A PROPOSED MACHINE-LEARNING AIDED FRAMEWORK FOR MISINFORMATION MANAGEMENT IN UGANDA USING WORD2VECTOR AND LONG SHORT-TERM MEMORY

CASE STUDY OF UGANDA COMMUNICATIONS COMMISSION

BY

TUMWEBAZE WILSON

(MIS) 2021-01-03187

REPORT SUBMITED IN PARTIAL FULFILMENT FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE IN INFORMATION SYSTEMS

TO

SCHOOL OF MATHEMATICS AND COMPUTING (SOMAC)

KAMPALA INTERNATIONAL UNIVERSITY

NOVEMBER 2023

DECLARATIONS

I **TUMWEBAZSE WILSON,** declare that this report is my original work and has never been submitted to any university, college or school for the award of a degree or diploma.

Signature……date…….

APPROVAL

As an academic advisor, I have been supervising this project, which is now being presented to the school of mathematics and computing of Kampala international university

Signature……. date…….

Assoc. prof. Kareyo Margaret

Signature……. date…….

Prof. Elly Amani Gamukama

DEDICATIONS

I dedicate this work to God Almighty, to my beloved parents who supported me from the start up to this stage. Also, to the family members and friends who supported me in one way or the other through the journey of life and academic pursuit.

ACKNOWLEDGEMENT

I use this medium to acknowledge my supervisors, Assoc. prof. Kareyo Margaret who is also the Assistant Deputy Vice Chancellor and prof. Elly Gamukama Dean of SOMAC. Am so grateful for the Time and commitment upon completion of this research thesis. To Associate Dean for Research SOMAC Dr. Businge Phelix who helped me in so many ways from the generation of the Idea and sharpening it, the Head of Department Computer Science Madam Akiteng Immaculate, Head of Department Information Technology Mr. Asiimwe John Patrick and the entire SOMAC community, I say thank you. Not forgetting Mr. Niyibizi Kenneth who is the middle player, Thank you. Mr. Migadde Elias and Mr. Tumusiime Joshua thank you for the support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Information quality is becoming an increasingly important issue as social networks have become the primary source for misinformation dissemination. Because of their ease of use, spreading behaviour, and low cost, social network platforms are leveraging news consumption. Many studies have been developed on methods to improve fake news classification, particularly on misinformation detection on social media, with promising results in recent years. In Uganda, information filtering was still done in a traditional way hence not efficient and effective. To address this challenge, this study aims at providing a framework that will manage misinformation using the machine learning algorithms like word2vec and LSTM. The study found that the most common type of misinformation is fake news and hate speech that is spread by Ordinally users and politician as the main agents that spread misinformation via social media platforms like TikTok and twitter respectively. I also extracted data from Twitter using Apache NIFI, elastic search and Kibana for data visualization. The dataset information included, tweets, the user details, retweets, tweet URL and date. Subsequently, I performed an offline analysis through the use of machine learning and deep learning techniques. I hope that this research work will provide useful insights for realizing ever more effective tools to counter misinformation and those who spread it intentionally.

CHAPTER 1

1.0 Introduction

In the past years, social media and other news sources that permit any client to deliver unconfirmed online substances have picked up notoriety. Social systems can breakdown physical obstructions by interfacing geologically scattered individuals, facilitating political and financial constraints. (Danielle, Caled1 & Silva1, 2021).

Since the advent of the Internet, people's access to news has radically changed. People used to rely on common media like radio and TV, which included less and fewer trusted news sources. People are becoming more familiar with online information sources, such as networking websites that let anyone to share anything without the need for "factchecking or publication judgements" (Alcott & Gentzkow, 2017). Many people are afraid that websites might publish false information while passing it off as real news. Due to the increasing prevalence of devices that can connect to the internet on a global scale and improved reliable internet speeds, many individuals are joining social media. Without a question, the majority of people use Facebook, and many of us obtain our news from social media. (Stephen, 2016; Erdoğmuş & Cicek, 2012).

This has made it much easier for false information to be created and spread, including rumors, spam, and fake news. Due to the openness of social network platforms and the ability for automation, false information can spread quickly to a broad population, posing a number of unprecedented issues.

1.1 Background of the study

1.1.1Historical background.

FILTERING

The Communication Decency Act (CDA) and later the Child Online Protection Act (COPA), both of which were found to be in violation of the first amendment of the United States of America, brought the problem of filtering to light in the mid-1990s (Hunter, 2000; Rosenberg, 2001). China began restricting internet access through the Temporary Regulation for the Management of Computer Information Network International Connection. Conceived in 1993, made public in 1996, and confirmed in 1997. The publication and amplification processes take place on a site like Facebook, and as the first of a number of internet-controlling policies adopted by China, these platforms have significant influence over this process. Australia has been using internet filtering for more than ten years (Conroy, 2006; NAIRN, 2007). On the basis of the principle of child protection, a number of systems have

been predicted, in part or in full, ranging from NetAlert (Coonan, 2007), Internet Service Providers (ISPs) are required by the present content filtering agreement to restrict the content on the Australian Communications Media Authority's (ACMA) website by the mandatory content filter proposed by Conroy (Conroy, 2012). The Turkish government has taken steps to censor Turkey's media landscape, including shutting down news organizations, imprisoning journalists, and blocking objectionable online information. Freedom House discovered that over 3300 news-related URLs were blacklisted in 2018. In regard to the Turkish government restoring access to Wikipedia, the Wikimedia Foundation, which owns and runs Wikipedia, filed a petition with the European Court of Human Rights (ECHR) in 2019. In the end, the constitutional court determined that prohibiting Wikipedia was unconstitutional.

A news release on the ongoing pattern of internet and social media shutdowns across Africa was released in 2019 by the Special Reporter on Freedom of Expression and Access to Information in Africa. Although this press release was more focused on internet outages, it served as a helpful reminder that social media and the internet have given Africans a voice, allowing them to discuss social, economic, and political issues at a greater depth than ever before. The government should not suppress this voice. There were increased tensions in the run-up to the 2020 presidential elections in Togo, with demonstrations against the authority of Gnassingbe Eyadema. Civil Society Organizations (CSOs) expressed their worries that the Togolese administration will impose internet access restrictions during the elections in a joint letter to the Togolese government. On Election Day, social media services like Facebook, WhatsApp, and Telegram were prohibited. Additionally, Ethiopia has historically been viewed as a troublesome state due to its usage of blocking and filtering. Numerous websites were blocked between 2012 and 2018. According to media outlets like the Electronic Frontier Foundation. The raised worry in 2018 about the rising trend of East African states adopting onerous rules addressing the internet and online platforms.

In 2013, when two media outlets faced temporary closure as a result of the publication of articles that implied Museveni was already preparing his son to take over as president, Uganda already had restrictive procedures in place to control media information. (Human Rights watch 2014). Yoweri Museveni, the president of Uganda, was sworn in for a fifth term during the presidential elections in May 2016 in an environment of social media suppression. Journalists and non-profit digital watchdogs in the nation claimed that Twitter, Facebook, and

WhatsApp were banned as of May 11, 2016. (Propa 2016). Similar circumstances occurred in February 2016, when officials for three days during the presidential election barred all access to social media and mobile money transfer, citing security concerns and a threat to public order and safety as justification. (Butagari, 2016). On the eve of the presidential elections held on January 14, 2021, Uganda also barred access to the internet and social media. On January 18, internet access was once again available, but social media is still unavailable. Virtual Private Networks have made social media accessible to millions of Ugandans. (Anguyo, 2021)

MISINFORMATION.

From a political smear campaign against the Roman general Mark Antony regarding his association with Cleopatra, which included slogans etched on coins, until 1439 AD, concerted deception efforts have been documented throughout recorded history. The printing press was created in 1439, allowing liars to disseminate lies even further. And between 1960, psychological study improved our understanding of belief, revealing, for instance, how individuals assess the reliability of a source and the kinds of messages that are most likely to persuade. Researchers also noticed that views persisted even when inaccurate information was corrected, so they started to experiment with persuasion-resisting strategies. Schwarz outlined five factors that people use when determining the veracity of information. The information's consistency with other known facts, the reliability of the sources, the opinion of others, its internal consistency, and the existence of supporting data are all factors. Facebook and Twitter, which were introduced in 2004 and 2006, respectively, made it possible for information to spread even more quickly and effectively. Online social networks satisfy several of the requirements identified by psychologists as constituting convincing arguments. For instance, friends and family members may praise and share messages that promote unverified claims. Norbert Schwarz, PhD, a psychologist who studies misinformation said Social media are practically built for spreading fake news. Petitioners, academics, and policymakers paid more attention to the topic of misinformation during the 2016 US elections (Allcott & Gentzkow, 2017; Silverman, 2016). Since then, misinformation has become more prevalent with the aim of undermining public institutions, escalating societal unrest, and interfering with political processes in impacted nations (Tan &Ang, 2017). Governments throughout the world have therefore been debating and advancing legislative measures to counter the damage news poses to social media and political stability. (Haciyakupoglu, HUI,

Suguna, Leong, &Abdul Rahman, 2018). In addition, growing pressure on digital giants like Google, Facebook, and Twitter to monitor misinformation spreaders has prompted them to pioneer efforts to solve this issue ((Drozdiak, 2018; Foo, 2018). Psychologists have increased their efforts to address false information from 2018 to the present, drawing on years of laboratory and field research on dispelling rumors. Debunking, preventive immunization, and prodding to evaluate the veracity of information are important strategies.

Because many colonial and post-colonial African states controlled the media and have been purveyors of propaganda and misinformation, Africans have lived with "non-truths" for decades (Mutsvairo and Bebawi 2019). False and misleading information has caused or assisted in a variety of harms to people and organizations across Africa in only the last few years, including vigilante violence and civil upheaval in Ethiopia. (Nur, 2019) and Nigeria (Adegoke, 2018) through the use of wrong medical treatments for Ebola (Ogala &Ibeh 2014).

Certain worry that the proliferation of regulations may prompt some governments to restrict free speech and the freedom of the press (Budoo-Schotz 2020; Selnes 2021). For instance, during recent elections in Tanzania and Uganda, the government blocked access to social media and the Internet under the pretext of preventing the spread of false information.

1.1.2 Conceptual Perspective

MISNFORMATION.

Misinformation is a phrase that is used frequently. Despite the concepts being somewhat simpler to distinguish, such as spam (a huge number of receivers), rumor (confirmed or unconfirmed), and fake news, disinformation is the most similar or confusing term (Campan, 2017). Whether the material was intentionally created to deceive is the key distinction between misinformation and disinformation. Misinformation frequently refers to unintended events, whereas disinformation describes situations when the information was purposefully designed to deceive. Throughout our discourse, we will use the word "misinformation" to describe any false or inaccurate information that is circulated on social media. For instance, misleading information presented as news is known as fake news. Rumor is unconfirmed information that may be true or false, spam is irrelevant information that is distributed to a large number of users, and erroneous information is not necessarily disinformation because it may be accidentally shared by innocent users. Accurate information that is intentionally

misleading is referred to as disinformation and is often distinguished from misinformation. Misinformation is when someone disseminates false information without realizing it, typically because their friends or others do so. (Campan, 2017). The social media system is composed of an algorithm that suggests specific news or information to a user based on the social media group to which he or she belongs, their past behaviour, and their network of friends. Because of this, when one friend views something, another friend is suggested to also view it, and the user is notified of this recommendation. The echo chamber effect, a phenomenon that greatly adds to this, is one such phenomenon.

Even if users are dubious about the veracity of the material, this recommendation system encourages them to share it. People who hold the same opinions or support the same political party will disseminate and divulge material without having it verified in order to further their political objectives. Mmisinformation. Relates to the intentional spread of misleading information to trick others. Depending on the intent, misinformation might occasionally turn into disinformation, but disinformation is always misinformation. (Ecker, 2017; Erku, 2021). It is disinformation, when a political figure is mentioned in an article with a factual error (Nikolov, 2020).

ICT FILTER (IV)

A filter is a program or portion of code used in computer programming that checks each input or output request against certain qualifications before processing or forwarding it. It is "pass-through" code that receives input data, processes it, maybe transforms it, and then sends it along to another program in a pipeline. A filter often doesn't perform any input/output operations on its own. Sometimes, filters are used to add or remove headers or control characters from data.

1.2 Statement of the problem

The openness and timeliness of social media have largely facilitated the creation and dissemination of misinformation such as spam, rumors, and fake news, disinformation among others which has affects the society by misleading the individuals, causing a lot of emotional pain and insecurity. Misinformation on social media has caused widespread alarm in recent years (Flynn et al., 2017; Lazer et al., 2018). Due to increased number of internet users and most especially social networks, the information is filtered or even denied access to most users by the use of ICT filters. Circumstances occurred in February 2016, when officials for three days during the presidential election barred all access to social media and mobile money

transfer, citing security concerns and a threat to public order and safety as justification. (Butagari, 2016). Additionally, during the general elections in 2021 social media platforms were shut down for a number of days and many organizations and individuals who do businesses online were affected even when they are not victims of misinformation (Innocent Anguyo, 2021). And this therefore calls for a framework to misinformation manage misinformation in Uganda.

1.3 Objectives of the study

1.3.1 General objective

The main objective is to design a framework for application of ICT filters in information management that can be adopted to combat misinformation in Uganda without affecting and interrupting the individual business.

1.3.2 Specific objectives

i. To establish the sources and types of misinformation.
ii. To identify the various ICT filters used in information management
iii. To design a framework for misinformation management.
iv. To test and validate a proposed framework for misinformation management.

1.4 Research questions

i. What are the sources and types of misinformation in Uganda?
ii. What are various ICT filters used in misinformation management?
iii. How will a framework for misinformation management be designed?
iv. How will a proposed framework for misinformation management be tested and validated?

1.5 Scope of the study

The study took 13 months starting from June 2022 to July 2023. This study focused on designing a misinformation management framework that can be used to manage misinformation in the society over social medial platforms and it didn't intend to develop a filtering system.

1.6 Significance of the study

The study will enable the researcher to acquire knowledge and skill of research.

The study will enable the government Uganda to apply the most effective and optimum ICT filters to manage misinformation for security purposes in the society.

The research will help the general public to always carry out their businesses online smoothly even when the filters are applied.

The study will help organizations help to increase security without sacrificing productivity by preventing employees from accessing unlawful or unproductive websites.

The study will assist educational institutions in securing themselves while blocking offensive and disruptive content.

CHAPTER 2: LITERATURE REVIEW

2.0 Introduction

This chapter describes the theoretical review, theoretical framework of the study, the different types of misinformation, the sources of misinformation, approaches to misinformation, information filtering techniques, tasks involved in misinformation detection, the related studies and the identified gaps.

2.1 Theoretical Review

2.1.1 Designing a conceptual framework for misinformation on social media (Peivand, Seyyed & Mohammad, 2021)

Most studies just consider the reach and content of fake news, omitting any potential impacts. Fake news, which we define as a type of disinformation, most likely results in perceptions with implications for politics (Lazer et al., 2017). Misperceptions are easily created, frequently upon initial exposure (Cook, Ecker, & Lewandowsky, 2015). Citizens may be less critical of content if they believe deception is a product of journalistic practice.

Additionally, research has demonstrated that political misinformation and deception continue to influence sentiments even after they have been addressed (Thorson, 2016). This could be explained by the fact that inaccurate information remains retrievable in memory and is thus still available when people attempt to explain how events are happening (Cook et al., 2015). In this way, even false news reports that fact-checkers have refuted may lead to political misunderstandings.

Also, it has been asserted that fake news plays on people's emotions (Bakir & McStay, 2018). This idea needs to be quantitatively and empirically tested first. Then, investigations might examine whether the impacts of fake news on misperceptions are enabled by cognitive and affective processing, in line with the study concentrating on the psychological processes that explain news effects.

Process of generating misinformation. The primary means of spreading false information were social media. The main subjects or typology of the incorrect information includes: disease data, remedies, preventative and safety measures, food advice, and virus transmission techniques. False claims, conspiracy theories, and pseudoscientific medical treatments are three examples of typical sorts of misinformation about COVID-19 that Murphy has identified (Murphy, 2020). It appears that during an epidemic, people have a great need to find both preventative and curative measures, as well as information on the likelihood and severity of the sickness (Gesser-Edelsburg, Diamant & Mesch, 2018).

Effects: According to specialists, the effects of spreading false information about the condition include social-mental effects, incorrect health effects, effects on the healthcare system, as well as economic and ethical effects. According to research by Lee et al., exposure to misinformation is linked to psychological distress, including symptoms of anxiety, depression, and posttraumatic stress disorder.

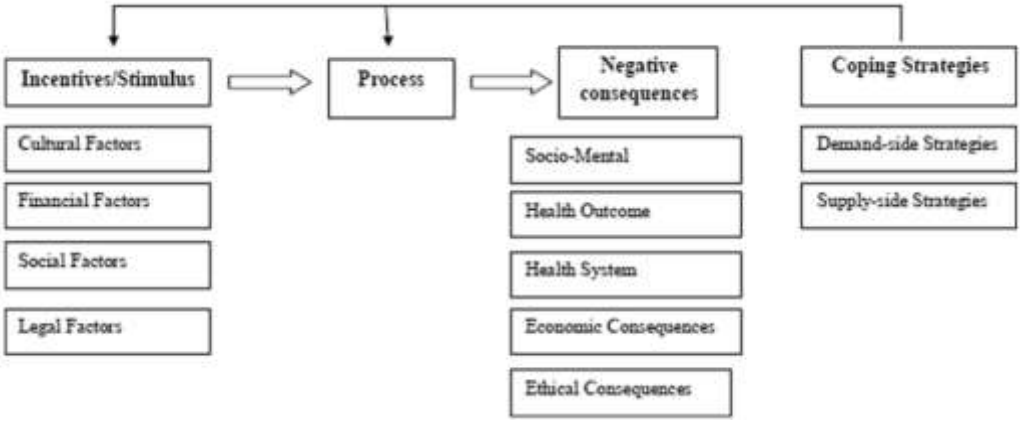


Figure 2.1. The thematic framework

2.1.2 A DATA-DRIVEN ANALYSIS OF THE ROLE OF INFLUENCERS IN THE SPREAD OF MISINFORMATION AND DISINFORMATION ON SELECTED SOCIAL MEDIA PLATFORMS

The theoretical framework that was suggested was based on social context. Based on the context in which the fake news was made and circulated, the trajectory and dissemination of false information on a social media platform by identified influencers at a certain time are explored. Either fake news is produced targeted towards a certain demographic on social media platforms, or shared by influencers, with the express purpose of changing attitudes and influencing members of the target demographic, group, or community politically and/or economically. The framework also takes into account the potential to influence public opinion against certain people by disparaging their reputations and seducing the public with false or misleading information about them. (Olatunji 2019)

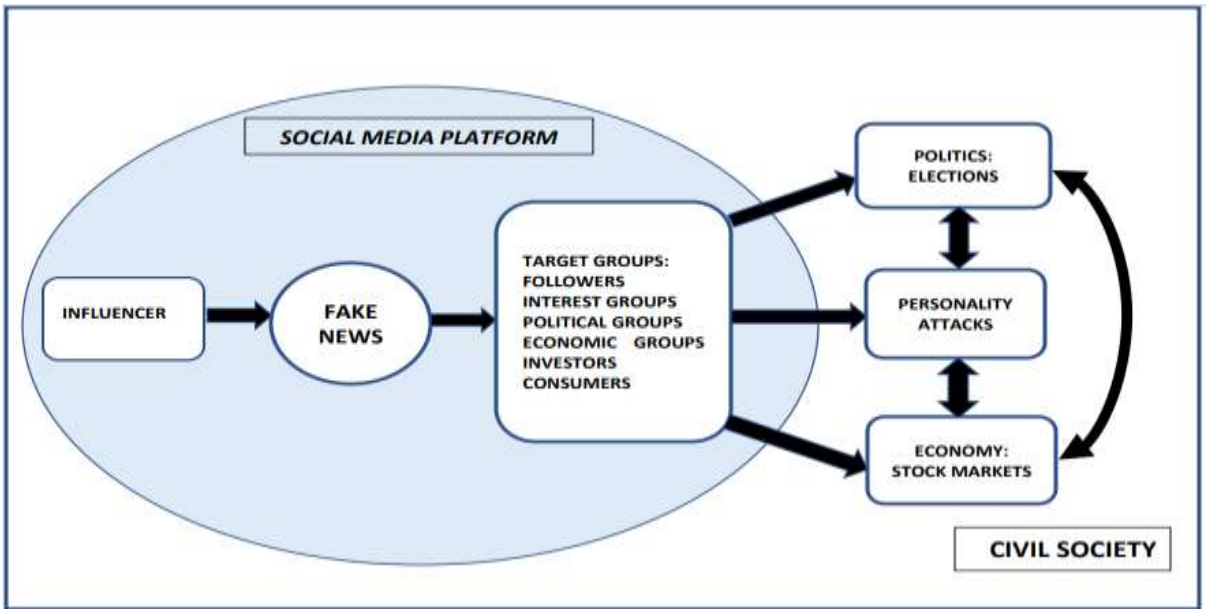


Figure 2.2. A data-driven analysis

2.3 Theoretical Framework

The theoretical framework **(figure 3)** was obtained after a review different theoretical reviews in 2.1 and adopted *figure 2*. It clearly categories different forms of misinformation and how true information can be stressed to form false information, distinction between misinformation and disinformation. The source/sender starts with the intent utilizing the weaknesses (motivators) that are available to disseminate the information. The information is regarded as misinformation or disinformation depending on the intent that is to say intentional or unintentional and then it will cause an impact to the public.

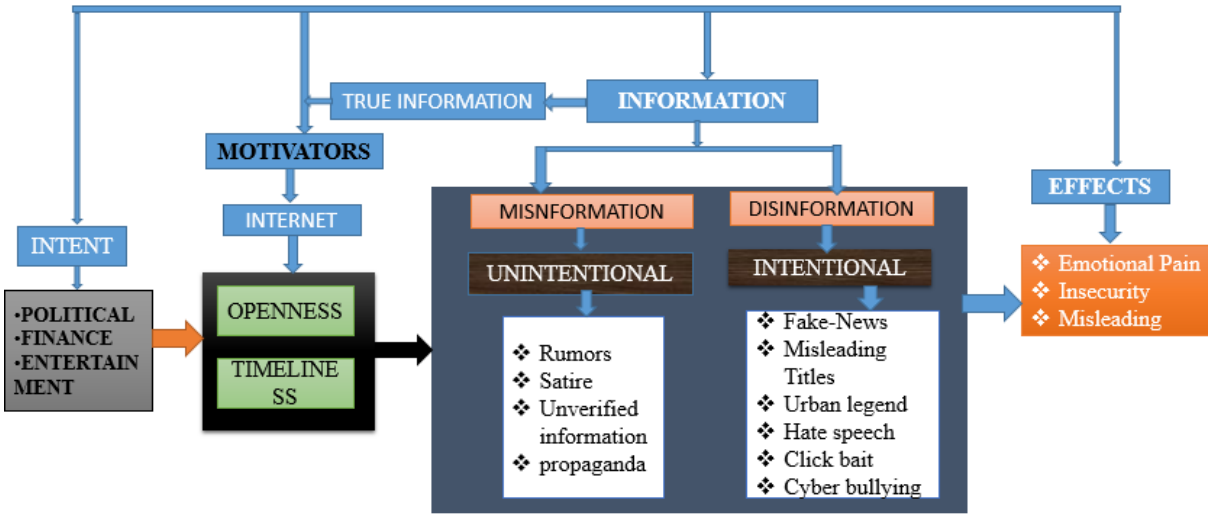


Figure 2.3. Theoretical framework

2.4 Types of Misinformation on Social Media.

To get a clear view of misinformation on social media, I reviewed different kind of misinformation.

Unintentionally spread misinformation: Some false information unintentionally betrays its intended audience. Because they rely on trusted information sources like their friends, work mates, or powerful social media users, even common and innocuous users may participate to the transmission. Instead of waiting to deceive their social network connections, they frequently strive to inform them about a specific issue or situation. One such instance is the pervasive misperception over Ebola.. (Liang, Fred, Kathleen, & Huan, 2016)

Intentionally spread misinformation: Some of it is disseminated with the goal to mislead its audience, which has recently sparked a lot of discussion about false information and fake news. Behind the popularity, there are frequently authors and organized groups with a specific intent working toward the same end. Conspiracies, rumors, and fake news that were popular during the 2016 presidential election are typical examples of intentionally disseminated false information. For instance, a fake news author named Paul Horner has taken credit for a number of stories that went viral in 2017 (Liang, Fred, Kathleen, & Huan, 2016)

Urban legends: Urban legends are purposely made-up stories about actual local events that convey false information. There are times when entertaining is the goal. (Liang, Fred, Kathleen, & Huan, 2016)

Fake news: Misinformation that is presented as news but is actually fake news. Recent events show that false information may be spread via social media and news media as propaganda. (Liang, Fred, Kathleen, & Huan, 2016)

Unverified information: Unverified information may occasionally be accurate and true. Information is regarded as unverified before it is verified, and those that are later discovered to be inaccurate are unquestionably regarded as disinformation. (Liang, Fred, Kathleen, & Huan, 2016)

Rumor: Can be True information but not yet verified. For instance, The alleged avian influenza infection that killed many ducks in Guangxi, China, is an example of a real rumor.

Before the authorities confirmed it to be genuine, it had been a true rumor. Another alleged case of avian influenza that was later proven to be untrue involved people who consumed well-cooked chicken. (Liang, Fred, Kathleen, & Huan, 2016)

Crowdturfing: A notion derived from astroturfing called "crowdturfing" refers to a campaign that hides its backers and sponsors to give the impression that it was started by ordinary citizens. Astroturfing that is crowdsourced is known as "crowdturfing," in which supporters recruit their ostensibly grassroots members online. The information advocated by crowdsourcing employees may be real, just as unconfirmed information or rumors, but their popularity has been artificially and unfairly boosted. Crowdturfing is responsible for several instances of false information that have detrimental impacts. Crowdturfing workers can be quickly employed on a number of internet marketplaces, including zhubajie, sandaha, and fiver. There have been allegations that specific politicians have been the targets of crowdsourcing. (Liang, Fred, Kathleen, & Huan, 2016)

Spam: This is unwanted communication that overburdens its recipients. It can be found on a number of channels, including social media, email, and instant messaging. (Liang, Fred, Kathleen, & Huan, 2016)

Hate speech: The term "hate speech" describes threatening and abusive social media posts that specifically target particular racial or ethnic groups. Hate speech against particular protected groups and the 2016 presidential election were found to interact dynamically, with the peak of hate speech happening on Election Day. (Liang, Fred, Kathleen, & Huan, 2016)

Cyberbullying: Cyberbullying is a sort of bullying that takes place online, most often on social media, and can involve any kind of misleading information, like rumors and hate speech. (Liang, Fred, Kathleen, & Huan, 2016)

Clickbait. This is a sensationalized piece of writing or headlines designed to grab readers' attention by appealing to their feelings (often anger or curiosity). As the name implies, clickbait's objective is to engage readers in order to generate ad revenue. Instead of spreading them thinly across numerous pages to maximize the amount of advertisements that may be provided to each user, it is typically devoid of facts or other helpful information. Clickbait may also spread incorrect information, if that wasn't bad enough before. Because this post

was poorly researched and written, readers may become outraged and share it with their social networks, which will spread false information to more people. (Lipschultz 2020).

***Misleading titles**:* Sensationalized writing or headlines like this one are used to draw readers in by appealing to their emotions (often anger or curiosity). As the term suggests, clickbait's goal is to interest readers in order to bring in money through advertisements. It is often empty of facts or other useful information, instead of distributing them sparsely across multiple pages to maximize the quantity of adverts that may be offered to each user. If clickbait wasn't terrible enough already, it may also propagate false information. Because this post was badly written and researched, readers may be offended and share it on social media, which will cause more people to become aware of the false information. (Lipschultz, 2020).

Satire: The humor of news satire comes from its deadpan, sardonic tone and imitation of real news sources. Satire like this runs the risk of being misinterpreted as reality because not every reader will get the irony. Despite not being intended to be manipulative, satire could, if interpreted incorrectly, have the same effect as false news. (Lipschultz, 2020).

2.5 Sources of Misinformation.

Although social media users' propagation of erroneous information has been the main target of much discussion about disinformation since 2016, research reveal that the sources of misinformation, how we find them, and the subjects they affect are all more diverse than is typically recognized (Cunliffe-Jones, 2020; Newman et al, 2020). This series' study on misinformation in Africa identifies a variety of sources for misinformation, including traditional media, politicians, public organizations, business executives, traditional and religious leaders, special interest groups, offline community networks, and everyday social media users (Cunliffe-Jones, 2022b). According to the study, false information spreads through a variety of various routes, including traditional media, social media platforms, messaging apps, speeches given in public forums or at governmental functions, traditional community networks, and drug labels. It also demonstrates how disinformation covers a wide range of subjects, including accidents and natural catastrophes, crime and justice, health, politics, the economy, and media (Cunliffe-Jones, 2022). Again, fact-checking organizations believe that it is crucial to understand who promotes disinformation, where it can be obtained, and the subjects it addresses in order to accurately identify misinformation as such (Cunliffe-Jones, 2022).

2.6 Tasks Involved In Misinformation Detection

There are certain related activities in mmisinformations detection that can speed up the identification process and somewhat enhance performance.

Stance detection*:* This involves determining whether an article refutes a particular claim. It tries to evaluate the consistency between an article and a claim rather than the truthfulness of information, which can aid in searching for evidence inside an article and extracting credibility features for the detection of disinformation (Ferreira & Vlachos, 2016). Recent studies have shown that misinformation tends to stir up debates more than facts do (Mendoza and Poblete, 2010), therefore there may be many blatantly opposed reactions to misinformation as it spreads (Dungs & Aker, 2018). As a result, the stance detection of responses can act as a supplementary credibility feature for the disinformation detection.

Abstractive summarization is a pertinent activity that can aid in the misinformation detection process. In particular, the summarization model can be used as a feature extractor before misinformation detection to find the main points of the input texts. To detect misinformation, for instance, Esmaeilzadeh used a text summarization model to first summarize an article before feeding the summarized sequences into an (Recurrent Neural Network) RNN-based neural network. Finally, higher performance is shown when the experimental results are compared to the task utilizing only the original texts.

Fact checking is the process of determining whether assertions made by public people, such as politicians, are true (Vlachos & Riedel, 2014). Since both processes strive to evaluate the veracity of statements, Fact checking and the identification of misinformation are frequently difficult to distinguish from one another. However, truth verification is more thorough, whereas disinformation detection typically concentrates on particular details. Fact checking can also be an essential component of the misinformation detection process when a piece of information contains claims that need to be examined to see if they are true or false..

Rumour detection is frequently mistaken with the detection of fake news since a rumor is a statement that, at the time it is posted, contains unverified information. The task of rumor detection is therefore described as classifying personal remarks as rumor or not (Zubiaga & Aker, 2018). Rumour detection is thus a crucial component of the misinformation detection process that may also be used to identify assertions that are worth verifying before classifying

them as true or untrue. This can lessen the influence that personal judgments or emotions may have when deciding which claims require additional verification. Identifying emotions in texts or user positions is the task of sentiment analysis. Since disinformation writers place a greater emphasis on how much they can wow their audience and how quickly the material spreads, the sentiment in real and misleading information may change. As a result, disinformation frequently either involves strong emotion that may readily resonant with the general population or makes contentious comments intended to arouse strong emotion in the audience. Thus, the emotion analysis through both the content and user comments can be utilized for disinformation identification.

2.7 Information Filtering Techniques

2.7.1 Email filtering methods/ techniques

DNS validation and sender authentication: The Simple Mail Transfer Protocol (SMTP) protocol in use today (Klenkin 2008) permits any server delivering mail to identify itself as any domain name it chooses without requiring any additional proof. Since this feature makes spam easier to spread, many different projects assert that they can handle the issue.

Sender policy framework: One of the DNS validation techniques is Sender Policy Framework (SPF) (Wong & Schlitt, 2006). It enables a domain to expressly approve the hosts that are permitted to use its domain name, with the ability for a receiving host to verify such approval. **Sender ID:** The Simple Mail Transfer Protocol (SMTP) server can detect whether an email address in a received message was used with the owner of the domain contained in that email address's authorization by looking up the sender ID (Lyon & Wong, 2006), (Lyon, 2006).

Purported Responsible Address (PRA): Sender ID establishes the concept of PRA by identifying this address from an email message's headers and then confirming that it is authorized, whereas SPF only validates the MAIL FROM address Domain keys identified Mail (DKIM): DKIM is a technique for linking a domain name to an email message that is based on the notion that a claim of responsibility is verified using a cryptographic signature and by directly requesting the relevant public key from the signer's domain.(Crocker, 2011).

DNS-based blacklisting and whitelisting. John Levin recently documented these (Levin 2010). This mechanism's goal is to publish lists of IP addresses for known spam-sending servers or reliable Simple Mail Transfer Protocol (SMTP) servers using the Domain Name System (DNS) (whitelists). Every address that is included in DNS whitelist, should have a

corresponding DNS record, which is made by flipping the IP address's octets and adding the DNSxL provider's domain name.

2.7.2 Greylisting

Evan Hariss was the first to describe this technique (Hariss, 2003). This technique enables Simple Mail Transfer Protocol (SMTP) servers to momentarily reject emails from unknown senders while anticipating that the distant server will attempt to send the email at a later time. The majority of spammers do not use RFC-compliant servers, thus when they obtain a soft-fail response code, they never attempt to transmit the message again. This approach takes advantage of this fact. For any incoming message, greylisting is dependent on three pieces of data (commonly referred to as a triplet).

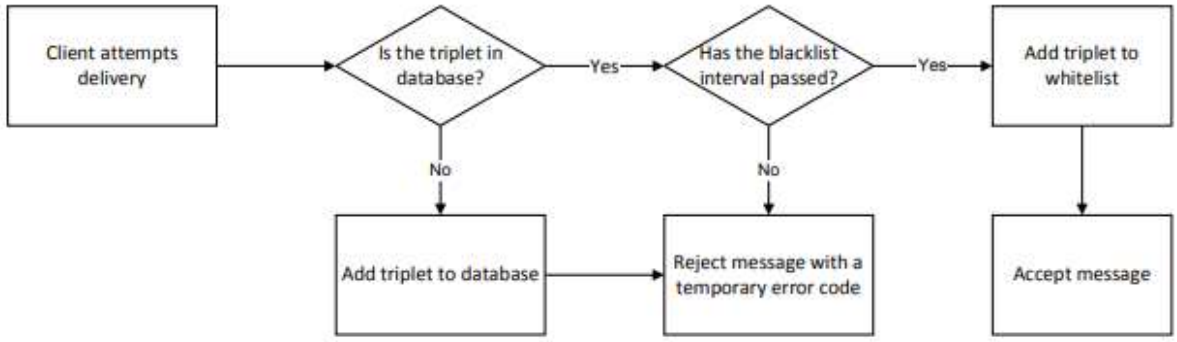


Figure 2.4. Grey listing technique

The relationship between the sender and the recipient is specifically described by the triplet. And it's handled according to the following rule: "If we haven't encountered these triplets before, then reject this delivery and any others that might arrive within a specific amount of time with a temporary failure" (Hariss 2003). Any Simple Mail Transfer Protocol (SMTP) server that complies with industry standards should make an attempt at retries after a predetermined amount of time.

2.7.3 Protocol defects/ Header-relay detection.

A thorough research revealed that spam filtering through protocol faults header-relay detection, even if various words were employed, are essentially the same (Trevio & Ekstrom, 2007). The technique uses tight adherence to (Request for Comments) RFC standards, message header analysis, and less than 1% false positives to identify over 90% of current spam with only a few milliseconds of (Central Processing Unit) CPU time needed per message. Due to the tight criteria that must be followed by authorized Simple Mail Transfer Protocol (SMTP) implementations, this method has a low percentage of false positives.

2.7.4 Communication Filters

This introduces the filters that are already implemented in parameter server. Lossless filters and lossy filters are two different categories for the filters. General distributed applications can use lossless filters. Here, quick compression methods and key caching are seen as lossless filters. **Lossless Filters:** Caching of keys (Lossless) A variety of (key, value) pairs must be sent when submitting parameters. It's possible that multiple communications share the same set of keys, with only the values changing. If so, the receiver can store the keys such that the sender only communicates data that are signed with the correct key. The network bandwidth can be doubled by not sending keys. Because of its quick speed, the lossless data compression library Snappy is employed by Snappy Compression Filter (Lossless), which is used in distributed systems. Snappy can effectively compress 0x0000 and 0x1111 by using the repeated substring to compress the data. A smaller dynamic range results in a higher compression rate when compressing numerical data. **Lossy Filters**: Local gradients are subject to a random skip filter that also skips some items. It reweights the remaining items by multiplying by 1 q, where q is the chance of a random skip. Significant filter modification only entries that have changed significantly since the last synchronization are pushed. (Yipei, 2015)

2.8 Approaches to Misinformation.

Complex challenges include the underlying causes, the spread, and the effects of fake news. There are several ways to approach the problem. people, researchers, and organizations have made attempts to do so. Lazer suggest an intervention that combines two approaches: giving people the tools to assess probable misinformation they come across and making structural adjustments to eliminate exposure to it. (Lazer, 2018)

Educating the public: To develop responsible citizens who uphold civil and democratic values and who are also capable of understanding the competing pressures of capitalist societies, such as the influence of lobby groups, political parties, and the obvious financial gain of creating online content that generates advertising revenue for the creator, media literacy and general education can be improved starting at the school level. We should take a moment to consider how the internet might have developed if it hadn't chosen to rely on advertising as a source of income. Education that is more narrowly focused focuses on information about using advertising as a source of income. The international federation of library Associations and institutions, for instance, published an infographic that implores

readers to look at the source, read past the title, or determine whether the information isn't meant to be amusing or sarcastic.

Analysing and curtailing the spread: Fake news rapidly spread and significantly more deeply via social networks than reliable news (Mustafaraj & Metaxas, 2017; Vosoughi, 2018). This might be as a result of its novelty, its ability to cause indignation or its function in reinforcing the reader's preexisting biases. The novelty and indignation may be the reasons why fake book's attempt to flag discredited false reports failed (Constine, 2018). Actually, more users shared flagged stories. Echo chambers or filter bubbles, where some people are exposed to only one point of view and find it easier to believe stories that support that point of view, contribute to the problem. Lazer encourage communication online across partisan or ideological lines as a result in their agenda for research and action. We also know that people tend to recall information that has been repeated, even when that repetition is part of a debunking or retracting of a myth. Stopping incorrect information in its tracks makes sense in order to avoid having it spread widely and stick in people's brains. (Lazer et al. 2017)

Manual checking: Online articles that contain misleading information, rumors, and fake news must be manually checked in order to stop the spread. There are two main categories of work that may be distinguished: using fact-checking websites and manually checking on particular social media platforms. However, they do have certain drawbacks. The first is that, similar to with schooling, the process places personal accountability on the individual. (Lazer et al. 2018) also made the point that individuals might not be inclined to fact-check a story if it supports their preconceived notions. (Berinsky, 2017; Ecker, 2017).

Automatic checking: The advantages of automating verification are obvious: It can be carried out on a large scale and spares moderators from having to at least go through objectionable stuff. Instead of looking at metadata like the source or dissemination rate, this type of automatic checking looks at the claims and content of the narrative itself. The goal of computational fact-checking is to identify unsupported statements in a report or rumor and compare them to credible sources. (Lazer, 2018)

Content-based Approaches: There have been studies that directly employ text data for various purposes, despite the fact that it is extremely difficult to extract relevant features from content information. For instance, some research concentrates on finding all postings connected to a piece of unknown information (Statrbird & Masddock, 2014). The targeted

postings in this line of inquiry are those that are extremely similar to or duplicates of an original post containing false information. The techniques can be highly beneficial in the later stages of spreading false information. Text-matching can be used to identify information that has been proven to be false or erroneous methods can be used to find all related posts. However, it is difficult for the techniques that catch inaccurate information that has been purposefully changed. The identification of false information has been researched using supervised learning techniques in order to push the boundaries of text matching techniques. They often gather postings and their labels from microblogging platforms like twitter and sina weibo before training a text classifier using the content and labels gathered (Yu, Li, & Liu 2017). These algorithms work on the premise that misinformation may contain particular keywords and/or keyword combinations, allowing a single post with enough misinformation signals to be identified.

Context-based Approaches: Geolocation and posting time are two examples of contextual information on social media networks. Usually, it is combined with other data to speed up detection, or it is directly vectorized and utilized as a feature on its own. There are additional studies that only take into account contextual information. For instance, Kwon suggest modeling the sporadic patterns of false information (Kwon & Cha, 2014). The authors contend that false information posts typically appear in bursts, in contrast to authentic posts, which are posted sparingly over time. The core premise is that distinct groups of accounts purposefully spread false information, which explains why they have diverse posting patterns. An earlier study revealed similar findings, including the fact that rumors occasionally experience spikes in popularity. (Adamic & Friggeri, 2014).

Propagation-based Approaches: Information diffusion is the study of how information spreads through social networks, and research in this area typically focuses on the people who publish and forward information, such as when attempting to foretell a message's final impact. It is exceedingly difficult to get valuable features from content for these new applications since deliberate disinformation spreaders may change it to make it look very real. Recent research focused on simulating the propagation messages in a social network to solve this issue. For instance, the framework TraceMiner categorizes message propagation channels based on the network embedding of social media users (Wu & Liu, 2018). Comparing the suggested method to content-based approaches, experimental findings on real-world datasets show that the proposed approach can offer a high degree of classification

accuracy. This makes sense given how noisy and scant content information may be on social media. Based on collections of content information, it is possible to determine how homogeneous each user's thematic interests are, and researchers have discovered that this homogeneity can be used as an extra feature to enhance supervised misinformation detection systems. In addition to propagation, user behaviors, and modeling information also allows for understanding characteristics of news being spread.

2.9 Misinformation Detection Tools

Misinformation has been considered to spread widely on internet in a variety of ways. In recent years, it has been possible to identify false information automatically depending on the form of the content, that is to say, text and graphics, audios, images and videos, the origin of the incorrect information, and the network of the origin.

Based mostly on content attributes, works of (Castillo 2011) investigated the information credibility on Twitter and developed semi-automatic classifiers to identify the credibility. According to their research, reputable tweets are typically lengthier and contain more URLs as compared to the tweets that are not credible. The TweetCred, web-based and a real-time, system that is accessible as a browser extension to evaluate the credibility of material on Twitter, was developed as a result of these investigations. Based on previously created classifiers, the system assigns a credibility value to each tweet; this credibility score is then verified by user feedback.

Other methods and works emphasize the use of network analysis techniques in addition to content analysis to identify false information. (Michael Conover and Jacob Ratkiewicz, 2011) The findings demonstrate that various diffusion patterns distinguish between false information and true memes, with false information patterns spreading more quickly and frequently being produced by machines rather than people (Adrien Friggeri, Lada & Adamic 2014). Contrarily, reliable information tends to come from a small group of users, receive a lot of re-posts, and spread through authors who have posted a lot of messages in the past and have a lot of friends.

Some of the methods to identify and show the spread of false information include Truthy, RumorLens, and Twitter trails. With the help of an interactive dashboard, users of these technologies can investigate how a rumor spreads in a semi-automatic manner. However,

they require the user to enter a specific rumor to investigate rather than automatically scanning the social media stream for false information. (Flynn, Nyhan, & Reifler. 2017)

Trying to solve this problem Shao and colleagues created the platform Hoaxy, which automatically scans social media to find and analyze internet hoaxes. Facebook recently unveiled new tools to assist stop the spread of false news articles, continuing a recent trend. In contrast to Hoaxy, Facebook tools accurately identify bogus news by combining content analysis with network analysis and user input. Continuous testing and development are being done on this system. However, present initiatives to tackle false information have come under fire because they don't do enough to stop platform abuse. Google's plan to combat disinformation also includes a link at the bottom of the snippet box where people can provide feedback. (Chengcheng, Giovanni & Menczer, 2016)

The Washington Post created a tool called TruthTeller, which summarizes political movies and checks their accuracy against a database utilizing PolitiFact and FactCheck.org. Which claims are true or false is made clear to the audience in the program. Misinformation also entails seeking user input and placing a link at the bottom of the snippet box. (Michael, Brendan & Reifler, 2013)

Po-Ching and Lin created an algorithm called “An Early Decision Algorithm” to the web content from a certain resource being filtered. This algorithm speeds up the procedures used to grant or restrict access to online pages during web filtering. DansGuardian sample testing is done using this method. The issue of lengthy delays from text classification algorithms was addressed in their research. However, it failed to incorporate most of the keywords and still keeps lists of URLs.

Teachers, parents, and children can use Neha Gupta and Saba Hilal's work to safely browse the internet and communicate with others. Through connecting their websites, rather than blocking or filtering them, our effort encourages youngsters to focus on websites that are educational and useful. No blocking is necessary to maintain a child's interest in educational matters, keep them secure from security dangers, and engage in psychological play with them.

Akebo and Almeida developed a comparative analysis of Mathews correlation coefficient prediction, Naive Bayes classifiers, and linear support vector machines. Then, it manages the

TREC05, TREC06, and TREC07 tests on a large data set. They contrast the viewpoint with other free and open-source anti-spam filters, including Spamong and Bogo.

Chena, Hammanai and colleagues created Web Grant System aimed at detecting and filtering adult content from the web. This method allows us to extract pertinent data from the web. It can extract photos, text, and URL names with the use of the data mining approach and evaluate them.

Zhifang Liao, Hui Li and Fei cai developed an algorithm that is used to deny unwanted information of social media websites. Users must manage their social media networking site walls with the use of this algorithm. Additionally, it has the ability to improve the effectiveness of filtering methods. In this algorithm, machine technology is utilized.

Mingliang, Wu, Weiming and Zhouyao created a revolutionary foundation web page filtering algorithm. It is basically used to remove offensive material and images from websites.

Sarifullah Khan, Sadaf Khurshid and Shariq Bashir have created a clever filtering methodology that uses text sentiment analysis and engineering method features to identify the content. A revolutionary content filtering technology is employed to ban undesirable websites. Text classification is carried out to categorize the positive and negative classes with the use of a machine learning system.

Ammar, Eman, Gupta and Samer have developed a technique to advert phishing email attacks. In essence, this is a client-based filtering method. In this method, supervised learning algorithms are used. With the aid of this method, we are able to recognize fresh email attacks.

Botometer identifies social bots and categorizes user accounts on online social media as bots or human. This categorization is based on a variety of user account profile traits, the hierarchical structures of online social networks, past behavial patterns, language, and attitudes. (Yang 2019).

Foller.me Social network users' profiles and tweets are analysed to reveal a variety of user attributes, such as name, location, language, join date, and time zone; data about tweets and tweet analysis. Understanding social media users' specific profiles is the major goal in order to validate social media content. (Sloan & Quan2016).

TinEye aanalyses user-generated content, such as images and videos, and determines whether something is real or phony (Middleton, 2017). Journalists in particular utilize this tool, along with other tools to analyse the contents generated by the users.

Rbutr is a system based on machine learning used to collect websites with content that has been refuted, contradicted, or otherwise contested elsewhere on the Internet. This program also offers sector-specific news and community rebuttal repositories for specific news sites that have a bad reputation (Mensio and Alani, 2019).

Ruchansky suggested a model that captures the source and integrates called CSI which includes the article content, user responses, and source use three aspects of fake news. In its initial module, a recurrent neural network is utilized to record the temporal patterns of the activities performed by the user on a certain piece of information. The next subsystem collects the source traits based on the activities performed by the user, the two are integrated to know whether is phony or not.

Rubin concentrated on detecting satire type of misinformation, five additional predictive features were included to an SVM-based system (Punctuation, Negative, Humor, Grammar Affect, and Absurdity). They conducted tests on 360 news articles and discovered that the best results were obtained when absurdity, grammar, and punctuation were combined.

Dhoju analyzed different distribution of certain features between reliable and unreliable health media by statistical analysis. They looked at the percentage of health-related articles, clickbait distribution patterns, gaps in days between (shared date of publication data), average numbers of photos, quotes, and links per story, word clouds, and the top 10 subjects in trustworthy and trustworthy health-related articles. They did not, however, continue to work on creating a classification model to distinguish between them.

Ghenai and Mejova build a model by investigating poor cancer treatments on twitter, to discover persons who are promoting dubious material. Users were first split into a control group and a rumor group. User in the control group once tweeted about cancer but it wasn't a rumor, while user in the rumor group once shared phony cancer cures. Then 35 features, such as linguistic, sentiment, readability, medical, and time features, were assessed. To find the important features, a logistic regression model with LASSO regularization was utilized.

2.10 RELATED STUDIES

2.10.1 Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking

The popularity of initiatives for fact-checking has increased as a means of preventing the spread of misleading information. This involved fact-checking organizations like FullFact and Snopes which have developed cutting-edge methods for spotting as well as debunking erroneous information as soon as it is created. The popularity of initiatives for fact-checking has increased as a means of preventing the circulation of misleading information. The Neural Semantic Matching Network (NSMN) of Chen (2017) modifies the Improved Sequential Inference Model (ESIM) by adding skip connections from the input to the matching layer and changing the output layer to just max-pool plus one affine layer with ReLU activation. Given a claim, they first extract prospective documents from the corpus, then they retrieve potential sentences from the documents that were selected as potential documents, and finally, the final stage categorizes the sentence into one of three groups. The claim and the sentences were encoded using Bidirectional LSTM (BiLSTM) and the embeddings GloVe (Pennington, 2014) and ELMO (Peters, 2018). However, these activities concentrate on static or slowly evolving domains and deal with topics that don't require specific knowledge to annotate. The aforementioned methods had some efficacy, however there were some problems with the accuracy of the outcomes.

A two-stage model, which they referred to as Model A and Model B, respectively, makes up the architecture. Gathering potential true information for a given claim is the aim of Model A. These justifications are then subjected to Model B's entailment analysis. Next, both Model A and Model B's training procedure and expected run-time behavior are discussed.

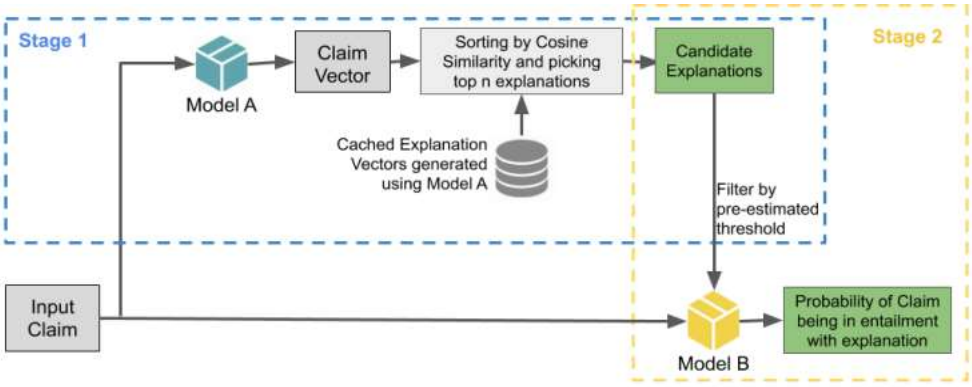


Figure 2.5. Block diagram of two stage model pipeline

Initially, they trained their Transformer model on a binary sentence entailment task, in which the two phrases of claim and explanation are input and separated a [SEP] tag. This allowed to

retrieve pertinent explanations. In order to ensure that there are equal numbers of positive and negative claim-explanation pairs, we produce negative claim-explanation pairs by random sampling. The model can grasp long-range correlations between the vector representations of claims and explanations of similar contexts Model B was also trained as model A. after obtaining the potential explanations for a given claim, Model B can be used to calculate the likelihood that the claim will be supported by each potential explanation. The criteria for Model B classification were determined using the statistic of mean probability score and standard deviation of aligning and non-aligning claim and explanation pairings in the validation set. Models A and B were trained and assessed using a variety of techniques based on traditional NLP techniques and more complex pre-trained Transformer models.

2.10.2 Integrated Monitoring, Searching, Checking, and Analytics of Factual Claims on Twitter

ClaimPortal consists of a web-based front-end Graphical User Interface, a MySQL database, an Elasticsearch 3 search engine, an API (Application Programming Interface), and several decoupled batch data processing parts. The system operates on two levels. The front-end display layer offers a number of filters that users can use to limit the amount of search results returned. The keyword search on tweets is powered by Elasticsearch, which also leverages database queries to offer additional filters. It also provides many graphs that are visualized. Tweet pre-processing, Elasticsearch batch insertion, claim type recognition, and locating relevant fact-checked claims for each tweet are all tasks carried out by the back-end data collection and computation layer (Hassan, 2017). Additionally, it uses the open ClaimBuster API to determine the check-worthiness of tweets. By periodically using the Twitter REST API, ClaimPortal keeps up with the most recent tweets.

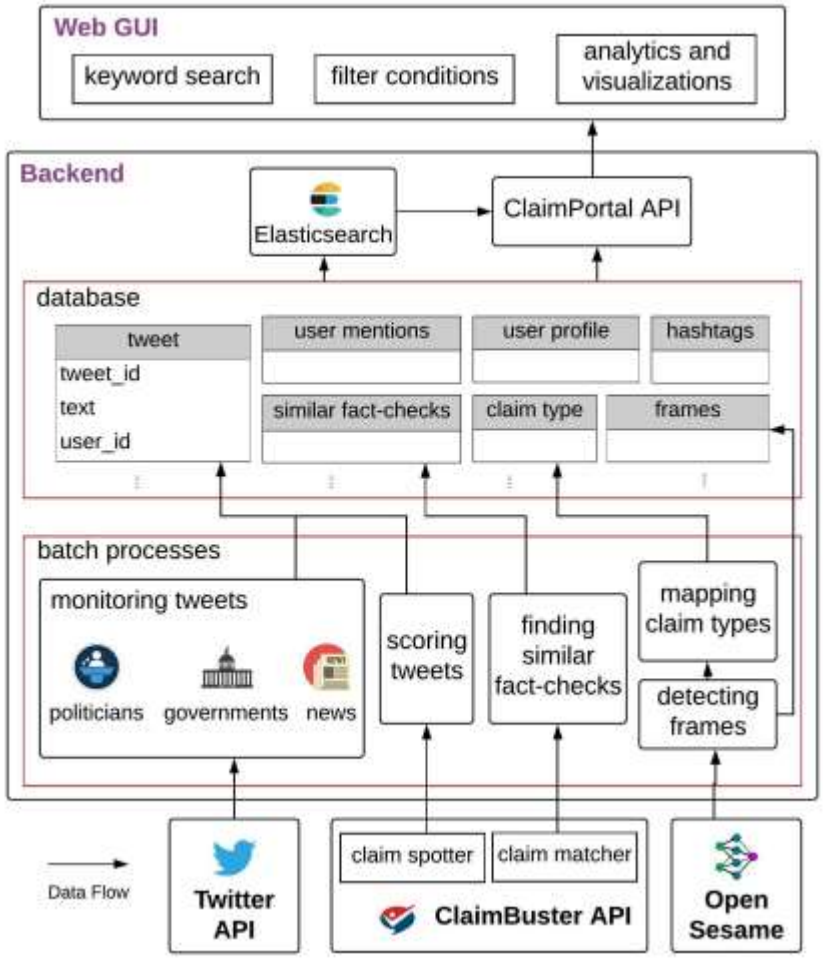


Figure2. 6. Claim Portal system architecture

Keeping, Monitoring, and Processing Tweets Now focusing on tweets with political overtones, ClaimPortal will soon be expanded to include all tweet kinds. We compiled a list of well-known Twitter usernames in American politics, including but not limited to governors, mayors, members of the U.S. Cabinet, other government officials, and political news media teams. Then, REST API was utilized to browse through each user's timeline and gather their tweets.

The back end of ClaimPortal concetrates on processing data and storing data. We have access to the required data thanks to the Twitter REST API. A web service called the claim portal API was created with Python and the Flask 4 micro-framework. It offers end points for importing tweets onto the GUI, looking up hashtags, and looking up people to apply from-user and user-mention filters on. The API searches the database for the resultant list of tweet IDs based on the user's keyword search and desired filters and delivers the list as a JSON response. There are numerous normalized tables in the MySQL database. The database keeps track of each tweet's text, creation date, and tweeter. The database also includes statistics on

retweets, quoted tweets, hashtags, URLs, and accounts that were mentioned in the tweets. Elasticsearch is used by ClaimPortal to facilitate keyword searching within the archived tweets.

Claim Spotter: Each tweet on ClaimPortal is assigned a checkworthiness score, which indicates if the tweet makes a factual claim whose veracity is significant to the general audience. This rating is generated by reviewing the ClaimBuster API, a potential tool for fact-checking created and regularly utilized by expert fact-checkers (Adair, 2019). A classification and rating model called ClaimBuster (Hassan, 2017; Jimenez & Li, 2018) was trained using a dataset comprising 8,000 words from previous U.S. presidential debates that had been human-labeled. For any given text, the ClaimBuster API gives a check-worthiness score. The score is given on a scale of 0 to 1, with 1 being the most checkworthy

ClaimPortal analyses tweets to identify the different types of factual claims that are being made, as well as who is making them, how frequently they are made, and whether or not they are accurate. Frame recognition a linguistic database for English called FrameNet contains 1,224 personally created semantic frames. Each frame contains details on the participants in a particular kind of event, situation, object, or relation as well as the language and semantic structure of those elements. The participants are semantic roles that are specific to a given frame and offer additional information (Swayamdipta, 2017)

Maps claims by kind Eight ACE event types were listed with their mapped frames in (Spiliopoulou, 2017): Business, Conflict, Contact, Justice, Life, Movement, Persons, and Transaction. We added four additional event types Comparison, Quantity, Stance, and Speech along with their respective frames to this list.

Claims Analyser: Claim matching is a crucial stage in the fact-checking workflow. It attempts to locate identical or related claims from a repository of past fact-checks given a factual assertion. The underlying assumption is that famous people consistently make incorrect statements. Politicians may avoid making blatantly incorrect assertions in order to avoid having their statements verified, but frequently they continue to make them even after they have been refuted. ClaimPortal makes use of the ClaimBuster API's claim matching feature. The Share-the-facts database makes up the fact-check repository. Fact-checking sources include PolitiFact, Snopes, factcheck.org, the Washington Post, and more, and there are ten fact-checks total. The system compares the tokens of a claim and a fact-check to determine

how similar they are. In order to search the repository based on token similarity, an Elasticsearch server is deployed.

Characteristics of the user interface using various criteria, a user can sort through the tweets using ClaimPortal. The following are the crucial filters.

1) Keyword search: this enables users to do a text search using terms such as "climate change".
2) Hashtags: Users can use hashtags like "#116thCongress" or "#2020" to further filter tweets.
3) Claim type; The user can look for the tweets that have a particular claim type, such as Conflict.
4) From: this searches for tweets published by a specific user account, such as "@elonmask".
5) Mentions: User mentions (i.e., using "@" to tag a user in a tweet, e.g., "@POTUS") can be used to further filter the search results.
6) ClaimPortal additionally provides a slider to filter results depending on a range of ClaimBuster scores. As the slider is changed, the resulting tweets are automatically updated.
7) Date range: The site also provides a date selector to filter tweets based on when they were originally posted.

2.10.3 Semi-automated fact-checking through semantic similarity and natural language inference

The fight against the spread of false information is being led by fact-checking organizations, who are working hard to verify new hoaxes and gather proof about the news that are spread through via many social media channels including sms like instant messaging services. Majority of the verification work done the firms is done manually, but due to the frequent publication of new posts and tweets, the results of this effort are hardly noticeable in OSNs. Members of these networks intentionally submit fraudulent statements without repercussions or spread misleading information without even realizing it. In this study, we take advantage of the most recent developments in NLP to create a multilingual, semantic-aware Transformer-based architecture for evaluating semantic similarity, fact-checking semi-automatically , and tracking information in online social media platforms. On the one hand, we offer an architecture that can assist the entire public in determining the validity of a

certain piece of information such as a tweet by context-aware automated comparison against a databases of verified information.

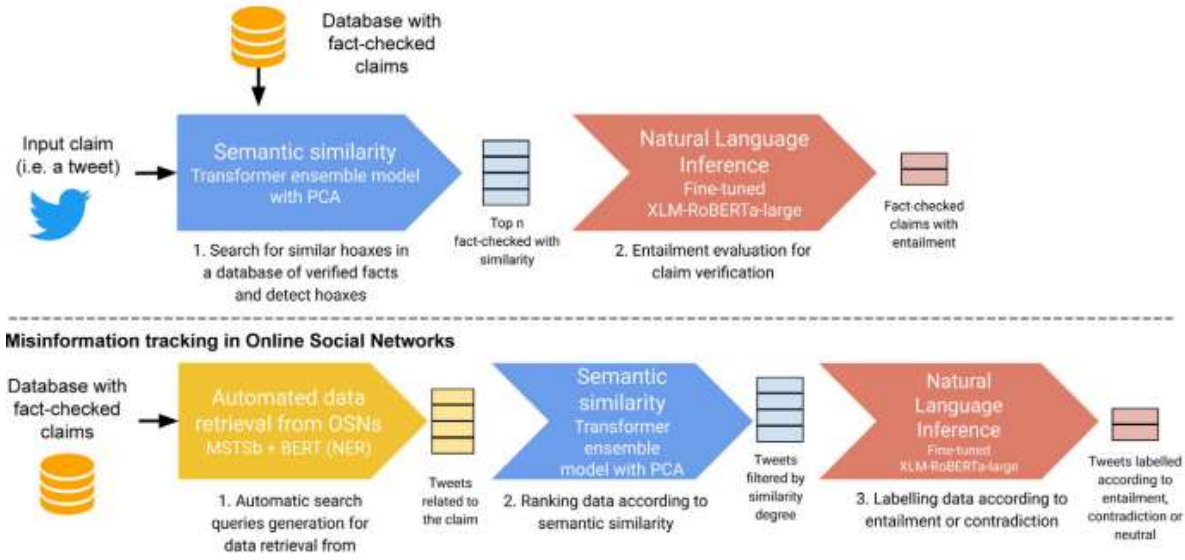


Figure 2.7. A semi-automated fact checking process

2.10.4 A Benchmark Dataset for Fake News Detection

The dataset that is related to India is not available (Dilip, Sharma1 & Sonal 2021). To create our dataset, we scraped authentic news from a number of reliable sources. The Parsehub scrapper was used as a tool used for website scraping, to produce a dataset. In the Indian dataset, there are 56,868 news items. We have preferred to get the news from fact-checked news sources, including Boomlive and Alt News, and manually review the label of each piece of news before classifying it. From the year 2013 through the year 2021, to compiled the news. We just choose the India news column in order to retrieve news about India, and also made a filter to omit others. The news were manually requested cross-verification of the dataset collected from a number of topic annotators. We have gathered data from a number of variables, including the news's title, date and time, source, link, image link, and label (true/false). LDA topic modeling is also used to include the news category. Five categories were created using LDA topic modeling: Election, Politics, COVID-19, Violence, and Miscellaneous. The planned working framework for the IFND dataset's generation is depicted in the figure below. Fake news is restricted after several fact-checking websites are deleted in order to gather the news. Consequently, a data augmentation technique is utilized to produce a biased dataset. Using LDA topic modeling, the news is sorted into many categories following data augmentation. The generated dataset then goes through pre-processing. Then, several deep-learning and machine-learning classifiers are used to text and picture analysis. For the purpose of identifying bogus news, a multi-modal technique is also used, combining

textual and visual features.

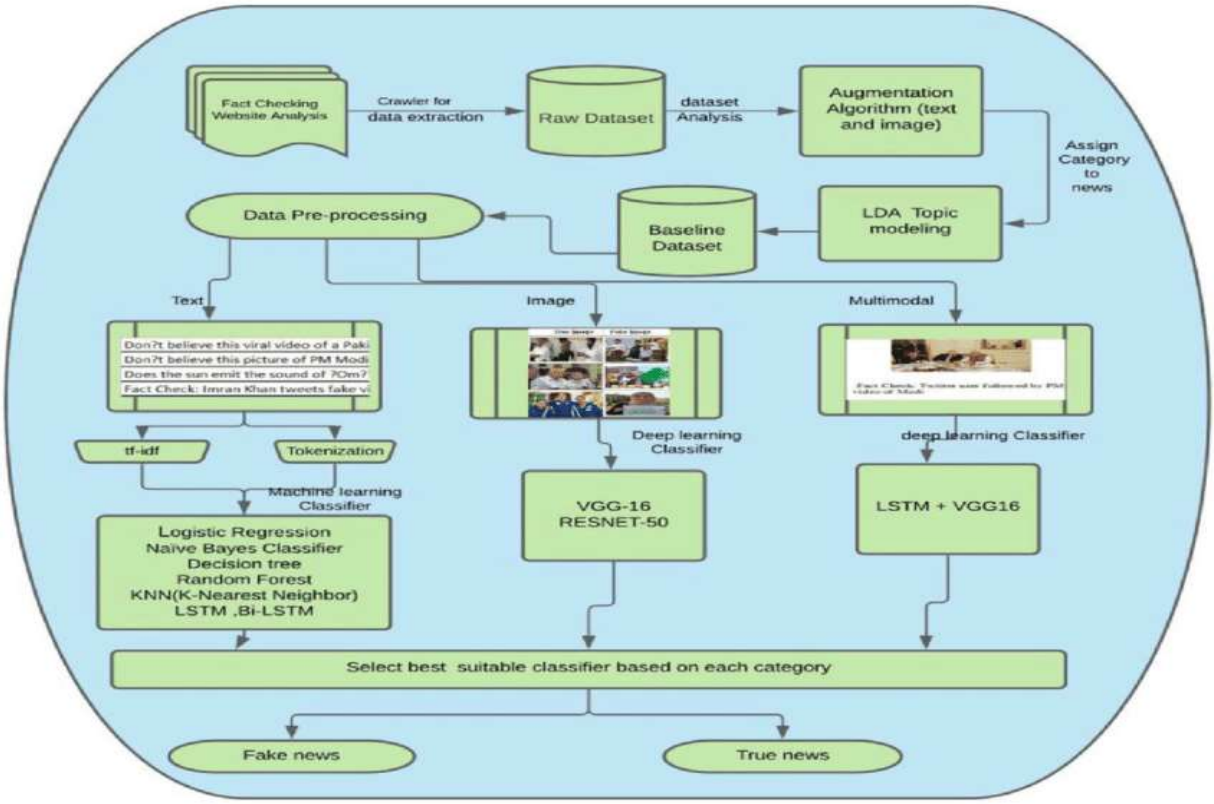


Figure 2.8. Proposed framework for IFND dataset

2.10.5 Gradient-Based Adversarial Training on Transformer Networks for Detecting Check-Worthy Factual Claims

The fact-checking structure shown above demonstrates how it is currently functioning. We keep an eye on statements made in a variety of places (such as Twitter, political debates, etc.), and for significant occasions like presidential debates, we can even scan live television closed-caption feeds. The claims that our claim monitoring system records are then scored by ClaimSpotter. The public can access ClaimSpotter through an API, which simply needs a free API key 21. To test and validate the models described in this study, we are making the deep learning models available to the general public and other academics. Every deep learning model has its own Nvidia GTX 1080Ti GPU (Graphical Processing Unit). Since every resource is connected to the same network, a server to server communication does not add much extra overhead. In addition, we also employ ElasticSearch in conjunction with a repository of verified claims. In our claimmatcher component, we assess the accuracy of any statements that have already undergone professional fact-checking. We can forward these assertions to our fact-checking component, which is still under development, if no prior fact-checks are identified for them. To test if knowledge bases (like Wolfram, Google, etc.) can

produce a precise judgment, our current method involves converting statements into questions. This method works well for general knowledge claims, but it is still very difficult to handle nuanced assertions that require domain-specific knowledge. Last but not least, we also offer newly rated Google search results that are organized according to the information found on the sites that the first search query delivers. The analysis is based on how closely each page's text context matches the original query using the Jaccard algorithm. Finally, during election cycles, we regularly release the results of the presidential debate checkworthiness scores on our website and tweet about the high-scoring statements.

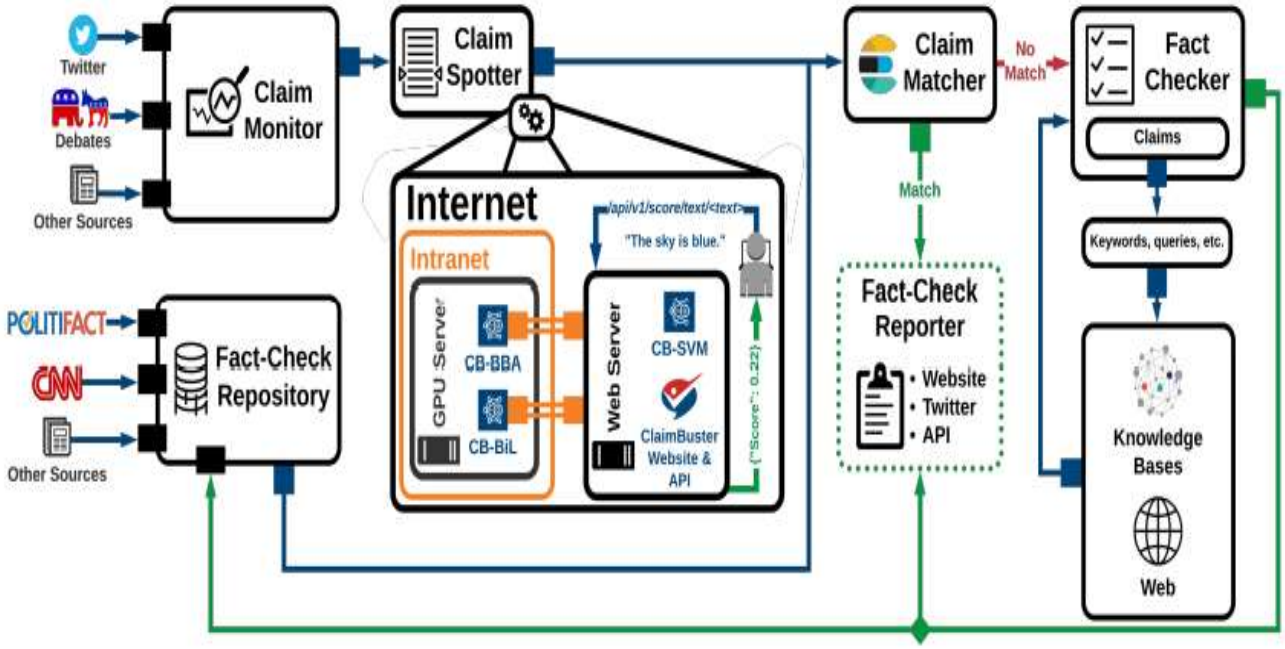


Figure 2.9. A diagram of fact checking method

2.10.6. A natural language processing framework for automated fact-checking.

According to Zhijiang, Michael and Andreas, Figure 13 demonstrates a three-stage NLP framework for automated fact-checking:

(i) claim detection to determine the claims that need to be verified;

(ii) evidence retrieval to locate the origin confirming or disputing the claim.

(iii) claim verification to determine the truth of the claim depending on the retrieved evidence.

claim detection is frequently handled independently, although evidence retrieval and claim verification are sometimes handled in a unified operation known as factual verification. There are two components to claim verification which are handled

independently or together: verdict prediction, in which claims are given labels indicating their veracity, and justification production, in which verdict justifications must be provided.

Claim Detection. The initial step in a non-supervised fact-checking, selects certain claims for verification. The idea of check-worthiness is frequently used in the context of detection. Check-worthy claims are ones that the general public can be interested in learning the real truth (Hassan et al. 2015)

Rumor detection is an additional instance. An unsubstantiated story or claim that is spreading (usually on social media) is referred to as a rumor. A stream of social media posts is a common source of information for rumor detection systems, and a binary classifier must decide whether or not each post is rumorous. (Zhang, 2021).

Evidence Retrieval Finding information beyond the claim—such as text, tables, knowledge bases, photographs, and pertinent metadata—is the goal of evidence retrieval. Some earlier attempts (Wang, 2017) only use the assertion itself as evidence. Without taking into account the condition of the world, relying on surface patterns of claims makes it impossible to spot well-presented false information, such as assertions made by machines (Schuster, 2020). This problem has been made worse by recent advances in natural language production, where machine-generated content is sometimes thought to be more reliable than human-written material (Zellers, 2019).

Verdict Prediction this is to assess the truthfulness of a claim given its identification and the pieces of evidence found to support it. The most straightforward strategy is binary classification, such as categorizing a proposition as true or false. When using evidence to support a claim, it is frequently preferable to use refuted by evidence rather than true/false, given the systems frequently do not evaluate the evidence on its own. In general, given the acknowledged constraints, it would be unwise to make such sweeping statements about the world. (Graves, 2018).

Justification Production In order to persuade readers of their interpretation of the data, fact-checkers must justify their judgments, which is a crucial component of journalistic fact-checking. A "backfire" effect, where belief in the false claim is reinforced, might result from debunking something simply by labeling it incorrect (Lewandowsky et al., 2012). The necessity for automated fact-checking, which might use black-box components, is considerably higher. These artifacts can have unforeseen, detrimental effects when

developers use black-box models whose decision-making mechanisms are opaque. (O'Neil, 2016). This issue may be solved by creating methods that explain model predictions (Lipton, 2018), and current research has concentrated on the creation of justifications survey of explainable claim verification. Since claim verification is frequently the stage of fact-checking that receives the most scrutiny, research to date has concentrated on justification production for that process. But for the other stages of our paradigm, explainability might likewise be desirable and necessary.

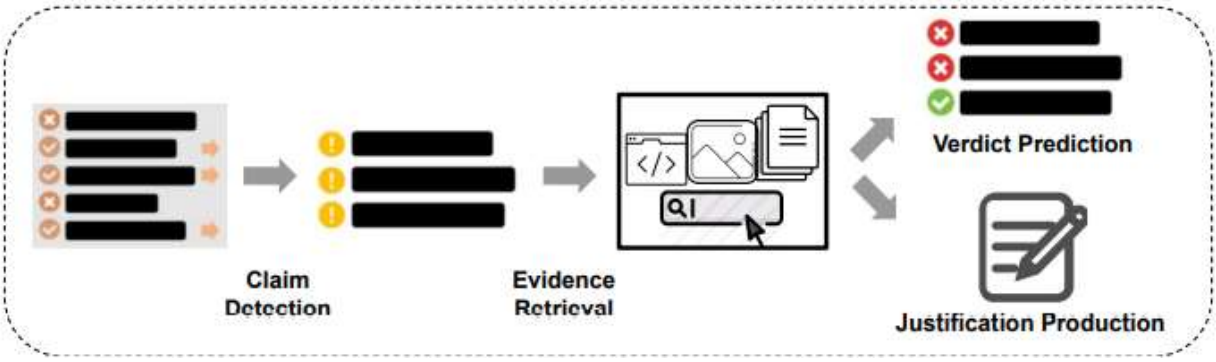


Figure 2.10. A natural language processing framework for automated fact-checking.

2.10.7 Workflow for training algorithms and classification of news articles.

Iftikhar, Muhammad & Yousaf developed a Framework as shown in Figure 14, they building depending on the already available research. In their proposed framework, they incorporating ensemble techniques with different linguistic feature sets to categorize articles from diverse areas as true or false. The uniqueness of our suggested strategy is provided by the ensemble approaches and Linguistic Inquiry and Word Count (LIWC) feature. Many reputable websites publish true news articles, while some others, are utilized for fact checking. For our trials, we chose three datasets with news from a variety of fields and a mixture of real and false stories. The datasets are downloadable online and were taken directly from the web. The second and third datasets are accessible to the general public via Kaggle, whereas the first dataset is the ISOT Fake News Dataset. Before being utilized as an input for training the models, the World Wide Web corpus is first preprocessed. Unwanted article variables including authors, date of publication, URL, and category are filtered out. Articles with no body text or a body of fewer than 20 words are also deleted. For consistency in format and structure, multicolumn articles are converted into single column articles. To achieve consistency in format and structure, these actions are applied to all datasets. After the data has been cleaned up and explored, the pertinent qualities have been chosen, and the next stage was to extract the

language features. In order to be used as an input for the training models, certain textual qualities had to be transformed into a numerical form. These characteristics include the proportion of words that express good or negative emotions, the proportion of stop words, punctuation, function words, informal language, and the proportion of adjectives, prepositions, and verbs that are employed in sentences. For consistency in format and structure, multicolumn articles are converted into single column articles. To achieve consistency in format and structure, these actions are applied to all datasets. After the data has been cleaned up and explored, the pertinent qualities have been chosen, and the next stage is to extract the language features. In order to be used as an input for the training models, certain textual qualities had to be transformed into a numerical form. Using the LIWC2015 program, which categorizes the text into many discrete and continuous variables, some of which are listed above, we were able to extract features from the corpus. From any given text, the LIWC program extracts 93 distinct features. No encoding of category variables is necessary because all of the features collected using the tool are numerical values. However, scaling is used to guarantee that the values of different features fall inside the range of (0, 1). This is required because some values, like percentages, have a range of 0 to 100, whilst other values, like word counts, can have any range. The various machine learning models are then trained using the input features. Each dataset is split into a training set and a testing set, each with a 70/30 split. In training and test situations, the articles are shuffled to provide a fair distribution of bogus and real articles. Different hyperparameters are used to train the learning algorithms in order to maximize accuracy for a particular dataset while maintaining an ideal variance to bias ratio (Bergstra & Bengio, 2012). Numerous ensemble techniques, some of which are novel to this research, are investigated to assess performance over numerous datasets, including bagging, boosting, and voting classifier. The first voting classifier is an ensemble of logistic regression, random forest, and KNN, while the second voting classifier is made up of logistic regression, linear SVM, and classification and regression trees (CART). **(**Iftikhar, Muhammad & Yousaf, 2020**)**

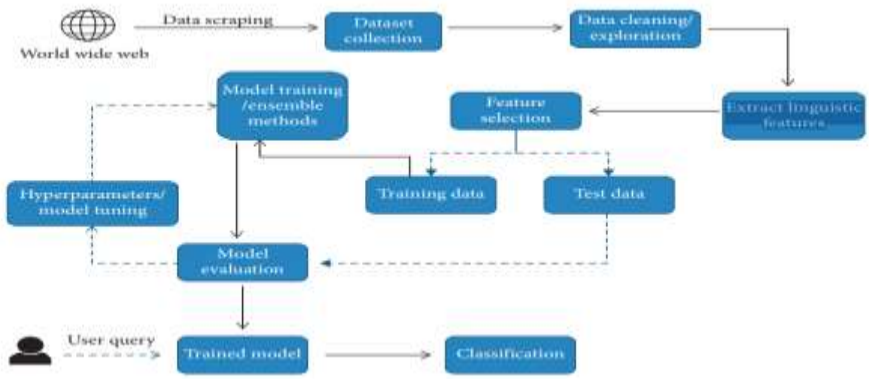


Figure 2.11. Workflow for training algorithms and classification of news articles

2.10.8 Filtering in the context of socio-technical systems for moderation

The illustration below demonstrates how a hybrid approach can incorporate both automatic and human moderation. The user-generated data can be processed in stages by automated filters. Certain items may be classified by the filter as unquestionably harmful, in which case they are automatically removed, or unquestionably not harmful, in which case they are made available online.0

Certain items might also be labeled as doubtful by the filter, which means they could be potentially hazardous. In this scenario, the items would be subject to human evaluation before being censored or published. After a user complains about automated removal, an item may additionally be subject to human review. Human review not only improves future performance by correcting specific historical outcomes of automatic filtering. In particular, if a machine learning approach is used, the new human assessments might be added to the filter's training set so that the system may update its model and learn to handle the new cases appropriately. (Giovanni & Andrea, 2020).

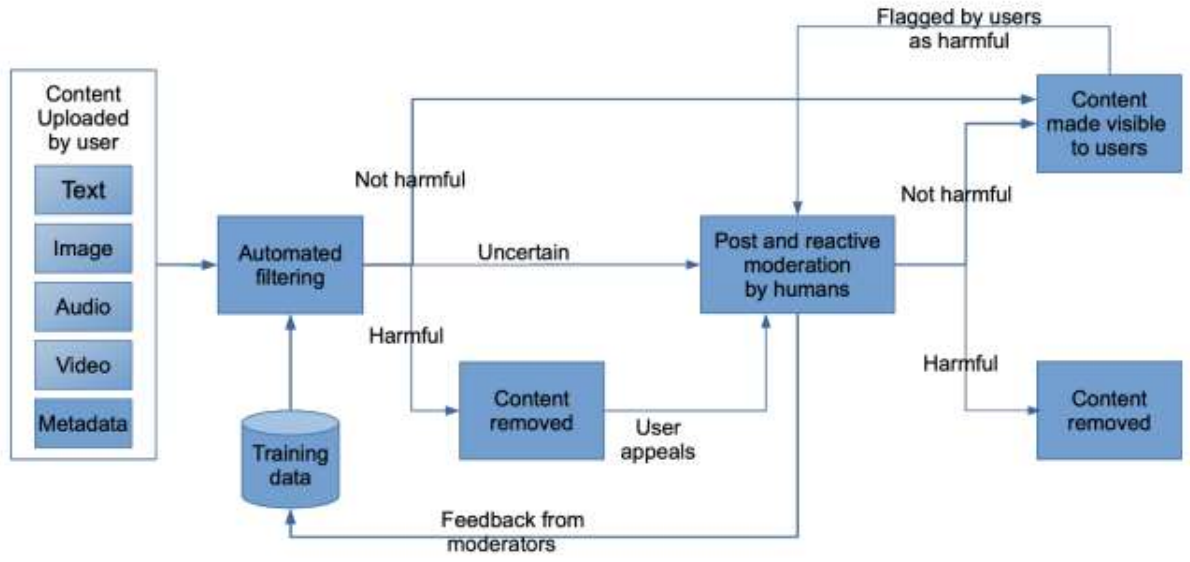


Figure 2.12. The integration of automated and human moderation

An automated filtering system shouldn't be viewed in isolation; rather, it should be seen as a part of social-technical systems, which combine human beings and technologies through organizational structures (such as rules, roles, task distribution, and workflow specification). Not just as the creators of the content, but also when flagging dangerous content or contesting an automated system's or human moderator's decision, users may play a significant role. (Giovanni & Andrea, 2020).

2.10.9 health-related misinformation detection

The schematic below depicts the suggested model's general structure. It has a training component and a detecting component. The input for the training phase is data gathered automatically or manually from all sources on Chinese internet domains, with labels of reliable health-related content. In order to achieve model training, we use two methods: Text-based methods and future-based methods. Our trained detection model is fed recently published health-related articles as input, and the result is an unreliable score indicating how likely it is that the articles are fraudulent. (YUE, KE, XIAOFEI, LINBO, & YONGHONG, 2019)

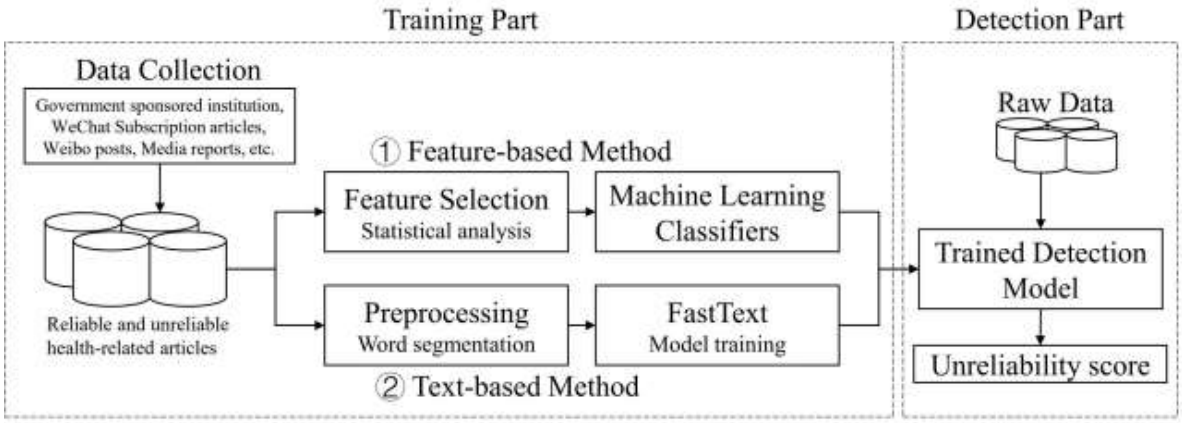


Figure 2.13. A framework of health-related Misinformation detection model

THE CIRCULATION OF INFORMATION

According to (Žiga, 2018) Editing, publishing, amplifying, and consuming news are all parts of the news process. The traditional news process and the online news process are the two categories into which he split the news production. Quality control, gatekeepers, and censors were located around the edit-publish-amplify stage in the conventional news process. The majority of information was produced by experts, but there was no easy way to verify the quality of internet news. Most of it hasn't been quality-checked. Only a small portion of content is produced by experts. Due to the internet, there have never been more opportunities to study, stay informed, and exchange views with others.

2.10.10 A Semi-Formal Model of the Circulation Of News

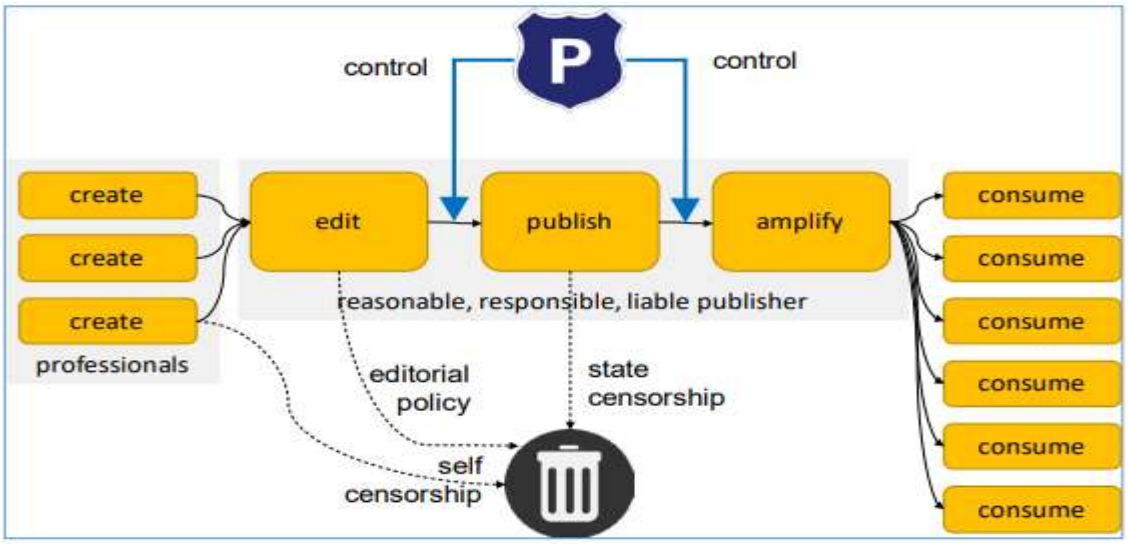


Figure 2.14. The integration of auto

When we refer to the "traditional news process," we mean the method that has traditionally been used in newspapers, magazines, publishing, radio, television, tapes, Compact Disks (CDs), and Digital versatile Disks (DVDs). The content is produced by a professional in the conventional news process. Media editors then alter this for both substance and language. Then, content that satisfies the necessary standards is published, perhaps by being printed in newspapers or broadcast on television. The publisher may decide to push some information harder than other news toward customers. For instance, you could advertise it on a newsstand or put it on the front page of a newspaper or the first few minutes of the evening news on TV. In the end, people consume the content. **(**Dr. Žiga 2018)

There is a very well-defined bottleneck or gateway in the process. The editing, publication, and amplification phases are overseen by a small number of persons. Edit, publish, and magnify stages can be controlled if there is a desire to regulate them. Reasonable editors maintain the standards of what is published even in the absence of outside influence and may eventually be held accountable for it. **(**Dr. Žiga 2018)

Internet news access

The World Wide Web, in particular, offers a variety of tools for disseminating ideas and news, and its evolution can be divided into three stages:

a) Websites and webpages first appeared in the 1990s. Technically speaking, publication required the author or publisher to set up a server online. Services that let users post on the internet without technical knowledge first appeared throughout the previous ten years. WordPress and Blogger provided a platform for texts services like SoundCloud and Spotify for audio, 500px and Flickr for images, YouTube, and Vimeo for videos.
b) Over the past few decades, these services have developed from providing only publication space for content into platforms that allow social interaction between content authors and consumers. Social media platforms like Facebook and Twitter serve as an extreme example of this. The main purpose of social media is not so much to publish original information as it is to share, suggest, and comment on stuff that is

available elsewhere.

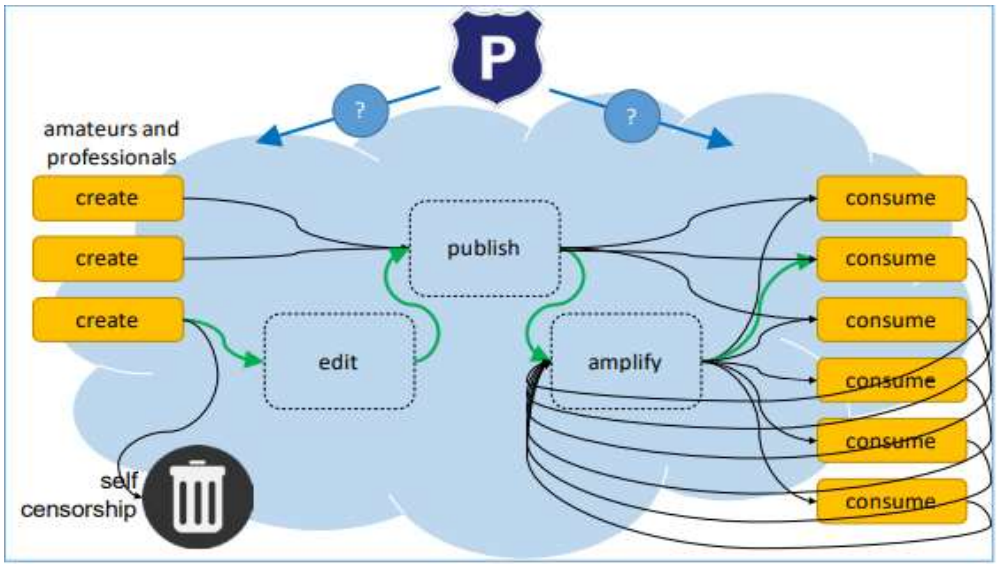


Figure 2.15. Internet news access

The internet used technology to replace manual publication, amplification, and editing. Anyone may make content; only professionals cannot. Editing and other quality checks are carried out by the authors themselves, or not. From basic web servers to hosting services to social networks that connect content creators and consumers, the internet offers a wide range of services where content can be produced. The information is likewise amplified by both, but primarily by consumers. In contrast to the conventional approach, there isn't a single place where the content traded may be monitored, managed, or assured of its quality. People act as their own gatekeepers rather than relying on a select group of experts. But because the stages of editing, publishing, and amplifying occur on a platform like Facebook, these platforms are in a strong position to affect this process. **(**Dr. Žiga 2018)

2.10.11 UCC FACT CHECKER

The Commission's system or procedure for information verification with the goal of encouraging factual reporting is called the UCC Fact-Checker. Before or after the content is published or distributed in another way, fact-checking might be done. Therefore, the UCC Fact-Checker disproves false information spread on social media and in traditional media. To instruct readers on how to flag dubious tales, a variety of materials are used.

The framework above was designed to improve the current UCC fact-checker. UCC currently has inefficiencies in its method of fact checking or information filtering. The user identifies a claim, takes a screenshot or forwards it onto the CERT (Computer Emergency Response Team) website or email or UCC hotline. The information is fact-checked and a

quick response is given to the user. The human in the backend of the fact checker contacts the respective groups or the individuals that can provide a fact about a given claim and then give a feedback to a user as soon as the facts are obtained from a credible source. This can be real or fake news or misinformation depending on the outcome of fact checking.

The process is as follows

1. Call the UCC Toll-free number, snap a screen shot, and submit it to WhatsApp, or forward it over email.
2. UG-CERT Requests will be delivered to the fake news checker via the aforementioned channels.
3. The team examines, authenticates, and verifies the reported information before publishing its results.
4. The requester receives a response or feedback as soon as feasible.
5. The results are categorised, and those that are relevant to the general public are publicized on the false news site and on the Consumer Affairs twitter account. Note: it is still under development.

RESPONSE: The length of time it takes to verify a message depends on its nature, but it typically takes thirty minutes. The response will state if the information was supplied is accurate, inaccurate, misleading, disputed, or out of bounds. Information pertaining to the Commission, ICT, and the sector is highlighted. Additional information, such as that on health, politics, business, education, and security, may be confirmed through intermediary contacts.

Figure of the current UCC misinformation fact checker

Figure 2.16. the current UCC fact-checking process

2.11 GAPS IDENTIFIED

1. All the models and frameworks reviewed, their mission is to stop misinformation and none is aiming at removing the misinformation after it has been created and in circulation. Well, that is the most challenging task. In this study, I will aim at having both false and true information in place. The framework should be able to detect the misinformation then publish real time true information on all available platforms so that if someone comes across that information, is viewing both the real and fake news. Thus, we amplify the speed of real information than false information.
2. Most of the models and frameworks tend to completely remove human intervention in misinformation management and fact checking but every day new information is created and we can't guarantee that that information is genuine neither false. Meaning that we have to keep training the machine to increase its accuracy.
3. The infrastructure set up of most countries are different so most of the models or frameworks can't be applied in Uganda.
4. In order to categorize messages using pattern or probabilistic analysis, the majority of spam filters today look at message headers and contents (Lieven, 2006). He acknowledges that the majority of spam may be effectively removed using this technique, but there are some significant disadvantages as well: Before they can analyze a message, content filters

need to receive the entire message. Due to the late classification, there is a large resource consumption, and spammers can quickly modify their messages to bypass content filters. The aforementioned tools they were created for particular puposes. Therefore, the proposed framework will address the problem by being generic that is to say. It should be able to capture misinformation of different kind for example health related misinformation, political related misinformation among others.

5. Digital media literacy can be an effective technique for limiting the spread of fake news. Disinformation might not be stopped until users of digital media are taught media and news literacy. (Sullivan, M.C. 2019).
6. According to Jenecek, Gansterer, and Kumar (jenecek at al., 2008), greylisting has some drawbacks. increases the time it takes for the mail to be delivered, Unreliable method for figuring out whether a session alludes to a previous delivery attempt Spammers can adapt and get around the filtering by sending conceivably new messages that match triplets that have already been whitelisted. Attempts to resend a temporarily rejected email may arrive from a different IP address and fail because large firms frequently utilize server farms to handle their outbound email traffic. In Content-based Approaches, Text-matching can be used to identify information that has been proven to be false or erroneous methods can be used to find all related posts. However, it is difficult for the techniques that catch inaccurate information that has been purposefully changed
7. In Propagation-based Approaches. It is exceedingly difficult to get valuable features from content for these new applications since deliberate disinformation spreaders may change it to make it look very real.
8. The majority of the verification work done is done manually, but due to the frequent publication of new posts and tweets, the results of this effort are hardly noticeable in OSNs. Members of these networks intentionally submit fraudulent statements without repercussions or spread misleading information without even realizing it.
9. UCC current fact checker only involves human beings and since human beings are slow because of the manual work, this delays the availability of real information in time and hence leading to the dominance of fake news on social media platforms. To have an environment that is totally free from misinformation is still a challenge in many countries including Uganda. And another challenging question is, if the fake news is released on one social media platform, how can it be pulled out before it circulates? “This is a complex task since we don’t have a central server that manages all social media platforms and user

accounts". Therefore, if we cannot eliminate misinformation, then we should learn how to leave with it.

CHAPTER THREE: METHODOLOGY

3.0 Introduction.

The methods used in this study to gather, arrange, and analyse data was emphasized in this chapter along with the presentation of the research findings and the rationale behind them. In keeping with this, the chapter describes how the selected research procedures and methods were applied to help address the research questions and subsequently achieve the research objectives.

3.1 Research Design

After determining the research problem, one of the challenges is typically how to construct the study design. A research design is typically expected to concentrate on what needs to be done, where it needs to be done, when it needs to be done, how it needs to be done, and what activities make up the means that make up a research design. A research design entails planning the collection and analysis of data in a way that tries to balance relevance to the study purpose with economy in the technique. (Kothari, 2014).

An investigation that uses qualitative, quantitative, and mixed approaches is known as research design. It offers particular guidelines for research techniques, also known as "inquiry strategies" (Denzin & Lincoln, 2011). Researchers and academics have noted that qualitative research relies on data that cannot be represented numerically (Keele, 2010). I used quantitative and qualitative research for this study in order to gather and process the data. In this instance, I conducted a literature study before conducting qualitative research to gather and analyze information from previous research findings. To find out the answers to the research questions and fulfill the particular goals in this study the mixed approach was adopted.

3.2 Sampling of Data

A researcher just needs to choose a small subset of the greater population in any study to accurately represent that population. This is due to the fact that it takes time and resources to reach out to and use every segment of the population in order to collect data for the study. The population is typically exceedingly huge, making it impossible to select any one component to use as study data. As a result, only a tiny sample that accurately represents the greater population was found in this study. This study set out to accomplish this by

employing a flexible sampling strategy that offered a foundation on which precise data could be acquired while taking into account the time and resources available. Nonprobability sampling approach was among the alternatives I had for sampling methods. According to Kothari (2004), Nonprobability sampling is a method where the researcher does not know in advance which element will be chosen as a real representation of the larger population and as a result, elements are only chosen based on their availability. Nonprobability sampling is frequently chosen over probability sampling in most studies and research. This is a result of the probability sampling technique's intricacy and complexities. With this in mind, I chose to identify sample elements from the population using a nonprobability sampling technique.

3.2.1 Target Population.

The following stage is to select the population, the sample frame, and the sample size after the study approach and data gathering techniques have been decided upon. This is crucial since choosing a sample size from the incorrect demographic can lead to inaccurate data. From this perspective, the study was directed at Uganda communications Commission (UCC) located in Kampala 44 Spring Rd central region, Uganda. But the study only focused on 5 employees who are directly interacting with the available factchecker (CERT members). 27 members under Media Challenge Initiative (MCI) fact-checker.

3.2.2 sample size

A sample is a subset of the population drawn from the research population in order to provide information that can be generalized to the complete population. Therefore, in order to determine a sufficient sample size from the anticipated population, the researcher utilized purposive sampling techniques on 5 Employees of Uganda communications commission (UCC)in the fact checking department and 12 members working with the factchecker at MCI (Media challenge Initiative), 5 bloggers and 5 journalists. The study only targeted the employees who only had direct interaction with the fact checking system or who had direct intact with the fact checking system. The survey was carried out online using google survey form.

3.3 Data Collection

Researchers and academictians distinguish between primary and secondary data as the two forms of information that can be employed to accomplish any study's goals (Sapsford & Jupp, 2006). In their investigations, researchers can employ either primary data or secondary data, or perhaps both. I combined primary and secondary data for this study.

3.3.1 Secondary data collection

The secondary data was first sort of information gathered for this investigation. The study outlined prior research and studies on the application of ICT filters in the control of misinformation. As a result, identified the researches that have been done on the topic, the current findings, as well as the research gaps. As a result, the conclusions drawn from the review of related literature aided in laying the groundwork for this investigation. Credible websites, academic journal databases, and other sources were used to get this data.

3.3.2 Collecting Primary Data

Researchers can use a variety of data collection techniques to get primary data. I create an online questionnaire using google form with closed-ended questions to answer my research objective I and II for the study. Since the study also target the general public, I found it easy and convenient to conduct online survey. It took 15 days to gather data from the targeted population. My decision to employ a survey and a questionnaire is driven by the fact that they are simpler to run, particularly in a study with constrained time and resources. This type of questionnaire is employed because it allows respondents to complete it whenever is most convenient for them, which lowers the cost of administering data gathering.

3.3.3 Interview

Here, the researcher conducted face-to-face interviews with the interviewee and interviewer with the express purpose of collecting data. I conducted formal interviews to get additional information in-depth, lessen resistance, provide instructions on how to complete the questionnaire, and collect personal information. In this case I used an interview guide that contained questions about which could contribute in answering objective number (iii) of my research. I interacted with the UCC research Team and this gave me more light on what I am supposed to do and areas of improvement and concentration.

3.3.4 Questionnaire

Questionnaire method is also employed and a questionnaire distributed among the respondents of targeted population. The questionnaire was issued to the respondents and given some time to enable thorough collection of data. This method is adopted because of the advantages it has and these are; it is effective in saving the researcher's time since the researcher may apply them in different department in a single day and later make collections on a greed date. Use of questionnaire as a method of data collection gives an opportunity of answering questions freely thus generates quality information than it would have been with other methods. An online questionnaire that was designed and distributed via media such as whatsApp and emails to the selected group of people. The selection was purposive targeting

users who can provide sufficient information about the study. The questionnaire target bloggers, journalists and the individuals that have interacted and participated in fact checking at UCC and MCI (Media challenge Initiative).

3.4 Analysis of the collected data

I arranged the information gathered after I received the duly completed surveys. I discarded All incomplete questionnaires were discarded and declared invalid. To organize the gathered data into descriptive statistics, SPSS (Statistical Package for the Social Sciences), Google forms as well as Monkey tools like excel sheets were utilized in the analysis phase. In order to make it simpler to analyse the gathered data, also these tools were used to build graphs.

3.5 Tools used in model development

Python programming language was used in model development and on Kaggle platform. This is because Kaggle has a lot of dataset that can be used to test and run the model developed. In each stage of building a machine learning model for misinformation management using LSTM and Word2Vec, various tools and libraries were used to facilitate the process. Among others, they include the following;

Web scraping libraries such as BeautifulSoup were used for collecting data from social media platforms and in this case twitter. APIs for accessing and retrieving data from platforms and in this case **REST API, NIFI** were adopted to scrap data from Twitter. The data was processed using **Natural Language Processing (NLP)** libraries, such as NLTK, for tokenization, stopword removal, and text normalization. Regular expressions for pattern matching and text cleaning, **Pandas** for data manipulation and cleaning, **Gensim** for text processing tasks like stemming or lemmatization. **Scikit-learn** to split the dataset into training, validation, and testing sets while ensuring class balance. **Gensim** for training **Word2Vec** embeddings on a large corpus of text, **TensorFlow** for integrating Word2Vec embeddings into the deep learning model, **LSTM** for building and training deep learning models, **GPUs** for accelerated model training. **Scikit-learn** for calculating standard evaluation metrics like accuracy, precision, recall, F1-score, and finally **Matplotlib** and **Seaborn** for data visualization.

3.6 Ethical considerations

This study was faced with a number of ethical concerns among the misuse of the collected data, anonymity, privacy and confidentiality. To begin with, the collection of primary research gave access to more than 50 emails of staff members from MCI (Media Challenge Initiative) and Uganda Communications commission. Therefore, to avoid interfering with

these peoples' privacy, I first requested them to participate in the research. For those who agreed, I only send out a maximum of two email reminders for them to fill out the questionnaire and send it back to me. Furthermore, I did not use their email address for any other purpose apart from this study, and did not divulge their contact information to third parties in order to maintain confidentiality. In addition, I did not mention the names of the participants in the duly filled questionnaires as a way of ensuring anonymity of the respondents. Lastly, to avoid misuse of the collected data, I only used it for the purpose of this study.

CHAPTER FOUR: DATA PRESENTATIONS ANALYSIS AND INTERPRETATION

4.0 Introduction

This chapter introduces the different finding, presentation and analysis of the data collected on the types of misinformation, the sources or creators of misinformation, the channel through which misinformation is disseminated, the social media platform that spreads misinformation most. On the framework part, it summarizes the how modules were developed tested and evaluated.

4.1 Objective i: To identify the types and sources of misinformation in Uganda

4.1.1 The sources of misinformation

	RK 1	RK 2	RK 3	RK 4	RK 5	TOTAL NO. OF RESPONDENTS
Traditional media	4	4	7	10	2	27
politicians	5	8	11	2	1	27
Academicians	2	0	1	2	22	27
Business leaders	2	6	6	12	1	27
Ordinary users	15	8	2	1	1	27

Table 4.1. sources of misinformation

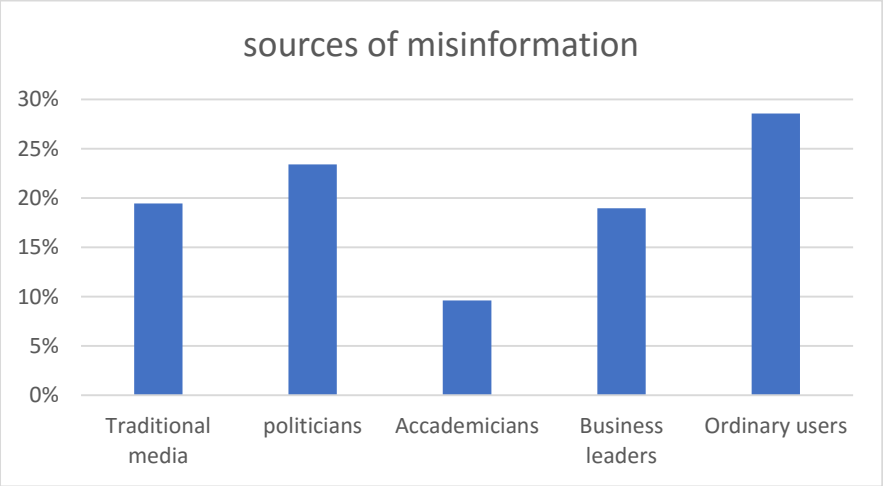
The frequency distribution table above shows the how the respondents ranked different misinformation creators in ascending order. Out of 27 respondents, 4 respondents ranked traditional media as 1, 4 ranked it as 2, while 7 respondents ranked it as 3, 10 ranked it as 4 and 2 respondents ranked it 5. Also 5 respondents ranked politicians' number 1 misinformation creators, 8 respondents ranked politicians as 2, 11 ranked politicians as 3 whereas 2 respondents ranked politicians as 4 and 1 respondent ranked politicians as 5. Academicians were ranked in the first position as the creator of misinformation by 2 respondents. while none of the respondents ranked academicians as 2, 1 respondent ranked academicians as 3, 2 respondents ranked academicians as number 4 while 22 respondents ranked academicians as number 5. However, 2 respondents ranked business leaders as misinformation creators on social media as number 1, 6 ranked business leaders as 2, 6 ranked business leaders as 3, 12 respondents ranked business leaders as 4 and 2 ranked business leaders as 5. Finally, 15 respondents ranked ordinary users as misinformation creators on social media as 1, 8 respondents ranked ordinary social media users as 2, 2 ranked ordinary users as 3, 1 ranked ordinary users as number 4 and 1 respondent ranked ordinary users in the fifth position.

Sources of misinformation scores in percentage

	TOTAL SCORES	TOTAL SCORES (%)
Traditional media	79	19%
Politicians	95	23%
Academicians	39	10%

Business leaders	77	19%
Ordinary users	116	29%
TOTAL	406	100%

The table above shows the total scores and the percentage total score for each misinformation creator. Traditional media 19% as misinformation creators, politicians scored 23% as misinformation creators, academicians scored 10% as creators of misinformation, business leaders scored 19% and ordinary users scored 29% as creators of misinformation.



The bar chart above represents the total percentage scores for each misinformation creators on social media. Ordinary social media users being the highest misinformation creators with 29%, followed by politicians with 23%, traditional media and business leaders in the third position with 19% and academician with 10% score being in the last position among the misinformation creators on social media.

4.1.2 Types of misinformation

	RK 1	RK 2	RK 3	RK 4	RK 5	TOTAL NO. OF REPONDENTS
Fake News	14	10	0	0	3	**27**
Hate Speech	8	14	2	2	1	**27**
Rumors	2	3	19	3	0	**27**
Misleading Titles	1	0	6	20	0	**27**
Satire	2	0	0	3	22	**27**

Table 4.2. types of misinformation

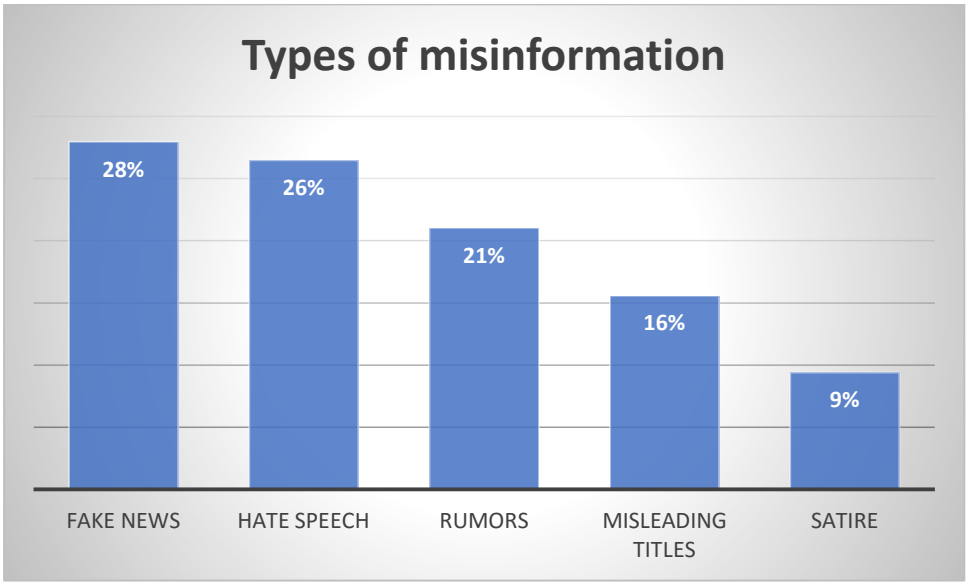
The table above shows the ranking of different types of misinformation that are encountered on social media. Out of 27 respondents, 14 ranked fake news as number 1 type of misinformation that spread on social media, 10 respondents ranked it as 2, while none of the respondents neither ranked it as 3 nor in the fourth position, 3 respondents ranked fake news as number 5. 8 respondents ranked hate speech as 1 type of misinformation spread on social media, 14 respondents ranked as 2, 2 ranked it as 3, also 2 ranked as 4 and 1 respondent ranked as 5. 2 respondents ranked Rumors as 1, 3 ranked it as 2, 19 respondents ranked as 3 while 3 ranked it as 4 and none of the respondents ranked as 5. 1 respondent ranked misleading titles as 1, none of the respondents ranked as 2, 6 respondents ranked misleading titles as 3 and 20 respondents ranked it as 4 and none of the respondents ranked it as 5. 2 respondents ranked satire as 1, none of the respondents ranked satire as 2 neither 3, 3

respondents ranked it as 4 while 22 respondents ranked it as 5 types of misinformation that it common on social media.

Types of misinformation scores in percentage

	TOTAL SCORE	TOTAL SCORE(%)
Fake News	**113**	28%
Hate Speech	**107**	26%
Rumours	**85**	21%
Misleading Titles	**63**	16%
Satire	**38**	9%
TOTAL	404	100%

The table shows the percentage scores for each types of misinformation that is encountered on social media. Fake news scored 28%, hate speech scored 26%, rumors scored 21%, misleading titles scored 16%, and satire scored 9%. This implies that fake news is the most common type of misinformation on social media with 28%, followed by hate speech with 26%, rumors in the third position with 21%, misleading titles in the fourth position with 16% and lastly satire with 9%.



4.2 The channels/platforms for misinformation

FOCTORS	Rk 1	Rk 2	Rk 3	Rk 4	TOTAL SCORE
Social Media	27	0	0	0	**27**
Public Events	1	9	17	0	**27**
Education Publications	1	1	3	22	**27**
Community Networks	1	17	7	2	**27**

Table 4.3. channels for misinformation

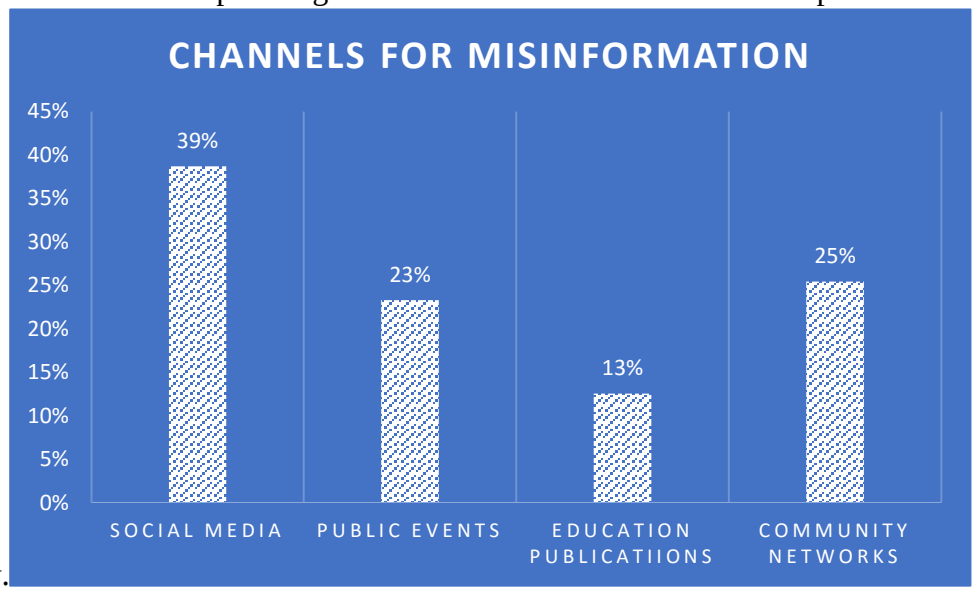
The frequency table above shows the ranking of different respondents on the channels or platforms that spread misinformation. All the 27 respondents ranked social media as 1 channel through which misinformation spread. 1 respondent ranked public events as 1

channel that spread misinformation, 9 respondents ranked public events as 2, 17 ranked public events as number 3 and none ranked public events as a channel that spread misinformation as number 4. Whereas 1 respondent ranked education publications as 1 channel that spread misinformation, 1 respondent ranked education publications as number 2, 3 ranked education publications as 3, and 22 ranked education publication 4 as a channel that spread misinformation.

Channels for misinformation in percentages

	TOTAL SCORE	TOTAL SCORE (%)
Social Media	108	39%
Public Events	65	23%
Education Publications	35	13%
Community Networks	71	25%
TOTAL	279	100%

The table shows the total percentage score for each channel/platform that spread misinformation. Social media scored 39%, public events scored 23%, education publications scored 13% while community networks cored 25%. Social media takes the lead as the channel that spread misinformation with 39% followed by community networks with 25%, public events in the third position with 23% and finally education publication being in the last position as a channel that spreading misinformation. The information is presented in a bar chart below.



4.2.1 Common social media platforms

	RK 1	RK 2	RK 3	RK 4	RK 5	RK 6	TOTAL NO. OF REPONDENTS
Facebook	3	7	5	4	5	3	**27**
Twitter	17	5	2	2	1	0	**27**
WhatsApp	0	7	8	7	3	2	**27**
YouTube	1	2	2	5	5	12	**27**

Instagram	4	1	3	2	11	6	**27**
TikTok	3	5	7	7	4	1	**27**

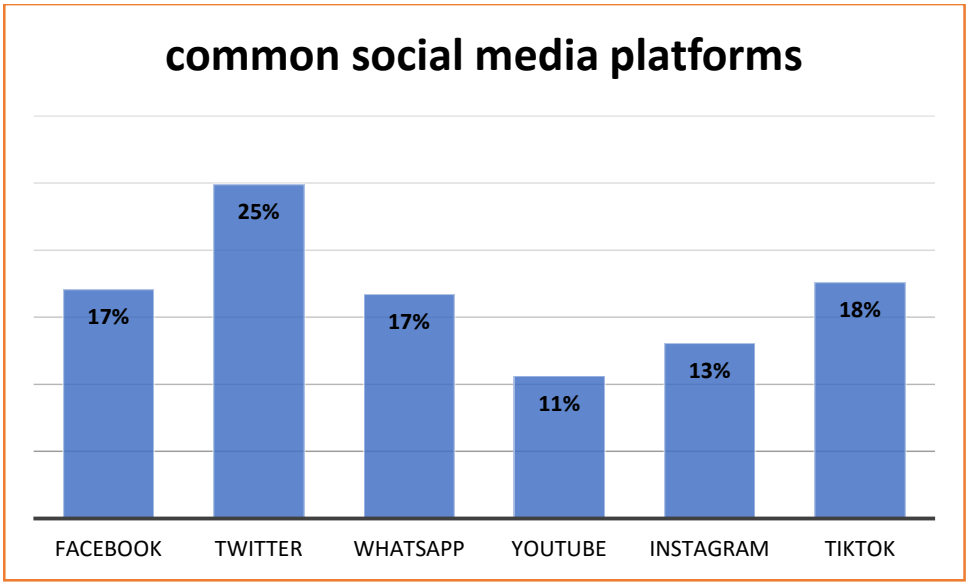
Table 4.4. social media platforms

The frequency table below shows the ranking of respondents on social media platforms that mostly spread misinformation. Out of 27 respondents, 3 ranked Facebook as 1 social media platform that spread misinformation, 7 ranked as 2, 5 respondents ranked it as 3, 4 ranked it as 4, 5 ranked it as 5and 6 respondents ranked it as 6. While 17 respondents ranked twitter as number 1, 5 ranked as 2, 2 ranked it as 3,2 ranked it as 4, 1 respondent ranked it as 5 and none of the respondents ranked as 6. None of the respondents ranked WhatsApp as number 1 social media platform that spread misinformation. 7 ranked WhatsApp as 2, 8 ranked it as number 3, 7 ranked it as number 4, 3 ranked as number 5 and 2 respondents ranked it as number 6. 1 respondent ranked YouTube as a social media platform that spread misinformation, 2 respondents ranked as 2, also 2 respondents ranked it 2, 5 respondents ranked as number 4 and five also ranked it as number5 and 12 respondents ranked social media platform that spread misinformation. 4 respondents ranked Instagram as a social media platform that spreads misinformation, 1 ranked it number 2, 3 respondents ranked it number, 2 respondents ranked as number 3, 11 ranked it number 5 and 6 respondents ranked as Instagram as number 6 whereas 3 respondents ranked TikTok as number 1,5 respondents ranked as 2, 7 ranked as number 3, 7ranked as number 4, 4 ranked it as number 5 and 1 respondent ranked as number 6.

Common social media platforms in percentages

	TOTAL SCORES	TOTAL SCORES (%)
Facebook	98	
Twitter	143	25%
WhatsApp	96	17%
YouTube	61	11%
Instagram	75	13%
TikTok	101	18%
TOTAL	574	100%

The table shows the total scores for each social media platform that spread misinformation. Facebook scored 17%, Twitter scored 25%, WhatsApp scored 17%, YouTube scored 11%, Instagram scored 13%, and TikTok scored 18%. Therefore, the leading social media platform that spreads misinformation is twitter with 25%, followed by TikTok with 18%, Facebook and WhatsApp in the third position with a tie of 17% and Instagram in the fifth position with 13% and YouTube being the social media platform that spreads misinformation the least with 11%. The information is represented on a bar chart below.



4.3.0 Objective iii: To design a proposed framework for misinformation management.

To design the framework in figure 17 below, I adopted the current UCC fact checking process framework as shown in 2.10.11 in chapter 2 figure 16. In the framework, there is a search engine that is to say elastic search, social media users, database, natural language processing algorithms, social media platforms, claims, claim hunter, channels through which the claims can reach and the human fact checker. However, the only part I will concentrate on is NLP (natural Language processing) since am lacking resources to enable me handle other parts of this framework. In the future, other researchers can implement them.

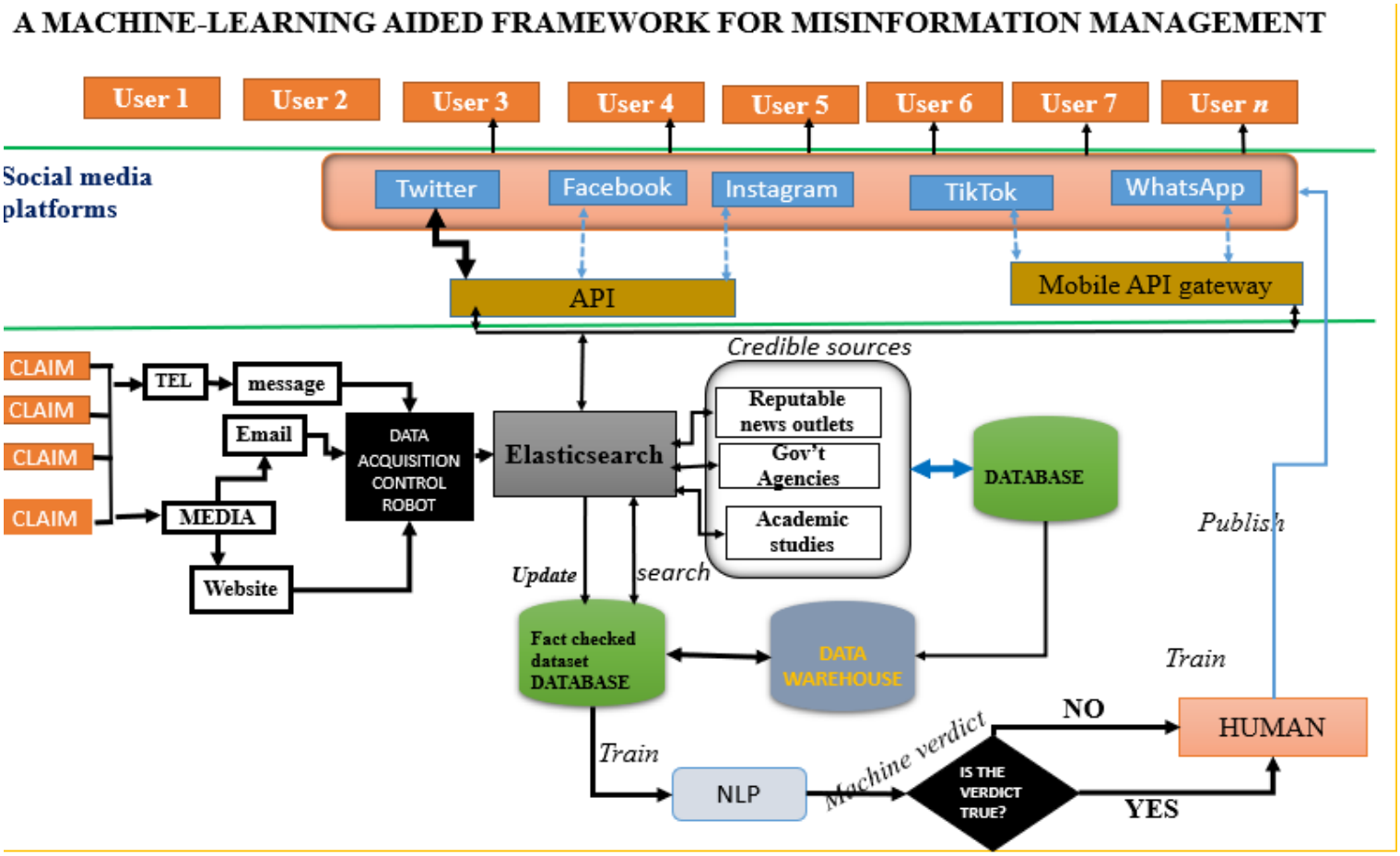


Figure 4.17. proposed machine learning aided framework

Elastic search: To provide a factual explanation in fact-checking, it is important to use credible sources, such as government agencies, academic studies, and reputable news outlets. The information should be cross-checked against multiple sources to ensure accuracy and to avoid potential biases. Overall, a factual explanation in fact-checking is essential to ensuring that accurate information is disseminated to the public, helping to prevent the spread of misinformation and promoting informed decision-making. The elastic search helps in deep searching for data in multiple sources and databases in different deployments, arranges visualizes it in a summarized format for easy analysis. The data is then used to train the machine in order to increase its accuracy.

Human: The human identifies a given data as fake or real and then keeps the entails in the database. After fact checking and annotating that piece of data and labelling it as fake or real news/information. The data therefore is used to train and update the natural language processing algorithm. The human also publishes the real information as soon as possible after being fact checker so that the public is guided.

User/claimer: The users also pray a big role in combating misinformation. The user

identifies a certain piece of information as fake or false. He/she takes a screenshot of the information and sends it to the UCC hotline number or on the UCC email. The piece of information is checked against the available fake news in the database

Claim Hunter: The human prays a big role in looking for claims from a certain social media account. In the above framework, I used twitter because most of the government official communicate to their official pages on twitter and people tend to misinterpret the information posted or remove or add their own word. Through the API, a user can access twitter and performs web scrapping/harvesting, or look for information that is suspected to be false and curries out fact check using the credited sources. The only target can be the accounts of individuals or groups that are well known in spreading misinformation or false news.

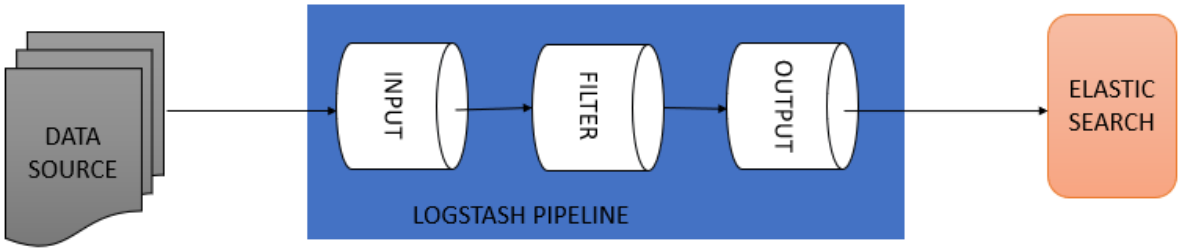
NLP (Natural Language Processing): NLP is a field of artificial intelligence that focuses on understanding and processing natural language text. By using NLP techniques, Twitter fact-checking tools can analyse tweets in real-time, identify false or misleading information, and provide accurate information to users. Some common NLP techniques used for fact-checking include sentiment analysis, named entity recognition, and topic modelling. The NLP (Natural Language Processing) in the framework is used to provide verdicts on the fact checked information. NLP (Natural Language Processing) can be used to combat fake news on Twitter in several ways. NLP techniques can be used to analyse the sentiment of tweets and identify whether they contain positive or negative language. This can be useful in identifying tweets that are spreading false information or propaganda. For example, tweets with negative sentiment that are promoting false information or attacking certain individuals or groups can be flagged for further review. Fact-Checking: NLP techniques can be used to compare the information in tweets with verified sources of information to identify false information. For example, if a tweet claims that a certain event occurred, NLP techniques can be used to compare the information in the tweet with news articles in the database or other verified sources to identify whether the information is true or false.

THE ELK STACK

The ELK Stack that is the combination of three major items working hand in hand. These are; Elastic search, Logstash, and Kibana. Elasticsearch is a high scalable index server that processes the data, Logstash collects, enriches, filters and forwards and transform the data and later Kibana visualizes the data. Elasticsearch is the heart of elastic slack, it is used to search, store and analyse the data.

Figure 4.2. The ELK stack

Connecting Elasticsearch to Twitter using the API can be a powerful way to analyze real-time social media data and gain insights into trends, sentiment, and user behavior. Elasticsearch is connected to Twitter using the Twitter API, which allows Elasticsearch to stream real-time tweets from Twitter and index them for search and analysis this involved Creating a Twitter developer account to access the Twitter API to obtain authentication credentials (i.e., API key, API secret key, access token, and access token secret), the elasticsearcher twitter plugin is installed and configured. The streamed data from twitter was then analysed and visualized on the elasticsearch dashboard. Elasticsearch has a powerful query language for searching data based on specific criteria such as keyword searches or more complex queries that involve aggregations, filters, and sorting. I also used Elasticsearch powerful set of tools for analyzing data, including aggregations, which allow you to summarize and group data based on specific criteria, and data visualization tools that allowed to create charts and graphs to better understand of the data.



APACHE NIFI ON THE FIRST LOUNCH

The figure below shows the sample of apache nifi a webabsed applicxation that louches extracts data from twitter and converts it into a JSON format that aacts aas an input to elasticsearch and later visualization with kibana. All kibana and elasticsearch run in apache nifi and nifi also run in java environment. The twitter API key, token, secret token, authenticatiokn token are very import to provide access to twitter implying that one must be having a developer account inorder to extract data.

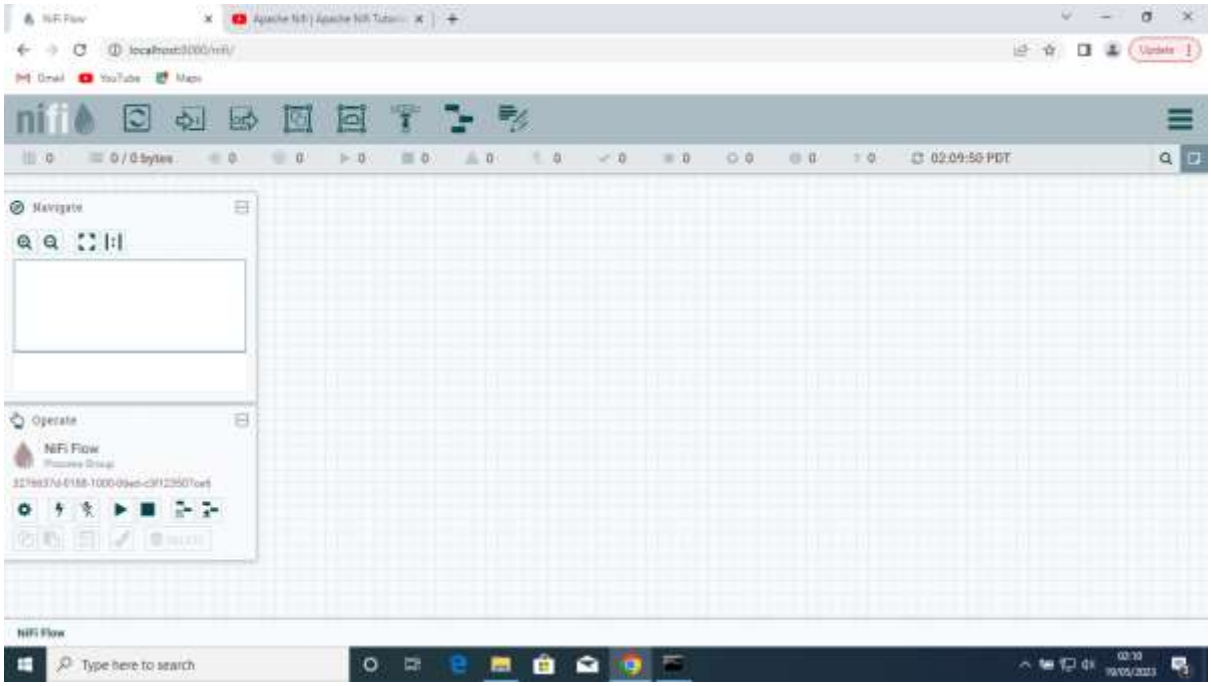


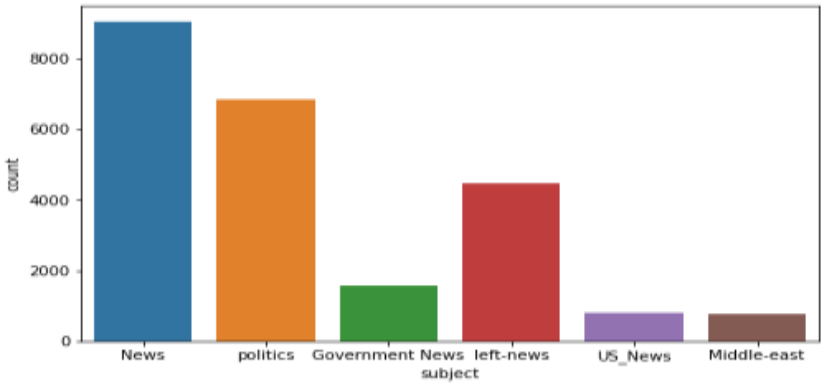
Figure 4.3. apache nifi on the first launch

The data is convereted into JSON data format that acts as inputs to the elasticsearch and later visualized using Kibana.

NLP (NATURAL LANGUAGE PROCESSING)

I obtained data from Kaggle that I used to train the train the machine and to generate a model classification and accuracy score of the model. 80% of the data was used to train the model and 20% was used to test the model. To build a machine learning model, I first import the necessary libraries for data pre-processing, machine learning, and evaluation then loaded the fact-checked data in form of a CSV file that I downloaded using the pandas library. Next, I pre-process the text data by converting it to lowercase, removing special characters and stop words, and tokenizing the text using word2vec tokenezer. Then split the pre-processed data into training and testing sets using the train_test_split function from scikit-learn. I then evaluated the performance of the model on the testing data using the predict method, and print a classification report that includes precision, recall, and accuracy scores for each class.

The data contained both fake news articles and true news articles which was about 2.8MB. see the figure below



The chart shows the dataset that I used to train the model. The news subjects included; news, politics, government news, US news and Middle East news. The largest part was normal news articles, followed by political news, followed by left-news, then government news and finally a tie of US-news and Middle East news. This implies that there is more false information on normal news and politics that on government news and Middle East.

```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 700, 100)          12224900
_________________________________________________________________
lstm (LSTM)                  (None, 128)               117248
_________________________________________________________________
dense (Dense)                (None, 1)                 129
=================================================================
Total params: 12,342,277
Trainable params: 117,377
Non-trainable params: 12,224,900
_________________________________________________________________
```

Table 4.5. model summary

The model has a total of 12,342,277 parameters. Out of these, 117,377 parameters are trainable, and 12,224,900 parameters are non-trainable. The trainable parameters are updated during training to make the model learn from the data, while the non-trainable parameters remain fixed throughout the training process.

embedding (Embedding): This is the first layer, which is an embedding layer. It is used to convert discrete categorical data (like words or tokens) into continuous vector representations. In this case, the input shape is (None, 700), where None represents the batch

size, and 700 represents the length of the input sequence. The output shape of this layer is (None, 700, 100), indicating that the input sequence of length 700 is embedded into a sequence of vectors, each of length 100. The layer has 12,224,900 parameters, which is the number of unique embeddings (vectors) created for each of the 700 input tokens.

LSTM: This is the second layer, which is a Long Short-Term Memory (LSTM) layer. LSTMs are a type of Recurrent Neural Network (RNN) that are well-suited for handling sequential data. The input shape for this layer is (None, 700, 100), which means it takes the embedded sequence of length 700 from the previous layer. The output shape is (None, 128), indicating that the LSTM layer produces an output vector of length 128 for each input sequence. The layer has 117,248 parameters, which are the weights and biases of the LSTM cells.

Dense: This is the third layer, which is a fully connected dense layer. The input shape for this layer is (None, 128), meaning it takes the output vectors of length 128 from the previous LSTM layer. The output shape is (None, 1), indicating that this layer produces a single scalar value as the final output (e.g., for binary classification). The layer has 129 parameters, which are the weights and biases for the dense layer.

In summary, this model consists of an embedding layer to convert input tokens into continuous vectors, followed by an LSTM layer to process the sequential data, and finally, a dense layer for the final prediction. The model has a total of 12,342,277 parameters, which includes the parameters from all three layers.

Total params: 12,342,277 This represents the total number of parameters in the model, which includes all the learnable parameters (trainable) and non-learnable parameters (non-trainable). Parameters are the internal variables that the model learns during the training process to make predictions on the given data. Trainable params: 117,377 This indicates the number of parameters that are trainable, meaning these are the variables that the model will learn and update during the training process. These trainable parameters are updated through backpropagation while optimizing the model on the training data. Non-trainable params: 12,224,900 This represents the number of non-trainable parameters, which are fixed and not updated during the training process. Non-trainable parameters could include things like pre-trained embeddings or fixed components of the model that are not modified during training.

Table 4.6. the accuracy scores for train and validation

```
model.fit(X_train, y_train, validation_split=0.3, epochs=6)

Train on 23570 samples, validate on 10102 samples
Epoch 1/6
23570/23570 [==============================] - 510s 22ms/sample - loss: 0.1299 - acc: 0.9526 - val_loss: 0.1
Epoch 2/6
23570/23570 [==============================] - 496s 21ms/sample - loss: 0.0712 - acc: 0.9765 - val_loss: 0.0
Epoch 3/6
23570/23570 [==============================] - 501s 21ms/sample - loss: 0.0549 - acc: 0.9798 - val_loss: 0.1
Epoch 4/6
23570/23570 [==============================] - 502s 21ms/sample - loss: 0.0289 - acc: 0.9905 - val_loss: 0.0
Epoch 5/6
23570/23570 [==============================] - 501s 21ms/sample - loss: 0.0234 - acc: 0.9922 - val_loss: 0.0
Epoch 6/6
23570/23570 [==============================] - 502s 21ms/sample - loss: 0.0158 - acc: 0.9945 - val_loss: 0.0
```

The figure shows the predictions of the model are accurate with 0.9930 and a loss of 0.0282 for the first epoch, the accuracy was increased by 0.0029 to 0.9959 accuracy and a loss of 0.0164 the loss was reduced by 0.0118 in the second run (epoch). in the third epoch, the model accuracy increased by 0.0001 to make 0.9960 with a loss of 0.0174. the loss increased by 0.0010. in the fourth epoch, the accuracy increased up to 0.9971 and that was the highest accuracy score with a loss of 0.0100 which is the lowest loss. In the fifth epoch, the accuracy declined to 0.9967 hence accuracy reduced by 0.0004 with an error or loss of 0.0150. in the last epoch, the accuracy increased to 0.9970 with a loss of 0.0110. thus, when the model accuracy predictions increase, the loss or error reduces. The average accuracy score is 0.9960 which is 99% with the average loss of 0.0163 which is 100%. If the average prediction is greater than 0.5, it is regarded as positive otherwise it is Negative. This implies that the prediction is positive for the above results.

Table4. 7. Classification performance evaluation

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5832
1	0.99	0.99	0.99	5393
accuracy			0.99	11225
macro avg	0.99	0.99	0.99	11225
weighted avg	0.99	0.99	0.99	11225

The figure above shows the classification performance evaluation results for a model, typically used in machine learning tasks, such as binary classification. The evaluation metrics are calculated on a dataset with a total of 11,225 samples.

Precision is the ratio of correctly predicted instances of class 0 to the total predicted instances of class 0. In this case, precision for class 0 is 0.99, which means 99% of the instances predicted as class 0 were actually class 0. For class 1, Precision is the ratio of correctly predicted instances of class 1 to the total predicted instances of class 1. In this case, precision for class 1 is also 0.99, indicating that 99% of the instances predicted as class 1 were actually class 1.

Under Recall for class 0: Recall (also called sensitivity or true positive rate) is the ratio of correctly predicted instances of class 0 to the total instances of class 0 in the dataset. Here, recall for class 0 is 0.99, indicating that 99% of the actual instances of class 0 were correctly predicted as class 0. For class 1: Recall for class 1 is also 0.99, indicating that 99% of the actual instances of class 1 were correctly predicted as class 1.

The F1-score is the harmonic mean of precision and recall. It is used to provide a balance between precision and recall when dealing with imbalanced datasets. In this case, the F1-score for both class 0 and class 1 is 0.99, indicating a balanced performance in terms of precision and recall for both classes.

Support represents the number of instances in each class in the dataset. For class 0, the support is 5899, and for class 1, the support is 5326.

Accuracy is the overall performance of the model and represents the ratio of correctly classified instances (both true positives and true negatives) to the total number of instances in the dataset. In this case, the overall accuracy is 0.99 (99%), meaning 99% of the instances in the dataset were correctly classified.

In conclusion, the model achieved high performance with an accuracy of 99%, and it shows excellent precision, recall, and F1-score for both classes, indicating a successful classification task.

4.4.1 Expert review:

This was the second method I used in evaluation of the framework. I approached experts from Kampala international university school of mathematics and computing who have extensive expertise in the field of artificial intelligence, machine learning and big data

analytics to evaluate the designed framework. I also interacted with the debunk team of MCI (media challenge initiative) and the CERT (computer Emergency Response Team) of UCC to validate the framework. The experts reviewed the framework design to provide feedback on its effectiveness and identify any potential gaps or weaknesses. After I presented the framework to the experts, I gathered their feedbacks through interviews and questionnaire. The sample questions were;

- What are your overall impressions of the framework?
- What are the strengths and weaknesses of the framework?
- Are there any important components or features missing from the framework?
- Do you think the framework is appropriate to solve the problem of misinformation in Uganda?
- Are there any potential unintended consequences of implementing the framework?

After the interviews, I compiled the feedback from the experts and analysed them to identify common themes or patterns. I summarized the results of the expert analysis in a report or presentation. Include the experts' feedback, analysis of the feedback, and any recommendations for improving the framework. By using expert analysis to evaluate a misinformation framework design, I was able to gain valuable insights into how the framework could be improved to better address the challenges of misinformation.

CHAPTER FIVE: DISCUSSIONS, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This chapter introduces the study key findings, conclusion to the study and finally the recommendations and future research.

5.1 Discussions of findings

The section summarizes the outcomes that resulted from the study aims as well as how the research questions assisted in achieving the goals. Also included are noteworthy responses to the research questions.

Ordinary social media users being the highest misinformation creators followed by politicians, traditional media and business leaders in the third position, academician being in the last position among the misinformation creators on social media.

The percentage scores for each types of misinformation that is encountered on social media are as follows; Fake news, hate speech, rumors, misleading titles, and satire. This implies that fake news is the most common type of misinformation on social media, followed by hate speech, rumors in the third position, misleading titles in the fourth position, and lastly satire.

Social media takes the lead as the channel that spread misinformation followed by community networks, public events in the third position and finally education publication being in the last position as a channel that spreading misinformation.

The leading social media platform that spreads misinformation is twitter, followed by TikTok, Facebook and WhatsApp in the third position Instagram in the fifth position and YouTube being the social media platform that spreads misinformation the least position.

The topic of misinformation filtering is complicated and always changing; it encompasses many different viewpoints and factors. The ability of automated algorithms to identify and filter out false information is crucial. However, there is ongoing discussion regarding their efficacy. The difficulties and limitations of automated systems, such as algorithmic biases, false positives and negatives, and the need for ongoing improvement, might be discussed.

It is unclear how to strike a compromise between protecting freedom of speech and promoting accuracy when disinformation is filtered. It can be interesting to talk about the moral issues and potential difficulties associated with choosing what information should be censored or marked.

Even while automated systems are useful, critical thinking and human judgment are still necessary for weeding out false information. Discussions regarding how journalists, fact-checkers, and other people verify material and make filtering judgments can reveal the advantages and disadvantages of both human and machine systems.

Collaboration amongst a variety of stakeholders, including technology companies, researchers, fact-checkers, journalists, and users, is necessary for misinformation filtering. The necessity for cooperation and coordination can be emphasized by talking about the value of collaboration, sharing resources and knowledge, and encouraging a group effort to battle misinformation.

The chart shows the dataset that I used to train the model. The news subjects included; news, politics, government news, US news and Middle East news. The largest part was normal news

articles, followed by political news, followed by left-news, then government news and finally a tie of US-news and Middle East news. This implies that there is more false information on normal news and politics that on government news and Middle East.

The predictions of the model are accurate in the way that they are consistent. when the model accuracy predictions increase, the loss or error reduces.

The trainable parameters of the model are updated during training to make the model learn from the data, while the non-trainable parameters remain fixed throughout the training process.

Accuracy is the overall performance of the model and represents the ratio of correctly classified instances (both true positives and true negatives) to the total number of instances in the dataset. In this case, the overall accuracy is (99%), meaning 99% of the instances in the dataset were correctly classified.

In conclusion, the model achieved high performance with an accuracy of 99%, and it shows excellent precision, recall, and F1-score for both classes, indicating a successful classification task.

5.2 Conclusions

Objective 1. The most common type of misinformation is fake news and hate speech according to the findings discussed in chapter four of the study

Objective 2.the main agents that spread misinformation is Ordinally users and politician via social media platforms like TikTok and twitter respectively.

It is impossible to fully eliminate misinformation. This is because computers still rely on the knowledge provided by human beings. Therefore, it takes a multifaceted, collaborative effort to battle misinformation, one that includes technology, human judgment, education, transparency, and ongoing development. By putting these ideas into practice, we can improve our ability to weed out false information and foster a society that is more informed and resilient. Methods for detecting false information should be continuously enhanced and improved. To do this, algorithmic biases must be addressed, false positives and negatives must be reduced, and misinformation strategies must be adapted as they change. Accuracy should be given priority in filtering, but overreach that stifles open discussion and different viewpoints should be avoided.

Objective 4. The model achieved high performance with an accuracy of 99%, and it shows excellent precision, recall, and F1-score for both classes, indicating a successful classification task. This therefore can be used to minimize misinformation on social media.

5.3 Recommendations and future research

5.3.1 Recommendation

Objective 1. Today's social media platforms frequently suffer from misinformation. Although it can't be entirely removed, it can be effectively managed or considerably diminished. In today's information-driven society, controlling misinformation is a crucial and difficult responsibility. However, we can contribute to building a more educated and resilient society by actively taking part in the fight against disinformation. Here are some suggestions for dealing with false information in Uganda.

Objective 1. As the body responsible for overseeing telecommunications in Uganda, UCC ought to enlighten members of the public and users of social media about the value of research, fact-

checking, and trustworthy sources of information. Assist people in learning how to assess the reliability of sources, comprehend biases, and spot false information. Users ought to have the ability to confirm information before sharing it.

Objective 2. Users should alert the platforms they are utilizing of any inaccurate information or deceptive content. Report any false information you come across to the platform's administrators, groups that fact-check content, or the appropriate authorities. This aids in lowering the exposure of such content. The CERT Team, MCI, and other fact-checking teams for news outlets, for instance, should be financially supported and more should be brought on board to combat disinformation in Uganda. Fact-checking organizations also play a critical role in discovering and dispelling false information.

Objective 2. The Governments might establish official social media channels and prepare how to use Digital Media before natural catastrophes to stop the spread of false information (Jayasekara, 2019). Regulators should be strict with those who spread incorrect information (Piazza, 2022). Responsible organizations should implement information regulations and professional ethics to stop the spread of false, detrimental news on social networking sites.

Objective 3. The easiest way to stop the spread of false information is to identify it as soon as possible and conduct targeted counter campaigns.

Objective 4. Similar to other nations like China, Uganda should create a backbone architecture that supports all social media platforms and adopt algorithms that automatically identify, block, or remove content that could be damaging to the public or society.

5.3.2 Future research

For the purpose of creating more potent tactics and technologies to stop the transmission of false or misleading information, future research on misinformation filtering is essential. Here are some prospective study areas to concentrate on in the future:

The remaining features on the framework can be developed and implemented for example the elastic search, the data acquisition control robot, the database and data warehouse.

More complex matching techniques for the entities must be created for use in the future. Some ontology matching methods that concentrate on the clever fusion of many filtering techniques may be used.

Examine real-time detection techniques that can spot and warn against false information as it spreads. This study may involve keeping an eye on news organizations, social media sites, and other internet resources to swiftly spot and eliminate erroneous or misleading material before it gets out of hand.

Future studies must investigate if individuals actually accept fake news as fact as well as whether misinformation disseminated in a pseudo-journalistic manner influences people’s perception differently or even more falsely than misinformation disseminated in non-journalistic formats.

REFERENCES

1. **(**Dr. Žiga 2018) Technology as Enabler of Fake News and a Potential Tool to Combat It.
2. A. Ghenai and Y. Mejova, ‘‘Fake cures: User-centric modeling of health misinformation in social media,’’ Nov. 2018.
3. Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades.. In Proceedings of the 2014 International Conference on Web and Social Media. ICWSM 2014
4. Allcott H and Gentzkow M (2017) Social media and fake news in the 2016 election. Journal of Economic Perspectives 31: 211–236.
5. Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction.
6. Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. ACL’14.
7. Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media, 2017
8. Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. arXiv preprint arXiv:1704.00656, 2017
9. Bechmann A and Nielbo KL (2018) Are we exposed to the same ‘‘news’’ in the News Feed? An empirical analysis of filter bubbles as information similarity for Danish Facebook users. Digital Journalism 6(8): 990–1002.
10. Berinsky AJ (2017) Rumors and health care reform: Experiments in political misinformation. British Journal of Political Science 47(2): 241–262.
11. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web. ACM
12. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter.
13. Chen A (2017) The human toll of protecting the Internet from the worst of humanity. New Yorker, 28 January.
14. Chen Y, Conroy NJ and Rubin VL (2015) Misleading online content: Recognizing clickbait as false news. In: Proceedings of the 2015 ACM on workshop on multimodal deception detection, ACM, pp. 15–19.
15. Ciampaglia GL, Shiralkar P, Rocha LM, et al. (2015) Computational fact checking from knowledge networks. PLoS One 10(6): e0128193.
16. Conroy, S. (2006). Coonan Out Of Touch On Porn Filtering.
17. Conroy, S. (2007a). Labor’s Plan for Cyber-safety. safety.pdf?4.KU86Jun1uWOhBfUk.PYtx_mU5_AXBm 65
18. Conroy, S. (2007b). Labor’s Plan for Cyber Safety. Election.
19. Conroy, S. (2012). Child abuse material blocked online, removing need for legislation [Press release].
20. Constine J (2018) Facebook shrinks fake news after warnings backfire. Tech Crunch, 28 April. Available at: https://tcrn. ch/2jb7gcp (accessed April 24, 2019)
21. Coonan, H. (2007, 10 August). NetAlert: Protecting Australian Families Online. Retrieved from http://www.minister.dcita.gov.au/coonan/media/media_releases/netalert_-_protecting_australian_families_online
22. Crocker D., Hansen T., Kucherawy M. (2011). DomainKeys Identified Mail (DKIM) Signatures. Retrieved from http://tools.ietf.org/rfc/rfc6376.txt
23. DJ Flynn, Brendan Nyhan, and Jason Reifler. 2017. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. Political Psychology 38, S1 (2017), 127–150
24. Ecker UK, Hogan JL and Lewandowsky S (2017) Reminders and repetition of misinformation: Helping or hindering its retraction? Journal of Applied Research in Memory and Cognition 6(2): 185–192.
25. Ellinika Hoaxes Tool (2019) https://www.ellinikahoaxes.gr/. Accessed 15 Nov 2019 Fakespot Analyzer Tool (2019) https://www.fakespot.com/.Accessed 14 Nov 2019
26. Farrel T, Mensio M, Burrel G, Picollo L, Alani H (2018) D3.2 Survey of misinformation detection methods. Co-Inform Project
27. Giovanni, S., & Andrea, L. (2020). The impact of algorithms for online content moderation

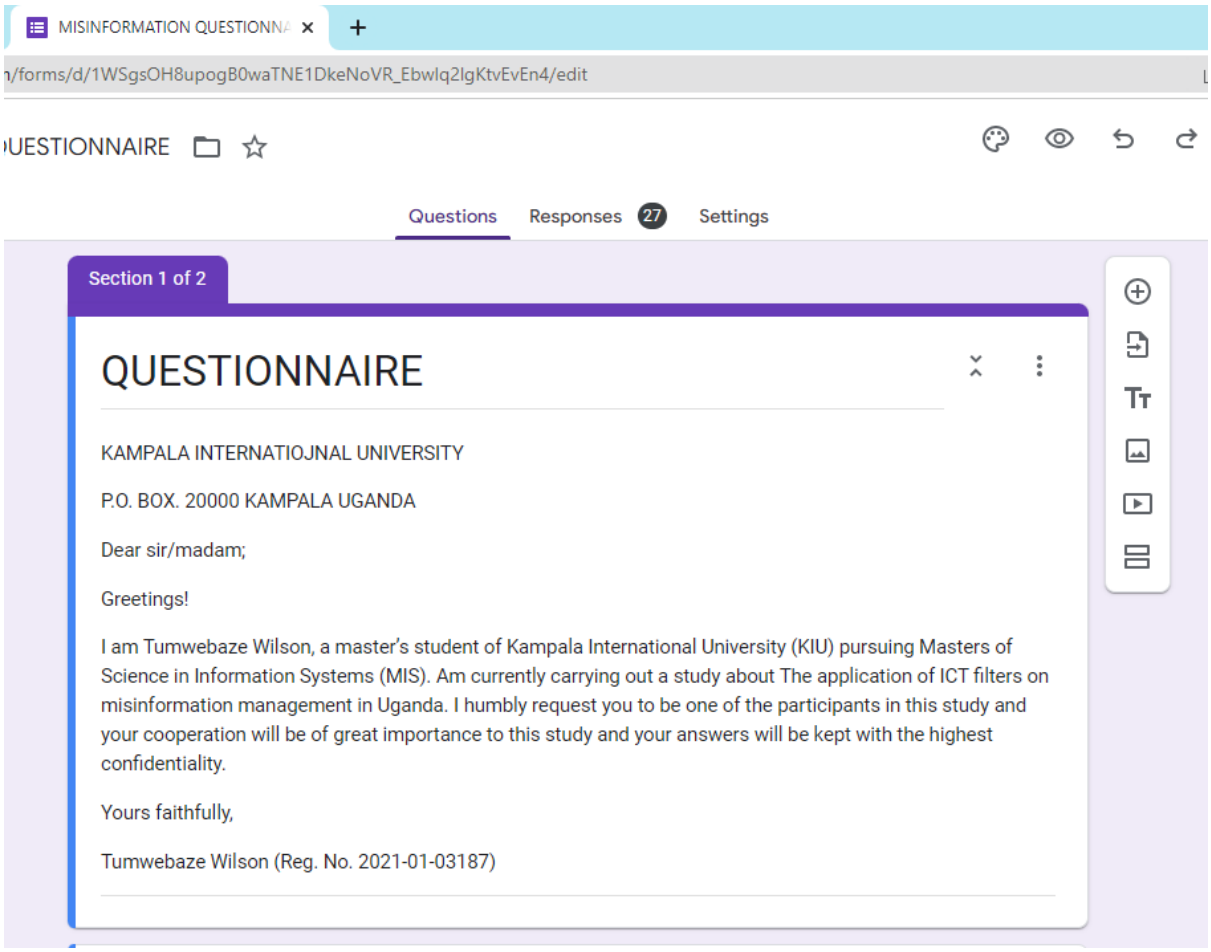
28. Greenhill KM and Oppenheim B (2017) Rumor has it: The adoption of unverified information in conflict zones. International Studies Quarterly 61(3): 660–676.
29. H. Samuel and O. Zaïane, ‘‘MedFact: Towards improving veracity of medical information in social media using applied machine learning,’’ in Canadian Conference on Artificial Intelligence. Toronto, ON, Canada: Springer, 2018, pp. 108–120.
30. Harris E. (2003). The Next Step in the Spam Control War: Greylisting. Retrieved from http://projects.puremagic.com/greylisting/whitepaper.html
31. J. C. S. Reis, A. Correia, and F. Murai, A. Veloso, and F. Benevenuto, ‘‘Supervised learning for fake news detection,’’ IEEE Intell. Syst., vol. 34, no. 2, pp. 76–81, Mar./Apr. 2019.
32. Jaradat I, Gencheva P, Barro´n-Ceden˜o A, et al. (2018) ClaimRank: Detecting check-worthy claims in Arabic and English. In: Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, New Orleans, LA, pp. 26–30.
33. Keele, R. (2010). Nursing research and evidence-based practice. Sudbury, MA: Jones & Bartlett Learning.
34. Klensin J. (2008). Simple Mail Transfer Protocol. Retrieved from http://tools.ietf.org/html/rfc5321
35. Kothari, C. R. (2004). Research methodology: Methods and techniques. New Delhi: New Age International.
36. L Wu, F Morstatter, X Hu, and H Liu. Chapter 5: Mining misinformation in social media, 2016.
37. L. Wu and H. Liu, ‘‘Tracing fake-news footprints: Characterizing social media messages by how they propagate,’’ in Proc. 11th ACM Int. Conf. Web Search Data Mining, New York, NY, USA, 2018, pp. 637–645, doi: 10.1145/3159652.3159677
38. Lazer D, Baum M, Grinberg N, et al. (2017) Combating fake news: An agenda for research and action. Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy, 2 May.
39. Lazer D, Baum MA, Benkler Y, et al. (2018) The science of fake news. Science 359(6380): 1094–1096.
40. Levine J. (2010). DNS Blacklists and Whitelists. Retrieved from http://tools.ietf.org/rfc/rfc5782.txt
41. Levine J. (2010). DNS Blacklists and Whitelists. Retrieved from http://tools.ietf.org/rfc/rfc5782.txt
42. Lewandowsky S, Ecker UK, Seifert CM, et al. (2012) Misinformation and its correction: Continued influence 12 Big Data & Society and successful debiasing. Psychological Science in the Public Interest 13(3): 106–131.
43. Lieven P. (2006). Pre-MX Spam Filtering with Adaptive Greylisting Based on Retry Patterns. Heinrich-Heine-Universität Düsseldorf.
44. Lyon J. (2006). Purported Responsible Address in E-mail Messages. Retrieved from http://tools.ietf.org/rfc/rfc4407.txt
45. Lyon J., Wong M. (2006). Sender ID: Authenticating E-mail. Retrieved from http://tools.ietf.org/rfc/rfc4406.txt
46. Matatov H, Bechhofer A, Aroyo L, Ofr, A, Naaman M (2018) DejaVu: a system for journalists to collaboratively address visual misinformation. In: Computation + Journalism Symposium. Miami
47. Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter
48. Michael D Cobb, Brendan Nyhan, and Jason Reifler. 2013. Beliefs don’t always persevere: How political figures are punished when positive information about them is discredited. Political Psychology 34, 3 (2013)
49. Middleton SE (2017) Reveal project deliverable D5.2.2-modality models for trust and credibility, https://revealproject.eu/wp-content/uploads/D5.2.2- Modality-models-for-trust-and-credibility_PU.pdf. Accessed 12 Nov 2019
50. Mohtarami M, Baly R, Glass J, et al. (2018) Automatic stance detection using end-to-end memory networks. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long papers), New Orleans, LA, pp. 767–776
51. Mustafaraj E and Metaxas PT (2017) The fake news spreading plague: Was it preventable? In: Proceedings of the 2017 ACM on web science conference, ACM, pp. 235–239.
52. N. Ruchansky, S. Seo, and Y. Liu, ‘‘Csi: A hybrid deep model for fake news detection,’’ in Proc. ACM Conf. Inf. Knowl. Manage., Nov. 2017, pp. 797–806.

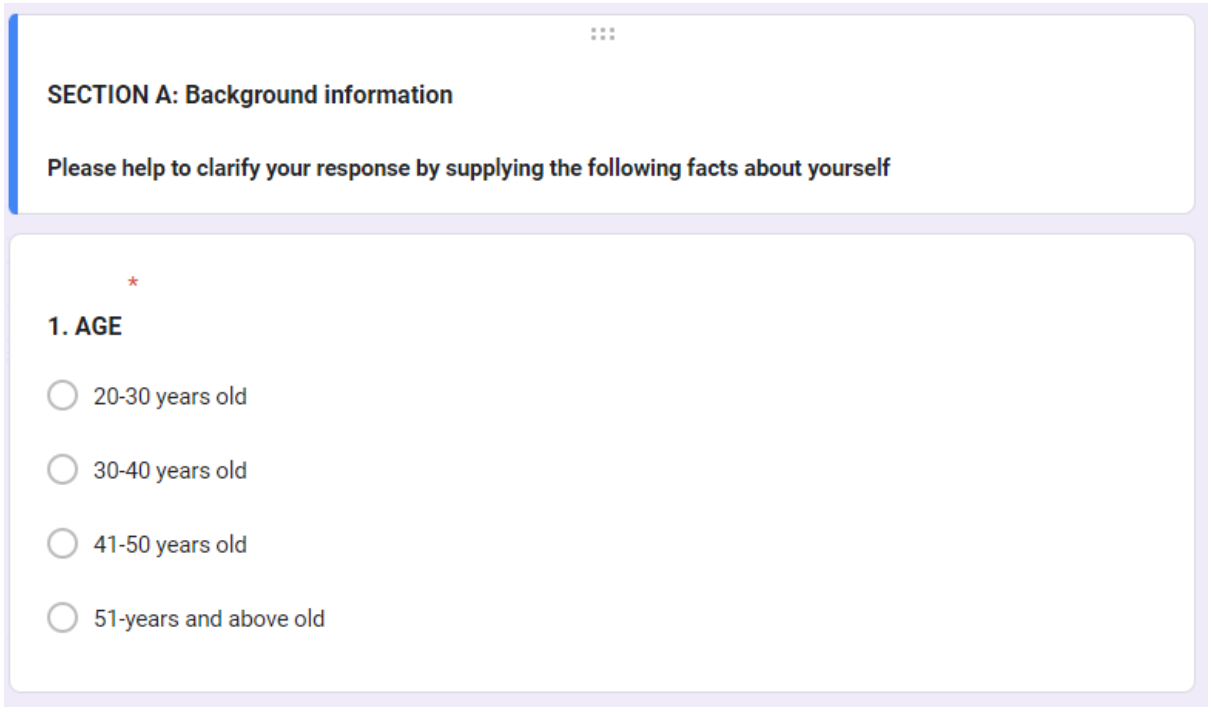
53. NewsGuard Tool (2019) https://www.newsguardtech.com/. Accessed 14 Nov 2019 Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Springer
54. of Domains in E-mail. Retrieved from http://tools.ietf.org/rfc/rfc4408.txt.
55. Potthast M, Kiesel J, Reinartz K, et al. (2018) A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: long papers), Melbourne, Australia, pp. 231–240.
56. S. Dhoju, M. M. U. Rony, M. A. Kabir, and N. Hassan, ‘‘Differences in health news from reliable and unreliable media,’’ in Proc. Companion World Wide Web Conf., New York, NY, USA, vol. 4, May 2019, pp. 981–987.
57. S. Kwon and M. Cha. Modeling bursty temporal pattern of rumors. In ICWSM, 2014.
58. S. Tschiatschek, A. Singla, M. G. Rodriguez, A. Merchant, and A. Krause, ‘‘Fake news detection in social networks via crowd signals,’’ in Proc. Companion Web Conf., Apr. 2018, pp. 517–524.
59. Stephens, Bret. 2017. The President Versus ‘Fake News,’ Again. URL accessed 30 August 2017:
60. S. Yu, M. Li, and F. Liu. Rumor identification with maximum entropy in micronet. Complexity, 2017, 2017.
61. Sapsford, R., & Jupp, V. (2006). Data collection and analysis. London: SAGE Publications Ltd.
62. Schifferes S, Newman N, Thurman N, Corney D, Göker A, Martin C (2014) Identifying and verifying news through social media. Digi Journalism 2(3):406–418
63. Sloan L, Quan-Haase A (2016) The SAGE Handbook of Social Media Research Methods. SAGE Publications Ltd, London
64. Treviño A., Ekstrom J. (2007). Spam Filtering Through Header Relay Detection. Retrieved from http://mel.byu.edu/spam/Trevino_EkstromRelay_Header_Analysis.pdf
65. V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, ‘‘Fake news or truth? using satirical cues to detect potentially misleading news,’’ in Proc. 2nd Workshop Comput. Approaches Deception Detection, Jun. 2016, pp. 7–17.
66. Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,
67. Vosoughi S, Roy D and Aral S (2018) The spread of true and false news online. Science 359(6380): 1146–1151.
68. Woods, M., Anderson, J., Guilbert, S., & Watkin, S. (2012). “The country (side) is angry”: Emotion and explanation in protest mobilization. Social & Cultural Geography, 13(6), 567–585.
69. Wong M., Schlitt W. (2006). Sender Policy Framework (SPF) for Authorizing Use Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In ICWSM, 2014.
70. Y. Long, Q. Lu, R. Xiang, M. Li, and C. Huang, ‘‘Fake news detection through multi-perspective speaker profiles,’’ in Proc. 8th Int. Joint Conf. Natural Lang. Process., vol. 2, Nov. 2017, pp. 252–256.
71. Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, ‘‘EANN: Event adversarial neural networks for multi-modal fake news detection,’’ in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2018, pp. 849–857.
72. Yang K, Varo O, Davis CA, Ferrara E, Flammini A, Menczer F (2019) Arming the public with artificial intelligence to counter social bots. arXiv.org e-Print archive. https://arxiv.org/. Retrieved 26 March 2020
73. Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs.
74. S. Lewandowsky, J. Cook, U. K. H. Ecker, D. Albarracín, M. A. Amazeen, P. Kendeou, D. Lombardi, E. J. Newman, G. Pennycook, E. Porter, D. G. Rand, D. N. Rapp, J. Reifler, J. Roozenbeek, P. Schmid, C. M. Seifert, G. M. Sinatra, B. Swire-Thompson, S. van der Linden, E. K. Vra
75. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The science of fake news. Science (80-.). (2018), doi:10.1126/science.aao2998.
76. S. van der Linden, E. Maibach, J. Cook, A. Leiserowitz, S. Lewandowsky, Inoculating against misinformation. Science (80-.). 358, 1141–1142 (2017)

77. Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions. Digital Journalism, 6(2), 154–175.
78. Lecheler, S., & Kruikemeier, S. (2016). Re-evaluating journalistic routines in a digital age: A review of research on the use of online sources. New Media & Society, 18(1),
79. Lecheler, S., Schuck, A., & De Vreese, C. (2013). Dealing with feelings: Positive and negative discrete emotions as mediators of news framing effects. Communications - The European Journal of Communication Research, 38(2), 189–209
80. Carey JM, Chi V, Flynn DJ, Nyhan B, Zeitzof T. The efects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. Sci Adv. 2020;6(5):7449. https://doi.org/10.1126/ sciadv.aaw7449
81. Bahrami MA, Nasiriani Kh, Dehghani A, Zarezade M, Kiani P. Counteracting online health misinformation: a qualitative study. Manage Strat Health Syst. 2019;4(3):230–9. https://doi.org/10.18502/mshsj.v4i3.2056.
82. Gesser-Edelsburg A, Diamant A, Hijazi R, Mesch GS. Correcting misinformation by health organizations during measles outbreaks: a controlled experiment. PLoS ONE. 2018;13(12):e0209505. https://doi.org/10.1371/ journal.pone.0209505
83. HLEG. (2018). A multi-dimensional approach to disinformation. Retrieved from https://ec.europa.eu/digital-single-market/ en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation
84. Flynn, D. J., Reifler, J., & Nyhan, B. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. Political Psychology, 38(S1), 127–150.
85. J. Roozenbeek, S. van der Linden, T. Nygren, Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. Harvard Kennedy Sch. Misinformation Rev. 1 (2020), doi:10.37016//mr-2020-008
86. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. Science (80-.). 359, 1146–1151 (2018).
87. M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The COVID-19 social media infodemic. Sci. Rep. 10, 16598 (2020). 36. U. K. H. Ecker, S.
88. Lewandowsky, M. Chadwick, Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. Cogn. Res. Princ. Implic. 5, 41 (2020). 37. S. Lewandowsky, U. K. H.
89. Ecker, C. M. Seifert, N. Schwarz, J. Cook, Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychol. Sci. Public Interes. 13, 106–131 (2012).
90. L. Fazio, N. M. Brashier, B. K. Payne, E. J. Marsh, Knowledge does not protect against illusory truth. J. Exp. Psychol. Gen. 144, 993–1002 (2015).
91. Wardle, C. (2017). Fake news. It's complicated. Retrieved from https://medium.com/1st-draft/fake-news-its-complicatedd0f773766c79
92. Tandoc, E. C. J., Lim, Z. W., & Ling, R. (2018). Defining "fake news". Digital Journalism, 6(2), 137–153.
93. Oxford Dictionaries. (2019). Fake. Retrieved from https://en.oxforddictionaries.com/definition/fake
94. Horne, B. D., & Adalı, S. (2017). This just in : Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Retrieved from https://arxiv.org/abs/1703.09398
95. Khaldarova, I., & Pantti, M. (2016). Fake news. Journalism Practice, 10(7), 891–901. Krämer, B. (2018). How journalism responds to right-wing populist criticism. In K. Otto & A. Köhler (Eds.), Trust in media and journalism (pp. 137–154). Wiesbaden: Springer VS.
96. Mustafaraj, E., & Metaxas, P. T. (2017). The fake news spreading plague: Was it preventable? Retrieved from http://arxiv.org/ abs/1703.06988
97. Tandoc, E. C. J., Lim, Z. W., & Ling, R. (2018). Defining "fake news". Digital Journalism, 6(2), 137–153. Thomas, D. (2017, April 6). Facebook to tackle fake news with educational campaign. BBC News.
98. McNair, B. (2017). Fake news: Falsehood, fabrication and fantasy in journalism. New York: Routledge.

99. Bale, J. M. (2007). Political realism: On distinguishing between bogus conspiracy theories and genuine conspiratorial politics. Patterns of Prejudice, 41(1), 45–60.
100. Douglas, K., Ang, C. S., & Deravi, F. (2017). Farewell to truth? Conspiracy theories and fake news on social media. The Psychologist, 30, 36–42.
101. Batchelor, O. Getting out the truth: The role of libraries in the fight against fake news. Ref. Serv. Rev. 2017, 45, 143–148.
102. Ahmed, W.; Lugovic, S. Social media analytics: Analysis and visualisation of news diffusion using NodeXL. Online Inf. Rev. 2019, 43, 149–160.
103. Wasike, J. Social media ethical issues: Role of a librarian. Libr. Hi Tech News 2013, 30, 8–16.
104. Gimpel, H.; Heger, S.; Olenberger, C.; Utz, L. The Effectiveness of Social Norms in Fighting Fake News on Social Media. J. Manag. Inf. Syst. 2021, 38, 196–221
105. Auberry, K. Increasing students' ability to identify fake news through information literacy education and content management systems. Ref. Libr. 2018, 59, 179–187.
106. Talwar, S.; Dhir, A.; Singh, D.; Virk, G.S.; Salo, J. Sharing of fake news on social media: Application of the honeycomb framework and the third-person effect hypothesis. J. Retail. Consum. Serv. 2020, 57, 102197.
107. Kim, A.; Moravec, P.L.; Dennis, A.R. Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings. J. Manag. Inf. Syst. 2019, 36, 931–968.
108. Razaghi, S.; Shokouhyar, S. Impacts of big data analytics management capabilities and supply chain integration on global sourcing: A survey on firm performance. Bottom Line 2021, 34, 198–223.
109. Lei, H. Modern information warfare: Analysis and policy recommendations. Foresight 2019, 21, 508–522.
110. Jayasekara, P.K. Role of Facebook as a disaster communication media. Int. J. Emerg. Serv. 2019, 8, 191–204
111. Piazza, J.A. Fake news: The effects of social media disinformation on domestic terrorism. Dyn. Asymmetric Confl. 2022, 15, 55–77
112. Si¸sman, B.; Yurtta¸s, U. An Empirical Study on Media Literacy from the Viewpoint of Media. Procedia-Soc. Behav. Sci. 2015, 174, 798–804.
113. Schuetz, S.W.; Sykes, T.A.; Venkatesh, V. Combating COVID-19 fake news on social media through fact checking: Antecedents and consequences. Eur. J. Inf. Syst. 2021, 30, 376–388
114. Fernandez, P. The technology behind fake news. Libr. Hi Tech News 2017, 34, 1–5.
115. Sullivan, M.C. Why librarians can't fight fake news. J. Libr. Inf. Sci. 2019, 51, 1146–1156
116. William Yang Wang. 2017. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection.
117. Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates.
118. Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims.
119. Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news.
120. Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news.
121. Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking.
122. Stephan Lewandowsky, Ullrich K.H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing.
123. Cathy O'Neil. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.
124. Zachary C. Lipton. 2018. The mythos of model interpretability.
125. gi
126. Boididou C, Andreadou K, Papadopoulos S, Dang-Nguyen DT, Boato G, Riegler M, Kompatsiaris Y (2015) Verifying multimedia use at MediaEval 2015.

APENDIX I: SAMPLE QUEUSTIONNAIRE





2. NATIONALITY *

Short answer text

3. JOB POSITION/TITLE *

Short answer text

After section 1 Continue to next section

SECTION B: QUESTIONS ON MISINFORMATION

Instruction: This section contains items on sources of misinformation, types of misinformation and where we find misinformation. Please rank the following statements according to your View in Ascending order

4. Who creates misinformation? *

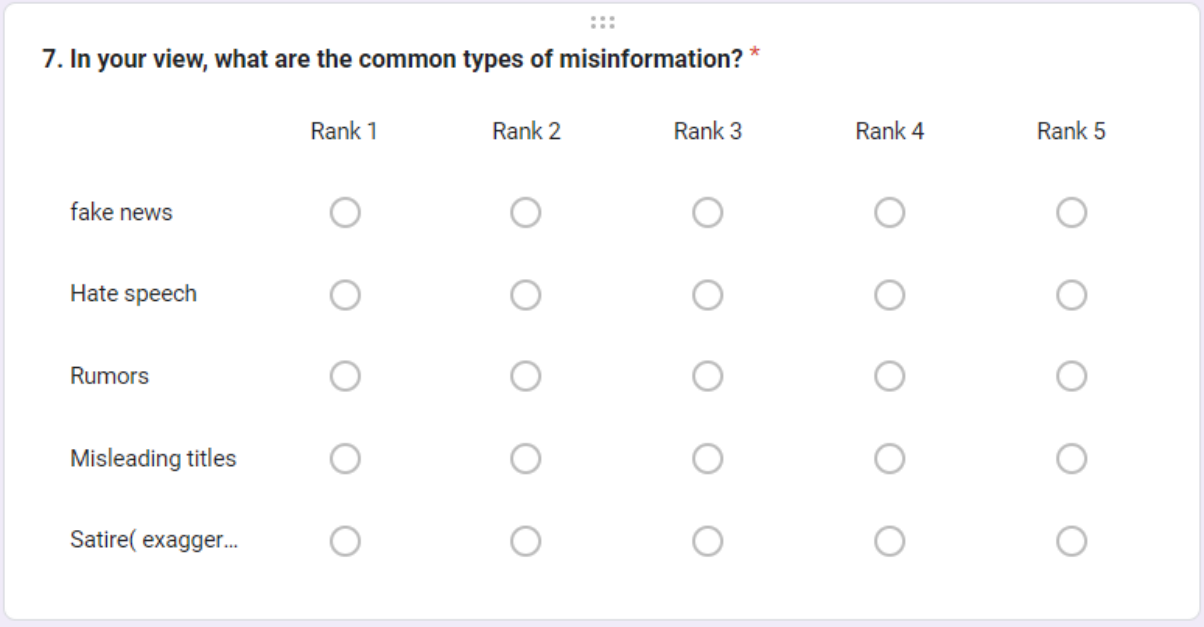
	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Traditional me...	○	○	○	○	○
politicians	○	○	○	○	○
Academicians	○	○	○	○	○
Business leade...	○	○	○	○	○

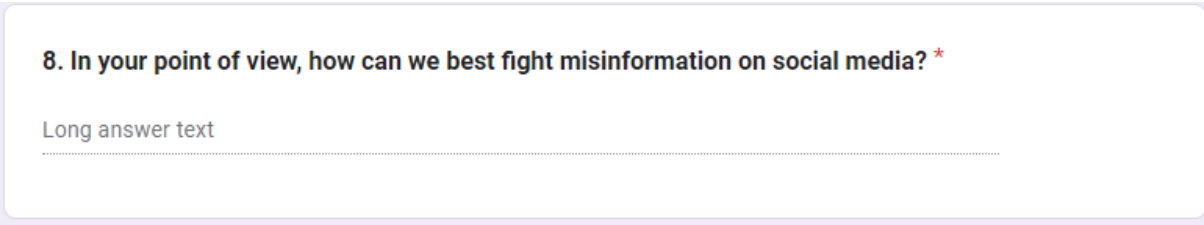
5.What are the platforms via which misinformation is mostly spread? *

	Rank 1	Rank 2	Rank 3	Rank 4
Social media platf...	○	○	○	○
Public events	○	○	○	○
Education publicat...	○	○	○	○
Community Netwo...	○	○	○	○

6. What social media platform spreads misinformation most? *

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
Facebook	○	○	○	○	○	○
Twitter	○	○	○	○	○	○
Whatsapp	○	○	○	○	○	○
Youtube	○	○	○	○	○	○
Instagram	○	○	○	○	○	○
Tiktok	○	○	○	○	○	○





APENDIX II: NLP CODE

IMPORTING LIBRARIES

```
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
import re
from wordcloud import WordCloud

from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Embedding, LSTM, Conv1D, MaxPool1D
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score
```

Exploring Fake News

```
fake = pd.read_csv("/kaggle/input/fake-and-real-news-dataset/Fake.csv")

fake.head()

#Counting by Subjects
for key,count in fake.subject.value_counts().iteritems():
```

```
    print(f"{key}:\t{count}")

#Getting Total Rows
print(f"Total Records:\t{fake.shape[0]}")

plt.figure(figsize=(8,5))
sns.countplot("subject", data=fake)
plt.show()

#Word Cloud
text = ''
for news in fake.text.values:
    text += f" {news}"
wordcloud = WordCloud(
    width = 3000,
    height = 2000,
    background_color = 'black',
    stopwords = set(nltk.corpus.stopwords.words("english"))).generate(text)
fig = plt.figure(
    figsize = (40, 30),
    facecolor = 'k',
    edgecolor = 'k')
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
del text
```

Exploring Real news

```
real = pd.read_csv("/kaggle/input/fake-and-real-news-dataset/True.csv")
real.head()

#First Creating list of index that do not have publication part
unknown_publishers = []
for index,row in enumerate(real.text.values):
    try:
        record = row.split(" -", maxsplit=1)
        #if no text part is present, following will give error
        record[1]
        #if len of piblication part is greater than 260
        #following will give error, ensuring no text having "-" in between is counted
        assert(len(record[0]) < 260)
    except:
        unknown_publishers.append(index)

#Thus we have list of indices where publisher is not mentioned
#lets check
real.iloc[unknown_publishers].text
real.iloc[8970]

#Seperating Publication info, from actual text
publisher = []
tmp_text = []
for index,row in enumerate(real.text.values):
    if index in unknown_publishers:
```

```
        #Add unknown of publisher not mentioned
        tmp_text.append(row)

        publisher.append("Unknown")
        continue
    record = row.split(" -", maxsplit=1)
    publisher.append(record[0])
    tmp_text.append(record[1])

#Replace existing text column with new text
#add seperate column for publication info
real["publisher"] = publisher
real["text"] = tmp_text

del publisher, tmp_text, record, unknown_publishers

real.head()
```

New column called "Publisher" has been added.

```
#checking for rows with empty text like row:8970
[index for index,text in enumerate(real.text.values) if str(text).strip() == '']
#seems only one :)

#dropping this record
real = real.drop(8970, axis=0)

# checking for the same in fake news
empty_fake_index = [index for index,text in enumerate(fake.text.values) if str(text).strip() == '']
print(f"No of empty rows: {len(empty_fake_index)}")
fake.iloc[empty_fake_index].tail()

#Looking at publication Information
# Checking if Some part of text has been included as publisher info... No such cases it seems :)

# for name,count in real.publisher.value_counts().iteritems():
#     print(f"Name: {name}\nCount: {count}\n")

#Getting Total Rows
print(f"Total Records:\t{real.shape[0]}")

#Counting by Subjects
for key,count in real.subject.value_counts().iteritems():
  print(f"{key}:\t{count}")
```

```
Total Records:  21416
politicsNews:   11271
worldnews:      10145
```

```
sns.countplot(x="subject", data=real)
plt.show()

#WordCloud For Real News
text = ''
for news in real.text.values:
    text += f" {news}"
wordcloud = WordCloud(
    width = 3000,
    height = 2000,
```

```
    background_color = 'black',
    stopwords = set(nltk.corpus.stopwords.words("english"))).generate(str(text))
fig = plt.figure(
    figsize = (40, 30),
    facecolor = 'k',
    edgecolor = 'k')
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
del text
```

Preprocessing Text

```
# Adding class Information
real["class"] = 1
fake["class"] = 0

#Combining Title and Text
real["text"] = real["title"] + " " + real["text"]
fake["text"] = fake["title"] + " " + fake["text"]

# Subject is diffrent for real and fake thus dropping it
# Aldo dropping Date, title and Publication Info of real
real = real.drop(["subject", "date","title",  "publisher"], axis=1)
fake = fake.drop(["subject", "date", "title"], axis=1)

#Combining both into new dataframe
data = real.append(fake, ignore_index=True)
del real, fake

# Download following if not downloaded in local machine


# nltk.download('stopwords')
# nltk.download('punkt')
```

Removing StopWords, Punctuations and single-character words

```
y = data["class"].values
#Converting X to format acceptable by gensim, removing annd punctuation stopwords in the process
X = []
stop_words = set(nltk.corpus.stopwords.words("english"))
tokenizer = nltk.tokenize.RegexpTokenizer(r'\w+')
for par in data["text"].values:
    tmp = []
    sentences = nltk.sent_tokenize(par)
    for sent in sentences:
        sent = sent.lower()
        tokens = tokenizer.tokenize(sent)
        filtered_words = [w.strip() for w in tokens if w not in stop_words and len(w) > 1]
        tmp.extend(filtered_words)
    X.append(tmp)

del data

import gensim
```

```
#Dimension of vectors we are generating
EMBEDDING_DIM = 100

#Creating Word Vectors by Word2Vec Method (takes time...)
w2v_model = gensim.models.Word2Vec(sentences=X, size=EMBEDDING_DIM, window=5, min_c
ount=1)

#vocab size
len(w2v_model.wv.vocab)

#We have now represented each of 122248 words by a 100dim vector.
```

Exploring Vectors

```
#see a sample vector for random word, lets say Corona
w2v_model["corona"[33]

w2v_model.wv.most_similar("uganda")

# Tokenizing Text -> Repsesenting each word by a number
# Mapping of orginal word to number is preserved in word_index property of tokenizer

#Tokenized applies basic processing like changing it yo lower case, explicitely setting that as False
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X)

X = tokenizer.texts_to_sequences(X)

# lets check the first 10 words of first news
#every word has been represented with a number
X[0][:10]

#Lets check few word to numerical replesentation
#Mapping is preserved in dictionary -> word_index property of instance
word_index = tokenizer.word_index
for word, num in word_index.items():
    print(f"{word} -> {num}")
    if num == 10:
        break

# For determining size of input...

# Making histogram for no of words in news shows that most news article are under 700 words.
# Lets keep each news small and truncate all news to 700 while tokenizing
plt.hist([len(x) for x in X], bins=500)
plt.show()

# Its heavily skewed. There are news with 5000 words? Lets truncate these outliers :)

nos = np.array([len(x) for x in X])
len(nos[nos  < 700])
# Out of 48k news, 44k have less than 700 words

#Lets keep all news to 700, add padding to news with less than 700 words and truncating long ones
maxlen = 700

#Making all news of size maxlen defined above
X = pad_sequences(X, maxlen=maxlen)
```

```
#all news has 700 words (in numerical form now). If they had less words, they have been padded with
0
# 0 is not associated to any word, as mapping of words started from 1
# 0 will also be used later, if unknows word is encountered in test set
len(X[0])

# Adding 1 because of reserved 0 index
# Embedding Layer creates one more vector for "UNKNOWN" words, or padded words (0s). This Vec
tor is filled with zeros.
# Thus our vocab size inceeases by 1
vocab_size = len(tokenizer.word_index) + 1

# Function to create weight matrix from word2vec gensim model
def get_weight_matrix(model, vocab):
    # total vocabulary size plus 0 for unknown words
    vocab_size = len(vocab) + 1
    # define weight matrix dimensions with all 0
    weight_matrix = np.zeros((vocab_size, EMBEDDING_DIM))
    # step vocab, store vectors using the Tokenizer's integer mapping
    for word, i in vocab.items():
        weight_matrix[i] = model[word]
    return weight_matrix

#Getting embedding vectors from word2vec and usings it as weights of non-
trainable keras embedding layer
embedding_vectors = get_weight_matrix(w2v_model, word_index)

#Defining Neural Network
model = Sequential()
#Non-trainable embeddidng layer
model.add(Embedding(vocab_size, output_dim=EMBEDDING_DIM, weights=[embedding_vectors],
 input_length=maxlen, trainable=False))
#LSTM
model.add(LSTM(units=128))
model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])

del embedding_vectors

model.summary()

#Train test split
X_train, X_test, y_train, y_test = train_test_split(X, y)

model.fit(X_train, y_train, validation_split=0.3, epochs=6)

#Prediction is in probability of news being real, so converting into classes
# Class 0 (Fake) if predicted prob < 0.5, else class 1 (Real)
y_pred = (model.predict(X_test) >= 0.5).astype("int")

accuracy_score(y_test, y_pred)

print(classification_report(y_test, y_pred))

del model
```