a column or row in a dataframe since we want to drop the label which is a column we provide the function with the name of the column and also the axis which represents whether we want to drop a column or a row. It can easily be seen from the outputs above that the label is first stored in a variable then it is dropped from the 'df' dataframe. This is basically done since when we are testing out models, for the testing set we don't want the labels with our news articles we want to check the accuracy of the models which we have produced.

```
[12]: X_train,X_test,Y_train,Y_test=train_test_split(df['text'],Y,test_size=0.25,random_state=53)
```

```
[13]: X_train
```

```
[13]: 6073 Divisions among Democrats are deepening over S...
288 In a Tuesday editorial, the paper's opinion ed...
3565 \n- My Name is Fate (@Destini41) October 29, 2...
4964 Washington (CNN) The future of health care in ...
3852 6 Great Halloween Costume Ideas For Duos Poste...
...
662 -Debby Borza stood before a wall of photos of ...
3261 Presumptive Republican nominee Donald Trump ha...
5883 December's job growth numbers are in, and they...
2933 In a wide-ranging discussion, Trump also said ...
797 Top officials of the Cruz campaign are convinc...
Name: text, Length: 4751, dtype: object
```

```
[35]: Y_train
```

```
[35]: 6073 REAL
288 REAL
3565 FAKE
4964 REAL
3852 FAKE
...
662 REAL
3261 REAL
5883 REAL
2933 REAL
797 REAL
Name: label, Length: 4751, dtype: object
```

The piece of code given above help us split our data into training and testing set in the ratio of 75 to 25. We have used `train_test_split` function which is the part of `sklearn` library. We provide the function with the data frames we want to split in our case 'df' and 'Y', the next parameter is the size of testing set which is given 0.25 which means the data is twenty five percent of data is testing and seventy five percent is for training.

When we have our data split up and clean, we need the data in a format in which the machine can understand the data. Therefore, we need the data in numerical form so that computer can easily understand, this is done by bag of words approach in which the text is converted into some numerical values and put in a vectorised form. Therefore, we use `count vectorizer` and `tfidf vectorizer`.

```
count_vectorizer = CountVectorizer(stop_words='english')
count_train = count_vectorizer.fit_transform(X_train.values.astype('U'))
count_test = count_vectorizer.transform(X_test.values.astype('U'))
```

In the piece of code given above, we make object of CountVectorizer class with stop words as the parameter which has the value “English” which means it will take out English stop words like a, an, the etc. We use the object we made to vectorise the variable x\_train and x\_test and store them in some variables.

```
print(count_train)
```

```
(0, 16485) 1
(0, 14865) 1
(0, 14526) 1
(0, 47544) 1
(0, 6812) 1
(0, 46523) 2
(0, 44879) 1
(0, 10161) 1
(0, 50479) 1
(0, 12572) 2
(0, 36471) 2
(0, 49797) 1
(0, 21678) 1
(0, 42789) 1
(0, 19167) 1
(0, 39171) 2
(0, 14856) 1
(0, 30917) 1
(0, 14981) 1
(0, 35445) 1
(0, 44384) 1
(0, 44216) 1
(0, 51630) 1
(0, 16218) 1
(0, 56634) 1
:
:
(4750, 13592) 1
(4750, 5187) 1
(4750, 20296) 1
(4750, 52566) 1
(4750, 33275) 1
(4750, 41524) 1
(4750, 7493) 1
(4750, 7961) 1
(4750, 14557) 1
(4750, 40845) 1
(4750, 12608) 2
(4750, 56302) 1
(4750, 52565) 1
(4750, 18067) 1
(4750, 21896) 1
(4750, 21682) 1
```

The vectorised form of x\_train is stored in count\_train which we can see above each line represents a unique word and the number represents the count of the word.

```
tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)
tfidf_train=tfidf_vectorizer.fit_transform(X_train.values.astype('U'))
tfidf_test=tfidf_vectorizer.transform(X_test.values.astype('U'))
```

Similar types of steps are used in the above code but for term frequency inverse document frequency vectorizer it is different from count vectorizer in many ways it doesn't just consider the count of the words but it also removes the words which appear too often in a document. It uses these both parameters to give us a weight for the words appearing.

```
print(tfidf_train)
```

```
(0, 47548) 0.12723133339397216
(0, 56634) 0.16405951224786605
(0, 16218) 0.26903902422228715
(0, 51638) 0.11620688174539322
(0, 44216) 0.24117958239972764
(0, 44384) 0.24117958239972764
(0, 35445) 0.13694440055887144
(0, 14981) 0.21459620580264593
(0, 30917) 0.11592757083287557
(0, 14856) 0.09409257856804112
(0, 39171) 0.18238517174107366
(0, 19167) 0.14183130738511285
(0, 42789) 0.12486848323215075
(0, 21678) 0.19011628968188365
(0, 49797) 0.24654656674117886
(0, 36471) 0.3509989989640679
(0, 12572) 0.2884696931533694
(0, 50479) 0.072919571696998
(0, 10161) 0.21797573150502753
(0, 44879) 0.13183649678877474
(0, 46523) 0.26139009167609195
(0, 6812) 0.1328413660919031
(0, 47544) 0.12046029300710442
(0, 14526) 0.2621319020857332
(0, 14865) 0.10783678215837471
:      :
(4750, 3626) 0.016826170249315594
(4750, 15599) 0.021537266230959067
(4750, 52564) 0.08035348582101966
(4750, 14410) 0.024248108328258577
(4750, 10974) 0.02082314839438641
(4750, 29226) 0.013776508726695022
(4750, 57485) 0.018826945053070802
(4750, 9192) 0.11417552815099273
(4750, 47342) 0.02178387330017116
(4750, 26295) 0.049085639245677826
(4750, 47552) 0.03583322897592721
(4750, 52819) 0.025209006589928883
(4750, 37080) 0.035728317922727945
(4750, 7570) 0.030233960020244994
(4750, 53167) 0.01722758208276934
(4750, 45342) 0.01696287486807236
```

The piece of code appearing above gives us the weights of the words we get according to the tfidf vectorizer.

```

def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):

    plt.imshow(cm, interpolation='nearest', cmap=cmap, aspect='auto')
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes, rotation=45)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('')

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="left",
                 verticalalignment="bottom",
                 color="black" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')

```

The python function above is used to construct confusion matrices for different models we are using in our project. Confusion matrix are used to compare our findings using the models with the labels which are already in the dataset.

When we have a problem which has two classes like fake and genuine news we can use confusion matrix to compare our results of models with original labelled data. When we are dealing with this kind of data we can get four different types of results:

- TN: when the models have predicted false and is actually false
- FN: when the models have predicted false but it is actually true
- TP: when the models have predicted true but it is actually true
- FP: when the models have predicted true but it is actually false

To get the confusion matrix, we go over all the predictions made by the model and count how many times each of those 4 types of outcomes occurs. Since to compare two different models it is often more convenient to have a single metric rather than several ones hence we compute two metrics from the confusion matrix, which we will later combine into one: True positive rate (TPR), aka sensitivity, hit rate, and recall is defined as,

$$\frac{TP}{TP + FN}$$

Intuitively this metric corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all the positive data points. False positive rate (FPR), aka. Fall-out is defined as:

$$\frac{FP}{FP + TN}$$

Intuitively this metric corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all the negative data points. In other words, the higher FPR, the more negative data points will be misclassified.

## MultinomialNB with tfidf vectorizer:

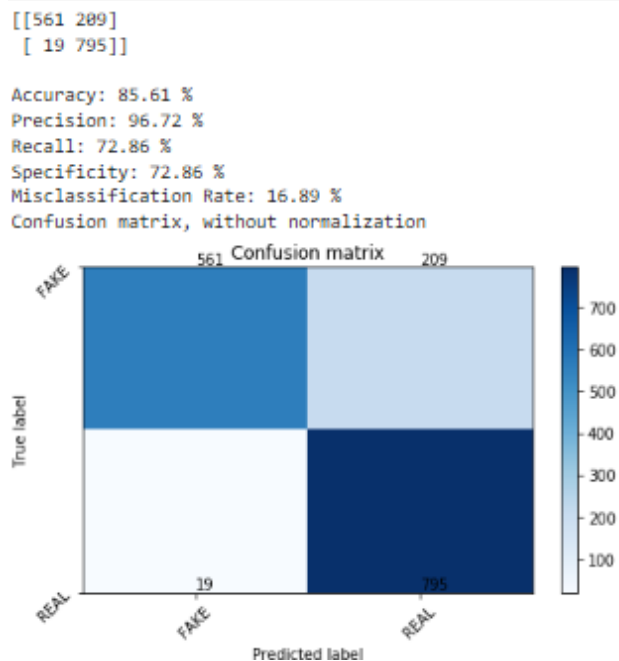
```
bnb=MultinomialNB(alpha=1.0)
y=bnb.fit(tfidf_train,Y_train)
y_predict=bnb.predict(tfidf_test)

cm = metrics.confusion_matrix(Y_test, y_predict, labels=['FAKE', 'REAL'])
print(cm)
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
cm = np.array(cm).tolist()

tn_fp, fn_tp = cm

tn, fp = tn_fp
fn, tp = fn_tp
print("Accuracy:",round(metrics.accuracy_score(Y_test, y_predict)*100, 2),'%')
print("Precision:",round(metrics.precision_score(Y_test, y_predict,pos_label='FAKE')*100, 2), '%')
print("Recall:",round(metrics.recall_score(Y_test, y_predict,pos_label='FAKE')*100, 2), '%')
print("Specificity:", round((tn/(tn+fp))*100, 2), '%')
print("Misclassification Rate:", round((fp+fn)/(tn+fp+fn+tn)*100, 2), '%')
print('')
print('Confusion matrix, without normalization')
```

The first algorithm we are using is Multinomial Naïve Bayes with training data which is vectorised using tfidf vectorizer and we are producing a confusion matrix and also the calculating the five parameters. The output can be seen in the figure given below, where we can see it approximately eighty five percent accurate.



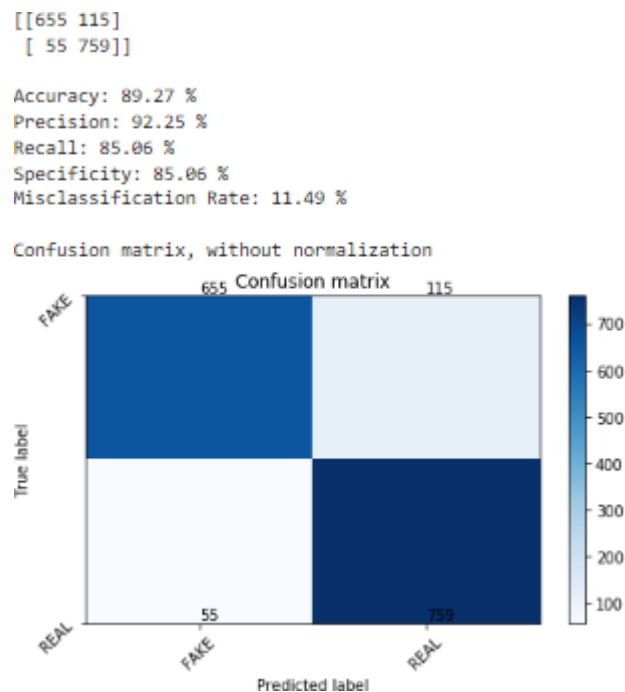


## MultinomialNB with Count Vectorizer:

The next algorithm we use is same but the way we have vectorised the data is different here. We are using count vectorizer here and producing similar type of output here.

```
bnb=MultinomialNB(alpha=1.0)
y=bnb.fit(count_train,Y_train)
y_predict=bnb.predict(count_test)
cm = metrics.confusion_matrix(Y_test, y_predict, labels=['FAKE', 'REAL'])
print(cm)
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
cm = np.array(cm).tolist()
tn_fp, fn_tp = cm
tn, fp = tn_fp
fn, tp = fn_tp
print("Accuracy:",round(metrics.accuracy_score(Y_test, y_predict)*100, 2), '%')
print("Precision:",round(metrics.precision_score(Y_test, y_predict,pos_label='FAKE')*100, 2), '%')
print("Recall:",round(metrics.recall_score(Y_test, y_predict,pos_label='FAKE')*100, 2), '%')
print("Specificity:", round((tn/(tn+fp))*100, 2), '%')
print("Misclassification Rate:", round((fp+fn)/(tn+fp+fn+tn)*100, 2), '%')
print('')
print('Confusion matrix, without normalization')
```

The output from the MultinomialNB with Count Vectorizer is given below as we can see its accuracy is around ninety percent.



## BernoulliNB with Count Vectorizer:

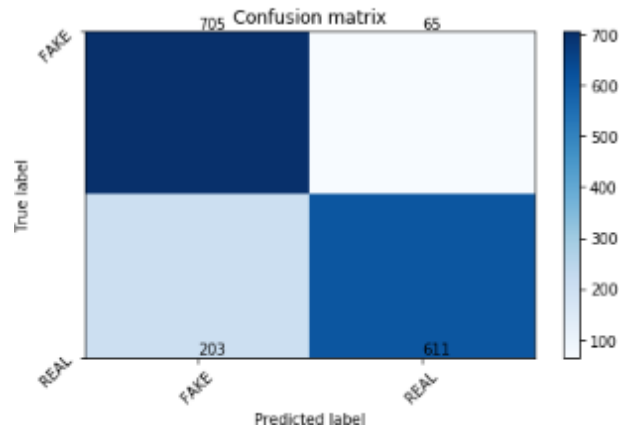
Here we are using Bernoulli NB with count vectorizer as input and the accuracy we are getting here is slightly lesser since it uses binary data i.e. about eighty tree percent.

```
bnb=BernoulliNB(alpha=1.0)
y=bnb.fit(count_train,Y_train)
y_predict=bnb.predict(count_test)

cm = metrics.confusion_matrix(Y_test, y_predict, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
print(cm)
cm = np.array(cm).tolist()
tn, fp, fn, tp = cm
tn, fp = tn - fp
fn, tp = fn - tp
print("Accuracy:", round(metrics.accuracy_score(Y_test, y_predict)*100, 2), '%')
print("Precision:", round(metrics.precision_score(Y_test, y_predict, pos_label='FAKE')*100, 2), '%')
print("Recall:", round(metrics.recall_score(Y_test, y_predict, pos_label='FAKE')*100, 2), '%')
print("Specificity:", round((tn/(tn+fp))*100, 2), '%')
print("Misclassification Rate:", round((fp+fn)/(tn+fp+fn+tn)*100, 2), '%')
print('')
print('Confusion matrix, without normalization')
```

```
[[705  65]
 [203 611]]
Accuracy: 83.08 %
Precision: 77.64 %
Recall: 91.56 %
Specificity: 91.56 %
Misclassification Rate: 15.97 %
```

Confusion matrix, without normalization



## Passive Aggressive Classifier:

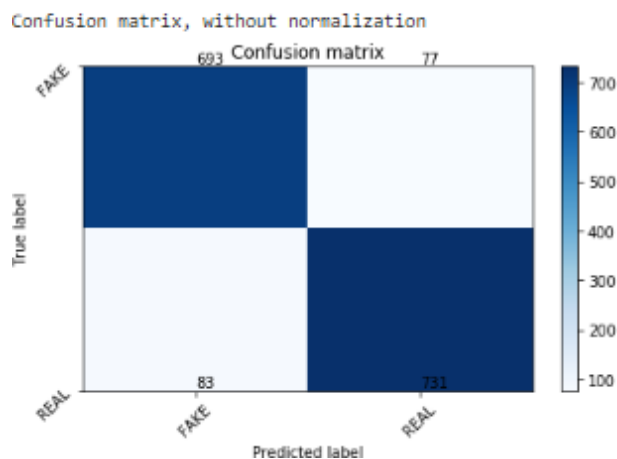
The last algorithm we are using is PAC with count vectorised data as the input. The accuracy we get here is eighty tree percent.

```
linear_clf = PassiveAggressiveClassifier(max_iter=50)

linear_clf.fit(count_train, Y_train)
pred = linear_clf.predict(count_test)
cm = metrics.confusion_matrix(Y_test, pred, labels=['FAKE', 'REAL'])
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
print(cm)
cm = np.array(cm).tolist()
tn_fp, fn_tp = cm
tn, fp = tn_fp
fn, tp = fn_tp

print("Accuracy:", round(metrics.accuracy_score(Y_test, y_predict)*100, 2), '%')
print("Precision:", round(metrics.precision_score(Y_test, y_predict, pos_label='FAKE')*100, 2), '%')
print("Recall:", round(metrics.recall_score(Y_test, y_predict, pos_label='FAKE')*100, 2), '%')
print("Specificity:", round((tn/(tn+fp))*100, 2), '%')
print("Misclassification Rate:", round((fp+fn)/(tn+fp+fn+tn)*100, 2), '%')
print('')
print('Confusion matrix, without normalization')
```

```
[[693  77]
 [ 83 731]]
Accuracy: 83.08 %
Precision: 77.64 %
Recall: 91.56 %
Specificity: 90.0 %
Misclassification Rate: 10.35 %
```



```

def most_informative_feature_for_binary_classification(vectorizer, classifier, n=100):

    class_labels = classifier.classes_
    feature_names = vectorizer.get_feature_names()
    topn_class1 = sorted(zip(classifier.coef_[0], feature_names))[:n]
    topn_class2 = sorted(zip(classifier.coef_[0], feature_names))[-n:]

    for coef, feat in topn_class1:
        print(class_labels[0], coef, feat)

    print()

    for coef, feat in reversed(topn_class2):
        print(class_labels[1], coef, feat)

most_informative_feature_for_binary_classification(tfidf_vectorizer, linear_clf, n=30)
lr_coef = np.array(linear_clf.coef_).tolist()
lr_coef = lr_coef[0]

lr_coef = pd.DataFrame(np.round(lr_coef, decimals=3),
                       feature_names, columns = ["penalized_regression_coefficients"])

lr_coef = lr_coef.sort_values(by = 'penalized_regression_coefficients',
                              ascending = False)

df_head = lr_coef.head(10)
df_tail = lr_coef.tail(10)

# merge back together
df_merged = pd.concat([df_head, df_tail], axis=0)

# plot the sorted dataframe
fig, ax = plt.subplots()
fig.set_size_inches(8, 6)
fig.suptitle('Coefficients!', size=14)
ax = sns.barplot(x = 'penalized_regression_coefficients', y= df_merged.index,
                data=df_merged)
ax.set(xlabel='Penalized Regression Coefficients')
plt.tight_layout(pad=3, w_pad=0, h_pad=0);

```

This piece of code is a really important code as the name of function suggests it gives us the most significant features for the classification we have done. We use this function for the tfidf vectorised data and check which words had the most significance in the case of fake news and real news according to the weights given to them by the tfidf vectorizer. The output can be seen below where the most significant words in fake and real articles are given with their respective weights.

FAKE	-0.6625007001095613	posts	REAL	0.6524262070168739	continue
FAKE	-0.5014764590248822	2016	REAL	0.5127117153267307	rush
FAKE	-0.49493107291954225	guest	REAL	0.44471700194453245	feuds
FAKE	-0.4671899684626758	losing	REAL	0.44129445852014604	held
FAKE	-0.4270307203455656	share	REAL	0.3542746927124335	march
FAKE	-0.3869922858173198	source	REAL	0.33410287136750816	trade
FAKE	-0.346085955277259	article	REAL	0.322692430728058	gop
FAKE	-0.33640689312035255	arrivals	REAL	0.32240575378827374	tuesday
FAKE	-0.3308635224613149	28	REAL	0.3186956319881803	ditch
FAKE	-0.3273262392471672	recent	REAL	0.3168850027563678	candidates
FAKE	-0.3268516456737015	october	REAL	0.314020205803923	friday
FAKE	-0.3200278819108136	comments	REAL	0.3109559671402788	momentum
FAKE	-0.3156866816143849	hillary	REAL	0.3086214068806922	easier
FAKE	-0.3106894128889011	print	REAL	0.29388883804837224	gaining
FAKE	-0.30993726131937965	com	REAL	0.29335083310889876	twitter
FAKE	-0.3027822366506724	nov	REAL	0.28363689849980905	jobs
FAKE	-0.28431680200964055	jewish	REAL	0.27646750629504013	exchanges
FAKE	-0.2814563277964204	establishment	REAL	0.2656799474894897	demolish
FAKE	-0.2798470177164572	08	REAL	0.26291604583526806	sen
FAKE	-0.27678222250794254	swipe	REAL	0.26226170986130687	demo
FAKE	-0.2625409898412498	oct	REAL	0.25872981212611634	marriage
FAKE	-0.2568608428805379	email	REAL	0.2556325989305017	ohio
FAKE	-0.2531364742063395	watch	REAL	0.24850486797519605	reform
FAKE	-0.253066222002227	video	REAL	0.24777104458218177	cabinet
FAKE	-0.24892932972807422	wikileaks	REAL	0.24676920466379368	232
FAKE	-0.24444404321606517	november	REAL	0.2464844558152495	131
FAKE	-0.24078064460269846	add	REAL	0.24087162255335995	takes
FAKE	-0.24028893024271084	corruption	REAL	0.23961035689599142	democrat
FAKE	-0.23989594842064105	believe	REAL	0.2384319138204723	winning
FAKE	-0.23344544587883895	11	REAL	0.23841909991179042	conference

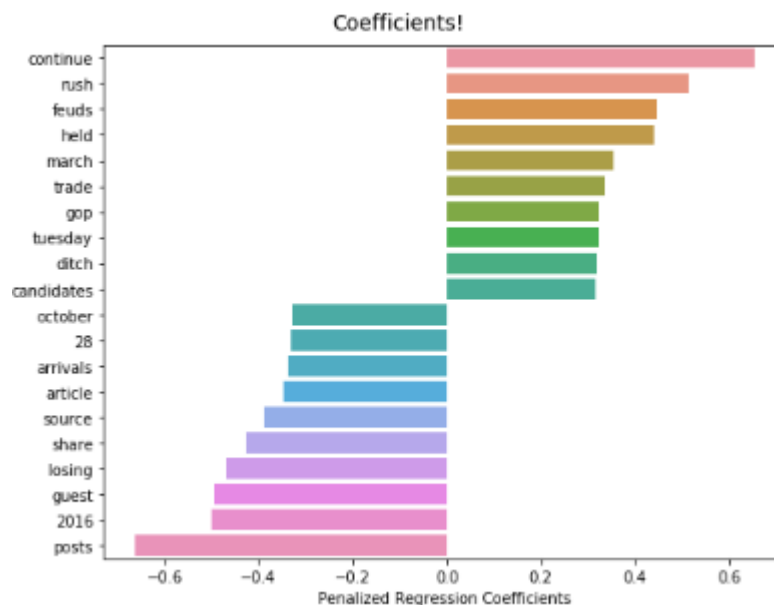


Fig. 28 Penalized Regression Coefficients

## Frontend:

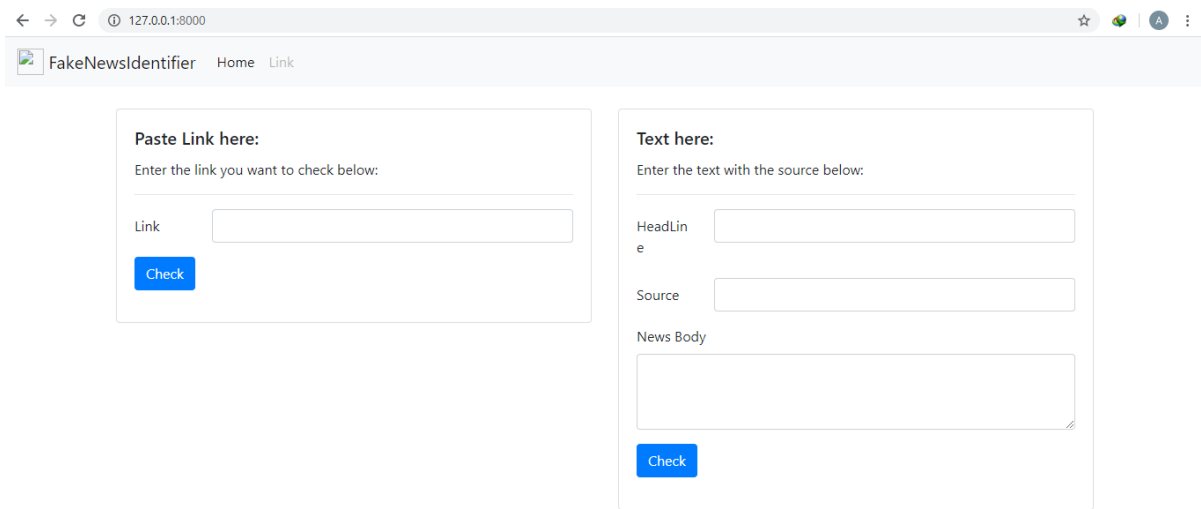


Fig. 29 Frontend of the project

The above figure shows a preview for the web app which is created on the Django frameworks. It can be used to check for whether if the news is fake or not. There are two ways it can be done: one is by putting the link of the news and the second is by putting in the news body, headline and from which source it is from. If the user uses the link form first a web scraping algorithm will scrap data such as news body, title and author and put in a csv file and the database. If the user used the second form then the data will be stored in the database. Then for both the forms when the data is put in the database the code models we have created will be able to check for the news article the user have used as input.

The code below is used to extract news article from different websites, it uses a library called newspaper from python libraries. It can extract features like news body, title, author etc. Below we are extracting only the news headline and the body because we don't need the other data. Then, after extracting data from that particular url it is stored in a csv file.

```

from newspaper import Article,fulltext
import requests
import pandas as pd
import os

url = input("Enter the url:")

article = Article(url)
article.download()
article.parse()

title = article.title
text = article.text

title1 = []
text1 = []
news_data = {}

title1.append(title)
text1.append(text)

news_data["title"] = title1
news_data["text"] = text1

news_df = pd.DataFrame(news_data)
#print(news_df)

if not os.path.isfile('news_data.csv'):
    news_df.to_csv('news_data.csv', index=False)
else:
    news_df.to_csv('news_data.csv', index=False, mode='a', header=False)

df_csv = pd.read_csv('news_data.csv')
print(df_csv)

```

In the code, we first need to put in the url which will come from the form on the web app then the functions from newspaper library are used to extract data from that url. At first the data is put in a list and then in a dictionary. Further, this dictionary is converted into a data frame then it is first checked if the file exists, if it exists the index is false i.e. it will not put the index and the mode is append i.e. the new data will be put in the end in csv file and the header will be false so that the title is not put again and if the file does not exist in the folder the file will be created with extension csv and the header will be set at the top.

As we can see below, we are inputting urls when we put in first url the data is put in csv file and at the end of the program the csv file is read and it contained only one news. Then we input other url and it is appended at the end of the csv file.

```

Enter the url:https://www.thehindu.com/news/international/chinese-president-xi-jinping-meets-pla-urges-battle-preparedness/article31688691.ece
      title                                     text
0  Chinese President Xi Jinping meets PLA, urges ...  Chinese President Xi Jinping on May 26 called ...

```

```

Enter the url:https://www.thehindu.com/news/international/security-law-not-a-threat-to-hong-kong/article31681221.ece
title text
0 Chinese President Xi Jinping meets PLA, urges ... Chinese President Xi Jinping on May 26 called ...
1 Security law not a threat to Hong Kong, says C... The city's chief executive, Carrie Lam, told r...

```

```

Enter the url:https://www.thehindu.com/news/national/other-states/cyclone-evacuees-and-migrants-under-one-roof/article31680675.ece
title text
0 Chinese President Xi Jinping meets PLA, urges ... Chinese President Xi Jinping on May 26 called ...
1 Security law not a threat to Hong Kong, says C... The city's chief executive, Carrie Lam, told r...
2 Cyclone evacuees and migrants under one roof Before cyclone Amphan struck on May 20, offici...

```

The spreadsheet below is the csv file where the news articles are being stored for further processing as we can see the news articles are being appended at the end of the csv files.

	A	B	C	D	E	F	G	H
1	title	text						
2	Chinese President Xi Jinping meets PLA, urges battle preparedness	Chinese President Xi Jinping on May 26 called on the military to think about worst-case scenarios and to scale up battle preparedness						
3	Security law not a threat to Hong Kong, says Carrie Lam	The city's chief executive, Carrie Lam, told reporters that there was no need for us to worry over the move being considered by China						
4	Cyclone evacuees and migrants under one roof	Before cyclone Amphan struck on May 20, officials of the Achintyanagar gram panchayat in Pathapatima block of West Bengal's South 24 Parganas district said						
5	Piyush Goyal blames Maharashtra for non-running of 55 trains	The political slugfest over running of Shramik Special trains continued with Railway Minister Piyush Goyal taking to Twitter to allege that the Maharashtra government is responsible						
6	Coronavirus lockdown   A day after domestic flights resumed, seats go abegging	Weak demand for air travel continued to haunt airlines on the second day of domestic flight operations on Tuesday as most flights saw only 25% seats filled						
7								
8								
9								
10								
11								

Fig. 30. csv file where the data is stored



## Chapter 5: Conclusion

Daily news plays quite an important role in every citizen's everyday life since it helps us know what is going on or happening in our country and all around the world example what has the elected government done, about new government policies and how they affect our lives etc. Basically it helps a citizen a lot in keeping the elected government in check on what they are doing. For this reason alone the news media is also known as the forth pillar of a democratic country. But this rise in the circulation of fake news wither on the traditional news source or the contemporary news sources like social media has tarnished this sacred pillar of democracy. There are many different factor behind this exponential rise in the circulation of misinformation or fake-news like for gaining upper hand in political races, sometimes they are used for personal financial gains or also used for increasing business like it was used in India by spreading misinformation about shortages of essential things like masks, hand sanitizers, gloves and other medical supplies were spread for the gain in their business. Since, it is not an easy task to stop the flow or circulation of this misinformation therefore an automatic system is quite essential which can help detect and notify the user about the news article they are reading i.e. whether it is fake or not.

The project we worked upon is able to solve this problem of spreading of fake news up to only some extent since right now we have used a fairly basic or simple approach to solve to this problem. For the project to be completed successfully we had to first collect a dataset of news which consisted both fake and real news, which was a very challenging task in itself since verifying the accuracy of a news article is quite difficult to predict that whether it is fake or not. Then the dataset of thirteen thousand news article we got was divided into two different sets which are training set and testing set respectively in the ratio of seventy to thirty in which the bigger set was used to train the model and the smaller set was used for testing of the classifier which was produced by the model. Then from the dataset itself we tried to find features like number of words or the frequency of words used in fake and real news and for tfidf vectorizer, tfidf weights were found for the words i.e. basically we converted the text in the news articles which the computer can't understand, into bunch of numbers to make the computer understand in some context. Then we used this data in our models like Naïve Byes and its different types like MultinomialNB and BernoulliNB and PAC to get different

classifiers. Then we used the remaining data for testing purposes. The results we got from them were:

	<b>MultinomialNB with TfidfVectorizer</b>	<b>MultinomialNB with CountVectorizer</b>	<b>BernoulliNB</b>	<b>Passive Aggressive Classifier</b>
<b>Accuracy%</b>	<b>85.6</b>	<b>89.2</b>	<b>83.08</b>	<b>83.08</b>
<b>Precision%</b>	<b>96.72</b>	<b>92.25</b>	<b>77.64</b>	<b>77.64</b>
<b>Recall</b>	<b>72.86</b>	<b>85.06</b>	<b>91.56</b>	<b>91.56</b>
<b>Specificity%</b>	<b>82.86</b>	<b>85.06</b>	<b>91.56</b>	<b>94.55</b>
<b>Misclassification%</b>	<b>16.89</b>	<b>11.49</b>	<b>15.97</b>	<b>6.0</b>

Table 2. Comparing performance of different models used

The results which we got from our classifiers are being compared on the basis of five different parameters which are accuracy, precision, specificity, misclassification and recall. In terms of being the most accurate, Multinomial naïve bayes with count vectorizer possess the highest accuracy percentage of approximately ninety and the least accurate was BernoulliNB and PAC at 83.08. The second parameter is the precision, the most precise classifier is Multinomial naïve bayes with tf-idf vectorizer and the least precise are similar to that of accuracy with 77.64 percent. In the case of third parameter the recall which is the sensitivity, it is the highest in both BernoulliNB and passive-aggressive classifier and it is lowest in the MultinomialNB with tf-idf vectorizer. The specificity percent is most in the PAC and lowest in the case of MultinomialNB with tfidf vectorizer. The last parameter is the misclassification is highest in the Multinomial naïve bayes with tf-idf vectorizer and the lowest on the passive aggressive classifier.

The approach we have used in our project for the transformation of the linguistic part of the news articles can be of great use in problem related to text like text classification. Though the way we are solving this problem is giving us an accuracy of ninety percent while testing, the attempt to solve this problem should not be just related to the usage of words in the news

articles but efforts to improve it with other properties like the interaction of people with this article. Also, the data set we used in the project includes real news from various real news from news firms like British broadcast company, CNN etc. instead of using short text which can also present disadvantage of model not be able to train properly.

### **Future Scope:**

Fake news is an ever growing problem due to the increase in users of internet and therefore increase in the users of social media. The approach we are following in our project is a fairly simple one in which we are using the frequency of words those were used in the news articles to find out which words appeared the most often in the false articles and which words appeared the most often in real news articles. But, this problem is vast and is a research field in itself for any future work purposes where many things can either be added in the current approach or entirely new approach from scratch can be used to solve menacing problem. The approach we are following in this project makes use of the news article texts linguistics and lexicon is fairly a good one but the approaches used must not be just limited to these extra characteristics of news articles, but also the interaction of a user with the news article i.e. how they got there, did they closed it immediately, whether they liked or shared the news or whether they put up a comment under it. Also, this can be used up with the GDU to draw relations between a page which has posted a news article and the article itself. New deep learning techniques like convolution and recurrent neural network can also be used. Some of these methods are discussed below:

### **Method based on the language:**

These methods can be used to add some extra features on our current approach of counting frequency of the words. This linguistic based approach uses the similar way of finding what is the different styles in which article are written, what type of lexicon are used and what is the emotion behind writing the article. They are based on finding out these above mentioned characteristics from the articles. Some of the methods which are based on the language of text are:

- **Bigrams and n-grams:** Some words are not able to convey the message individually they have to be considered together i.e. they provide more weight together in a news

article example the words ‘thank’ and ‘you’ are appearing individually in our word counts but since they are quite often used together they must be considered as one words and be given count as when it appears in the news article. Similarly there are higher level n-grams like trigrams etc.

- **POS tags:** Part-of-speech tags are very important and are used to put up tags with words which tells what its part-of-speech is. They work by putting a tag on individual words on the basis of how they are used syntactically example adverbs, pronouns, prepositions, nouns, verbs etc. Many other research works have found out that the count in particular text body of part of speech tag can be used to find where do they belong or particularly which topic they represent like an article in context of a meeting, medical, consulting each possess unique usage of words.
- **Probabilistic CFG:** Context free grammar is very important while representing the sentence in a linguistic structure with the leaf node representing words and the middle nodes representing part of speech tags like adverbs, verbs, nouns, pronouns etc. depending on how the sentence has been made up it can possess many different representations.

**Gated diffusive unit:** This is a really good approach to draw a relation between the news article to its author or the publisher. Initially it also uses similar steps to check what type of language is being used in the articles and by the authors which is down the road used to draw relation between these two entities which are news and author. A third entity can also be used which is ‘subject’ that basically represents the subject of the article.

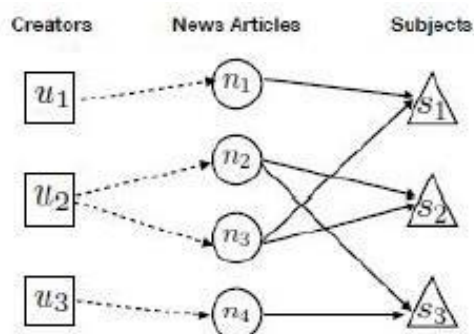


Fig. 31. Relationships of Articles, Creators and Subjects

**Methods based on deep learning:** Methods based on deep learning are really good to be used on text based dataset since they are able to drastically decrease language based analytics using feature extraction which is done automatically, it can extract both simple features and more compound characteristics which are not easy to define.

- **CNN:** They are quite often being used in text based problems i.e. NLP based problems which also include text classification and linguistic analysis. It can be used for fake news detection based more on the content of the news article.
- **RNN** uses the words embedded according to the sequence of words, it takes one letter at a time, using at these individual steps the information of the current token to find its state which can't be seen easily which also helps with info about the last word. This state we get is considered as the feature which are taken out by the recurrent neural network for given text body.

## References

- News use Across social media platforms 2016  
By Jeffrey Gottfried and Elisa Shearer at Pew Research Center
- Social Media and Fake News in the 2016 Election by Hunt Allcott and Matthew Gentzkow
- Automatic Detection of Fake News  
By Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea  
Computer Science and Engineering, University of Michigan  
Department of Psychology, University of Amsterdam
- Fake News Detection Using Naive Bayes Classifier  
By Mykhailo Granik, Volodymyr Mesyura (Computer Science Department) Vinnytsia National Technical University, Vinnytsia, Ukraine  
<http://ieeexplore.ieee.org/document/8100379/>
- Combating Fake News: A Survey on Identification and Mitigation Techniques by  
Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Yan Liu (University of Southern California)  
Ming Zhang (Peking University)
- Fake News: A Survey of Research, Detection Methods, and Opportunities by Xinyi Zhou, Reza Zafarani (Syracuse University, USA)
- FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network  
By Jiawei Zhang, Bowen Dong, Philip S. Yu  
IFM Lab, Department of Computer Science, Florida State University, FL, USA  
BDSC Lab, Department of Computer Science, University of Illinois at Chicago, IL, USA
- Fake News Detection on Social Media: A Data Mining Perspective  
By Kai Shuy, Amy Slivaz, Suhang Wangy, Jiliang Tang , and Huan Liuy Arizona State University, Tempe, AZ, USA  
Charles River Analytics, Cambridge, MA, USA  
Computer Science & Engineering, Michigan State University, East Lansing, MI, USA

- Automatically Identifying Fake News in Popular Twitter Threads by Cody Buntain  
Available: <http://ieeexplore.ieee.org/abstract/document/7100738/>
- Article what is 'fake news,' and how can you spot it?  
<https://www.theglobeandmail.com/community/digital-lab/fakenews-quiz-how-to-spot/article33821986/>
- Wikipedia article about Naïve Bayes. Available:  
[https://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)
- A proposed way of implementation.  
<https://www.datacamp.com/community/tutorials/scikit-learn-fakenews>
- <https://www.statista.com/statistics/278407/number-of-social-network-users-in-india/>

## Group 82

### ORIGINALITY REPORT

5%

SIMILARITY INDEX

1%

INTERNET SOURCES

1%

PUBLICATIONS

5%

STUDENT PAPERS

### PRIMARY SOURCES

1

Submitted to University of Mumbai

Student Paper

3%

2

Submitted to University of Greenwich

Student Paper

1%

3

Submitted to Troy University

Student Paper

<1%

4

Submitted to Netaji Subhas Institute of

Technology  
Student Paper

<1%

5

[link.springer.com](https://link.springer.com)

Internet Source

<1%

6

James Albert Cornel, Carl Christian Pablo, Jan

Arnold Marzan, Vince Julius Mercado et al.  
"Cyberbullying Detection for Online Games Chat  
Logs using Deep Learning", 2019 IEEE 11th  
International Conference on Humanoid,  
Nanotechnology, Information Technology,  
Communication and Control, Environment, and  
Management ( HNICEM ), 2019

Publication

<1%



---

7	Submitted to University of Bristol Student Paper	<1%
8	Submitted to AlHussein Technical University Student Paper	<1%
9	Priyanka Meel, Dinesh Kumar Vishwakarma. "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities", Expert Systems with Applications, 2020 Publication	<1%
10	arxiv.org Internet Source	<1%
11	Submitted to City University Student Paper	<1%
12	kamokun.com Internet Source	<1%
13	Yong Fang, Jian Gao, Cheng Huang, Hua Peng, Runpu Wu. "Self Multi-Head Attention-based Convolutional Neural Networks for fake news detection", PLOS ONE, 2019 Publication	<1%
14	Submitted to University of Leeds Student Paper	<1%

---

Submitted to Parkway South High School

15

Student Paper

<1%

---

16

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu. "Fake News Detection on Social Media", ACM SIGKDD Explorations Newsletter, 2017

Publication

<1%

---

17

Submitted to Federal University of Technology

Student Paper

<1%

---

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off

# JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

## PLAGIARISM VERIFICATION REPORT

Date: 24/07/2020

Type of Document (Tick):  PhD Thesis  M.Tech Dissertation/ Report  B.Tech Project Report  Paper

Name: Anmol Rana Department: CSE Enrolment No: 161359

Contact No. 7018025259 E-mail. raanaa.anmol@gmail.com

Name of the Supervisor: Dr. Hemraj Saini

Title of the Thesis/Dissertation/Project Report/Paper: FAKE NEWS IDENTIFICATION

### UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/ revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

#### Complete Thesis/Report Pages Detail:

- Total No. of Pages = 71
- Total No. of Preliminary pages = 61
- Total No. of pages accommodate bibliography/references = 63

(Signature of Student)

### FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 5(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.



(Signature of Guide/Supervisor)

Signature of HOD

### FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	<ul style="list-style-type: none"><li>• All Preliminary Pages</li><li>• Bibliography/Images/Quotes</li><li>• 14 Words String</li></ul>		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by  
Name & Signature

Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at [plagcheck.juit@gmail.com](mailto:plagcheck.juit@gmail.com)