A COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR FAKE NEWS DETECTION

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF DEGREE
OF
MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted by:

VISHAL DOHARE
2K20/SWE/25

Under the supervision of

MS SHWETA MEENA
(Assistant Professor)

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2022

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I, Vishal Dohare, Roll No. 2K20/SWE/25 student of M. Tech (Software Engineering) hereby declare that the project Dissertation titled "A Comparative Analysis of Machine Learning Techniques For Fake News Detection" which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Vishal Dohare

VISHAL DOHARE

Place: Delhi

Date: 31/05/2022

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"A Comparative Analysis of Machine Learning Techniques For Fake News Detection"** which is submitted by Vishal Dohare, 2K20/SWE/25 Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

31/5/2022

Place: Delhi
Date: 31/5/2022

MS SHWETA MEENA
SUPERVISOR
Assistant Professor
Department of Software Engineering
Delhi Technological University
(Formerly Delhi College Of Engineering)
Bawana Road, Delhi - 110042

ACKNOWLEDGMENT

The success of this project requires the assistance and input of numerous people and the organization. I am grateful to everyone who helped in shaping the result of the project.

I express my sincere thanks to **Ms Shweta Meena**, my project guide, for providing me with the opportunity to undertake this project under her guidance. Her constant support and encouragement have made me realize that it is the process of learning which weighs more than the end result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout with new ideas, provided information necessary and pushed me to complete the work.

I also thank all my fellow students and my family for their continued support.

VISHAL DOHARE

ABSTRACT

Differentiating between false and true information has become challenging due to the easy availability and rapid increase of data presented on social media platforms.The convenience with data may be exchanged has contributed to the fast spread of data piracy. The authenticity of social networking sites is questioned in this field where incorrect news is widely shared. We have introduced some new characteristics and evaluate the accuracy of current approaches and features for automatically detecting fake news. In addition to assessing the essential characteristics given in this study for fake news detection, we proposed a new set of features and examined the prediction effectiveness of current approaches and attributes for accurate classification of false news. and our findings show some surprising details about the utility and significance of traits in spam detection. We hope people will be able to distinguish between counterfeit and genuine news.

INDEX

LIST OF FIGURES

LIST OF ABBREVIATIONS

TFIDF – Term Frequency Inverse Document Frequency

SVM – Support Vector Machine

TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

GBC:gradient boosting classifier

ML: machine learning

NLP:natural language processing

CHAPTER 1

INTRODUCTION

1.1 Introduction to machine learning

We are living in the data era, in which everything around us is connected to a data source and what we do is stored virtually. The Internet of Things (IoT) data, military data, smart urban data, corporate data, cell phone information, social networks, patient records like COVID-19 data, and other types of data abound in today's electronic environment. The information might be organized, semi-organized, or unorganized which is growing day by day. Finding useful information from the data that is unorganized and semi organized we need a machine that is smart enough to identify useful insides from the information and can automatically learn from it. is required so that we can utilize the information in a better way that can improve our day to day life. Artificial intelligence (AI) and, in especially, machine learning (ML), have exploded in popularity over the years in the context of information processing and computation, allowing programs to perform intelligently. ML gives systems the capacity to learn and improve based on their experiences without having to be explicitly written. "A learning algorithm can be explanatory, in which case the data is used to describe what happened; predictive, in which case the data is used to predict what would happen; or normative, in which case the data is used to make recommendations about what steps to take."
“Machine learning is a type of AI technology that is described as a machine's capacity to mimic natural intelligence. A.i. systems are utilized to complete complicated jobs in a comparable pattern to how people solve issues.”

1.2 Types of machine learning

There are three types of machine learning supervised, unsupervised, reinforcement

1.2.1 Supervised Learning

supervised ML models are trained with labeled sets of data, promoting learning and improvement over time. For instance, an algorithm might be taught using photographs of animals and other objects, all of which would be classified by living beings, and the computer would learn how to recognise image of animals on its own. The most popular sort of machine learning nowadays is supervised machine learning.

1.2.2 Unsupervised Learning

A software that uses unsupervised machine learning looks for patterns in unlabeled data. Unsupervised machine learning can uncover patterns or trends that individuals aren't aware of. An unsupervised machine learning software, for example, may sift through internet sales data to identify distinct sorts of customers buying stuff.

1.2.3 Reinforcement Learning

The reinforcement of active learning is comparable to that of a human person; for example, a youngster understands different stuff through his daily encounters. Playing a game in which the environment is the game, the motions of an agent at every step define states, and the agent's aim is to acquire a high score is an example of reinforcement learning. Agents get information in the form of punishments and incentives.

1.3 Introduction to computer security

To identify hazards, extract threat traits, and encode those qualities into the system to detect threats, traditional security software requires a large amount of human labour. This time-consuming method might be made more efficient using machine learning techniques. As a result, a number of academics have investigated various machine learning algorithms to increase the speed and accuracy of attack detection. The Internet's arrival has completely transformed modern civilization. The huge amount of data generated by the Internet, as well as new data-driven decision-making tools, have contributed to the Internet's success. Data analysis and the underlying principles of machine learning, which extract useful information from seemingly unstructured volumes of data, are critical to online advertising, recommendation systems, consumer profiling, and many other Internet-related enterprises. Unfortunately, the Internet's widespread usage has fueled misuse and the growth of sophisticated cyber-crime. It has allowed criminals to establish long-term enterprises based on the exploitation of security flaws. Attackers use unique exploits techniques to avoid being detected by security systems, placing a lot of burden on cybersecurity firms. The attackers are motivated to find new ways to elude detection as a result of the vendors' success. The kitty game between both the security field and the buried world of cybercrime highlights a basic scientific concern with analysis of data and machine learning approaches: they were built on the assumption of "trustworthy" data and did not explicitly account for the possibility of data manipulation by adversaries. Data-driven security instruments have been demonstrated to be readily cracked in several investigations [19], raising the question of whether machine learning approaches can be used in hostile contexts at all [20]. Recent advances in learning technique [21] and growing experience with its use in security practise [22] have highlighted the need for a better understanding of machine learning's security implications. These results prompted the Perspectives Paper "Machine Learning algorithms for Computer Security". The goal of the presentations and discussions performed during this session was to produce appraisals of current approaches as well as to highlight unresolved challenges and research objectives. The work was also a significant step in shaping the scientific community in the subject of safe machine learning, which is still in its early stages. It brought together experts from a variety of fields, including machine learning and security, as well as spam filtering, internet advertising, and computer forensics. The primary conclusions of the article are summarised in this manifesto, which also includes an outline of future scientific breakthroughs in safe machine learning.

Spam detection look for highly unsafe emails sent out by intruders or businesses. Intruders regularly send emails promising to offer a helpful service or to protect you from potential threat, but they are typically junk designed to get you to click on a link that instals malicious software or leads you to a dangerous website. Spam can also include relatively innocuous material, but it can clog up your inbox, taking up precious storage space and making it more difficult to distinguish between vital, helpful emails. Spam filters are capable of identifying spam emails.

Machine learning enables our digital shopping safer and more efficient by detecting fraud activities. Whenever we execute an internet banking, there are several ways for deception to happen, such as the use of bogus accounts, bogus ids, and the robbery of cash in the process of the transaction. The Multilayer Feedforward Networks allows us to identify this by assessing if the transaction is real or illegitimate. Each legitimate transaction's result is converted into a set of hashed, which would then be used as the inputs for the next round. Each actual transaction has a pattern that differs from the fraud transaction, allowing it to be identified and making our online purchasing safer.

1.4 Introduction to fake news detection

Many people from all around the world have been intrigued by the designation of any news report, article, or blog as fake or legitimate[1]. Several studies have been conducted to determine the impact of faked and false news on the general public and people's reactions to such news. Falsified news, sometimes known as faked pos news, is any phone textual or non-textual information created with the intent of convincing readers to believe something that is not real. Fake news has proliferated rapidly over the last decade, most notably during the 2016 US presidential election[2].

In this study, many machine learning approaches such as logistic regression, decision tree, and random forest are used for the assess performance across several data. Increased internet dissemination of misleading information has resulted in a host of problems, including in politics, sports, health, and science. False news has ramifications in the stock exchanges, A rumour might have far-reaching consequences, perhaps bringing the market to a standstill. We used six different machine learning algorithms to find out whether the news is true or false and the algorithms are logistic regression, random forest, Gradient Boosting Classifier (GBC), decision tree, Support Vector Machine (SVM) and naive bayes. We have extracted the feature from the dataset and use the countVectorizer for tokenization and count of words. It essentially turns a sequence of textual information into a token vector. It is represented into vector form. It is to create a bag that contains a word and their frequency or count for each word. For this we used the python library CountVectorizer.

We used Term frequency-inverse document frequency as TF-IDF. This is a method for determining the total number of words in a set of documents. Each word is frequently given a score to signify its importance in the document and dataset[3]. The logarithm of the total count of documents divided by the count of documents, files containing the phrase yields the inverse document frequency. Based on these two libraries, we count the number of words in the documents and provide features for training and testing. We trained our model using the machine learning algorithms for best results and then we used a confusion matrix that tells the performance of the model. The classifiers were trained according to the following guidelines: Individual models are trained with the best parameters, then tested by choosing the best output label from all models[4].

"The purpose of this research is to develop a predictability classifier that can determine if a user's claim is true or not." The "Fake News Detection System" project employs ML and NLP algorithms. ML is a field of study that focuses on applying computer approaches to translate empirical data into usable models. The machine learning discipline arose from the merger of traditional statistics and artificial intelligence groups[5]. NLP is a field of artificial intelligence (AI) concerned with the ability of computers to comprehend text and spoken words in the same way as humans do, and especially, how to design systems that can analyse and assess enormous amounts of raw language input [6].

These days, social media is utilised to quickly distribute misleading information. "A lie spreads halfway around the globe even before truth can put its trousers on," Wiston Churchill famously said. Because of the high number of active users on social media, rumours and false information spread like wildfire. The reaction to such news can be a deciding element in determining whether the news is "false" or "genuine."

1.5 Definition of fake news and its types

Fake news is information that is purposefully and verifiably untrue and has the potential to deceive viewers or readers. This concept has two crucial dimensions: "intention" and "authenticity." For starters, bogus news spreads unverifiable nonsense. Second, fake news is manufactured with the purpose of misleading the audience. This term has been frequently used in recent studies [11]. Fake news may be divided into three categories in general. We may place those forms of news in the first group - "Actual Fake News," which are fake and made up by the article's creator. The second category, dubbed "fake news that is genuinely satire," exists solely to .entertain rather than to deceive its viewers. As a result, purposely false and misleading fake news differs from evident satire or parody"Poorly reported news that serves an agenda" is the third category. This style of news contains some accurate information, but it is mostly intended for political propaganda. To make their study easier, several academics have simplified the many categories of fake news. For example, according to [1], there are a few sorts of news that cannot be classified as "fake":

(1) Satire news with sufficient context.

(2) Inadvertently generated false information.

(3) Conspiracy theories are difficult to categorise as factual or incorrect.

This study [1] discussed the "characterization" and "detection" parts of the false news detection issue.

Fake news can be believed by people who believe their interpretations are the only correct ones. They believe that individuals who disagree with them are illogical and prejudiced[15]. People who like to hear news that confirms their current beliefs and viewpoints are also prone to prejudice[16],Others, on the other hand, are civic minded and prefer to be on the safe zone by absorbing and judging news in accordance with social norms, even if the material conveyed is fake.

The two fundamental foundations of Fake news in traditional media are psychological and social human behavioural tendencies. Malicious fake accounts, in addition to these two criteria, are the basis of misinformation [1].

1.5 Impact of Fake News and Detection Techniques

According to Wang et al. in their article, the scourge of false news causes not only a mistrust of the news stories but also political upheaval. People's decisions about who to support for in election are influenced by fake news. As per experts at the Oxford University, fake news was popular and spread swiftly with the growth of social media bots in the run-up to the 2016 US Presidential election [16]. A social bot is a social media account that has been designed to create content and engage with people or malevolent bots [6]. According to studies, these bots greatly affected election internet debates [1]. Journalists have a harder time reporting important news items because of fake news[7]. According to a Buzzfeed analysis, the top twenty fake news stories on Facebook about the 2016 US National election got more attention than the top twenty campaign articles from Nineteen major news organisations[8]. Fake news is usually the cause of death. People have been physically assaulted as a result of false information circulated on social media. Because of fake news, Rohingya individuals in Myanmar have been imprisoned, incarcerated, and in some cases raped and killed. These efforts appear to have sparked real-life worries, affecting civic involvement and community discussions. There are different techniques that used for fake news detection.

1.5.1 Style Based Detection

The manner the information is given to the consumers is the subject of style-based detection. Although fake news is rarely published by journalists, the manner of writing may change [9].In [35], the author used PCFG (Probabilistic Context Free Grammars) to translate sentences into rules such as lexicalized/unlexicalized production rules and grandparent rules, which characterise syntactic structure and may be used to identify dishonesty. Another article [22] used deep network models, such as Convolutional neural networks (CNN), to verify the accuracy of news. Fake stories frequently depict extreme conduct in behalf of a political party. Hyper-partisan writing styles are the name for this form of writing..To recognize which type of writing style, linguistic features might be used. Several article titles provide just enough content to entice readers to visit a certain webpage or video. Click-bait headlines [1] are a form of eye-catching titles or online links that can be a source of hoaxes. Style-based approaches also include methods that use word embedding characteristics to locate tokens with a greater posterior distribution in two groups.

used the Naive Bayes technique to choose tokens that were most descriptive of the categorization and employed them in deep learning and logistic regression. Although writing styles can help detect false news, they appear to be less effective since fake news might be written in a manner that is close to that of legitimate news [10].

1.5.2 Stance based detection:

This approach evaluates how a set of reliable sources or a series of social media posts feel about a claim: accept, reject, neutrality, or irrelevant.To determine the stance of the articles, the authors employed syntactic and related characteristics fed via a multi-layer perceptron (MLP) with one hidden layer [10].They used countrywide research studies to hard-code the reputation score feature of diverse sources.

1.5.3 Based Detection on Social Media:

Visual Images that have been digitally manipulated are moving like wildfire on social networks. Photoshop may now be used freely to alter photos sufficiently to deceive others into believing they are viewing the actual thing. A large variety of approaches for detecting manipulation in movies [40] and photos have emerged from the field of multimedia forensics. Yet, there are multiple causes why all these strategies are unlikely to work on social networking sites photographs, according to. On the internet, there are a few simple tools for detecting photo-shopped photographs, such as Google's reverse image search, Get image metadata, and so on. has identified a number of visual and statistical indicators that may be used to determine whether or not media is legitimate.

CHAPTER 2

RELATED WORK

According to several studies, there are four fundamental categories of fake news detection methods: knowledge-based, style-based, stance-based, and visual-based. This section explains research into all of these sorts of detection approaches, as well as a few more notable studies that have gotten more attention. It also discusses some of the key elements that have lately been employed in numerous research studies to assess news reliability. The feature extraction stage of machine learning is critical. It seeks to fact-check statements made in news articles using outside sources. Public web and knowledge base are two common external sources. The assertions are compared against open web sources in terms of results and regularity [18], while the Knowledge graph is utilised to see if the claims can be deduced from the facts in the graph or not.Several fact-checking sites (such as AltNews, Snopes, Smhoaxslayer, and Boomlive) rely on specialists to manually verify the authenticity of news. Automated fact- 6 checking, which is connected with categorization of sentences into non-factual, unimportant factual, and check-worthy factual assertions, is an issue with this technique.

CHAPTER 3

METHODOLOGY

The categorization approach is presented in this section. A technique for identifying bogus articles is developed using this approach. The dataset is classified using supervised machine learning in this manner. Fake news frequently elicits strong emotions and spreads quickly on social media. The believability of the news is determined using a response-based approach that considers the gathered comments on tweets/posts. There were two stages to this project. I used the higher prior distribution approach on the paper's content and titles in the first phase. Despite the fact that I saw improved accuracy results, I thought this strategy to be ineffective since fake news can exist in a well-written piece. In the second step, I suggested a technique for more correctly classifying false news by examining the reaction to such news pieces."

The following five sub-phases were used to perform the task:

1) Data Collection
2) Data preprocessing
3) Evaluation of different model performance based on extracted features
4) Improving performance
5) Discussion and results

The dataset collecting phase is the initial stage in this classification challenge, followed by preprocessing, feature selection, dataset training and testing, and lastly running the classifiers. The suggested system technique is depicted in Figure 1. The approach is centred on employing algorithms to run numerous tests on a dataset.

The Sci-kit, sklearn, pyplot, seaborn libraries were used to construct this project in Python. Python includes a large number of packages and extensions that are useful for Machine Learning. The Sci-Kit and sklearn [6] libraries are the best site for machine learning techniques, since it contains practically all types of ML methods for Python, allowing for easy and rapid assessment of Ml techniques in the jupyter notebook.

4.1 Flowchart

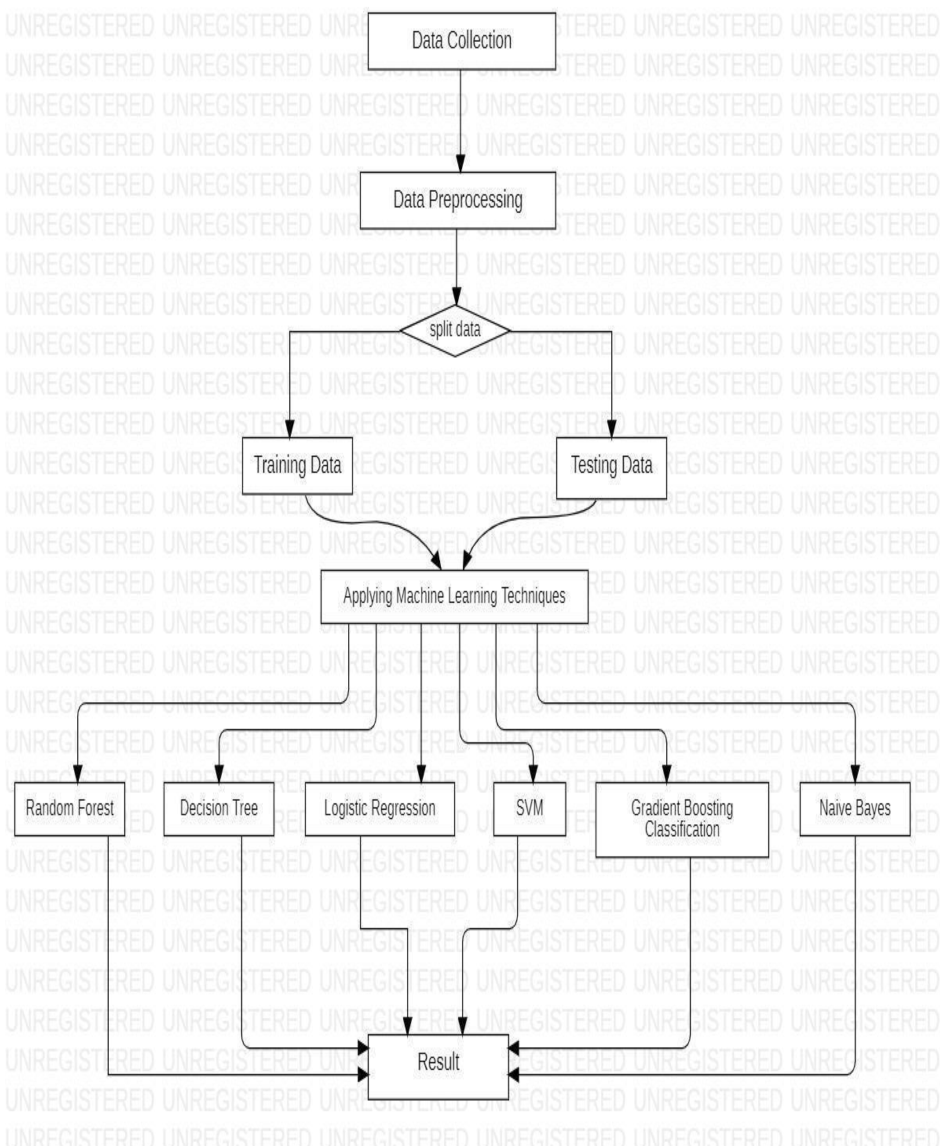


Fig 4.1 flowchart

The strategy of identifying false news is as follows during the model construction process: The first phase is data gathering, followed by crude noise reduction, features synthesis, and feature selection using NLP (Natural Language Processing). After that, partition the dataset and use machine learning methods to develop the recommended classification algorithm. After the dataset has been properly preprocessed in the system, a message is created to allow the algorithms to be applied to the trained section.

CHAPTER 4

IMPLEMENTATION

4.1 Data collection and Exploration

We may get online news from a number of sources, such as social networking sites, search engines, news agency websites, and fact-checking websites. On the Internet, there are a number publicly available databases for categorising false news, including BuzzfeedNews, LIAR, BS Detector, CREDBANK [1], and others. In several study studies, these statistics have been used to judge the veracity of news. In the following sections, the sources of the dataset used in this investigation are briefly discussed.

4.1.1 Proposed method data collecting

I discovered that neither of the publicly accessible datasets provided Responses for the suggested approach, which is dependent on user responses. I gathered the necessary information from the social media sites Twitter and kaggle. This data collection procedure has two primary components.

1) Collecting both fake and true news

2) exploring the data sets

4.1.1.1 Collecting both fake and true news

I used fact-checking websites for this. Some of the agencies that are real and acknowledged for breaking fake stories include AltNews.com, Smhoaxslayer.com, and Boomlive.com. I looked there at pieces they published refuting the fake news. I only sought for data that was relevant to the dataset's creation. The relevant data included Tweets and Facebook posts from various people, which were flagged as fake by fact-checking organisations. During the initial phase, all of the urls for fake Twitter and Facebook postings were gathered. This was the less difficult assignment. I gathered posts/tweets from a few well-known news organisations, journalists, and even some verified users and organisations. I chose the news because it had powerful emotions (both happy and bad), was demanding more attention, and was true. As a result, the dataset developed resembled both fake and real news in terms of attracting attention. Of course, this was an important stage in evaluating the model's success because unfavourable responses to news might lead consumers to feel it is fake.

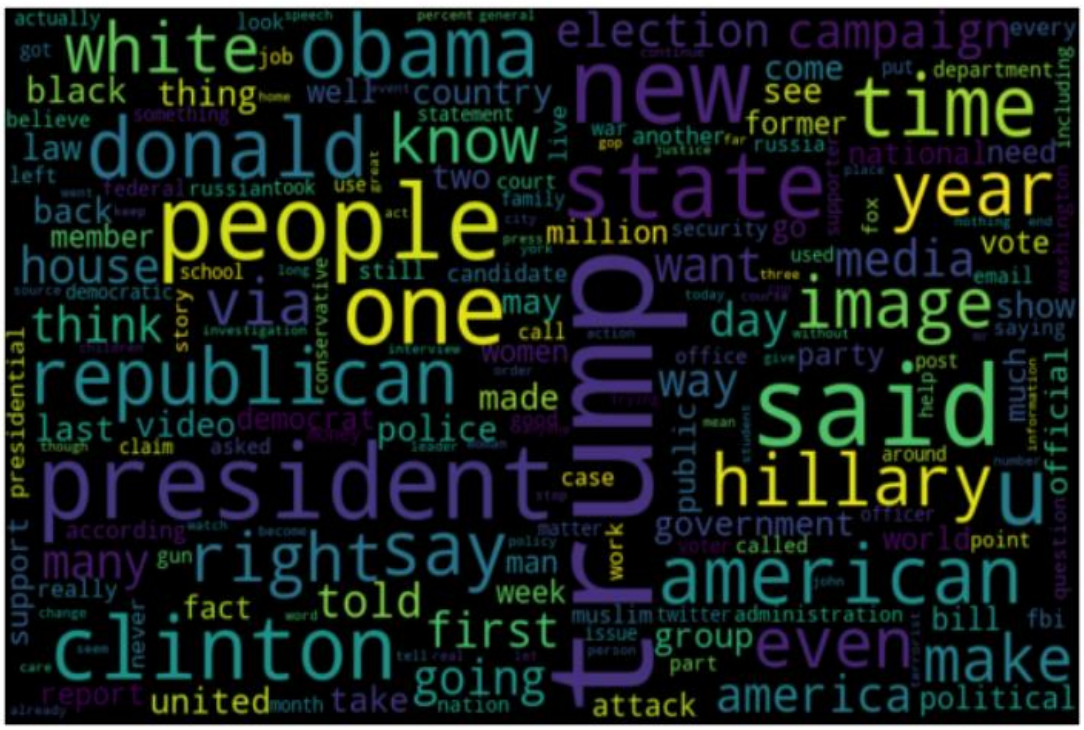
4.1.1.2 Exploring the data sets

In this study I have collected two open source datasets from kaggle one dataset contains real news and the second dataset contains the fake news in two different csv files and then merge both files for our training and testing models. There are four columns: title text subject , date and target. Fig 4.1 is the false news dataset, which includes all true news data

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

Fig 4.1 true dataset

While exploring the data set we have count the number of word in each dataset in the fig show the count of term and most frequently occurring terms in the data of real news data set and count of most number of times a particular term occurs in data



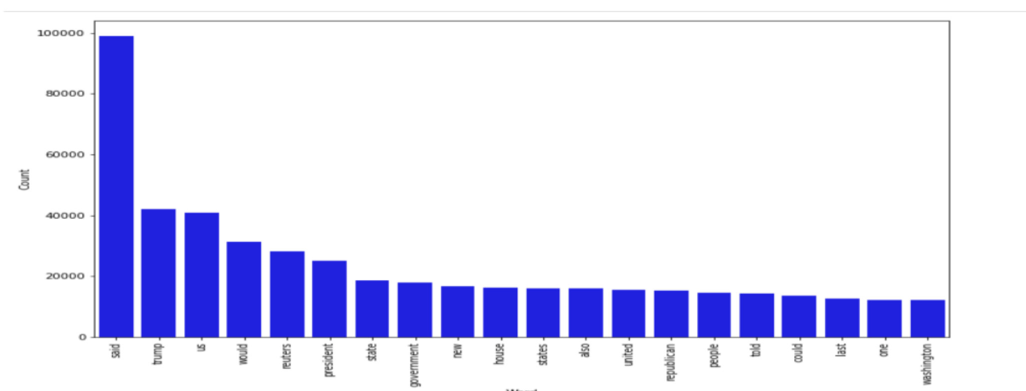


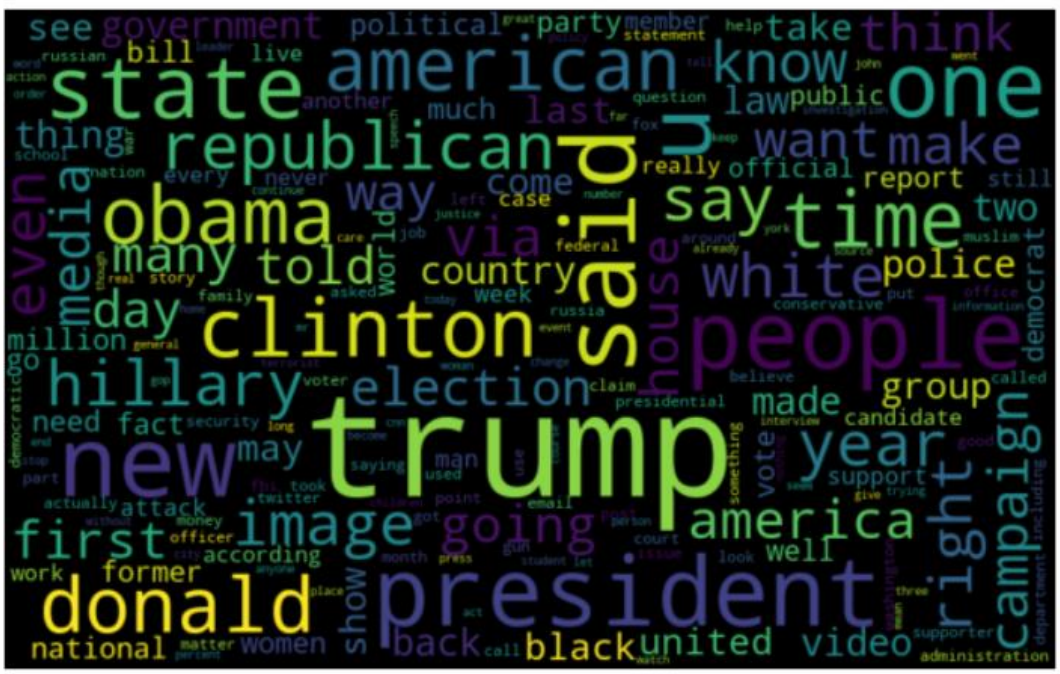
Fig 4.2 most frequent occurring words

Fig. 4.3 shows the overview of true dataset that contains all the false news articles

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

Fig 4.3 false dataset

While exploring false news data set we have count the number of word in each dataset in the fig show the count of term and most frequently occurring terms in the data of real news data set and count of most number of times a particular term occurs in data



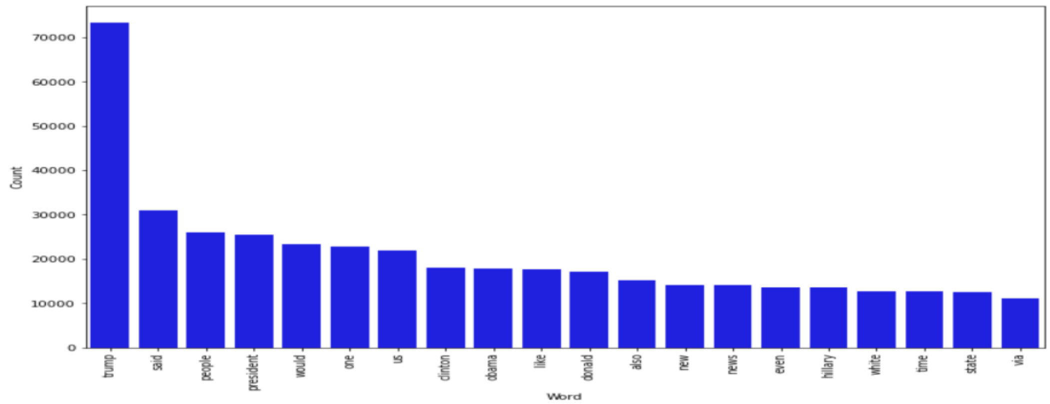


Fig 4.4 most frequent occurring words in fake dataset

after that we concatenate both the dataset into one dataset and add a label for false and true news, preprocess the data and removing all the stopword tha data shown in Fig.3 is used for train the model After that we have explore on the basis of subject to find about which type of news data contains the we have found that news are belongs to various categories.

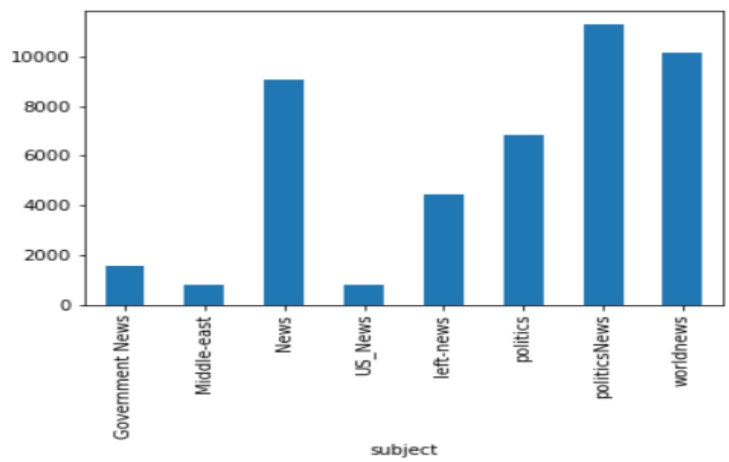


Fig 4.5 categorical representation of dataset

To get the comments for each of the collated articles, I utilised Python Web Scraping tools Selenium and Beautiful Soup. Selenium may be used to extract the server version of the page content. Beautiful Soup, on the other hand, is unable to do so since it scrapes information from the website's client version. As a result, the relevant data was scraped using Selenium and Beautiful soup. To avoid the material from being too long or too short, I picked the first five to six pages of filled comments. The language of the replies gathered was limited to English for simplicity. Facebook offers a feature called "Translate all" that automatically transforms all comments to English.

4.2 Data Preprocessing

The amount of social media information is unorganized conversation with mistakes, profanity, and terrible language, among other things. Before predictive modelling may be utilised to get greater insights, the data must be cleaned. This required some basic pre-processing of the News training examples. This stage included the following:

4.2.1 Punctuation Removal:

The message has been stripped of any punctuation. Python's string library includes certain before the characters, like '!"#$ % &'()*+,-./:;?@[] '|'.

4.2.2 Lower case conversion:

To avoid numerous duplication with the same words, the first phase was to convert the information to lowercase letters. The content are changed into same case, ideally lowercase letters, one of the most typical preparation stages. However, because lower casing might result in data lost in the some NLP situations, this procedure is not required every time you worked on an NLP issue. For instance, when we are carrying out a task that involves a people's feelings, capital letter phrases might indicate dissatisfaction or enthusiasm.

For example the output after convert into lower case

clean_msg	msg_lower
Go until jurong point crazy Available only in bugis n great world la e buffet Cine there got amore wat	go until jurong point crazy available only in bugis n great world la e buffet cine there got amore wat
Ok lar Joking wif u oni	ok lar joking wif u oni
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005 Text FA to 87121 to receive entry questionstd txt rateTCs apply 08452810075over18s	free entry in 2 a wkly comp to win fa cup final tkts 21st may 2005 text fa to 87121 to receive entry questionstd txt ratetcs apply 08452810075over18s
U dun say so early hor U c already then say	u dun say so early hor u c already then say

Fig.4.6 conversion of data from upper to lower

4.2.3 Tokenization:

It consists of breaking down a document into a series of words or groups of words, such as bigrams and trigrams. Segmentation was performed in order to acquire rate vector numbers for such symbols.

The output after looks like

msg_lower	msg_tokenied
go until jurong point crazy available only in bugis n great world la e buffet cine there got amore wat	[go, until, jurong, point, crazy, available, only, in, bugis, n, great, world, la, e, buffet, cine, there, got, amore, wat]
ok lar joking wif u oni	[ok, lar, joking, wif, u, oni]
free entry in 2 a wkly comp to win fa cup final tkts 21st may 2005 text fa to 87121 to receive entry questionstd txt ratetcs apply 08452810075over18s	[free, entry, in, 2, a, wkly, comp, to, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, to, 87121, to, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]
u dun say so early hor u c already then say	[u, dun, say, so, early, hor, u, c, already, then, say]

Fig.4.6 conversion of data text to token

4.2.4 Stop-words removal:

Stop-words are the most frequently used terms in a sample space, hence they must be removed. For example, an, that, by, upon, in, you’ve, you’ll, you’d, your, yours, yourself, yourselves, he, most, other, some, such, no, nor, are some of the stop words in NLP. These frequently determine a message's organization rather than its environment. These might perform badly if they had been viewed as a feature. As a result, as part of the text cleaning procedure, stop-words were eliminated from the training data.

After removing stop word corpus looks like this:

msg_tokenied	no_stopwords
[go, until, jurong, point, crazy, vailable, only, in, bugis, n, great, world, la, e, buffet, cine, there, got, amore, wat]	[go, jurong, point, crazy, available, bugis, n, great, world, la, e, buffet, cine, got, amore, wat]
[ok, lar, joking, wif, u, oni]	[ok, lar, joking, wif, u, oni]
[free, entry, in, 2, a, wkly, comp, to, win, fa, cup, final, tkts, 21st, nay, 2005, text, fa, to, 87121, to, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]
[u, dun, say, so, early, hor, u, c, already, then, say]	[u, dun, say, early, hor, u, c, already, say]

Fig.4.7 conversion of data text to token

4.2.5 Stemming:

The word standardisation phase refers to the stemming or decrease of terms to their original form. For example, terms like 'programmer,' 'programming,' and 'programme' will be shortened to 'programme.' The difficulty with stemming, on the other hand, is that it stems words to the point where their basic form loses meaning or is not reduced to a valid English term. After stemming the words, the corpus looks like this:

no_stopwords	msg_stemmed
[go, jurong, point, crazy, available, bugis, n, great, world, la, e, buffet, cine, got, amore, wat]	[go, jurong, point, crazi, avail, bugi, n, great, world, la, e, buffet, cine, got, amor, wat]
[ok, lar, joking, wif, u, oni]	[ok, lar, joke, wif, u, oni]
[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]	[free, entri, 2, wkli, comp, win, fa, cup, final, tkt, 21st, may, 2005, text, fa, 87121, receiv, entri, questionstd, txt, ratetc, appli, 08452810075over18]
[u, dun, say, early, hor, u, c, already, say]	[u, dun, say, earli, hor, u, c, alreadi, say]

Fig.4.8 stemming of data

4.2.6 Lemmatization:

It breaks down the phrases into their fundamental phrases. It uses morphology approach to determine the root word with the help of a lexicon. lemmatization was used to increase the values of number of root vectors in this study.

no_stopwords	msg_stemmed	msg_lemmatized
[go, jurong, point, crazy, available, bugis, n, great, world, la, e, buffet, cine, got, amore, wat]	[go, jurong, point, crazi, avail, bugi, n, great, world, la, e, buffet, cine, got, amor, wat]	[go, jurong, point, crazy, available, bugis, n, great, world, la, e, buffet, cine, got, amore, wat]
[ok, lar, joking, wif, u, oni]	[ok, lar, joke, wif, u, oni]	[ok, lar, joking, wif, u, oni]
[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]	[free, entri, 2, wkli, comp, win, fa, cup, final, tkt, 21st, may, 2005, text, fa, 87121, receiv, entri, questionstd, txt, ratetc, appli, 08452810075over18]	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]
[u, dun, say, early, hor, u, c, already, say]	[u, dun, say, earli, hor, u, c, alreadi, say]	[u, dun, say, early, hor, u, c, already, say]

Fig.4.9 lemmatization of data

Before the data could be analysed, it had to go through text pre-processing. The sample space for features in a noise-free corpus is smaller, resulting in greater accuracy.

4.3 Generation of features

Number of terms, frequency of cases appear, rate of different terms, n-grams, and other criteria can be extracted from text texts. We will be able to read text and perform tasks such as segmentation, classification, and other tasks by generating word representations that capture their values, meaningful linkages, and the many environments in which they are utilised. Text is converted into numbers or matrices that may be input into the system using word embedding methods.

4.3.1 Word Embedding:

Among the most common ways to express textual content is by word embedding. This could capture a phrase's context in a text, meaning and grammatical similarities, relationships with other keywords, and so on.A dictionaries are used to convert a term to a vectors in a word document. The information were trained using the frequency-based word2vec vectors shown below. Lexical traits are also divided into categories.

4.3.2 Count Vector as feature generation:

The data is displayed as a column, with rows representing the library articles, columns containing a word from the textual data, and cells indicating the frequency of that phrase in each article. It's used to convert a phrase into a vector based on its frequency (of each word that appears in the full text). When working with a huge number of such papers and translating each word into matrix, this is really beneficial. The vocabulary is built from the corpus's unique tokens or words. Consider the following text, which appears in documents collection .

D1: Strong heatwave is going on in the capital

D2: People and animals suffered from extreme heatwaves.

D3: This is the fifth heatwave in the capital

Distinct terms could be added to the dictionary. The following are the distinct terms found: [strong, heatwave, capital, people, animal, suffered, extreme]

No of Documents=3

No of unique terms=7

The incidence of each phrase in each document is represented by the Counting Matric.

The count matrics would be of size M=3*7 that shown in table below

Table 1 Showing word document matrix

	strong	heatwave	capital	people	animal	suffered	extreme
D1	1	1	1	0	0	0	1
D2	0	1	0	1	1	1	0
D3	0	1	1	0	0	0	0

For the appropriate word in the Matrix M, a column can be called a term vector. [1] is the word vector for "heatwave." From the greatest to the lowest frequency in the Corpus, the Count vector outputs all those tokens. Heatwave, for example, has the highest incidence in the Corpus, hence it ranks first in the dictionary's word list. This characteristic was employed in the suggested strategy to offer machine learning models a notion of which terms are often used by internet users when they encounter Fake or Real news. Below shown fig are from the dataset matrics

	abadi	abandon	abba	abc	abc news	abe	abedin	abil	abl	abort	...	zero	zika	zika fund	zimbabw	zimbabw mnangagwa	zimbabw mugab	zone	zor	zuckerberg	zuma
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Fig.5.2.1 count vector corpus

4.3.3 TF-IDF vectors as a feature:

The comparative relevance of a phrase in the article and throughout the collection is represented by TF-IDF value.

Word Frequency (TF) is a calculation that determines how frequently a term appears in a document. Because document sizes vary, a phrase may appear more frequently in a large document than in a small one. As a result, the document's length frequently divides Term frequency.

$$TF(t,d) = \frac{Number\ of\ times\ t\ occurs\ in\ document\ 'd'}{Total\ word\ count\ of\ document\ 'd'}$$

IDF stood for Inverse Document Frequency, which asserts that if a phrase occurs in every paper, it is of minimal use. The words "such as a," "as," "some," "on," "of," and others appear often in documents yet have little significance. IDF decreases the value of these phrases while elevating the relevance of unusual terms. The more valuable IDF becomes, the more unique the phrase becomes.

$$IDF(t) = log\left(\frac{Total\ number\ of\ document}{Number\ of\ document\ of\ term\ 't'\ in\ it}\right)$$

Term Frequency-Inverse Document Frequency (TF-IDF): The TF-IDF penalises the most frequently happening words by giving them less weight, while giving high weight to keywords that are included in the correct subset of the sample of feature and have a high occurring in a specific sample space. Term Frequency and Inverse Document Frequency combine to create it.

$$TFIDF(t, d) = TF(t, d) * IDF(t)$$

The text categorization feature TF-IDF is commonly utilised. TF-IDF Vectors may also be computed at multiple levels, such as at the word and N-gram levels, which I did in my project. TFIDF is based on the idea that terms which are both too common and too infrequent in a corpus are not statistically relevant for discovering a pattern. The Logarithmic component in tfidf penalises terms in the sample that are overly common or unusual by assigning them low tfidf scores. A greater tfidf number indicates that the terms in the sample are more important, whereas a smaller value indicates that they are less important.

(i) TF-IDF at the word level: derives the value for each phrase in a manuscript.

(ii) TF-IDF at the Level is needed: Evaluates a scoring rate for the combination of N words in distinct texts.

This is how corpus of tf idf looks like

	abadi	abandon	abba	abc	abc news	abe	abedin	abil	abl	abort	...	zero	zika	zika fund	zimbabw	zimbabw mnangagwa	zimbabw mugab	zone	zor	zuckerberg	zuma
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig.5.2.2 tfidf corpus

4.4 Splitting the dataset

Vector characteristics in tests The following strategies were used to perform Count Arrays and Tf-Idf arrays. All of the models were determined to be correct. I used the K-fold cross validation method to improve the models' performance. I employed text classification to categorise the body of articles in two available datasets in the first part of my research. In the second step, an experiment was conducted on the responses obtained on a set of Fake news and Real news. Using this cross validation approach, the data was randomly split into k-folds. The model's effectiveness was evaluated using the kth fold, which was built using (k-1) folds. This procedure was repeated until each of the k-folds could be utilised as a test set. For this research, I employed 3-fold cross validation, with 67 percent of the data used for training and the remaining 33 percent for testing.

4.5 Feature selection:

When creating a prediction models, selection is the method of minimising the number of independent variables. The set of input data should be reduced to lower the computation complexity of modelling and, in some situations, to increase the model's performance. The data attributes you have used to build your machine learning algorithm have a significant impact on the results you may get. I have taken 5000 feature for the building models and predicting the output. For selecting feature i have used Chi-square.

4.5.1 Chi-square:

The absence of independence between term(t) and class is measured (c). If t and c are independent, it has a natural value of zero. If it's higher, the term will be affected. For low-frequency words, it is unreliable. For categorical characteristics in a dataset, the Chi-square test is utilised. We calculate the Chi-square between every characteristic and the objective and choose the characteristics with the highest Chi-square values. The variables must be categorical, sampled separately, and values must have an expected frequency larger than 5 to properly apply the chi-squared to examine the association among different aspects in the sample and the target output.

4.6 Algorithms used for classification:

The classifiers in this section are trained and then evaluated against the dataset. Many classifiers were researched in order to predict the text's class. To mention a few, I investigated Naive Bayes Classifiers, Logistic Regression, Decision Trees, Support Vector Machines, Random Forests, and Gradient Boosting. These classifiers were created using the Python programme Sci-Kit Learn.

4.6.1 Support Vector Machine:

SVM is a classification and regression supervised machine learning technique. It's most often used for classification, although it can also be used for regression. SVM locates a hyperplane that divides the various sorts of data. A line in two-dimensional space is all that this hyper-plane is. Select the optimal hyperplane for dividing the data after that. As a result, you've probably noticed that by default, SVM can only perform binary classification (i.e., choose between two classes). However, there are a variety of approaches that can be employed to address multi-class issues [11].

4.6.2 Logistic regression

It provides a simple equation to categorise issues into two or more classes, a The logistic regression classifies content data from the large number of characteristics and a binary output such as right or wrong, real article or fraudulent article. We changed model parameters to get the best results for each dataset, and we experimented with several values before deciding on the best model validity. The logistic regression assumption function has the following mathematical expression:

$$h_{\theta}(X) = \frac{1}{1+e^{-(\beta_0+\beta x_1)}}$$

A sigmoid activation function is used to translate the outcome of logistic regression into a probability value, with the purpose of minimising the cost function to achieve the best chance. The following is how the cost function is calculated:

$$\text{Cost}\left(h_{\theta}(x),\ y\right) = \begin{cases} \log\left(h_{\theta}(x)\right), & y = 1, \\ -\log\left(1 - h_{\theta}(x)\right), & y = 0. \end{cases}$$

4.6.3 Decision Tree

The most successful and extensively used classification and decision-making strategy is the decision tree. Every branch marks the completion of the experiment, and each terminal nodes the class label. On this foundation, we form node roots.

"When do we finish growing our tree?" Real-world datasets often include a huge number of features, which leads to a large number of splits and a gigantic tree. Such trees take longer to grow and may result in overfitting. For example, the tree's accuracy on the training dataset will be good, but its accuracy on the test dataset will be poor[10].

4.6.4 Random Forest

Random Forest is built on ensemble learning, which is an approach for addressing complicated problems by combining numerous classifiers and enhancing the performance of the model. Random Forest is a learner that averages the outcomes of numerous decision trees on different subsets of a dataset to enhance the dataset's anticipated accuracy. Instead, then relying on a decision tree, it collects forecasts from every tree and predicts the final output based on the majority of votes. It becomes more accurate as the forest grows denser The random forest method has the following steps:

1. From a data collection of k records, random forest selects n random records at random.
2. A distinctive decision tree is created for each sample.
3. Each decision tree produces a result
4. For classifiers, the final result is based on a popular vote or averaging, as suitable.

4.6.5 Naive Bayes:

The naive Bayes approach is a supervised learning strategy based on Bayes' theorem to deal with the difficulty of classification **.** It's most typically used for tasks that need a large amount of training data, such as text categorization. The naive Bayes approach is a supervised learning strategy to overcome the difficulty of classification based on Bayes' theorem. This is a twin system. That is, it generates predictions based on the likelihood that a particular event will occur. It's called Naive because the existence of one feature has no effect on the presentation of others[12]. It's termed an apple if the colour, shape, and flavour of a red, spherical, and sweet fruit are utilised to identify it. it is called an apple. As a consequence, each feature helps to determining whether or not the object is an apple without relying on the others. Because it is based on Bayes' Theorem, it is called Bayes.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

4.6.6 Gradient Boosting Classifier:

Each prediction in gradient boosting tries to improve on the one before it by lowering mistakes. Gradient boosting is a revolutionary concept since it fits a new predictor to the error variance from the previous prediction rather than matching a predictor to the data at each iteration[13]. In practise, gradient boosting classification works. In order to make early predictions on the data, the approach will collect the log of the target feature's probability. A popular method of determining this is to divide the count values (True = 1) by the count of False values (False = 0).

1. It calculates the residuals, or the observed value minus the predicted value, for each occurrence in the training set.
2. It then builds a new decision tree that tries to predict the previously calculated remainder. However, in compared to gradient boosting regression, here is where things become a little trickier. Table 1. Countvectorizer performance metrics

4.7 Metrics used to access the Performance of Model

In this part, I've focused at some of the most important indicators for evaluating the success of a classification model. Those indicators show how successfully our model can identify and assess outcomes. This project made use of the metrics described below.

4.7.1 Classification Accuracy:

It is the most commonly used parameter in categorization task evaluation. The number of correct answers is divided by the total number of projections to arrive at this figure. But, this statistic by itself is insufficient to determine if the algorithm is excellent or not. It is appropriate when each class has an equal number of observers.

4.7.2 Confusion Matrix:

A confusion distribution is a table that displays how well a classification algorithm (or "classification algorithm") works on a set of test data with actual values generated. The confusion matrix is simple in itself, but the related jargon can be confusing. The Error Matrix, which is a tabular depiction of the algorithm's effectiveness, is another name for it. It's a unique type of probability table with two aspects: "real" on the x-axis and "expected" on the y-axis. The amount of predicted by the algorithm is represented in the table's columns. Let's start with more fundamental terminology, such as whole numbers.

Table 4.1 Confusion Matrix

Total Instances		Predicted	
		Yes	No
Actual	Yes	*True Positive*	*False Negative*
	No	*True Negative*	*True Negative*

True positives (TP): are situations in which we forecast positive and they actually are.

True negatives (TN): are situations in which we forecast negative and they are negatives

False positives (FP): are actually incorrect predictions like we predicted positive but they are false. It is called a type I error.

False negatives (FN): In this we predicted positive but actually they are negative. It is called a type II error.

When dealing on classification issues, Scikit-learn provides a handy report that produces accuracy, recall, F1 score, and supports for every category.

4.7.3 Precision:

Precision is defined as the ratio of precisely predicted positive cases to total expected positive cases. A low False Positive rate indicates a high level of accuracy.

$$precision = \frac{TP}{TP+FP}$$

4.7.4 Recall:

The ratio of accurately forecasted positive examples to all examples inside the real class is called recall - Yes

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

4.7.5 F1-Score:

In this computation, precision and recall are weighted. As a consequence, both false positives and false negatives are accounted for. F1 score is usually more useful than accuracy when there is an uneven class distribution. When the occurrence and costs of false positives and false negatives are equal, accuracy performs the best. Precision and Recall should be evaluated if the cost of false positives and false negatives differs considerably.

$$F1 - Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}$$

CHAPTER 5

RESULTS AND ANALYSIS

ML classifiers are evaluated using a variety of measures. Accuracy measures the difference between expected and actual labels. Data retrieval is done with precision. Precision is computed as a percentage of the model's total positive predictions. The percentage of good results that are relevant. True positive rate or True negative rate is referred to as recall. F1-score is made up of precision and recall.

Figure 5.1 depicts the performance evaluation of naive bayes classifiers using a count vector and several metrics. Figure 5.2 illustrates how to use TF-IDF to analyse the quality of ML classifiers using various metrics.

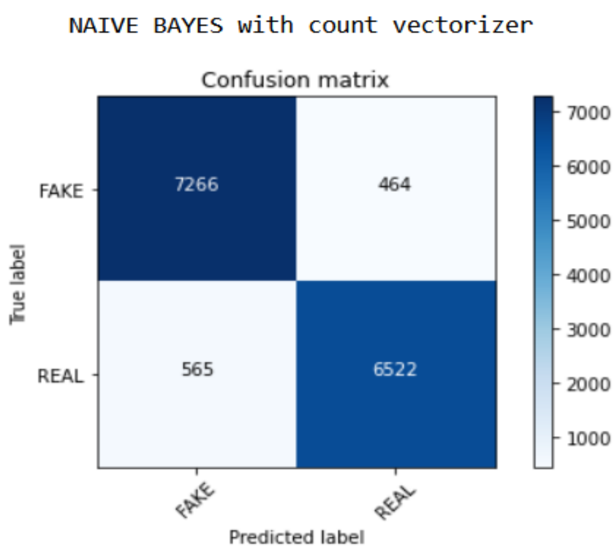


Fig 5.1 naive bayes confusion matrix with count vectorizer

In the above confusion matrix true positive denote the value that are true and model predicted correct in this case TP=6522, true negative is the value has false and model also predicted false i.e TN=7266 ,false positive is the actual value is false and the model predicted is true, false negative is the value is true but model is predicted is false. False negative and false positive are the wrong prediction made by the model.

TP=6522, TN=7266, FP=464, FN=565

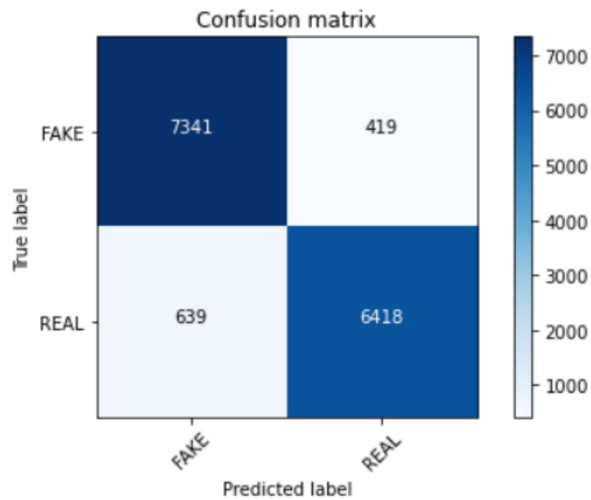


Fig 5.2 naive bayes confusion matrix with TFIDF

TP=6418, TN=7341, FP=419, FN=639

With these values we can conclude that the naive bayes perform better with count vector than TF-IDF. It has accuracy 92.86% which is less than 93.06%.

Figure 5.3 depicts the performance evaluation of naive logistic regression classifiers using a count vector and several metrics. Figure 5.4 illustrates how to use TF-IDF to analyse the quality of ML classifiers using various metrics.

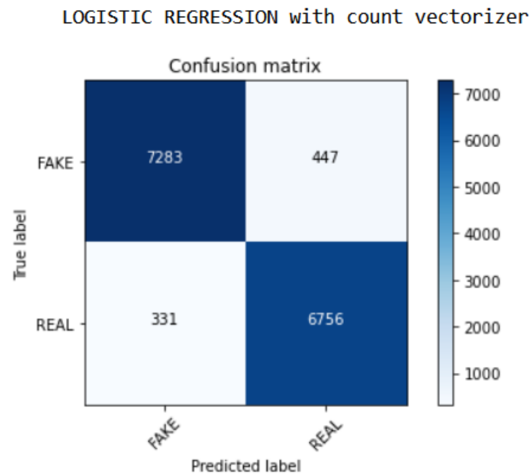


Fig 5.3 logistic regression confusion matrix with count vectorizer

TP=6756, TN=7283, FP=331,FN=447

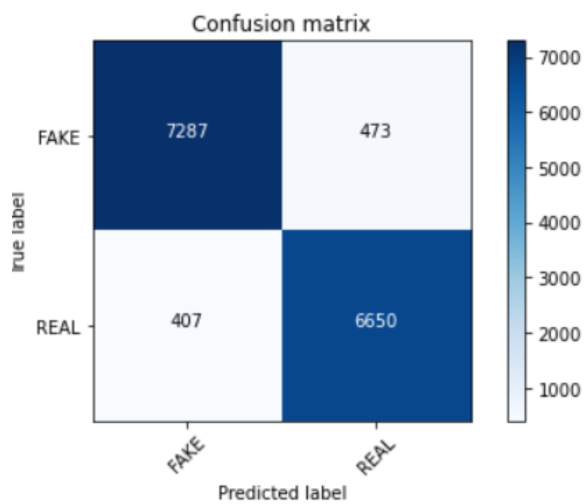


Fig 5.4 logistic regression confusion matrix with TFIDF

TP=6650, TN=7287, FP=407, FN=473

With these values we can conclude that logistic regression perform better with count vector than TF-IDF. It has accuracy 94.06% which is less than 94.76%.

Figure 5.5 depicts the performance evaluation of decision tree classifiers using a count vector and several metrics. Figure 5.6 illustrates how to use TF-IDF to analyse the quality of ML classifiers using various metrics.

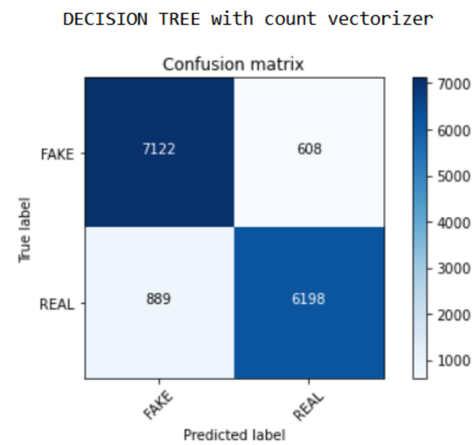


Fig 5.5 decision tree confusion matrix with count vector

TP=6198, TN=7122, FP=889, FN=608

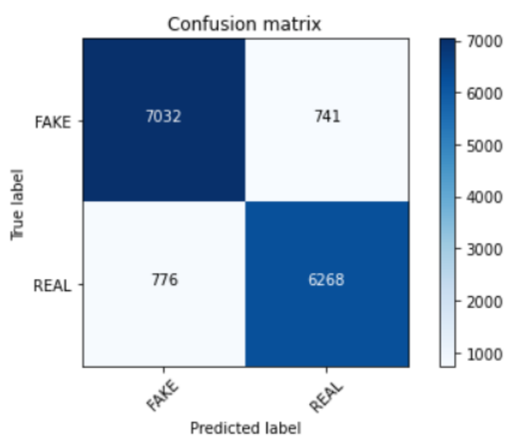


Fig 5.6 decision tree confusion matrix with TFIDF

TP=6268, TN=7032, FP=776, FN=741

With these values we can conclude that logistic regression perform better with count vector than TF-IDF. It has accuracy 89.9% which is greater than 89.76%.

Figure 5.7 depicts the performance evaluation of random forest classifiers using a count vector and several metrics. Figure 5.8 illustrates how to use TF-IDF to analyse the quality of ML classifiers using various metrics.

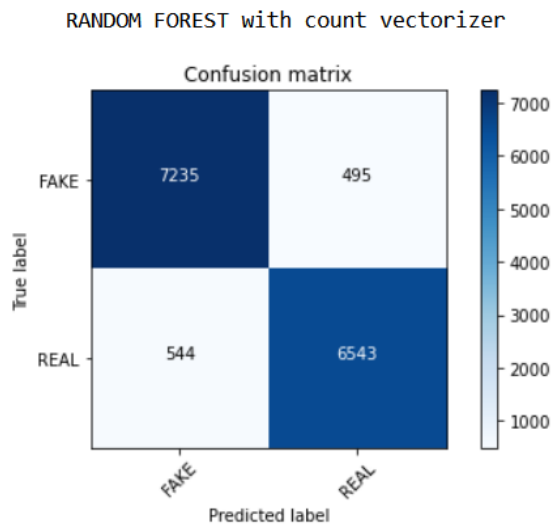


Fig 5.7 random torest confusion matrix with count vector

TP=6543, TN=7235, FP=544, FN=495

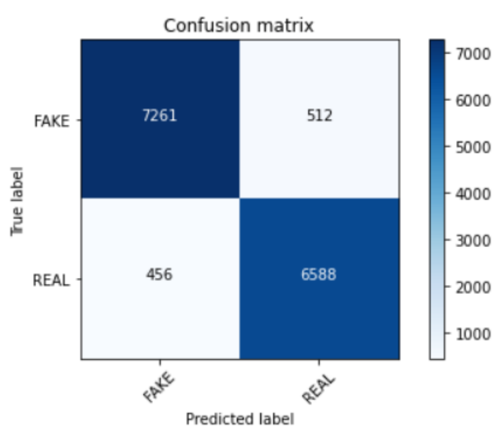


Fig 5.8 random forest confusion matrix with count vector

TP=6588, TN=7261, FP=456, FN=512

With these values we can conclude that random forest perform better with count vector than TF-IDF. It has accuracy 92.69% which is greater than 93.47%.

Figure 5.9 depicts the performance evaluation of SVM classifiers using a count vector and several metrics. Figure 5.10 illustrates how to use TF-IDF to analyse the quality of ML classifiers using various metrics.

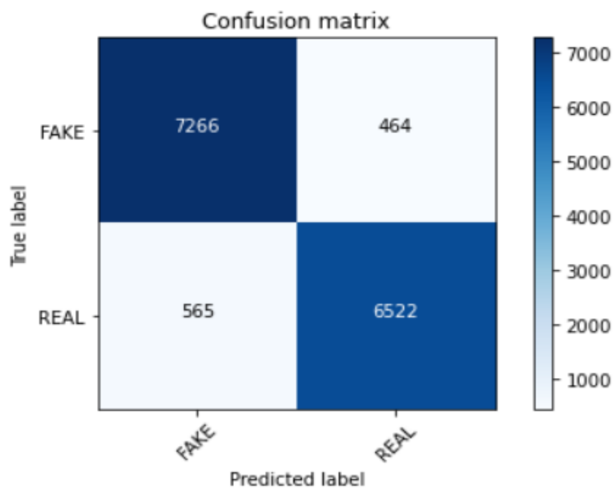


Fig 5.9 random forest confusion matrix with count vector

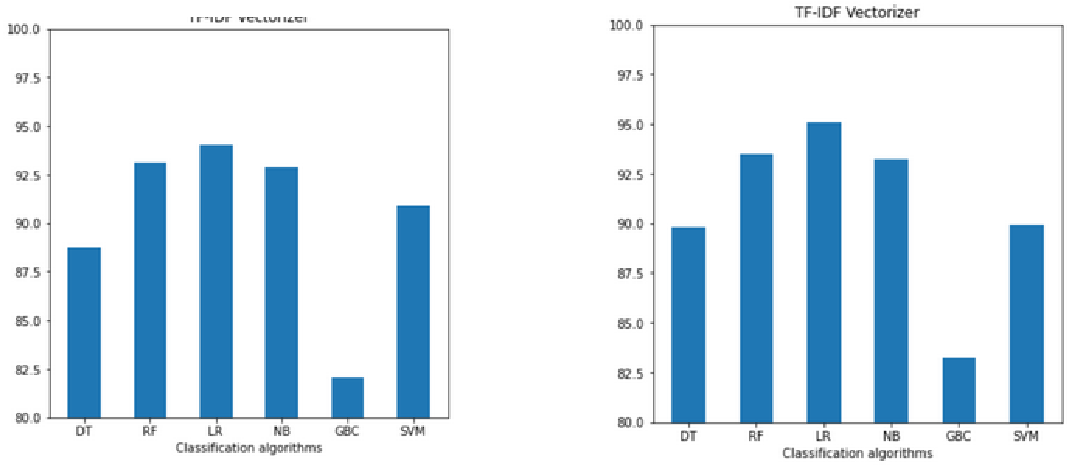


Fig.5.10 count vectorizer results

Fig 5.10 denote the accuracy of all the classification methods in which logistic regression performs well for both of the classification models. The accuracy of this model is 94.67% and 95.8% respectively for count vector and tfidf.

Since the classification model based on the binary output format for binary output logistic regression perform better for its binary classification properties which uses sigmoid function for its classification.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This study uses two levels of inspection to detect fake news: description and experience. In the first level on social networks, the fundamental ideas of misleading news are covered.During the research phase, the current tactics for identifying fake news using various supervised learning algorithms are explored. The study employs models and modeling techniques based on vocal attributes that differ from other current models for the mentioned false news detection approaches, which are focused on text analysis. The study's main goal is to find linguistic characteristics that separate fake news from real news. Manually categorising news requires extensive subject knowledge but also it able to detect irregularities in the text. This study uses machine learning algorithm and collective techniques to identify false news reports. We gathered material for our inquiry on the internet, which included current events. The test's purpose is to recognize linguistic characteristics that separate fake news from real news. Certain models have been shown to be more realistic than others. We used a range of performance indicators to compare the results of each strategy. In order to detect bogus news, researchers must address a number of serious challenges. For example, identifying essential features of news transmission is a critical step in limiting the spread of fake news. The primary sources of misleading news propagation might be identified using machine learning techniques. We have two alternative ways for text processing in this study, and the final findings are based on merging both approaches. We discovered that the decision tree has better results than other models, and the gradient boosting classifier worked well as well.

REFERENCES

[1] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, ”Automatic Online Fake News Detection Combining Content and Social Signals,” 22nd Conference of Open Innovations Association, 2018, pp. 272 - 279

[2] M. Granik and V. Mesyura, ”Fake news detection using naive Bayes classifier,” 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering , 2017, pp. 900 - 903.

[3] S. Bird, E. Klein, and E. Loper, ”Natural language processing with Python: analyzing text with the natural language toolkit,” OReilly Media, Inc., 2009.

[4] H. Wasserman, “Fake news from africa: Panics, politics and paradigms,” Journalism, vol. 21, no. 1, pp. 3–16, 2020, retrieved: 2018, September22.

[5] https://en.wikipedia.org/wiki/Machine_learning

[6] https://en.wikipedia.org/wiki/Natural_language_processing

[7] Carlos Merlo (2017), "Millonario negocio FAKE NEWS", Univision Noticias

[8] Chang, Juju; Lefferman, Jake; Pedersen, Claire; Martz, Geoff (November 29, 2016). "When Fake News Stories Make Real News Headlines". Nightline. ABC News.

[9] https://www.cjr.org/analysis/facebook-rohingya-myanmar-fake-news.php

[10] https://blog.paperspace.com/fake-news-detection/

[11] Aayush Ranjan, Fake News Detection Using Machine Learning, Department Of Computer Science & Engineering Delhi Technological University, July 2018

[12]“https://www.javatpoint.com/logistic-regression-in-machinelearning.”

[13] S. Kogan, T. J. Moskowitz, and M. Niessner, “Fake News: Evidence from Financial Markets,” 2019

[14] M. Zimdar. (2016) False, misleading, clickbait-y, and satirical ”news” sources. Retrieved:September 22,2018.

[15] T. Hofmann, B. Sch¨olkopf, and A. J. Smola, “Kernel methods in machine learning,” The Annals of Statistics, vol. 36, no. 3,pp. 1171–1220, 2008.

[16] “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection, William Yang Wang

[17] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In AC'12.

[18] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. "Detecting hoaxes, frauds, and deception in writing style online". In IEEE Symposium on Security and Privacy, pages 461–475, 2012.

[19] Magnus Almgren and Erland Jonsson. "Using active learning in intrusion detection". In IEEE Computer Security Foundations Workshop, pages 88–98, 2004.

[20] Dana Angluin and Philip Laird. "Learning from noisy examples. Machine Learning", 2(4):434– 470, 1988.

[21] Arthur Asuncion and David J. Newman. UCI machine learning repository, http://www.ics.uci.edu/∼mlearn/ML Repository.html, 2007.

[22] AV-TEST. Malware Statistics. http://www.av-test.org/en/statistics/malware/.