

**FAKE NEWS DETECTION SYSTEM USING MACHINE LEARNING BASED ON
SOMALI NEWS**

**AWIL AHMED SULEIMAN
ZAKARIE SAEED AWALE
ABDIRAHMAN ABDULLAHI SIAD
HASSAN ABDI SALAD
IBRAHIM MOHAMED IBRAHIM**

**SUBMISSION OF GRADUATION PROJECT FOR PARTIAL
FULFILLMENT OF THE DEGREE OF BACHELOR OF
COMPUTER APPLICATIONS**

**JAMHURIYA UNIVERSITY OF SCIENCE AND
TECHNOLOGY (JUST)
FACULTY OF COMPUTER & INFORMATION
TECHNOLOGY**

AUGUST 2022

JAMHRURIYA UNIVERSITY OF SCIENCE AND TECHNOLOGY (JUST)

Original Literary Work Declaration

Name of Candidate 1: **Awil ahmed Suleiman** ID No: C117458

Name of Candidate 2: **Zakarie Saciid Awale** ID No: C118257

Name of Candidate 3: **Hassan Abdi Salad** ID No: C118788

Name of Candidate 4: **Abdirahman Abdullahi Siyad** ID No: C118297

Name of Degree: **Bachelor of Computer Application**

Title of Project Paper/Research Report/Dissertation/Thesis (“this Work”):

Fake News Detection Using Machine Learning based on Somali news

Field of Study: **Computer Applications**

We the undersigned, do solemnly and sincerely declare that:

(1) We are the authors/writers of this Work;

(2) This Work is original;

(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyrighted work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;

(4) We do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyrighted work;

(5) We hereby assign all and every right in the copyright to this Work to Jamhuriya University of Science and Technology (“JUST”), who henceforth shall be the owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of JUST having been first had and obtained;

(6) We are fully aware that if in the course of making this Work, we have infringed any copyright whether intentionally or otherwise, we may be subject to legal action or any other action as may be determined by JUST.

Candidate 1’s Signature: _____ Candidate 2’s Signature: _____

Candidate 3’s Signature: _____ Candidate 4’s Signature: _____

Date: _____

Subscribed and solemnly declared before,

Supervisor’s Signature: _____ Date: _____

Name: _____ Designation: _____

Dedication

we dedicate our dissertation work to our family and many friends. A special feeling of gratitude to our loving parents, whose words of encouragement and push for tenacity ring in our ears.

Abstract

In the past few years, fake news has taken over the media landscape. As a result, there is a desire to accurately and directly identify fake news and true news as appropriate action can be done. One of the places where this kind of information is disseminated is mostly social networking sites. We can determine whether a post is fake or not to a significant extent by using the Three machine learning algorithms that has the ability to classify where the news as true or false, we will choose the one with the highest accuracy score and they are SVM with an its accuracy score is 92%, naive Bayes with an its accuracy score is 93% and we select logistic regression because it is the one with the highest score Algorithm we used and is Its accuracy score is 96.95% because it is the first time that this research has been done in the Somali language. We have collected the true ones from BCC and VOA which contain 1809 posts and the false ones which contain 1599 posts. Algorithms show best performance for full data storage

Keywords: Fake news detection, Social Networks, Naïve Bayes Classifier, SVM and

LogisticRegression

Acknowledgements

First, we express our heartiest thanks and gratefulness to almighty Allah for His divine blessing makes us possible to complete the final year project successfully. secondly, we like to thank our supervisor, **Eng. Abdullahi Mohamed Hassan Beerey** that gives us idea, guidance until completes our projects. Thirdly we like to thank for our parents for endless support in every time and their encouragement. finally, we thank our Eng.Hanad who corrected our research errors and also gives as tips and tricks. we also like to thank our grateful research coordinator **Eng. Sharmake Ali Kahie** who support us encouragement and guidelines by any time and also gives as Fake News Detection Using Machine Learning based on Somali news, we would like to thank our entire course mate in Jamhuriya University of Science and Technology (“JUST”), who took part in this discuss while completing the course work. Finally, we must acknowledge with due respect the constant support and patients of our parents

Table of Contents

Dedication	ii
Abstract	iii
Acknowledgements	iv
List of Figures	viii
List of Tables.....	ix
CHAPTER I: INTRODUCTION	1
1.0 Background of the Study.....	1
1.1 Problem Statement	3
1.2 Objectives.....	4
1.3 Research Question.....	4
1.4 Scope of The Study	4
1.5 Significance of the Study	5
1.6 Organization of The Study	5
CHAPTER II: LITERATURE REVIEW.....	7
2.1 Introduction	7
2.2 History of fake news	8
2.3 How does fake news spread?	12
2.4 How to Prevent the Spread of the Fake News?.....	14
2.5 Why machine learning is required to detect the fake news?.....	16
2.6 Traditional Media and Fake News	17
2.7 Text Classification	18
2.8 Basic concept of machine learning	19
2.8.1 Types of machine learning	21
2.8.1.1 Supervised machine Learning.....	21
2.8.1.2 Unsupervised machine Learning	22
2.9 Fake news on social media.....	23
2.10 Related Work	25

2.11 Research Gap	26
2.11.1 Gap	26
CHAPTER III: RESEARCH METHODOLOGY	27
3.0 Introduction	27
3.1 System Description	27
3.2 System Architecture	28
3.2.1 Extracting the Training Data	28
3.2.2 Preprocessing Data	29
3.2.3 Data cleaning	29
3.2.4 Feature Extraction	31
3.2.5 Test Train Split	31
3.2.6 Models	31
3.2.7 Data Collection	32
3.3 System Features	32
3.4 System Development Environment.....	33
3.5 System Requirement Specification	34
3.5.1 Hardware Requirement.....	34
3.5.2 Software Requirement	34
CHAPTER IV: SYSTEM ANALYSIS AND DESIGN.....	36
4.1 Introduction	36
4.2 System Analysis	36
4.3 Proposed System	36
4.4 System Requirements.....	37
4.4.1 Functional Requirements.....	37
4.4.2 Non-Functional Requirements.....	38
4.5 System Design.....	38
4.5.1 Entity Relationship Diagram:	38
4.6 Dataset Design	39
CHAPTER V: IMPLEMENTATION AND TESTING	41
5.0 Introduction	41

5.2 Overview of the implementation environment	41
5.3 Snapshots of the system	41
5.3.1 Front-end	41
5.3.2 Back-end.....	43
5.4 Count Words	44
5.4.1 Fake news	45
5.4.2 Real News.....	45
5.4.3 Unicode Words.....	46
Chapter VI: Conclusion and Future Work	48
6.1 Introduction.....	48
6.2 Conclusion	48
6.3 Discussion	49
6.4 Recommendation.....	49
6.4 Future Work	50
References	51

List of Figures

Figure 3.1 System Architecture.....	28
Figure 4.1 Use Case	39
Figure 4.2 Dataset True.....	39
Figure 4.3 Dataset False.....	40
Brief Description.....	40
Figure 5.1 Home Page.....	42
Figure 5.2 Prediction.....	43
Figure 5. 3 Confusion Matrix for each model.....	44
Figure 5. 4 Fake News Word	45
Figure 5. 5 True News Words.....	45
Figure 5. 6 Unicode Words	46
Figure 5.7 Training.....	46
Figure 5.9 Result	47

List of Tables

Table 2.1 Research Gap	26
Table 3.1 Hardware Requirement	34
Table 3.2 Software Requirement.....	35

Chapter I: Introduction

1.0 Background of the Study

The internet is one of the most essential innovations, and it is used by a vast number of people. These people use it for a variety of reasons these individuals have access to a variety of social media channels. Through these internet platforms, any user can make a message or spread the news. As a result, some users try to spread the word. Through these platforms, bogus news is spread (Alim Et al., 2021)

Modern technology allows anyone with a cell phone or computer to publish anything online, regardless of subject matter expertise. It is becoming increasingly difficult for people to discriminate between true and false news (De Wet & Marivate, 2021). We have described “fake news” as deliberately verified news that may cause readers to be misled (Allcott & Gentzkow, 2017).

Machine learning has played a critical role in information classification, but with significant drawbacks (Manzoor et al., 2019). This survey we will focus on to distinguish fake news and real news. By using Machine learning approaches. Machine learning refers to the process of creating computer algorithms that can be limited to human intelligence. (El-Naqa & Murphy, 2015).

Fake news has quickly become a social media problem, with people using it to spread fake news or rumors to influence their behavior. It has been confirmed that the dissemination of false information had a significant impact on the 2016 US presidential election (Lorent & Itoo, 2018) Fake news is a new word, but it is not a new phenomenon. However, advances in technology and the transmission of information in many forms of media have accelerated the spread of fake news today. As a result, the impact of fake news has increased

dramatically in the recent years, and action must be taken to prevent it from happening again in the future (Agrawal et al., 2020).

People are notoriously bad at detecting lies, and they are often unaware that they are being tricked. Users of social media are frequently uninformed that there are postings, tweets, papers, and other written evidence that are only intended to influence the opinions and judgments of others. Information manipulation is a poorly understood issue that is rarely mentioned, particularly when a friend spreads false information. (Stahl, 2018)

Combating false information is a demanding and exhausting task. Because of its tremendous impact on the political environment, fake news is having a never-before-seen impact on people's lives. As a result of the agreement, activities aimed at automatically detecting fake information have gained traction, generating a lot of academic interest. Even still, the bulk of methodologies have flaws when it comes to building solutions for English and other languages (Busioc et al., 2020).

There are a lot of websites that will provide wrong information. They actively aim to bring forth intended publicity, deceptions and lies under the garb of true news. Their major role is to keep control over the information in order for others to trust it. Various examples of comparable websites may be found all around the world. As a result, people's thoughts are influenced by erroneous information. Researchers believe that a variety of man-made intellectual ability evaluations can aid in the identification of bogus information, according to the research (Jain et al., 2019).

A specific group on social media has muddied the facts by blending believable and ludicrous material. That is the truth, and it will be classed accordingly. On the other side, the appearance of fake news poses a major threat to people's lives and property. There is misinformation that the distributor believes to be true, and there is propaganda that the

distributor knows is false. but he deliberately lies in the spread of fake news. (Aphiwongsophon & Chongstitvatana, 2018).

The major purpose of social media has been to connect and communicate comparable goals among connected friends and social groups online. Nowadays, the focus of worry has shifted to social media. In terms of design, Facebook and other social media platforms are vastly different from previous media technology (Sirajudeen et al., 2005).

False information that is widely disseminated has the potential to have terrible implications for both society and the individual. As a result, spotting fake news on social media has recently been a popular topic that has gotten a lot of attention. Fake news detection on social media has unique characteristics and challenges that render traditional news media detection systems ineffective or inapplicable. First, fake news is carefully crafted to persuade readers to believe false information, making it difficult and time-consuming to detect solely based on news content. As a result, we'll need to incorporate auxiliary data like client social media activities.to aid in our decision (Shu et al., 2017)

1.1 Problem Statement

Fake news can be used to spread propaganda against a person, a society, an organization, or a political party. Human being is unable to detect all this fake news. There is need for machine learning which can differentiates and can detect fake news automatically (Alim Et al., 2021)

Spreading fake news is a major challenge facing all over the world in recent years due to the rise of technology and internet access The People who spread fake news through social media are trying to gain popularity among the people or to make money misleading the public The most common places to spread fake news are Facebook and YouTube where many are connected Especially in our country there is no machine to

distinguish between fake news and factual news and everyone has written on social media what they think people are interesting and most of the people reported false information. Fake news is a great problem in our country. Due to the political differences between the two groups, the majority of these groups spread fake news on social media to deceive the public to address, we proposed the solution for the above issue by implement a System which enables the user to distinguishing between fake and actual news.

1.2 Objectives

The objectives of this research are to:

- I.** To develop a machine learning model to detect Somalia fake news web- sites/content
- II.** To implement a machine learning system that separates fake and true news.

1.3 Research Question

The Questions of this research are to:

- I.** How to develop a machine learning model to detect Somalia fake news web- sites/content?
- II.** How to implement Somalia a machine learning system that separates fake and true news?

1.4 Scope of The Study

Fake news has become a major challenge for the Somalia community and all over the world in the recent years due to the use of modern technology. Therefore, our research focuses on machine learning which enables people to distinguish between fake news and factual information This system is used for all languages in the world but currently our research is limited to Somali language and will only check Somali news

1.5 Significance of the Study

as we all know, fake news has a significant influence on society, and it is difficult for humans to discriminate between true and false information. As a result, the relevance of this research is to construct a system that can distinguish between fake and true information, as well as a means to detect fake news. This machine checks whether the information on the worldwide website is correct.

1.6 Organization of The Study

Chapter I: Introduction - This chapter will introduce our research study, discuss the background of the study, problem definition, research objectives, questions, and Significance, of study that to research this study.

Chapter II: Literature Review- This chapter focuses on the previous literature about fake news, and real news and researches related to this topic

Chapter III: Methodology - This chapter discusses the methodological development of this project, which includes the techniques that were being used, also will discuss the system's description, overview & features, development environment, Hardware, software requirements, and the best choice device that we are selecting to develop this project.

Chapter IV: Analysis and design – This chapter discusses how the system works. Which analyses the current system status including the manual system and the attached software and the vulnerability of that system among the security and accuracy of the information, this chapter also discusses the solution given by this work to solve all possible problems in the current system.

Chapter V: Implementation – this chapter confers about the design and development of a system by using software tools and hardware devices. It also displays the most important

code which makes a fundamental impact on the system functionality and screenshot about the system interface.

Chapter VI: Conclusion and Recommendation - This chapter clears the overall summary of this project based on the objectives and findings, the drawback of our system and recommendations for future works that important will to improve

CHAPTER II: LITERATURE REVIEW

2.1 Introduction

Fake news has been spreading on social media for several years, yet there is no agreed-upon definition of the term "fake news." Appropriate research on false news identification is needed to better guide future directions clarifications are necessary. Fake news detection on social media is yet in the initial age of growth and there are still many troublesome issues that need to be investigated additional. It's critical to talk about prospective research avenues for improving false news detection and mitigation(Shu et al., 2017).

The main goal is to identify bogus news, which is a standard text classification problem with a simple solution. It is necessary to develop a model that can distinguish between "real" and "fake" news. This has ramifications on social networking sites like Facebook, Instagram, Twitter, and instant messaging apps like WhatsApp where bogus news gains traction and spreads across the country and beyond the world. The proposed system aids in determining the veracity of news. If the news is false, the user is sent to the appropriate news article.(Jain et al. 2019).

When a major event occurs, many individuals use social networking sites to talk online. They seek out, relate, and discuss current events as part of their daily lives. Some types of news, such as natural disasters or climate change, are unpredictably unpredictable. When unexpected occurrences occur, fake news is aired, causing uncertainty owing to the nature of the events. Who knows the true facts about the incident, but the majority of people believe forward news from trusted friends or relatives? Fake news is created by disinformation, misunderstanding, or unreal content that is spread by reliable sources(Aphiwongsophon & Chongstitvatana, 2018).

As of today, the Internet allows easy access to a plethora of undesired, false, and inaccurate statements that can be generated by anyone, making it an excellent platform for dissemination. Fake news is currently a major issue all over the world. It's become normal to see celebrities and even representatives of the government spread false information in order to influence people's actions, whether deliberately or subconsciously. Fake news frequently takes the shape of news media material, but not the organizational process or aim(Botha & Pieterse, 2020)

2.2 History of fake news

Misleading news isn't a new concept. "The use of propaganda is ancient, yet have never has there been the technology to broadcast it so efficiently," writes Natalie Nougayrède, a journalist for the Guardian. When looking at and publishing on modern manifestations of what has been dubbed a "21st-century information disease," it is critical to grasp the historical backdrop Since at least the time Antony met Cleopatra in Roman era, misinformation, deception, and propaganda have been a part of human communication. Octavian launched a propaganda effort against Antony in order to destroy his name "Brief, snappy phrases inscribed on coins in the style of archaic Twitter" were used. These statements depicted Antony as a womanizer and an alcoholic, claiming that he had been corrupted by his affair with Cleopatra and had become Cleopatra's slave. "Fake media had enabled Octavian to hack the republican system once and for all," said Octavian, who later became Augustus, the first Roman Soldier.(Posetti & Matthews, 2018)

Rameses the Great propagated lies and disinformation depicting the War of Kadesh as a decisive win when the war officially ended in a standstill in the 13th century BC. Many other examples of fake news arose over the ages, but its usage as a form of propaganda during both the first and second world wars became popular during the 1900s. However,

the emergence of the Internet in the late 1990s resulted in an increase in the quantity and accessibility of information, allowing fake news to spread rapidly(Botha & Pieterse, 2020)

This definition of fake news would be put to the test in light of the Russian government-funded news channel RT (formerly Russia Today) and the news agency Sputnik's protection of the 2016 'Lisa case.' Both have been tried to accuse of trying to produce fake news in order to disintegrate Western societies by trying to sow doubt about the dignity and capabilities of Western institutions. Numerous researchers share this view with the US intelligence services, the European Parliament, and the French President. Lisa, a 13-year-old Russian-German girl, went missing in Berlin for roughly 30 hours in January 2016. The Russian state television channel One (also known as Pervij kanal) was the first to broadcast that Lisa had been abducted and raped by foreigners, according to her aunt. Sputnik's German-language edition reported the same story. Lisa's relatives, according to both outlets(Baade, 2018)

The World Economic Forum recognized "digital networks" as a worldwide threat in 2013, describing how the viral spread of false information can have major real-world implications. The propagation of unsubstantiated content, for example, can harm the reputations of politicians, businesses, and organizations. It can even destabilize social order by instilling fear of security concerns or disease outbreaks. Due to the dissemination of misinformation on social media, violence and robbery erupted in London and other places throughout England in 2011. The next year, Lord McAlpine, a famous retired politician, was falsely accused of child sexual abuse on Twitter. With false news reports circulating during the US presidential election and obvious misrepresentation surrounding the United Kingdom's vote to quit the European Union(Andy Yee, 2017)

Servicio Mundial, S.A.", a feature news and opinion firm with an apparently commercial face, was another secret Joint Information Office organization. It was deemed "an intrinsic part" of the propaganda apparatus since it was funded and handled by Franco-British agents. It was led by Alexis Loustau, a Mexican of French ancestry whose mission was to provide stories and images to the press in the countryside that were friendly to the Allies and, later, the "Free French Party". He employed well-known Mexican reporters to produce pre-packaged reports to hide his propaganda objectives. is used when the author is French or British(José Luis, 2017)

The concept of fake news is not new. Although related terminology such as fake news has been available since the 16th century, and the ability of news to skew public opinion for political or financial gain has long been acknowledged, it appears to have emerged in the late 19th century. Yellow journalism was a term used in the early 19th century to characterize exaggerated or outright invented reports, and it was linked to profit goals by news agencies, much like fake news is today. Yellow journalism is credited with fuelling the fury that led to the Spanish-American War, and it was likely the predecessor of tabloid news.(Mason et al., 2018)

The blockade went into effect in June 2017. Tensions first erupted on May 23, 2017, when the state-run Qatar News Agency published controversial claims. Emir Sheikh Tamim bin Hamad Al Thani, Qatar's head of state, is said to have uttered these remarks. The remarks reaffirmed Qatar's good relations with a number of other countries and groups, including Iran and the Muslim Brotherhood. Hamas and the Muslim Brotherhood The sentiments contrasted with the Gulf Cooperation Council's (GCC) traditional international policy, which held such groups and countries in low regard—at least publicly. Al Thani also purportedly alluded to Iran's role as a regional force, an apparent jab at Saudi King Salman

bin Abdul-Aziz and US President Donald Trump's Hamas and the Muslim Brotherhood. The sentiments contrasted with the Gulf Cooperation Council's (GCC) traditional international policy, which held such groups and countries in low regard—at least publicly. Al Thani also purportedly alluded to Iran's role as a regional force, an apparent jab at Saudi King Salman bin Abdul-Aziz and US President Donald Trump's attempts to isolate Iran during the May 20–21 summit in Riyadh. Although some of these comments may appear benign, they were regarded as Qatar deviating from GCC foreign policy when viewed through the lens of GCC animosity against Iran. to isolate Iran during the May 20–21 summit in Riyadh. Although some of these comments may appear benign, they were regarded as Qatar deviating from GCC foreign policy when viewed through the lens of GCC animosity against Iran. Qatar denied Al Thani made such statements, claimed the country's social media accounts and official news network had been hacked(Jones, 2019)

Despite the fact that fake news has a long history and that it is not a completely new phenomenon, the creation of a new information environment and behavior known as post-truth cannot be denied. Overconsumption of information, fueled by the internet, has resulted in a "post-truth" society, in which individuals absorb material that confirms their pre-existing ideas and ideologies rather than attempting the hard work process of determining the truth(Revez & Corujo, 2021).

On October 20, 2016, Buzz Feed reported on a 21st-century media phenomenon That was both unsettling and transformative: the avalanche of "fake news" during a presidential election. Buzz Feed's revelation drew little notice at initially. However, three weeks later, Donald Trump was elected president. Then came CEO Mark Zuckerberg's declaration that it was "a pretty insane idea" to say that Facebook was responsible for Trump's election, which unlocked the floodgates. Over the next two months, hundreds of articles and

editorials sounded the alarm that fake news was the Gotterdammerung of democratic countries in the Information Age(Gorbach, 2018).

Fake news has increased and spread as a result of recent political events. Humans are inconsistent, if not outright terrible detectors of fake news, as evidenced by the pervasive effects of the ubiquitous onset of fake news. With As a result, efforts have been made to automate the process of detecting bogus news(O'Brien, 2018).

Anyone in today's world can publish content on the internet. Unfortunately, fake news attracts a lot of attention on the internet, especially through social networking sites. People are misled, and they don't think twice about sending such erroneous information to the farthest reaches of the system(Jain et al., 2019)

2.3 How does fake news spread?

Despite the development of several fact-checking tools in academia and business, false news continues to spread on social media These systems primarily focus on fact-checking, but they frequently overlook internet users, who are the major causes of misleading transmission. In the recent years, the widespread dissemination of biased news, politicized tales, incorrect claims, and misleading information has prompted social concerns. Many investigations said that falsified news may have influenced voters' misperceptions about political candidates, causing stock prices to be manipulated.(Vo & Lee, 2020)

Digital communication has aided in the removal of time and geographical constraints to information exchange and presentation. Although all of its benefits quicker communication has resulted in widespread dissemination of fake information The fatal COVID19 epidemic is presently sweeping the globe, and misleading news about the disease, its treatments, prevention, and causes has been rapidly disseminated to millions of people. During such critical times, the spread of fake news and misinformation may have serious implications,

causing widespread fear and exacerbating the pandemic's threat. Limiting the spread of false news and ensuring that accurate information is conveyed to the public is therefore critical (Vijjali et al., 2020)

While fake news is not a new phenomenon, the internet information ecology is especially conducive to spreading false information. Because of the low cost of creating fake websites and the enormous amount of software-controlled profiles or pages known as social robots, social media may be readily used to spread propaganda.(Shao et al., 2018)

False information on social media sites has become a confrontational public issue, as these platforms provide third parties with a variety of digital tools and strategies to spread disinformation in order to further self-serving economic and political interests, as well as alter and polarize popular opinion. We investigate misinformation tactics on social media sites.(Ng & Taeihagh, 2021)

Fake news articles can be found on a variety of websites. For example, certain websites, such as denverguardian.com, are specifically devoted to publishing purposefully falsified and misleading content. These sites' names are frequently intended to seem similar to those of reputable news organizations. Other satirical websites, such as wtoe5news.com, have articles that may be misconstrued as factual if read out of context. Other websites, such as endingthefed.com, publish a combination of true and misleading material. Fake news websites have a short lifespan, and several that were influential in the run-up to the 2016 election are no longer active.(Allcott & Gentzkow, 2017)

Social media has been dubbed "the lifeblood of false news" because it allows anybody to simply and cheaply broadcast pandemic fake news to large audiences. Concerns concerning the spreading of false news center on social media's ubiquity as well as the simple dissemination of information that social media platforms allow owing to their technical

capabilities. These features of social media have opened up a new channel for propaganda on the Internet, which will be used to propagate misinformation with growing complexity.(Leeder, 2019)

2.4 How to Prevent the Spread of the Fake News?

The development of media technology and social networks over the last two decades has made it possible to produce and broadcast information more quickly and widely than ever before. However, technology has also enabled the quick and widespread dissemination of deliberate misinformation and fake news. In the last several years,(Kedar, 2020) As the spread of disinformation online grows, particularly in media platforms such as social media feeds, news websites, and newspaper articles, the identification of fake news has increasingly attracted the interest of the general public and researchers.(Pérez-Rosas et al., 2017).

To fight the spread of false news, journalists must comply with strict criteria, according to researchers and several news organizations and politicians have collaborated to devise strategies to combat the spread of fake news. Manually identifying fake news sites is a subjective and time-consuming operation that necessitates topic knowledge. Due to the breadth of social media platforms, which include a wealth of information, and the lack of a perfect method to check the reliability of accounts, it is practically difficult to completely prevent the spread of fake news on social media(Ansar & Goswami, 2021)

The recognition of satirical news is critical for preventing the spread of fake news on the Internet. Machine learning methods like SVM and hierarchical neural networks, as well as hand-engineered features, are used in existing ways to capture news satire, but they don't look at the distinction between sentences and documents. This research provides a robust,

hierarchical neural network based strategy for satire recognition that can detect satire at both the sentence and article level.(Sarkar et al., 2018).

We improve the fact-checking technique employed by major online social networking sites to limit the spread of disinformation. Any article in a user's feed can be reported as false information, and if a story gets enough flags, it is submitted to a third party for fact verification. If a third-party determines that a story is false, it is marked as challenged and may show lower in the readers' feeds. Because third-party fact-checking is expensive, we must choose which articles to fact-check and when to do so—decide how many flags are sufficient.(Kim et al., 2017)

The research community began devoting its efforts to the challenge of "false" information around the turn of the decade, then to "rumor" detection, or disinformation, and, most recently, to the identification and suppression of fake news. discovered a set of three factors that define the spread of information to forecast the truth of rumors: language style used to describe rumors, characteristics of persons participating in spreading information, and networking propagation dynamics. Their model (created with Hidden Markov Models) was said to be capable of properly predicting the truth of 75% of rumors speedier than any other public source, even journalists and law enforcement officers.(Figueira & Oliveira, 2017)

The proposed approach is a multi-layered evaluative method that will be implemented as an app, with any online material being connected with a tag and a description of the facts about the content. For a greater understanding of the methodologies described, a proof of concept is given. This has supported the development of potential initiatives that certain major Microblogging sites may take to prevent the spread of fake news.(Sirajudeen et al., 2005)

We show preliminary results from our studies using machine learning algorithms to detect bogus news in this research. In specifically, we investigated and created methods and tools for detecting fake news, as well as providing a methodology and creating an algorithm for reporting and detecting bogus media articles. The information will be utilized to build a machine learning technique that will distinguish the articles as bogus or factual. External sources of information, such as reader input, will be used to train the machine learning model. In contrast, we provided a static dataset in our software, and no feedback would be used to build the machine.(Al Asaad & Erascu, 2018)

2.5 Why machine learning is required to detect the fake news?

The majority of mobile phone users prefer to read news on social media rather than on the net. The news is published on news sites, which also serve as a source of identification. The challenge is how to verify news and information shared through social media platforms such as WhatsApp groups, Facebook Pages, Twitter, and other microblogging and social sites. It is detrimental to society to take in tales and try to be a news source.(Jain et al., 2019)

Several experts claim that machine learning and artificial intelligence can help solve the problem of misinformation. Because hardware is cheap and larger datasets are given, machine learning algorithms have lately started to enhance research on a variety of categorization challenges, such as picture identification and speech identification(Agarwalla et al., 2019). Machine learning has been used to detect bogus news headlines automatically.(Manzoor et al., 2019)

As a result, stifling bogus news is a must. Only when a person understands the full article of a subject can he determine whether or not the information is false. It's a challenging endeavor because most readers don't know the whole story and simply believe the

misinformation without verifying it. Because a person cannot control bogus news, the question of how to limit it emerges. Machine learning is the solution. Misinformation can be detected with the use of machine learning. When somebody posts bogus information, machine learning algorithms will examine the information of the post and identify it as bogus. Various academics are attempting to discover the most effective machine learning classifier for detecting bogus news.(Alim Et al., 2021)

2.6 Traditional Media and Fake News

Fake News in Traditional Media: Newspapers, radio, and television are examples of traditional media. When someone comes across a bit of news, determining if it is authentic is tough since everyone has the tendency to feel that their view is always correct. As a result, journalists frequently target this "first-hand view" without verifying its veracity in order to spark the reader's interest. Furthermore, correcting incorrect information takes longer than delivering true news to individuals. Another characteristic of traditional media fake news is social acceptability.(Nyow & Chua, 2019)

Traditional media are traditional forms of information and communication that have been used by many worldwide groups and cultures from the beginning of time. Although they depict communication channels for, by, and of the common people of a culture or area, folk media are some of the most dynamic manifestations of traditional media. Modern media, in contrast to traditional communication, refer to recent mass communication or current communication linked to a newly produced or improved technology. No, modern media is not one of them. (Debashis, 2009).

Traditional media are those which give out identical messages in a one-way channel to huge, similar audiences. In the development of media contents, social media differs from traditional media in that social media content has been more widely disseminated across the

public, rather than being confined to media professionals, and there is greater accessibility, quality, and engagement. Advertisement, entertainment, and news are just a few of the areas where social media and traditional media have a relationship. Although social media is here to stay, the research predicts that conventional media will continue to attempt to complement social media.(Apuke, 2017)

Nowadays, social media has replaced traditional news channels as the primary method for disseminating information. The grounds for this substitution are that: I getting news through social media is less expensive; and sharing, reviewing, and debating with other readers on social media is quicker.(Nyow & Chua, 2019)

2.7 Text Classification

Text information is growing in importance as the Internet continues to develop, and textual information evaluation will become increasingly important. Text categorization is a main technology in text information analyzation. We can classify text data using an algorithmic text classification system, allowing individuals to better discover, organize, and evaluate text information resources, therefore developing a successful text classification model is critical. In the last few years,(Liu et al., 2010).

Automatic Text Classification is a machine learning approach that automatically assigns a text content to a set of pre-defined classifications. Important terms or traits collected from the text material are frequently used to classify the document. It's a supervised machine learning problem because the categories are pre-defined.(K. Dalal & A. Zaveri, 2011). Because of the massive volumes of text data generated in a range of social network, online, and other information-centric systems, the subject of text mining has gotten a lot of attention in recent years. Unstructured data is the simplest type of data to produce in any

setting. As a result, there has been a huge need for techniques and algorithms that can handle a wide range of text applications.(Aggarwal & Zhai, 2012)

However, we discovered that the underlying classification model has a significant impact on the performance of a text classifier. Initially, the inherent large dimensional with tens of thousands of words, even for a moderate-sized text collection, makes many learning techniques unacceptably computationally costly and rapidly increases the overfitting problem. Second, polysemy occurs when a single word can have several meanings and synonyms are employed to express the same topic. interfere with the formation of proper categorization functions, making this a tough assignment (Tao Liu et al., 2004)

Many redundant words, such as punctuation marks, spelling mistakes, and terminology, may be found in most text and document data sets. Noise and superfluous features can degrade the performance of many algorithms, particularly statistical and probabilistic learning algorithms. We'll go over several strategies and methods for text cleanup and pre-processing text data sets in this part.(Kowsari et al., 2019)

Text classification has become a necessary due to the large amount of text documents we deal with on a daily basis. Topic-based text classification and text genre-based document classification are the two types of text classification. The most of the information used to categorize genres comes from the internet, as well as communities, message boards, and broadcast or printed news. They have a diversity of media, favored terminology, and sentence styles, even within the same issue, since they are inter. In particular, the information is diverse..(Ikonomakis et al., 2005)

2.8 Basic concept of machine learning

During the previous two decades, Machine Learning has become one of the pillars of information technology, and with it, a highly important, yet often overlooked, component

of our lives. There's reason to expect that intelligent data analysis will become more widespread as the amount of data available grows. As a prerequisite for technological advancement, it is becoming increasingly common.(Krzysztof R. Apt, 2003).

Machine learning is a concept that refers to improving future performance by learning from existing experience (in this example, historical data). This field's main concentration is on automated learning methods. Learning is the automated adjustment or enhancement of an algorithm based on previous "experiences" without the need for human intervention. Machine Learning and Statistics combined; the result was phenomenal: Machine Learning. Computer science focuses on creating machines that solve specific problems, as well as determining whether or not issues can be solved at all. Data inference, hypothesis modeling, and determining the dependability of results are the three basic approaches used by Statistics.(K. Das et al., 2007)

Machine Learning (ML) Whenever the computer program is given a set of tasks to do, it is claimed that the machine has learned from its experience if its measured performance in these activities improves as it obtains more experience. As a result, the machine makes judgments and forecasts based on facts. Consider a computer software that learns to detect/predict cancer using a patient's medical investigation reports Beginning with the basic idea of Machine Learning is an excellent place to start for this article. A computer model is given to execute some activities in Machine Learning, and it is claimed that the machine has learned from its experience if its measured performance in these tasks improves as it obtains more experience in performing these jobs.(Ray, 2019).

Machine learning is the research of how to use models to control human learning activities, as well as the study of computer self-improvement techniques for acquiring new knowledge and skills, identifying current information, and continuously improving performance.

Machine learning, as comparison to human learning, learns quickly, accumulates more information, and spreads the outcomes of learning more easily. As a result, every advancement made by humans in the field of machine learning would improve the capabilities of computers, having an influence on human civilization.(Wang et al., 2009)

2.8.1 Types of machine learning

Machine learning tasks are classified into two main categories, that is, supervised and unsupervised learning, based on the learning system's learning signal in supervised learning, data is provided with example inputs and outputs, with the goal of developing a general rule that maps inputs to outputs. In other circumstances, inputs are only partially available, with some desired outputs missing or only provided as feedback to actions in a dynamic context. The gained knowledge (trained model) is employed in the supervised scenario to for the test data, forecast the missing outputs (labels). However, there is no such thing as unsupervised learning. With the data being unlabeled, there is a separation between training and test(Liakos et al., 2018)

2.8.1.1 Supervised machine Learning

Supervised learning approaches build prediction models by learning from a large number of training instances, each of which contains a label representing the outcome of the ground-truth. Despite the fact that existing methodologies have had considerable success, it is worth noting that owing to the high cost of the data-labeling process, it is impossible to get sufficient supervision information such as entirely ground-truth labels in many activities. As a result, machine-learning algorithms should be able to function with little or no supervision.(Schrider & Kern, 2018)

The quest for algorithms that reason from externally given cases to generate broad theories, which accurate predictions about future situations is known as supervised machine learning. To put it another way, the purpose of supervised learning is to construct a compact model of the distributions of class labels in terms of predictor characteristics. The resultant classifier is then used to give class labels to the testing examples with known prediction feature values but unknown class label values. Various supervised machine learning classification algorithms are described in this article.(Kotsiantis, 2007)

Machine learning has performed in a variety of tasks, specifically supervised learning tasks like classification algorithms. Models are often trained using a training data set with a large amount of training examples.(Zhou, 2018).

2.8.1.2 Unsupervised machine Learning

Finally, in unsupervised learning, the system just accepts data without receiving unsupervised target outputs or rewards from its surroundings. Given that the machine receives no feedback from its surroundings, it may seem difficult to understand what it might possibly learn. However, based on the idea that the machine's purpose is to generate representation of the input that can be utilized for decision making, predicting future inputs, and effectively transferring the inputs to another machine, a formal framework for unsupervised learning may be developed.(Zoubin , 2004)

Unsupervised machine learning approaches make it easier to analyze raw information, allowing for the generation of analytical findings from unsupervised learning. Recent advancements in hierarchical learning, clustering algorithms, factor analysis, latent models, and outlier identification have significantly improved the state of the art in unsupervised machine learning approaches(Usama et al., 2019). Unsupervised techniques, on the other hand, divide a dataset into multiple groups based on the algorithm's strength of grouping.

Unsupervised approaches that have meaningful outcomes in detecting the unobserved abnormality includes innovation and outlier detection algorithms.(S. Das et al., 2020)

2.9 Fake news on social media

The use of social media for news consumption has two sides. On the one hand, consumers seek out and consume news via social media because of its low price, quick access, and rapid distribution of information. On the other side, it facilitates the widespread dissemination of "fake news," which is low-quality news that contains intentionally misleading material. The widespread dissemination of fake news has the potential to have tremendously detrimental consequences for both individuals and communities. Social media has proven to be a powerful resource for spreading fake news. There are also some new patterns that can be used to detect bogus news on social media. A fundamental grasp of the state-of-the-art false news detection methods can be gained by analyzing current fake news detection approaches in different social media settings.(Shu et al., 2017)

Several early definitions of social media have been proposed, both within and throughout fields such as public relations, information systems, and main stream media. One of the goals of this project is to develop a new, broad but accurate, a causal definition of social media. We believe it is important to distinguish between a social medium and a medium that promotes social behavior. We differentiate social media as an unique subset of media tools that share a common set of qualities and behaviors, where the attributes for disparate individuals and organizations to help in the creation of the content they are going to consume provide inherent value far greater than what each personal site feature can provide. As a result, we have explicitly(Carr & Hayes, 2015)

When comparing to other kinds of social media such as websites and workgroups, social networks clearly have the upper hand. From 2009, social networks have risen to

prominence as the major means by which active Internet users communicate with one another. Since 2009, official corporation and brand internet sites have seen a decline in traffic. This reduction, according to the Wave research group, may be due in part to the development of social media marketing by brands as a more prevalent marketing strategy.(Hutton & Fosdick, 2011)

Leaders make decisions in the context of an immediate global media cycle, impacted for anyone with the potential to activate an emotion in the audience and biases via social media. Individuals constantly post events that occur around them on social media in practically every minute of modern warfare, where the timeliness of reporting can lead to information superiority. The writer, David Patrikarakos, a Britain investigative reporter who covers war and world relations, shows how social media has altered the landscape of warfare in the twenty-first century by shifting the power of institutional media outlets to the individual, whom he refers to as "homo-digitalis," or the hyper-empowered individual. Patrikarakos claims that everybody with internet connection is a potential terrorist can be used as a fighter.(Ron Chernow, 2017)

People generate stuff, share it, bookmark it, and network at a rapid rate on social media, which has grown as a category of online conversation. On the academic side, Facebook, Myspace, Twitter, and are examples. Social media is rapidly transforming public debate in society and defining trends and objectives in themes ranging from the environmental and politics to technology and the entertainment sector, thanks to its simplicity of use, speed, and reach. We decided to explore social media's ability to anticipate real-world outcomes because it might be viewed as a type of collective knowledge. Unexpectedly, we observed that a society's chatter may be able to increase quantitative predictions that exceed those made using artificial markets.(Asur & Huberman, 2013)

2.10 Related Work

Several initiatives have been undertaken to identify bogus news. The University of Pretoria in South Africa. Students Harm de Wet and his teammates have produced a study report on false news by South African students. In the twentieth century, they conducted their study on paper and through local social media. Use of the internet is growing, as are the number of postings and papers being recorded. To detect bogus material, they used a number of strategies and tools, including technological Logistic Regression models, education, and machine intelligence. (De Wet & Marivate, 2021)

Manisha Gahirwal and his teammates at the Vivekanand Educational Society Institute of Technology in Mumbai, Maharashtra. A research study on the detecting of false information was released by 2018. They used Random Forest model, variety, and novel filters to find bogus information and show that they were able to achieve highest accuracy. (Gahirwal, 2018)

A study article on the identification of fake news has been released by Kelly Stahl of California State University Stanislaus. He uses the Naïve Bayes Classifier, Vector Machine Support, and Text Analytics to detect bogus news from social media as an accurate technique to recognize fake stories. (Stahl, 2018).

A study article on the identification of disinformation was released in 2018 by Simon Lorent of the University of Liege. With the use of machine learning Naïve-Bayes SVM embedding, he designed a method for detecting bogus news that works far better for this job. community. (Lorent & Ito, 2018)

2.11 Research Gap

Table 2.1 Research Gap

Article	Accuracy	Implementation Method
(Granik & Mesyura, 2017)	74%	Naive Bayes
Aphiwongsophon & Chongstitvatana 2018)	94.67%	Naive Bayes, SVM,
(Jain et al., 2019)	93.50%	Naive Bayes, SVM,

2.11.1 Gap

Although this is the first time, we are going to detecting the Somali fake news on somalia language so this will be a huge challenge for us to this project. This is only reason we can't find any full research paper in detecting somalia news. As the grammar of somalia language is too different to English grammar so to identify the somalia fake news we have to build a new model that can detect somalia fake news easily. This task is not easy like to detecting English fake news.as well as, collection of data for this project it was so difficult to us because the fake and real news has no other exceptional identity to detect them. As a result, we will need to do more research to find all the different types of data.

CHAPTER III: RESEARCH METHODOLOGY

3.0 Introduction

The main purpose of this unit is to describe the use of machine learning by distinguishing between fake and factual information using the supervised machine learning to determine the appropriate model for distinguishing between false and factual information.

This chapter separates the appropriate model by distinguishing between false and true information by collecting data and then converting it into a database which can be understood by the machine and has a multi-model machine that divides into two major categories namely Supervisor machine learning and unsupervised machine learning and we use supervised machine learning.

In general, the components are composed and I will discuss everything we usually use in our system such as Hardware, software and dataset used for search. and analysis to design a system capable of sorting the data obtained by this model we will benefit from distinguishing false and true information by seeking and using various algorithms such as supervised and the like.

3.1 System Description

This machine learning-based fake information detection system can recognize exaggerated and false information and presents the findings of the user interface, which is based on webapp flask and makes it easy for the user to detect false information. The system of the webapp flask is able to notify the user whether or not previous suspicious information is accurate news or fake news

3.2 System Architecture

In this section we will explain the structure of our system and how it will work. This system consists of data collection that is sent to the model by verifying the extracted data and then displayed on the dashboard using the Python flask.

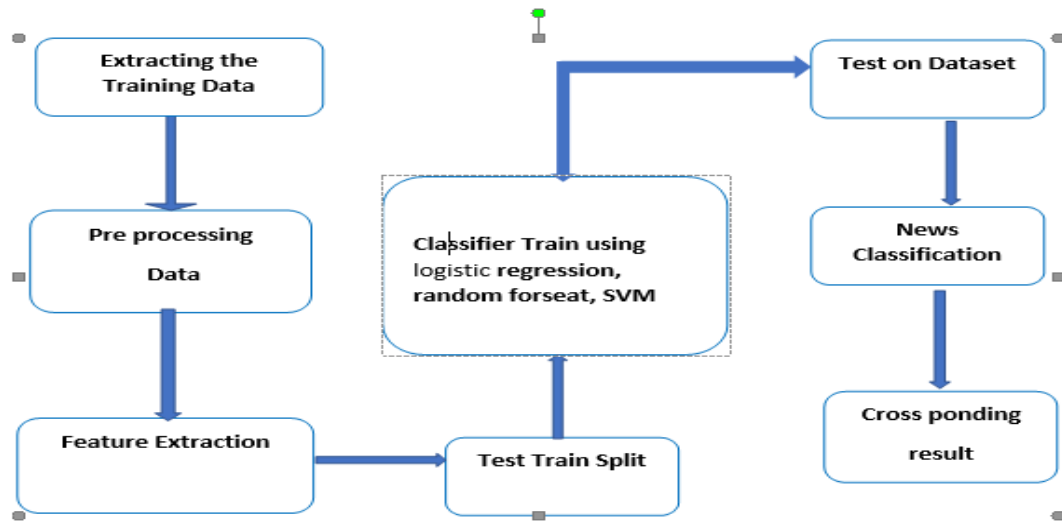


Figure 3.1 System Architecture

3.2.1 Extracting the Training Data

Extraction of training data presents a series of private information. Most of these dangers are avoided in our study since we target, whose training data is publicly available. However, because our assaults might be used against anybody, we also talk about the possible effects of future assaults, ML on models that may be trained on personal information.

Natural Language Processing (NLP) is the capacity of a computer software to interpret spoken and written human language, often known as natural language. It's a part of AI (artificial intelligence) (AI).

3.2.2 Preprocessing Data

The data preprocessing can often have a significant impact A supervised ML algorithm's extension ability. One of the most challenging tasks in deductive ML is the reduction of background occurrences. Data preprocessing is the process of preparing and "cleaning" text data so that machines can examine it. Preprocessing transforms data into a usable format and emphasizes text characteristics that an algorithm can use.

3.2.3 Data cleaning

In any machine learning project, this is a vital stage. In tabular data, you may investigate your data using a variety of statistical analysis and data visualization approaches to find data cleaning activities you might wish to do. Before moving on to the more advanced methodologies, you should definitely undertake some fundamental data cleaning activities on any machine learning project. These are so fundamental that even seasoned machine learning practitioners sometimes forget them, yet they are so important that if they are skipped, models may break or produce unduly optimistic performance results.

Data cleaning is a highly critical phase in every machine learning model, but especially so for NLP. Without the cleaning procedure, the dataset is frequently a jumble of words that the machine is unable to comprehend. We'll go through the procedures involved in cleaning data in a typical machine learning text pipeline

I. Stop Words: Once we've broken down text into tokens, it's often evident that not all phrases provide the same amount of information, if any at all, for prediction. Stop words are common terms that contain little (or no) relevant information. The elimination of stop words is popular advice and practice for numerous NLP tasks, however the work is more complicated than many resources lead you to assume. In this chapter, we'll look at what a

stop word list is, how it differentiates from those other lists, and how it affects your preprocessing workflow.

We now have a list of terms that do not contain any punctuation. Let's move forward and get rid of the term "halt." Termination words are meaningless words that have no bearing on the ability to determine if a text is truthful or untrue. It will be used to halt the creation of terms based on the Somali language. As we have said before stop words are sentences that are meaningless and every language has stop words and we want to make stop words based on the Somali language not done before and now we want to do it to some extent

II. Punctuation There are various punctuations in the title text. Punctuation is rarely used since it adds no value or meaning to the NLP model. There are 32 punctuations in the "string" library. Punctuation is regarded separately from word and numeric tokens as a token. Commas (,) and apostrophes (') are considered as their own tokens when bounding punctuation is used.

III. Tokenization is the process of breaking down a phrase, sentence, paragraph, or even an entire text document into smaller components like individual words or phrases. Tokens are the names given to each of these smaller units. Words, numerals, or punctuation marks might be used as tokens. Splitting strings into a list of terms is known as tokenization. To separate the data, we'll utilize Regular Expressions, often known as regex. A search pattern may be described using Regex

IV. Lemmatize/ Stem The process of reducing a word to its root form is known as stemming and lemmatizing. The basic goal is to minimize the number of variants of the same term, hence lowering the number of words in the model. The distinction between stemming and lemmatizing is that stemming removes the last letter of a word without considering its context. Lemmatizing, on the other hand, takes into account the context of the word and

shortens it to its basic form depending on the dictionary meaning. When opposed to Lemmatizing, stemming is a speedier procedure. As a result, there is a trade-off between speed and precision.

3.2.4 Feature Extraction

Refers to the act of converting raw data into numerical features that may be handled while keeping the original data set's content. It produces better outcomes than applying machine learning to raw data directly. The extraction of characteristics results in the identification of the most important elements that contribute to the detection of false news. 70% of the dataset is utilized for training and the remaining 30% is used to test the classification model using k-fold cross validation in the suggested technique.

3.2.5 Test Train Split

When machine learning algorithms are used to generate predictions on data that was not used to train the model, this approach is used to estimate their performance.

3.2.6 Models

I. Decision Tree is a Supervised learning Although it may be used to solve both classification and regression issues, it is most commonly employed to solve classification difficulties. Internal nodes contain dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier.

II. Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning It's a method for predicting a categorical dependent variable from a set of independent

III. Random forest is a supervised machine learning Regression and classification issues are solved using this method. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers.

IV. Naïve Bayes algorithm is a supervised learning algorithm is a basic and effective classification technique that aids in the development of rapid machine learning models capable of making quick predictions.

V. Support Vector Machine (SVM) is a supervised machine learning algorithm This method may be applied to both classification and regression problems. It is, however, mostly employed in categorization difficulties.

3.2.7 Data Collection

Dataset is defined as a collection of data that a computer treats as a single entity. This implies that a dataset can contain a lot of different bits of data, yet it can be used to train an algorithm to uncover predictable patterns within the dataset as a whole. As we mentioned earlier in this chapter, we will utilize a dataset, which is a collection of data, and we will begin collecting dataset one about fake news and factual information about the dataset because our Somali language does not have a dataset already prepared. We'll collect and analyze factual news from the BBC and VOA websites, while we will generate fake news from social media posts Our target will be groups or individuals who publish false information on their social media accounts.

3.3 System Features

There are a variety of characteristics that can improve the effectiveness of a Machine Learning model on any given job. Data correlation is one of the feature selection strategies, and it has a significant influence on the model's performance. This will save the Machine Learning model a lot of time and effort while preprocessing and cleaning the data. The data

properties used to train the Machine Learning model have a significant influence on the model's efficiency. The model output will be lowered due to the irrelevant characteristics that are supplied.

A textbox component of our system allows the user to paste the text he wants to examine for accuracy; the text is then accepted by the model, and an alert appears indicating whether the model data given is accurate or untrue.

3.4 System Development Environment

In this study, we will need to development environment tool that will assist us in proposing this system, as well as a set of processes and programming tools that will provide an interface and a convenient view of the development process, which will include writing code, testing it, and packaging the build for deployment.

The Development Environment includes:

I. Python flask

II. Visual Code (Vs Code).

Flask is a Python based web application framework. Armin Ronacher, who led a team of worldwide Python aficionados known as Pooeco, created it. The Werkzeug WSGI toolkit and the Jinja2 template engine are the foundations of Flask. Both are Pocco initiatives.

Flask is a web framework and a Python module that makes it simple to create web applications. It's a microframework with a minimal and extensible core: it's a microframework without an ORM (Object Relational Manager) or similar functionality.

Visual Studio Code comes with a lightning-fast source code editor that's ideal for everyday usage. VS Code's syntax highlighting, bracket-matching, auto-indentation, box-selection, and snippets let you be more productive faster with support for hundreds of languages.

3.5 System Requirement Specification

In order for the system's implementation stage to be effective, several things must be considered. Each category of software and hardware has its own set of standards. Software refers to the management of a data processing system, which consists of a collection of computer programs, procedures, and documents.

3.5.1 Hardware Requirement

This section explains the necessary hardware requirements for running our program, as shown in the table below.

Table 3.1 Hardware Requirement

Device	Description
Processor	Intel(R) Core (TM) i5-4300U CPU @ 1.90GHz (4 CPUs), ~2.5GHz
RAM	8GB
Display	2.7 inch
System Type	64-bit Operating System
Operating System	Windows 10 Pro

3.5.2 Software Requirement

This section explains the minimum software requirements for using our program, as shows in the table below.

Table 3.2 Software Requirement

Software	Minimum requirment	Reason
python flask	Version 3.7	To build System front end
Microsoft Excel	Version 2019	To use dataset
Microsoft Visual Studio	Version 17.1	To build System code

CHAPTER IV: SYSTEM ANALYSIS AND DESIGN

4.1 Introduction

This chapter provides a system for analyzing and designing and implementing a system that separates fake news and factual news. This part we will discuss the current system problems for distinguishing between fake news and factual news, there for this section will talk about a brief description of the system and its design and the requirements the system needs to work

4.2 System Analysis

System analysis involves the study of machine learning methods of sorting out fake news and the most important face facts is to collect dataset based on the somalia language after teaching the models by removing any Data cleaning as mentioned Chapter three after those models have been tested and any data cleaning will be done, the interface will be streamlined for users to use this system

4.3 Proposed System

This system is designed to prevent the spread of fake news and has two functions including detecting fake and factual news. The accuracy of false news is being investigated in this study. A lot of negative things may be avoided if fake news could be detected early. The data was first and principally acquired through the internet. To prepare and clean the acquired data, data preprocessing methods were used. The data was then made more comprehensive. Finally, when it is presented to the algorithm for prediction, the algorithm produces a result. Our expected performance will be estimated using this technique. When it got to the algorithms, there have been three algorithms we used to predict and prevent bogus news.

4.4 System Requirements

In this section we will discuss the criteria for distinguishing between fake and factual information and classifying the requirements into two categories and it is working part and part not working requirements. However, there is nothing wrong with the requirements of this study

4.4.1 Functional Requirements

A functional requirement is one that specifies how an action or activity should be carried out. The following are the functional criteria that the proposed system must meet:

- **User:** A user is a person who utilizes something, and it is nearly usually used in connection to that object.
- **Input as Data:** Input refers to any data that is delivered to a computer or software application. The process of delivering information to the computer is also known as data entry since the information delivered is also considered data.
- **Data preprocessing:** which is part of data preparation, refers to any sort of processing done on raw data in order to prepare it for further processing.
- **Data extraction:** is the process of gathering or obtaining various sorts of data from a number of sources, many of which are unstructured or poorly organized.
- **Data segmentation:** is the act of splitting and grouping comparable data based on predetermined parameters so that it may be used more effectively in marketing and operations.
- **Classification model:** takes some data and produces an output that categorizes it into one of many categories

4.4.2 Non-Functional Requirements

The requirement of non-functionally are:

- **Security:** the system should have security to ensure the secureness of information.
- **Accessibility:** the system is available in the internet and can be accessed at any time from any place through internet connection.
- **Privacy:** The system can only be used by administrator and authorized users.
- **User Friendly:** The system is simple and interesting.

4.5 System Design

In this part, we'll go through machine learning model system design. System design is the process of defining pieces of a system, such as modules, architecture, components, and their interfaces, as well as data, depending on the requirements. Design architecture, Design interface, and Design databases are all phases in the system design process.

4.5.1 Entity Relationship Diagram:

In our project, the ER model stands for Entity Relationship Diagram. The entities are: Dataset, Preprocessing, Train Machine, Machine Learning Algorithm, Develop Module and Classifier. Dataset has two characteristics: training data and test data. Data pre - processing includes the features such as feature extraction, stop word removal, and steaming.

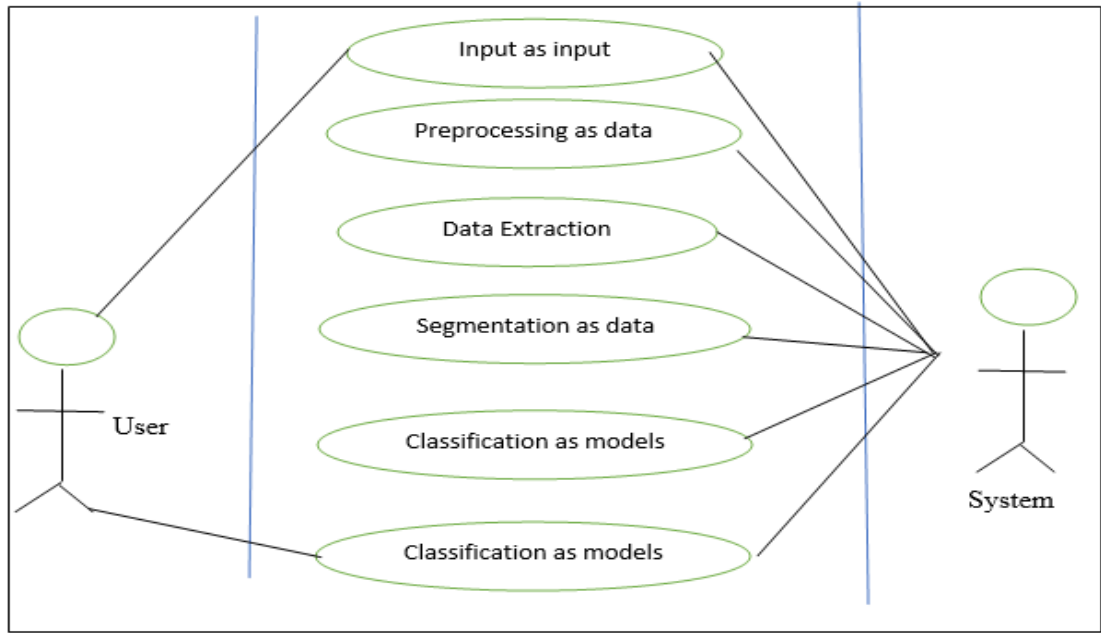


Figure 4.1 Use Case

4.6 Dataset Design

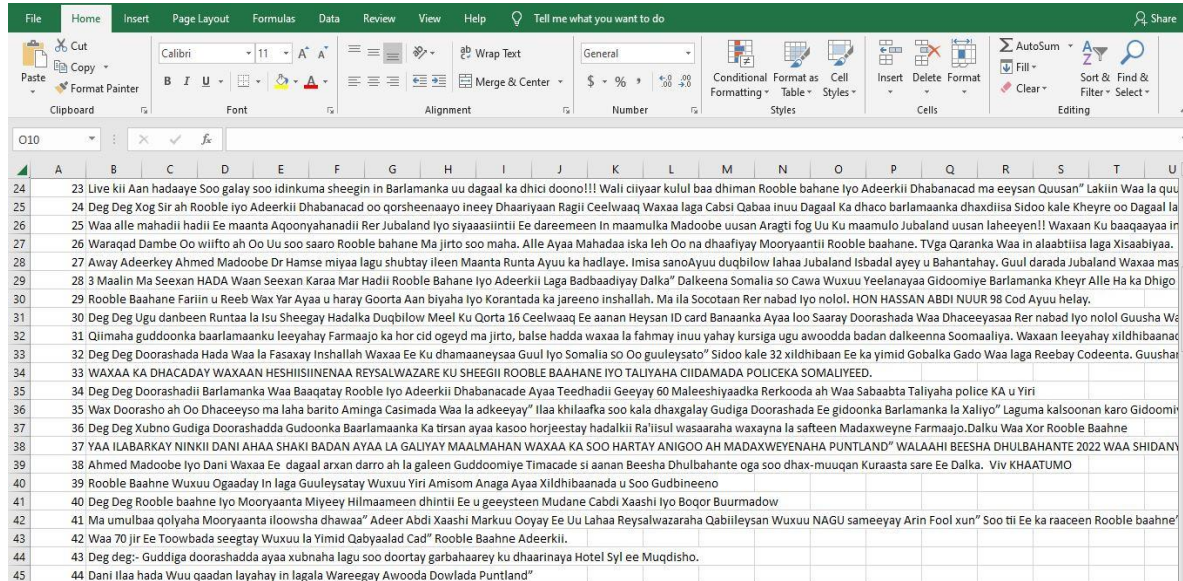
A dataset is a group of cells on an Excel worksheet that contain data that can be analyzed. To make Analysis process work with your data, you must apply a few simple guidelines when structuring data on an Excel worksheet: To clearly characterize the data, use a title. So that we will not use dataset that already prepared we will create a new database based on the Somali language.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	boy																			
2	1	MAAMULKA XASAN SHEIKH WAXAAN U BIXIYAY #DOWLADA_UNKAA_LEH	MARKAAN ARKAY QAABKA UU U HADALAY MOORYAANKAAN EE QABYAALADU	DISHAY EE CUQDADA LAFAHA KA FADHIISATAY BAL A																	
3	2	Manshallah Madaxweyne Xasan Sheikh Mahamuud Walahi Waa Nin wax badan soo bartay inta Uu Mucaaradka ahaa	Abdirahman Abdishakur Waxaa Lagu Aamusihiyay Inuu Kaamam fara badan ka furto dul																		
4	3	Madaxweynaha Qaranka Mudane Xasan Sheikh Mahamuud Waxaan Talo Ku Siinaayaa ugu Horeen inuu Reform Ku Sameeyo TVGA Iyo Radio Qaranka. Si Sumacadii Ee laheeyed hayada Warfaafinta Qaranka																			
5	4	Madaxweynaha la doortay Mudane Xasan soo tuu yiri aargoosi dhici maayo	Waa maxay siyaasiga dhakhsa la beegsaday	Mudane Madaxweyne Waxaa kugula talinaaadaad Shacabka Somaliyeed Khudba																	
6	5	Lacagta \$9.6 million ah Waxaa amaray in la siiyo Imaaradka Reysalwazarahi Hore Rooble	waxaana bank ka soo qaaday wasiirkiisii hore Mohamed Nuur Iyo Safiirka Imaaradka	Madaxweyne Xasan Sheikh																	
7	6	Ahmed Madoobe Maadaama uu iga Guuleysatay Oo Ee dhimatay rajadii ahayd Kismaayo ayaan ka saarayaa	Hal sharuud baan Ku xirayaa Haji Ahmed sharuudaa waxaa weeye Reysalwazare Rer Jubaland ah																		
8	7	MADAXWEYNE FARMAJO MARKA UU HASSAN SHEIKH MOHAMUUD LA HADALAAAY WUU QOOLAAAY LAKIIN MARKA UU SADIID DANI ISKU DAYAY INUU HASSAN LA HADLO WAXAA LA TUSAY WAJI CARO LEH																			
9	8	WAR WEYNE OO FARAXAD AH INSHALLAAH RER NABAD IYO NOLOL IYO SOMALIWEYN IS DIYAARIYA. GUUSHA WAA MADAXWEYNE FARMAJO																			
10	9	Maanta Alle ayaan u soomay si Uu ducada Iiga Aqabalo	Waxaana Ku duceeysanayay Inuu Alle Dib xukunka ugu soo celiyo Mudane Madaxweyne Mohamed Cabdullahi Farmajo Oo Ee Shacabka Somaliyeed																		
11	10	MOORYAANTA IYO DANLEEY CAAYIDII WAA KA DHAMAATAY FACEBOOK WAA KU DAWAKHEEN SABAABTOO AH MA YAQAAANAN CIDA EE TAAGEERAYAAN. HALKA NABAD IYO NOLOL MUSHARAXOODA UU YAHAY																			
12	11	Madaxweyne Farmaajo oo Khudbaddiisa Musharaxnimo Ujeediyey Labada Aqal ee Baarlamaanka JFS	Madaxweynaha Jamhuuriyadda Federaalka Soomaaliya Mudane Maxamed Cabdullaahi Farmaajo, aya																		
13	12	Deg Deg Danleey Iyo Imaaradku Waxaa Ee Doonayaan Iney Dilaan Madaxweyne Farmaajo Xog Sir Madaxweynaha Qaranka Ayey Dhaheen Waa Inuu																			
14	13	Inshallah Madaxweyne Mohamed Cabdullaahi Farmaajo Isagoo Aamisan Oo Aan Booto Iyo Qabqab Ku Jirin Ayuu Guuleysan Doonaa	Bootada Iyo Iswaalka Waxaa Faraha Looga qaaday Danleeyda u shaqeey																		
15	14	Madaxweyne FARMAJO Walee Inuu Libaax yahay Mucaaradkii Marka Ee dagaayad u Waayeen Waxaa Ee Ku Mashquuleen Kabaha Uu Wato	Light ka Cameraa Ayaa wax kale u noqday	Walee Kabahaa Aya																	
16	15	Deg Deg Xog Xasaasi Ah Muqalo Sir Ah Sadiid Dani Iyo Xasan Sheikh Maxamuud Iaga Qabtay Beesha Caalamka Ee Xalane Oo Sheegtay in MD Farmajo 6 Qodob kaga Guuleysatay Mucaaradka Kala u Jeedada																			
17	16	Dad baa Xassan ka dhigaayo shaqsi Cusub!!!!	Waa ninkii Ku yiri Ciidamada ninkaan \$100 dooneeynin irida Ha ka dago	Waa ninkii gabadhaha Somaliyeed u beec geeyay Dalalka carabta.	Waa Ninkii d																
18	17	Suaal Mooryaanti Maxaa Ee Ku Cayayaan Xassan Ali Kheyre	Hassan Ali Nin Afgaaban buu ahaa Qabiilka Iyo Mooryaanimu Shuqul Iyo Shaqo Kuma uusan laheeyn	Mooryaaneey Dhiigadiina iska walaqaada.																	
19	18	Hadaadan arkin Maalmaha soo socda iyagoo gudoomiyaha cusub ee baarlamaanka caaynaayo oo dhahaayo farmaajo ayuu jeebka u galay beenaan idiin sheegay Qof kasta oo sharciga Ilaaliyo farmaajo inuu j																			
20	19	Gudoomiyaha Golaha Shacabka Aaden Madoobe oo kale wareegay Rooble Amniga Doorashada Dalka	maanta laga bilaabo islamarkaana kulan leeyahay taliyayaasha Amniga																		
21	20	Danleey Iyo Qaran dumis Iyo Mooryaan Sawirkan Wuxuu idinka bahanyahay inaad Fasirtaan Waa hadii aad Dowladnimada Wax ka taqaanan!!!	Yaa Ila arko fadhii xumada Rooble Baahane?																		
22	21	Madaxweyne FARMAJO Iyo Gudoomiyaha Barlamanka Somalia Mudane Sheikh Adan Madoobe Ayaa Kulamay	Ilaahow Qof Aan Wax Ogeen Ha Caadaabin	Mooryaanti Igu Sawir Qaracan Ha Is Dooxaan Ma																	

Figure 4.2 Dataset True

Brief Description

This dataset contains around 1500 true news headlines from the year 2021 to 2022 obtained from websites. The model trained on this dataset could be used to identify true news articles.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
24		23	Live kii Aan hadaaye Soo galay soo idinkuma sheegin in Barlamanka uu dagaal ka dhici doono!!! Wali ciyaar kulul baa dhiman Rooble bahane Iyo Adeerkii Dhabanac ma eeysan Quusan" Lakiin Waa la quu																		
25		24	Deg Deg Xog Sir ah Rooble Iyo Adeerkii Dhabanac oo qorsheenaayo inee Dhaariyaan Ragii Ceelwaaq Waxaa laga Cabsi Qabaa inuu Dagaal Ka dhaco barlamaanka dhaxdiisa Sidoo kale Kheyre oo Dagaal la																		
26		25	Waa alle mahadii hadii Ee maanta Aqoonyahanadii Rer Jubaland Iyo siyaasiintii Ee dareemeen In maamulka Madoobe uusan Aragti fog Uu Ku maamulo Jubaland uusan laheeyen!! Waxaan Ku baaqayaa in																		
27		26	Waraaq Dambe Oo wiifto ah Oo Uu soo saaro Rooble bahane Ma jirto soo maha. Alle Ayaa Mahadaa iska leh Oo na dhaafiyay Mooryaantii Rooble baahane. TVga Qaranka Waa in alaabtisa laga Xisaabiya.																		
28		27	Away Adeerkey Ahmed Madoobe Dr Hamse miyaa lagu shubtay ileen Maanta Runta Ayuu ka hadlaye. Imisa sano Ayuu duqbilow lahaa Jubaland Isbadal ayeey u Bahantahay. Guul darada Jubaland Waxaa mas																		
29		28	3 Maalin Ma Seexan HADA Waan Seexan Karaa Mar Hadii Rooble Bahane Iyo Adeerkii Laga Badbaadiyay Dalka" Dalkeena Somalia so Cawa Wuxuu Yeelanayaa Gidoomiye Barlamanka Kheyre Alle Ha ka Dhigo																		
30		29	Rooble Baahane Fariin u Reeb Wax Yar Ayaa u haray Goorta Aan biyaha Iyo Korantada ka jareeno inshallah. Ma ila Socotaan Rer nabad Iyo nolol. HON HASSAN ABDI NUUR 98 Cod Ayuu helay.																		
31		30	Deg Deg Ugu danbeen Runtaa la Isu Sheegay Hadalka Duqbilow Meel Ku Qorta 16 Ceelwaaq Ee aanan Heysan ID card Banaanka Ayaa loo Saaray Doorashada Waa Dhaceeyasaa Rer nabad Iyo nolol Guusha Wa																		
32		31	Qilmaha guddoonka baarlamaanku leeyahay Farmaajo ka hor cid ogeyd ma jirto, balse hadda waxaa la fahmay inuu yahay kursiga ugu awoodda badan dalkeenna Soomaaliya. Waxaan leeyahay xildhibaanac																		
33		32	Deg Deg Doorashada Hada Waa la Fasaxay Inshallah Waxaa Ee Ku dhamaaneysaa Guul Iyo Somalia so Oo guuleysato" Sidoo kale 32 xildhibaan Ee ka yimid Gobalka Gado Waa laga Reebay Codeenta. Guusha																		
34		33	WAXAA KA DHACADAY WAXAAN HESHIISIINENAA REYSALWAZARE KU SHEEGII ROOBLE BAAHANE IYO TALIIYAHA CIIDAMADA POLICEKA SOMALIYEED.																		
35		34	Deg Deg Doorashadii Barlamanka Waa Baaqatay Rooble Iyo Adeerkii Dhabanacade Ayaa Teedhadii Geeyay 60 Maleeshiyaadka Rerkooda ah Waa Sabaabta Taliyaha police KA u Yiri																		
36		35	Wax Doorasho ah Oo Dhaceeyso ma laha barito Aminga Casimada Waa la adkeeyay" Ilaa khilaafka soo kala dhaxgalay Gudiga Doorashada Ee gidoonka Barlamanka la Xaliyo" Laguma kalsoonan karo Gidoomi																		
37		36	Deg Deg Xubno Gudiga Doorashada Gudoonka Baarlamaanka Ka tirsan ayaa kasoo horjeestay hadalkii Ra'iisul wasaaraha waxayna la saftaan Madaxweyne Farmaajo. Dalku Waa Xor Rooble Baahne																		
38		37	YAA ILABARKAY NIINKII DANI AHAA SHAKI BADAN AYAA LA GALIYAY MAALMAHAN WAXAA KA SOO HARTAY ANIGOO AH MADAXWEYENAHNA PUNTLAND" WALAAHI BEESHA DHULBAHANTE 2022 WAA SHIDAN																		
39		38	Ahmed Madoobe Iyo Dani Waxaa Ee dagaal arxan darro ah la galeen Guddoomiye Timacade si aanan Beesha Dhulbahante oga soo dhax-muuqan Kuraasta sare Ee Dalka. Viv KHAATUMO																		
40		39	Rooble Baahne Wuxuu Ogaaday In laga Guuleysatay Wuxuu Yiri Amisom Anaga Ayaa Xildhibaanada u Soo Gudbineeno																		
41		40	Deg Deg Rooble baahne Iyo Mooryaanta Miyeey Hilmaameen dhintii Ee u geeysteen Mudane Cabdi Xaashi Iyo Boqor Buurmadow																		
42		41	Ma umulbaa qolyaha Mooryaanta ilowsha dhawaa" Adeer Abdi Xaashi Markuu Oooyay Ee Uu Lahaa Reysalwazaraha Qabiileysan Wuxuu NAGU sameeyay Arin Fool xun" Soo tii Ee ka raaceen Rooble baahne																		
43		42	Waa 70 Jir Ee Toowbada seegtay Wuxuu la Yimid Qabyaalad Cad" Rooble Baahne Adeerkii.																		
44		43	Deg deg:- Guddiga doorashada ayaa xubnaha lagu soo doortay garbahaarey ku dhaarinaya Hotel Syl ee Muqdisho.																		
45		44	Dani Ilaa hada Wuu qaadan layahay in lagala Wareegay Awooda Dowlada Puntland"																		

Figure 4.3 Dataset False

Brief Description

This dataset contains around 1300 fake news columns from the year 2021 to 2022 obtained from social media. The model trained on this dataset could be used to identify fake news articles

CHAPTER V: IMPLEMENTATION AND TESTING

5.0 Introduction

This chapter looks at how research is really carried out, which is a vital aspect of the study because it is tangible. This part covers crucial topics including an overview of the implementation environment, system diagrams, and a description of system features and how they function. Finally, this chapter will describe the real outcome of one of the procedures.

5.2 Overview of the implementation environment

The main objectives of this system are to predict Fake news and factual news using machine learning and modern technology web technology and monitoring and analysis

the correct use of the network user.

Our system for implementing machine learning and webserver. The graphical user interface software component is the python flask at the front end. We also use python so that the Models of Data Collection program must be implemented Dataset as the back-end.

5.3 Snapshots of the system

This system has two major components one is Front-end and back-end of each section I will explain it separately and include her pictures

5.3.1 Front-end

The area of web development that focuses on what users view on their end is known as front end development. It requires transferring the code created by back-end developers into a graphical interface and ensuring that the data is displayed in an understandable manner. All you'd see on a website or web application without Front End Development are unreadable scripts. People without a coding experience, on the other hand, may quickly

understand and use web applications and webpages due to Front-End developers. Everything you see on Google Apps, Canva, Facebook, and other web services is the result of collaboration between back-end and front-end engineers.



Figure 5.1 Home Page

Brief Description

A home page is a webpage that serves as the starting point of website. It is the default webpage that loads when you visit a web address that only contains a domain name



SOMALI FAKE NEWS DETECTION

Home Prediction Get Started

FAKE NEWS

PREDICTION FORM

Siiyeenka warka uu dadka awood uunah in ay kula saaraan marka runta ah iyo warka beenta ah

Enter News:

Predict

Figure 5.2 Prediction

Brief Description

This section is a data analysis section that has one textbox user is required to enter the data and test it immediately and the result is either factual or false information.

5.3.2 Back-end

The back-end refers to the sections of the code that allow it to perform but are not visible to the user. The back end of a computer system stores and accesses the majority of data and operational procedures. One or more programming languages are usually used in the code. The back end, often known as the data access layer of software or hardware, contains any functionality that requires digital access and control.

5.3.2.1 Evaluating the accuracy of classification models

Evaluating the accuracy of a classification model starts with summarizing the results into the following groups:

- i. **True positives (TP):** Observations predicted to be part of a class, that are actually part of the class.
- ii. **True negatives (TN):** Observations predicted not to be part of a class, that are actually not part of the class.
- iii. **False positives (FP):** Observations predicted to be part of a class, that are actually not part of the class.
- iv. **False negatives (FN):** Observations predicted not to be part of a class, that are actually part of the class.

Confusion Matrix A confusion matrix is a table that allows you to visualize the performance of a classification model. You can also use the information in it to calculate measures that can help you determine the usefulness of the model.

Here is how you would arrange a 2×2 confusion matrix in abstract terms

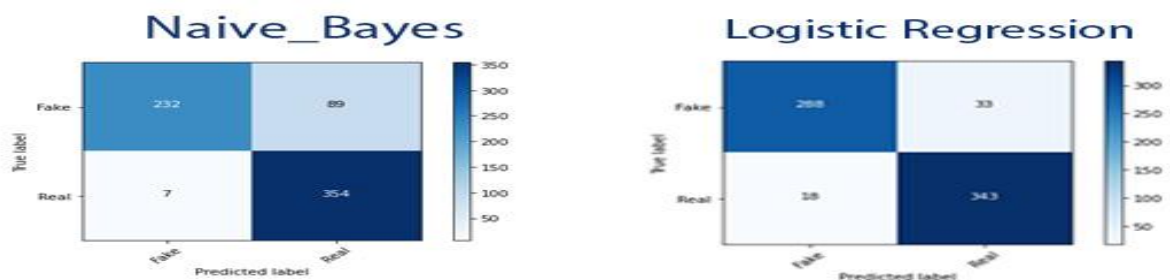


Figure 5. 3 Confusion Matrix for each model

5.4 Count Words

In this section, we will explain the most frequent words in this system, which are the most popular keywords words for the system to repeated mostly in the dataset we have collected, these words we will divide them into the two categories: Actual news and fake news

5.4.1 Fake news

the most frequently used words fake news dataset in the system and we will be presented as graph

```
In [67]: # Most frequent words in fake news  
counter(data[data["target"] == "fake"], "news", 20)
```

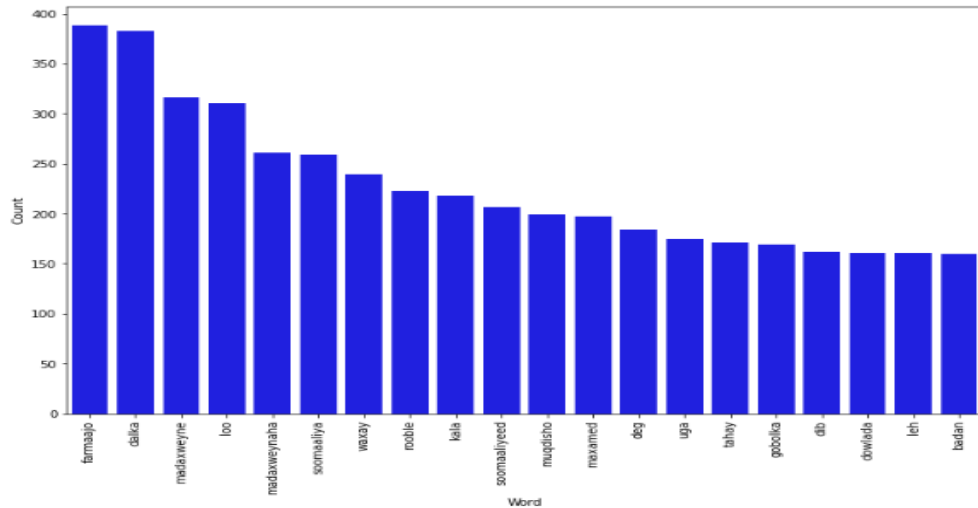


Figure 5. 4 Fake News Word

5.4.2 Real News

the most frequently used words Actual news dataset in the system and we will show as graph below

```
In [102]: # Most frequent words in real news  
counter(data[data["target"] == "true"], "news", 20)
```

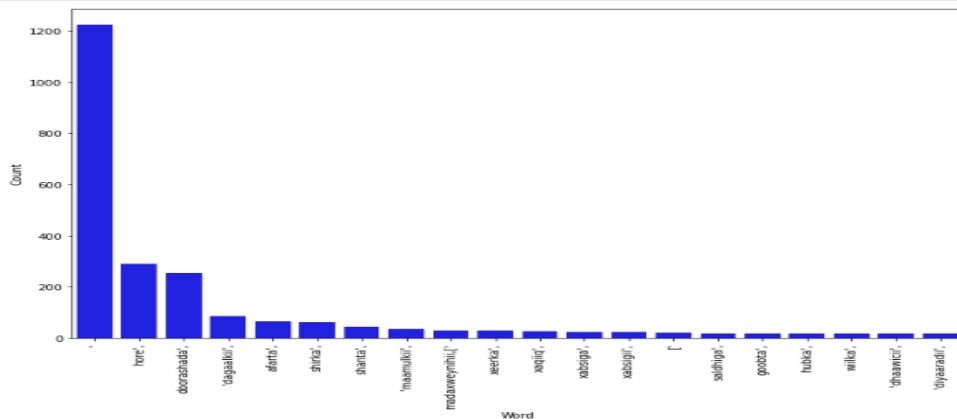


Figure 5. 5 True News Words

5.4.3 Unicode Words

in this section we explain both of the fake news and actual news the most frequent words in the dataset that the system does not repeats is called Unicode often by removing most frequently repeated words the system these words we present the picture below

```
In [99]: data['news'].unique().tolist() #wuxu so bandhiga ereyada unique a

Out[99]: ["sano,markii,koowaad,ayaanay,alshabaab,billaabin,['bishii', 'bisha'],ramadaam,dadka,soomaaliyeed,dilaan,taas,badalkeeda,caa
wa,qofkii,koowaad,['dil', 'dileen'],dowlada,maqlay,['ahaansho', 'ahaadeen'],['wiilka', 'wiilkii'],afuray,gurigiisa,fadhiya
y,sidoo,['dhaawaca', 'dhaawicii', 'dhaawacyada'],ogahay,xaaladda,nolosha,muqdisho,fikir,qariban,badow,saaro,waxaana,dhici,k
arta,goaankaan,gaareen,dad,qurbaha,daahay,badan,kala,socon,nolosha,dadka",
'dagnada,afgooye,mesha,badan,keeno,barandhada,iibiyo,suuqyada,dalka',
'wararka,magaalada,beledweyne,gobolka,hiiraan,sheegaya,xalay,weerar,qaadeen,['saldhiga', 'saldhigii'],ciidan,yaal,xaafad
a,howladdaag,magaalada,beledweyne,dagaalka,dhinteem,carruur,walaalo,saddex,isla,qoyskana,dhaawacmeen",
"maxkamadda,caalmiga,icj,diiday,doodda,weyn,kenya,doodaysay,jiro,['xadka', 'xadkii'],dhexeeya,soomaaliya,kenya,barbar,xar
iiqa,lookka",
'musharax,madaxweyne,daahir,maxamuud,geelle,dhoweyn,ballaaran,loo,sameynayo,socotaa,garoonka,diyaaradaha,magaalada,muqdish
o',
'madaxweyne,cabdullahi,yusuf,axmed,aun,abtirsada,daaroodmajeerteen,dowladiisii,waxay,muqdisho,dhistay,gole,deegaan,maamulk
a,caasimaddu,noqdo,doorta,leh,maqaan,madaxbanaan,madaxweyne,shariif,shiikh,axmed,hawiyemudulood,abtirsada,isla,markii,xafiis
ka,yimidba,kala,diray,golihii,deegaanka,caasimadda,dhistay,dowladii,horreysay,madaxweyne,yusuf,aun,isaguna,dhisin,madaxweyn
e,xasan,shiikh,maxamuud,isna,hawiyemudulood,abtirsada,cad,horay,dhicin,dibna,dhici,doonin,diiday,muqdisho,loo,sameeyo,gole,d
eegaan,maqaan,gaar,waliba,xubno,matala,yeelato,aqalkii,sare,sameeyay,xilligiisii,isaga,cararaya,kacdoon,siyaasadeed,markaas,
heystay,qabailada,muqdisho,badan,madaxweyne,maxamed,cabdullahi,farmaajo,daaroodmareexaan,sheegay,['shirka', 'shirkii'],fa
dhiyeen,aqal,ummada,soomaaliyeed,caalamkaba,lagama,maarmaan,tahay,caasimaddu,maqaan,matalaad,rasmi,yeelato,isaga,gaar,codsad
ay,waajibnimada,arrinkaas,sokoom,isaga,gaar,loogu,sharfo,arrintaasi,dhacdo,muddo,xileedkiisa,maadaama,yahay,madaxweynaha,kal
a...
```

Figure 5. 6 Unicode Words

LogisticRegression

```
In [368]: from sklearn.linear_model import LogisticRegression
          model = LogisticRegression()

In [369]: model.fit(X_train, y_train)

Out[369]: LogisticRegression()

In [370]: X_train_prediction = model.predict(X_train)

In [371]: training_data_accuracy = accuracy_score(X_train_prediction, y_train)

In [372]: print(f'Accuracy score of training data :{round(training_data_accuracy*100,2)}%')
Accuracy score of training data :96.92%

In [373]: X_test_prediction = model.predict(X_test)
          testing_data_accuracy = accuracy_score(X_test_prediction, y_test)

In [374]: print(f'Accuracy score of testing data :{round(testing_data_accuracy*100,2)}%')
Accuracy score of testing data :92.52%
```

Figure 5.7 Training

Brief Description

Training is the process of being conditioned or taught to do something, or is the process of learning and being conditioned

Predicted label

```
In [383]: # As Logistic is able to provide best results - SVM will be used to check the news liability

def fake_news_det(data):
    input_data = {"news": [data]}
    new_def_test = pd.DataFrame(input_data)
    #new_def_test["news"] = new_def_test["news"].apply(wordopt)
    new_x_test = new_def_test["news"]
    #print(new_x_test)
    vectorized_input_data = cv.transform(new_x_test)
    prediction = model.predict(vectorized_input_data)

    if prediction == 1:
        print("Not a Fake News")
    else:
        print("Fake News")
```

Figure 5.8 Prediction Model

Brief Description

This section shows the prediction of the model level and that you have saved the stop words we described earlier.

```
In [30]: fake_news_det("Guddiga Maamulka doorashada Hirshabelle SEIT ayaa shaaciyay jadwalka doorashada 7 kursi oo ka mid ah kuraasta deeg")

Not a Fake News
Not a Fake News
Not a Fake News
```

Figure 5.9 Result

Brief Description

This is the last part and this is the part that was used to describe the results

Chapter VI: Conclusion and Future Work

6.1 Introduction

This section clarifies the research's conclusion after six months of exploration; it describes key points such as the research's conclusion, the achievement of the research objectives the were previously mentioned in chapter one of the research, guidelines, and future work for those who plan to conduct similar work.

6.2 Conclusion

In our research we develop Machine learning based Fake News in Somalia News, in this digital era the news spreads faster than anything but due to deceptive news many unexpected situations occur. We basically used 3 models' logistic regression, SVM and Naïve Bayes model which gives us 3 different accuracies. But the high accuracy we can gain by logistic regression model. The augmentation of Somali fake news and its extension on social media has become a main anxiety due to its caliber to make demolishing dominance. Different machine learning intercourse have been effort to identify Somali fake news However, for our survey, we developed a machine learning model known as the Naïve Bayes model. Through a SVM we used a binary classifier, more precisely a logistic regression. We have a model that is accurate to 96.75 percent. We also presented an output of confusion matrix to determine the insight of fake and true news, which shows true positives, true negatives, false positives, and false negatives as 288, 33, 343, and 18 respectively. Any user may quickly determine the accuracy of the information and determine whether the information is real or incorrect. This model provides us the highest accuracy score while also being quick. As a result, the logistic regression model is the best model for detecting fake news.

6.3 Discussion

In our research we are developing machine learning for fake news in Somalia, nowadays news spreads faster than anything but fake news there are many unexpected situations. We basically used 3 models 'logistic regression, SVM and Naive Bayes model which gives us 3 different accuracies. But we can gain high accuracy with the logistic regression model when we built three models and then we took the logistic regression which is the one with the highest occurrence score in the training dataset and we collected the fake one from the social media and the real one. Collected by BCC and VOA

We have learned from this study that fake news can be prevented using machine learning

We have learned a lot about how the machine works and the algorithms of the machine

One of the main challenges we faced was the collection of the dataset, which we had a lot of trouble with because there is no library to find Somali news also, we had biggest problems training data the Somali language to the machine because the machine knows only international languages.

Our research is limited to only political news so we suggest in the near future we will build a program to classify all news like sports news Although this is the first time that this research has been done in the Somali language, we suggest that researchers behind us who want to contribute to this research they have to create a library to store fake news because it does not exist in our country. library to collect news

6.4 Recommendation

Based on our experience, we suggest that more research be done on fake news detection.in this study; we have focused on political news to know that the news is true or fake through the use of modern technology. We recommend that to continue from there and that this should be a start point. we would suggest two points to the future researchers:

- I. The machine learning that we have described above it can be updated any time in the algorithms that we have used, maybe A lot of algorithms become more accurate and better in terms of predictions the ones we used before, so that we recommend if its available new algorithms, they should be used by the future researchers.
- II. Although this is the first time that this research has been done in the Somali language, we suggest that researchers behind us who want to contribute to this research they have to create a library to store fake news because it does not exist in our country. library to collect news

6.4 Future Work

We simply used three models, and we got three different accuracy values from them. We can see that the logistic regression model, which is a machine learning model, gives us the best accuracy rate. We only implemented the model that displays the accuracy score; however, in the future, we want to build a website where any user can input specific news from any social media, online news portal, or other website to our website and verify whether the news is true or false, which will be very effective in reducing confusion caused by fake news. Users will be able to determine whether a new article is accurate or false using the website's options. The user can easily enter the website, copy and paste the news, and determine whether it is true or not. This will be the first website because while there is a false news detector for English, there is none for somalia. As a result, this website can quickly determine whether a newspaper article is accurate or false, hence reducing misunderstanding. Our research is limited to only political news so we suggest in the near future we will build a program to classify all news like sports news

References

- Agarwalla, K., Nandan, S., Nair, V. A., & Hema, D. D. (2019). *Fake News Detection using Machine Learning and Natural Language Processing*. 7(6), 4.
- Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining Text Data*. Springer US.
<https://doi.org/10.1007/978-1-4614-3223-4>
- Al Asaad, B., & Erascu, M. (2018). A Tool for Fake News Detection. *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 379–386. <https://doi.org/10.1109/SYNASC.2018.00064>
- Alim Et al. (2021). Detecting Fake News using Machine Learning: A Systematic Literature Review. *Psychology and Education Journal*, 58(1), 1932–1939.
<https://doi.org/10.17762/pae.v58i1.1046>
- Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), 211–236.
<https://doi.org/10.1257/jep.31.2.211>
- Andy Yee. (2017). *Yee-2017-post-truth-politics-and-fake-news-in-asia (2)-converted.pdf*.
See discussions, stats, and author profiles for this publication at:
<https://www.researchgate.net/publication/318673840>
- Ansar, W., & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights*, 1(2), 100052.
<https://doi.org/10.1016/j.jjime.2021.100052>

- Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting Fake News with Machine Learning Method. *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 528–531. <https://doi.org/10.1109/ECTICon.2018.8620051>
- Apuke, O. D. (2017). Social and Traditional Mainstream Media of Communication: Synergy and Variance Perspective. *Online Journal of Communication and Media Technologies*, 7(4). <https://doi.org/10.29333/ojcmt/2614>
- Asur, S., & Huberman, B. A. (2013). Predicting the Future with Social Media. *Applied Energy*, 112, 1536–1543. <https://doi.org/10.1016/j.apenergy.2013.03.027>
- Baade, B. (2018). Fake News and International Law. *European Journal of International Law*, 29(4), 1357–1376. <https://doi.org/10.1093/ejil/chy071>
- Botha, J., & Pieterse, H. (2020). *Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security*. 10.
- Carr, C. T., & Hayes, R. A. (2015). Social Media: Defining, Developing, and Divining. *Atlantic Journal of Communication*, 23(1), 46–65. <https://doi.org/10.1080/15456870.2015.972282>
- Das, K., Behera, R. N., & Tech, B. (2007). *A Survey on Machine Learning: Concept, Algorithms and Applications*. 5(2), 10.
- Das, S., Venugopal, D., & Shiva, S. (2020). A Holistic Approach for Detecting DDoS Attacks by Using Ensemble Unsupervised Machine Learning. In K. Arai, S.

- Kapoor, & R. Bhatia (Eds.), *Advances in Information and Communication* (Vol. 1130, pp. 721–738). Springer International Publishing. https://doi.org/10.1007/978-3-030-39442-4_53
- De Wet, H., & Marivate, V. (2021). Is it Fake? News Disinformation Detection on South African News Websites. *ArXiv:2108.02941 [Cs]*. <http://arxiv.org/abs/2108.02941>
- Debashis “Deb” Aikat. (2009). Traditional and Modern Media. *JOURNALISM AND MASS COMMUNICATION*, 5.
- Figueira, Á., & Oliveira, L. (2017). The current state of fake news: Challenges and opportunities. *Procedia Computer Science*, 121, 817–825. <https://doi.org/10.1016/j.procs.2017.11.106>
- Gahirwal, M. (2018). Fake News Detection. *Nternational Journal of Advance Research*, 35(2), 3.
- Gorbach, J. (2018). Not Your Grandpa’s Hoax: A Comparative History of Fake News. *American Journalism*, 35(2), 236–249. <https://doi.org/10.1080/08821127.2018.1457915>
- Granik, M., & Mesyura, V. (2017). *Fake News Detection Using Naive Bayes Classifier*. 4.
- Hutton, G., & Fosdick, M. (2011). The Globalization of Social Media: Consumer Relationships with Brands Evolve in the Digital Space. *Journal of Advertising Research*, 51(4), 564–570. <https://doi.org/10.2501/JAR-51-4-564-570>
- Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). *Text Classification Using Machine Learning Techniques*. 9.

- Jain, A., Shakya, A., Khatter, H., & Gupta, A. K. (2019). A smart System for Fake News Detection Using Machine Learning. *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 1–4. <https://doi.org/10.1109/ICICT46931.2019.8977659>
- Jones, M. O. (2019). *Propaganda, Fake News, and Fake Trends: The Weaponization of Twitter Bots in the Gulf Crisis*. 27.
- José Luis. (2017). *Activities of the British Ministry of Information in Mexico during the Second World War (1939-1945)*. Conference: Information and its Communication in WartimeAt: University of London, Senate House. Activities of the British Ministry of Information in Mexico during the Second World War (1939-1945)
- K. Dalal, M., & A. Zaveri, M. (2011). Automatic Text Classification: A Technical Review. *International Journal of Computer Applications*, 28(2), 37–40. <https://doi.org/10.5120/3358-4633>
- Kedar, H. E. (2020). Fake News in Media Art: Fake News as a Media Art Practice Vs. Fake News in Politics. *Postdigital Science and Education*, 2(1), 132–146. <https://doi.org/10.1007/s42438-019-00053-y>
- Kim, J., Tabibian, B., Oh, A., Schoelkopf, B., & Gomez-Rodriguez, M. (2017). Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. *ArXiv:1711.09918 [Cs, Stat]*. <http://arxiv.org/abs/1711.09918>
- Kotsiantis, S. B. (2007). *Supervised Machine Learning: A Review of Classification Techniques*. 20.

- Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. *Information*, 10(4), 150. <https://doi.org/10.3390/info10040150>
- Krzysztof R. Apt. (2003). *Principles of constraint programming*. Cambridge University Press.
- Leeder, C. (2019). How college students evaluate and share “fake news” stories. *Library & Information Science Research*, 41(3), 100967. <https://doi.org/10.1016/j.lisr.2019.100967>
- Liakos, K., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, 18(8), 2674. <https://doi.org/10.3390/s18082674>
- Liu, Z., Lv, X., Liu, K., & Shi, S. (2010). Study on SVM Compared with the other Text Classification Methods. *2010 Second International Workshop on Education Technology and Computer Science*, 219–222. <https://doi.org/10.1109/ETCS.2010.248>
- Lorent, S., & Itoo, A. (2018). *Fake News Detection Using Machine Learning*. 91.
- Manzoor, S. I., Singla, J., & Nikita. (2019). Fake News Detection Using Machine Learning approaches: A systematic Review. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 230–234. <https://doi.org/10.1109/ICOEI.2019.8862770>

- Mason, L. E., Krutka, D., & Stoddard, J. (2018). Media Literacy, Democracy, and the Challenge of Fake News. *Journal of Media Literacy Education*, 10(2), 1–10. <https://doi.org/10.23860/JMLE-2018-10-2-1>
- Ng, L. H. X., & Taeihagh, A. (2021). How does fake news spread? Understanding pathways of disinformation spread through APIs. *Policy & Internet*, poi3.268. <https://doi.org/10.1002/poi3.268>
- Nyow, N. X., & Chua, H. N. (2019). Detecting Fake News with Tweets' Properties. *2019 IEEE Conference on Application, Information and Network Security (AINS)*, 24–29. <https://doi.org/10.1109/AINS47559.2019.8968706>
- O'Brien, N. (2018). *Machine Learning for Detection of Fake News*. 56.
- Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic Detection of Fake News. *ArXiv:1708.07104 [Cs]*. <http://arxiv.org/abs/1708.07104>
- Posetti, J., & Matthews, A. (2018). *A short guide to the history of 'fake news' and disinformation*. 20.
- Ray, S. (2019). *A Quick Review of Machine Learning Algorithms*. 5.
- Revez, J., & Corujo, L. (2021). Librarians against fake news: A systematic literature review of library practices (Jan. 2018–Sept. 2020). *The Journal of Academic Librarianship*, 47(2), 102304. <https://doi.org/10.1016/j.acalib.2020.102304>
- Ron Chernow. (2017). *How Social Media.pdf*. Grant By Ron Chernow NY: Penguin Press, 2017, 1,104 pages Reviewed by Arthur I. Cyr

- Sarkar, S. D., Yang, F., & Mukherjee, A. (2018). *Attending Sentences to detect Satirical Fake News*. 10.
- Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, 34(4), 301–312.
<https://doi.org/10.1016/j.tig.2017.12.005>
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1), 4787. <https://doi.org/10.1038/s41467-018-06930-7>
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ArXiv:1708.01967 [Cs]*.
<http://arxiv.org/abs/1708.01967>
- Sirajudeen, S. M., Azmi, N. F. A., & Abubakar, A. I. (2005). ONLINE FAKE NEWS DETECTION ALGORITHM. . . Vol., 9.
- Stahl, K. (2018). *Fake news detection in social media*B.S. Candidate, Department of Mathematics and Department of Computer Sciences, California State University Stanislaus, 1 University Circle, Turlock,. 6.
- Tao Liu, Zheng Chen, Benyu Zhang, Wei-ying Ma, & Gongyi Wu. (2004). Improving Text Classification using Local Latent Semantic Indexing. *Fourth IEEE International Conference on Data Mining (ICDM'04)*, 162–169.
<https://doi.org/10.1109/ICDM.2004.10096>

- Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. A., Elkhatab, Y., Hussain, A., & Al-Fuqaha, A. (2019). Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. *IEEE Access*, 7, 65579–65615. <https://doi.org/10.1109/ACCESS.2019.2916648>
- Vijjali, R., Potluri, P., Kumar, S., & Teki, S. (2020). Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking. *ArXiv:2011.13253 [Cs]*. <http://arxiv.org/abs/2011.13253>
- Vo, N., & Lee, K. (2020). Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. *ArXiv:2010.03159 [Cs]*. <http://arxiv.org/abs/2010.03159>
- Wang, H., Ma, C., & Zhou, L. (2009). A Brief Review of Machine Learning and Its Application. *2009 International Conference on Information Engineering and Computer Science*, 1–4. <https://doi.org/10.1109/ICIECS.2009.5362936>
- Youngkyung Seo, Deokjin Seo, Chang-Sung Jeong. (2018). *FaNDeR Fake News Detection Model Using.pdf*.
- Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44–53. <https://doi.org/10.1093/nsr/nwx106>
- Zoubin Ghahramani†. (2004). *Machine learning.pdf*.

Appendix A: import

```
import pandas as pd
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
from sklearn.feature_extraction.text import CountVectorizer
```

```
from sklearn.feature_extraction.text import TfidfTransformer
```

```
from sklearn import feature_extraction, linear_model, model_selection, preprocessing
```

```
from sklearn.metrics import accuracy_score
```

```
from sklearn.model_selection import train_test_split
```

```
from sklearn.pipeline import Pipeline
```

```
import re
```

```
import nltk
```

Appendix B: Shuffle the data

```
from sklearn.utils import shuffle
```

```
data = shuffle(data)
```

```
data = data.reset_index(drop=True)
```

Appendix C: Removing stop words

```
stop_words_to_lower = []
```

```
stopwords = [
```

```
"Waxaan","wuxuu","iyo","ku","oo","aad","aan","een","ee","soo","ka","uu","ay","ey","mar  
ka","waxaa","waxa","wax","in","ah","ayo","mr","u","isu","iyo","waa","ayaa","mid","isku",  
"taasi","la","Muxuu","maxay","inta","uun","uma","sidi","ugu","mar","kasoo","si","hor","  
ma","balse","e","waxayna","inuu","sii","is","miyuu","U","inay","ayuu","ke","jira","jirtey",  
"kale","lagu","laga","kaliya","jeer","looga","qaab","cusub","labada","ayey","ayay","sida",  
waayey","mida","jirta","xa","doonaa","dona"]
```

```
for element in range(len(stopwords)):
```

```
    stop_words_to_lower.append(stopwords[element].lower())
```

Appendix D: corpus message

```
for i in range(0, len(data)):
```

```
    message = re.sub('[^a-zA-Z]', ' ', str(data['news'][i]))
```

```
    message = message.lower()
```

```
    message = message.split()
```

```
    message = [word for word in message if not word in stop_words_to_lower]
```

```
    message = ' '.join(message)
```

```
corpus_message.append(message)
```

Appendix E: Most frequent words counter

```
from nltk import tokenize
```

```
token_space = tokenize.WhitespaceTokenizer()
```

```
def counter(text, column_text, quantity):
```

```
    all_words = ''.join([text for text in text[column_text]])
```

```
    token_phrase = token_space.tokenize(all_words)
```

```
    frequency = nltk.FreqDist(token_phrase)
```

```
    df_frequency = pd.DataFrame({"Word": list(frequency.keys()),
```

```
                                "Frequency": list(frequency.values())})
```

```
    df_frequency = df_frequency.nlargest(columns = "Frequency", n = quantity)
```

```
    plt.figure(figsize=(12,8))
```

```
    ax = sns.barplot(data = df_frequency, x = "Word", y = "Frequency", color = 'blue')
```

```
    ax.set(ylabel = "Count")
```

```
    plt.xticks(rotation='vertical')
```

```
plt.show()
```

```
data['news']=corpus_message
```

Appendix F: confusion matrix

```
from sklearn import metrics
```

```
import itertools
```

```
def plot_confusion_matrix(cm, classes,
```

```
                           normalize=False,
```

```
                           title='Confusion matrix',
```

```
                           cmap=plt.cm.Blues):
```

```
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
```

```
    plt.title(title)
```

```
    plt.colorbar()
```

```
    tick_marks = np.arange(len(classes))
```

```
    plt.xticks(tick_marks, classes, rotation=45)
```

```
    plt.yticks(tick_marks, classes)
```



```

if normalize:

    cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]

    print("Normalized confusion matrix")

else:

    print('Confusion matrix, without normalization')


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="center",

             color="white" if cm[i, j] > thresh else "black")

plt.tight_layout()

plt.ylabel('True label')

plt.xlabel('Predicted label')

```

Appendix G: model results

```

def fake_news_det(data):

    input_data = {"news":[data]}

    new_def_test = pd.DataFrame(input_data)

    #new_def_test["news"] = new_def_test["news"].apply(wordopt)

    new_x_test = new_def_test["news"]

    #print(new_x_test)

    vectorized_input_data = cv.transform(new_x_test)

    prediction = model.predict(vectorized_input_data)

    if prediction == 1:

        print("Not a Fake News")

    else:

        print("Fake News")

```

Appendix H: unique Words

```
data['news'].unique().tolist()
```

Appendix I: results

```
def fake_news_det(data):
```

```

input_data = {"news":[data]}

new_def_test = pd.DataFrame(input_data)

#new_def_test["news"] = new_def_test["news"].apply(wordopt)

new_x_test = new_def_test["news"]

#print(new_x_test)

vectorized_input_data = cv.transform(new_x_test)

prediction = model.predict(vectorized_input_data)

if prediction == 1:

    print("Not a Fake News")

else:

    print("Fake News")

```

Appendix I: Front-end

```

from flask import Flask, escape, request, render_template

import pickle

from sklearn.metrics import accuracy_score

```

```
# TfidfVectorizer

vector = pickle.load(open("vectorizer.pkl", 'rb'))


# saved model LogisticRegression

model = pickle.load(open("finalized_model.pkl", 'rb'))


app = Flask(__name__)


@app.route('/')

def home():

    return render_template("index.html")


@app.route('/prediction', methods=['GET', 'POST'])

def prediction():

    if request.method == "POST":

        news = str(request.form['news'])

        print(news)
```

```

predict = model.predict(vector.transform([news]))[0]

print(predict)


if predict==1:

    predict="warkan wa war sax ah"

else:

    predict="warkan wa war been abuur ah"


    return render_template("prediction.html", prediction_text="News headline is ->
{ }".format(predict))


else:

    return render_template("prediction.html")


if __name__ == '__main__':

    app.debug = True

```

```
app.run()
```