Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Detecting Manipulation on Social Media Platforms

Holly McEvoy

19334663

April 17, 2023

A Final Year Project submitted in partial fulfilment of the requirements for the degree of BA (Computer Science)

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Holly McEvoy

April 17th, 2023

Abstract

This paper investigates the possibility of manipulation on social media platforms, focusing on Twitter and its owner, Elon Musk. With the growing influence of social media on public opinion, it is crucial to ensure the impartiality and accuracy of information shared on these platforms. This study designs and implements a tweet monitoring and sentiment analysis system, called 'Doubtfire', to detect potential manipulation by analysing trends in deleted tweets in real-time. The monitor focuses on tweets containing keywords related to Elon Musk and his inventions, such as Tesla and SpaceX, to gather a comprehensive data sample. The results of the study provide valuable insights into the extent of Elon Musk's potential influence on Twitter and shed light on the importance of impartiality and integrity in social media information sharing.

Acknowledgements

I would like to express my gratitude to my supervisor, Mr. Owen Conlan, and PhD candidate, Mr. Dipto Barman, for their invaluable guidance and support throughout the duration of my final year project. Their contributions have made this project an exceptionally gratifying and captivating one.

I am deeply thankful to my parents for their unwavering encouragement and backing over the course of the last 22 years. Their unwavering support has helped me through this course.

I would also like to extend my appreciation to my siblings, Emma, and Rory, for all the advice they have given me.

Furthermore, I would like to acknowledge my brother, Jamie, for helping and supporting me throughout the past four years.

I am also grateful to the friends I have made throughout my time at Trinity for fostering a positive and enjoyable atmosphere.

Lastly, I want to thank Bread 41 for fuelling me with caffeine throughout this project and for all the coffee breaks we took there.

Contents

Table of Contents

List of Figures

1. Introduction

1.1. Research Motivation

In a world where information is widely available and easily accessible, it is becoming increasingly difficult to distinguish between credible and unreliable sources. Fake News and misinformation can be spread rapidly through social media and other channels, leading to confusion, misinformation, and even harm to individuals and society. This project is motivated by the need to address this pressing issue and to contribute to the development of solutions that will help promote transparency, accuracy, and truth in information dissemination.

The term 'Fake News' has gained widespread recognition and has become a frequent topic of discussion, particularly after its usage during the 2016 American presidential election by former President Donald Trump. This term refers to a wide range of false or misleading information spread deliberately or unintentionally through various media sources (Allcott and Gentzkow, 2016). Fake News has been shown to have a greater impact on individual’s opinions and beliefs compared to accurate information, even when the misinformation is later corrected. This phenomenon is known as the "illusory truth effect" and it highlights the power of repetition in shaping people's perceptions and reinforcing false beliefs (Guess and Lyons, 2020). Research has shown that Fake News can have significant negative impacts on individuals and society, leading to the need for a deeper understanding and effective strategies to address this growing problem.

The COVID-19 pandemic led to a surge in the spread of Fake News and increased social media usage. As people sought to stay informed about the rapidly evolving situation, social media platforms became a primary source of information. However, the abundance of false information and misinformation circulating on these platforms added to the confusion and fear (Germani and Biller-Andorno, 2021). The ease of spreading information on social media, combined with the desire for likes and engagement, led to the spread of false claims, conspiracy theories, and misleading information (Shahi, Dirkson and Majchrzak, 2021). This not only created panic among the public but also hindered efforts to control the spread of the virus. The pandemic has emphasised the importance of fact-checking and verified information, highlighting the need for individuals to be critical consumers of information, especially during times of crisis.

Twitter is a major player in the social media landscape, with a global user base of over 396.5 million people (Shepherd, 2022). However, this platform has also been identified as a significant contributor to the spread of Fake News. Studies have shown that false information on Twitter can travel faster and farther than accurate information (Dizikes, 2018), hence it being a focus point of this study. The COVID-19 pandemic resulted in a marked increase in Twitter usage (Haman, 2020). People around the world sought up-to-date information and connections during a challenging and uncertain time, and Twitter proved to be a valuable resource. Twitter was used by health organisations, governments, journalists, and individuals alike to share information, exchange thoughts and experiences, and keep the public informed. The platform saw a substantial rise in its daily user base and a substantial increase in the volume of content being generated and consumed. The pandemic accentuated the vital role that social media, particularly Twitter, can play in times of crisis, by bringing people together and ensuring the dissemination of accurate and timely information.

On October 27, 2022, Elon Musk's acquisition of Twitter for $44 billion made him the company's owner and CEO (Southern, 2022). The acquisition faced widespread criticism, as the public and Twitter users raised concerns about the potential impact on freedom of expression. Some employees of Twitter staged a walkout and resigned, expressing their concerns about Musk's leadership style and his approach to running the company. News reports also emerged about Musk's employees sleeping in their offices, indicating tight deadlines and a new demanding work environment has been imposed by the new CEO (Barrabi, 2022). The acquisition sparked discussions about the potential for abuse of power and the impact it may have on free speech on Twitter. The concerns raised by the public and Twitter employees highlight the importance of protecting user rights and freedom of expression. This is especially relevant in the ownership and operation of social media platforms.

The acquisition of Twitter by Elon Musk and the resulting concerns about his leadership and the potential abuse of power were the focus of this project. This project examines the ways in which Musk's control over Twitter could be used to manipulate the information and content that users see on the platform, and the potential impact on freedom of expression and user rights.

1.2. Research Question

This research question aims to explore the potential impact on free speech and the rights of Twitter users due to the possible manipulation of content on the platform. The research question explored in this project is:

'Is it possible to detect manipulation on Twitter by analysing trends in deleted tweets?'

This project aims to shed light on the issue of manipulation on Twitter and raise awareness about the potential for false information to be spread on the platform. By calling out manipulation on Twitter, the project aims to promote critical thinking and encourage users to question the authenticity of the information they encounter online. The goal is to raise awareness about the need for fact-checking and to encourage users to seek out reliable sources of information. This project also aims to highlight the importance of protecting user rights and freedom of expression in the ownership and operation of social media platforms (Vese, 2021).

1.3. Objectives

The research question can be broken down into six objectives:

O1 - Examining the Phenomenon of Fake News and Parody.

- ❖ This will be done by examining previous research and literature.

O2 - Investigating Censorship and Manipulation on Social Media.

- ❖ This will be done by reviewing the available literature on censorship and manipulation in social media.

O3 - Developing a Monitoring System to Identify Manipulation.

- ❖ This will be done using Python and the library Twint.

O4 - Analysing Deleted Tweets for Trends and Patterns of Manipulation.

- ❖ This will be done using sentiment analysis.

O5 - Evaluating the Extent of Manipulation on Twitter.

- ❖ This will be done by evaluating the results of the previous objective.

O6 - Identifying Opportunities for Further Research.

- ❖ This will be done by evaluating the research results and identifying further steps that can be taken to combat and detect manipulation on social media.

1.4. Approach

The following approaches were taken to conduct this project.

1.4.1. Research

Cross Sectional Research

The research method employed in this study was cross-sectional research. This type of research involves collecting data from a sample of individuals at one point in time (Setia, 2016). It is an observational study that allows for the comparison of differences across demographic variables such as age, gender, etc.

In this study, cross-sectional research was chosen because it was essential to examine how many tweets related to Elon Musk were being deleted at a single point in time. This design allowed for a comprehensive examination of the issue, providing a snapshot of the data and insights into potential patterns or trends that may indicate manipulation.

The importance of cross-sectional research lies in its ability to provide a comprehensive understanding of a particular issue or phenomenon at a specific point in time. This design is useful for answering questions about the prevalence or frequency of a particular phenomenon and for identifying differences across demographic variables. In the context of this study, cross-sectional research was critical in providing a thorough examination of the extent to which negative tweets are being deleted, and it allowed for a more accurate understanding of the issue at hand.

1.4.2. Technical

The monitor for detecting deleted tweets in the ‘Doubtfire’ system was built using the Python programming language and imported the Twint library. Python is a high-level, interpreted programming language that is widely used for various applications, including data analysis and web development. The Twint library, on the other hand, is an advanced Twitter scraping tool that allows for the extraction of data from Twitter without using Twitter's API.

The use of Python and Twint enabled the creation of a robust and efficient monitor for detecting deleted tweets. The Twint library provided the necessary functions for scraping Twitter data, including the ability to search for specific tweets, extract tweet information, and filter the results based on specific criteria. The combination of Python and Twint allowed for the automation of the data collection process, making it

easier to collect substantial amounts of data in a short amount of time. Twint was also particularly important in this project as it does not make use of Twitter's API, making it less affected by any decision Musk made.

In this project, the monitor was programmed to search for tweets containing keywords related to Elon Musk and his personal beliefs and close associations. The monitor would then extract information about the tweets, including the tweet text, creation date, and user information. The extracted data would then be analysed to determine if there were any tweets that had been deleted and to assess the sentiment of those deleted tweets.

Moreover, sentiment analysis was performed using Python, using the data received from the monitor and the collection of tweets scraped by Twint. Vader was used to ensure that sarcasm and irony were considered. This allowed for the analysis of tweets and the categorisation of them into negative, positive, and neutral categories based on their sentiment.

The use of Python, Vader and Twint in this project allowed for the creation of a reliable and efficient monitor and sentiment analysis for detecting trends in deleted tweets. The ability to automate the data collection allowed for the examination of a large amount of data in a short amount of time. The study's findings could provide valuable insights into the potential manipulation of social media, and the use of a technical approach like this ensured the accuracy and validity of the data collected and analysed.

This approach ensured that O3, O4 and O5 were met.

1.5. Report Overview

This chapter lays out the driving force, research inquiry, goals, and methodology of the project. Building upon the foundation established in this chapter, the subsequent chapters delve deeper into the background, rationale, and strategy behind the development and execution of the monitor and sentiment analysis.

Chapter 2 delves into the core concepts of the research. It offers clear definitions for terms such as ‘Fake News’, ‘parody’ and other relevant concepts in the realm of social media manipulation. The purpose of this chapter is to equip the reader with a thorough understanding of the issue of Fake News and manipulation on social media. It provides

a comprehensive overview of the problem and ensures the reader is fully informed before proceeding with the rest of the project.

Chapter 3 outlines the methodology used in creating the design of the required system. The process started with the classification and definition of the system's requirements, aimed at identifying the specific tasks that the system was required to perform. Based on these requirements, relevant libraries were selected from existing systems, evaluated for compatibility, performance, and their ability to meet the design goals. The result of this process was a comprehensive system design that served as a blueprint for the implementation phase in a subsequent chapter.

Chapter 4 focuses on the implementation of the 'Doubtfire' tweet monitoring and sentiment analysis system. The chapter details the implementation of various features, the technologies employed, and the ‘Doubtfire’ system architecture. It provides insights into the technical aspects of the project and how the design blueprint from Chapter 3 was brought to life. This chapter presents a comprehensive overview of the technical implementation, highlighting the key aspects of the system.

Chapter 5 focuses on the evaluation of the 'Doubtfire' tweet monitoring and sentiment analysis system. It includes an analysis of the trends in deleted tweets for all the keywords being monitored. The analysis of the trends in deleted tweets provides valuable insights into the effectiveness of the system in detecting and monitoring manipulated content on social media. It also includes a technical evaluation of the ‘Doubtfire’ system.

Chapter 6 focuses on the limitations of the 'Doubtfire' tweet monitoring and sentiment analysis system and outlines potential areas for future work and improvement. It provides a critical analysis of the system and identifies areas that can be improved upon. It discusses the ‘War on Information’ and how this study is important to help combat this war. The opportunities for future work discussed in this chapter include additional features, improved performance, and increased accuracy.

Chapter 7 summarises the main findings and outcomes of the project. This chapter reviews the material presented throughout the paper, including the research question, objectives, methodology, implementation, evaluation, and limitations. It assesses the extent to which the project's objectives were met and evaluates the overall success of the 'Doubtfire' tweet monitoring system.

2. Literature Review

2.1. Chapter Introduction

This chapter provides a comprehensive overview of the techniques, concepts, and ideas considered when designing the proposed system. It defines key terms such as 'Fake News', misinformation/disinformation, parody, satire, manipulation, and other relevant terminology to ensure that the reader has a clear understanding of the significance of this project and the need for the proposed system. This serves as a foundation for the discussion of the proposed system's design and implementation. By providing a clear and thorough understanding of the context in which the proposed system operates, we aim to enable the reader to fully appreciate the need for such a tool and the potential impact it can have in the fight against manipulation on social media platforms.

2.2. The Focus of Fake News

Fake News can be defined “to be news articles that are intentionally and verifiably false, and could mislead readers” (Allcott and Gentzkow, 2017). Fake News can have dire consequences as it can mislead people, spread misinformation, and cause confusion and mistrust in news sources. It can also be used to manipulate public opinion, influence elections, and disrupt societies. The problem of Fake News has been present for centuries, with a notable example being the ‘Great Moon Hoax’ of 1835, when a newspaper claimed there was a civilization living on the moon (Bossaller et al., 2019).

Fake News can be compared to a game of broken telephone, a popular children's game. Just as a message can become distorted and changed as it is passed along through a group of people, Fake News can also become distorted and changed as it is passed along through various sources. Starting with a small amount of false information, it can quickly spread and be repeated by multiple sources, becoming increasingly distorted and difficult to distinguish from the truth. Like the game of broken telephone, by the time it reaches the end, it bears little resemblance to the original message, and can cause confusion and misunderstanding.

In the era of digital media, it is easier for Fake News to spread widely and quickly, making it important for individuals to be able to critically evaluate the information they consume. Today, more than 4.59 billion people worldwide use social media (Oberlo,

2023), making it easier for Fake News outlets to target popular platforms such as Twitter, Facebook, and Instagram. A prime example of this was during the 2016 American presidential election, when a satire/fantasy website, WTOE 5 News created a news story claiming that Pope Francis endorsed Donald Trump as president. This article was shared on Facebook over a million times, despite the website categorising itself as a 'fantasy news' site although it did not mention this in the article (Allcott and Gentzkow, 2017).

Figure 2.1: Pope Francis endorses Donald Trump for President (WTOE 5 News, 2016)

Fake News is a growing concern that has led to various negative consequences, such as harassment and violence (Fisher, Woodrow Cox, and Hermann, 2016). It can be used to spread propaganda and manipulate people's beliefs and actions. While it can be difficult to act against Fake News due to the blurred definition and its association with freedom of speech (Funke and Flamini, 2014), it is important to address its harmful effects on society. Moreover, it is crucial for individuals to be able to critically evaluate the information they consume and to be able to identify Fake News and misinformation/disinformation.

2.2.1. Misinformation vs Disinformation

The reason it is so challenging to police Fake News is because of the complex and dynamic nature of misinformation and disinformation. The distinction between these two forms of false information is crucial as it highlights the different motivations, methods, and consequences of their spread.

Misinformation

Misinformation refers to false or inaccurate information that is not intentionally spread (Vraga and Bode, 2020). It can occur due to a lack of understanding, a

misinterpretation of facts, or a mistake in reporting. Misinformation can be spread through various channels, including social media, news outlets, and word of mouth. It can also be spread by individuals with good intentions, who may not realise that the information they are sharing is false. Misinformation can be difficult to combat as it is often spread by well-intentioned individuals who do not realise that the information, they are sharing is false (Wu et al., 2019).

Disinformation

Disinformation can be defined as the intentional creation and dissemination of false or manipulated information with the goal of deceiving and misleading audiences for the purpose of causing harm, advancing personal, political, or commercial interests (Buchanan and Benson, 2019). This can take many forms, from the creation and spread of Fake News articles to the manipulation of digital media through deepfake technology. Disinformation can be used to influence public opinion and political decisions, spread propaganda, and disrupt societies. It can also be used for personal gain, such as financial fraud or reputation management. The deliberate nature of disinformation makes it a powerful tool for those who seek to manipulate and deceive. Disinformation campaigns can be sophisticated and well-funded, making it difficult for individuals to distinguish between real and Fake News. Examples of this were seen during the Brexit campaigns in 2017 (Höller, 2021).

The ability to police Fake News is further complicated by the fact that it can be spread across a wide range of platforms, from traditional news sources to social media, making it difficult to track and combat. In addition, determining the intent and motivation behind the spread of false information can be difficult. Furthermore, the blurred definition of Fake News and the association with freedom of speech can make it difficult to define and act against without infringing on individuals' rights.

2.2.2. Censorship and Fake News

To understand the harmful risks of manipulation on social media we first need to discuss censorship and the role it plays in the spread of Fake News.

Censorship

Censorship refers to the suppression of speech, communication, or other information deemed objectionable, harmful, sensitive, or inconvenient by governments, private institutions, or controlling bodies (ACLU, 2019). It can take various forms such as

blocking access to certain websites, limiting access to certain media outlets, or removing certain content from the internet. Censorship can be done for several reasons, such as maintaining social order, protecting national security, or preventing misinformation. Unfortunately, censorship also can have negative consequences when it is used to suppress dissenting voices, limit access to information, and control public opinion.

Censorship can have detrimental effects on the freedom of expression and the free flow of information. It can stifle creativity, limit access to knowledge, and impede scientific and technological progress (Rojas, Shah, and Faber, 1996). Additionally, it can be used to conceal corruption, abuse of power, and other forms of misconduct. Furthermore, censorship can be used to limit the ability of individuals to hold others accountable, by preventing the public from accessing information about an individual's actions and decisions. Censorship can also be used to silence whistle-blowers, journalists, and other sources of essential information. An example of this includes the deletion of tweets on social media platforms, based on personal beliefs or views, which can be seen as censorship and might allow for immoral actions to go unnoticed and unchallenged.

2.2.2.1. Intentional vs Unintentional Censorship

Censorship can take many forms and can be done both intentionally and unintentionally. Intentional censorship refers to the deliberate suppression of information or speech, often with the goal of controlling or manipulating public opinion. An example of this is the use of keyword filtering censorship in China, where certain apps and websites are blocked or restricted to control speech and prevent the spread of dissenting ideas or negative news about China (Ruan et al., 2016). This type of censorship can have severe negative consequences, such as spreading propaganda and false narratives.

On the other hand, unintentional censorship can occur through economic policies, legislation, and technologies that inadvertently limit certain forms of speech, while granting support and rewards to other forms of speech (Deibert, 2003). This type of censorship can happen because of a lack of awareness or understanding of the potential consequences of these policies, legislation, and technologies. It is important to be aware of both intentional and unintentional censorship as both can have detrimental effects on freedom of expression and the free flow of information.

2.3. The Purpose of Parody

Parody is one of the main types of Fake News. It is a key element in the creation of this study, making it important to understand the definition and types of parody.

2.3.1. Defining Parody

There are several types of parody, including satirical parody, spoof parody and parody news. Satirical parody uses humour and irony to mock and criticise political figures, social issues, or current events. Spoof parody imitates a particular genre, style, or format, often in a comedic or exaggerated manner. Parody news involves the creation of Fake News stories, often with the intention of making a comic or satirical point. Parodic Remake is a re-creation of a pre-existing work or content, usually with a comedic or satirical twist. Parodic Advertisements re-create a pre-existing advertisement, usually with a comedic or satirical twist (Condren et al., 2008).

The use of satirical parody in news often employs the emotion of "disgust" to create a sense of humour and irony (Ghanem, Rosso and Rangel, 2020). By exaggerating or mocking a particular situation or individual, satirical news aims to bring attention to social or political issues and encourage critical thinking. The use of disgust as an emotion in satire highlights the absurdity or hypocrisy of a particular issue or person, adding an element of comedy and making the commentary more memorable and impactful.

2.3.2. Uses of Parody

Parody is a form of satire that uses humour, irony, and exaggeration to criticise and mock elements of society (Herrero-Diz, Conde-Jiménez, and Reyes de Cózar, 2020). Unlike disinformation or misinformation, the main purpose of parody is not to deceive or harm, but to provide a form of social commentary and critique. The key aspect of parody is that it is done without the intention to harm, it is a way to express criticism and make a point without resorting to harmful or malicious methods.

2.4. The Meaning of Manipulation

The manipulation of social media has been a persistent issue since the launch of Facebook in 2004 (Good, 2012). Manipulation involves controlling or influencing people or things in a dishonest and often undetected way (Fitzpatrick, 2018). This

form of manipulation on social media can be considered a violation of freedom of speech (Dušan V., 2021) and can lead to the spread of misinformation through censorship.

2.4.1. Targeted Audience

Manipulation on social media can be targeted to influence what certain groups of people see or do not see. Studies have revealed that troll accounts on social media often target specific individuals, communities, or organisations with hate speech or various forms of online harassment (Bradshaw and Howard, 2018). The purpose of this targeted harassment is often to discourage these groups from using social media, thereby reducing the number of people who speak against the troll's beliefs. Additionally, a study conducted in May of 2022 found that journalists commonly delete their tweets and social media posts to "control their reputation and livelihood" (Ringel and Davidson, 2022). This too can be considered a form of manipulation, as it is targeted at a specific audience, and by deleting tweets that may be detrimental to their reputation, it can be considered as spreading misinformation as it does not provide a full understanding of the individual's beliefs and shortcomings.

2.4.2. Harmful or Harmless

One of the most pressing questions related to manipulation on social media is whether it is truly harmful. The answer is affirmative. Manipulation on social media can have a range of negative effects, such as the ability of domestic and foreign actors to weaponize social media and manipulate democratic elections, as well as the ability of individuals in positions of power to evade accountability for their actions (McDermott and Hatemi, 2020). Examples of the effects of manipulation on social media include the 2016 presidential election, where Russian-backed misinformation campaigns targeted hundreds of millions of U.S. citizens, potentially influencing the outcome of the election in favour of Donald Trump (Aral and Eckles, 2019). Additionally, the tendency for accounts that post Fake News or misinformation to delete their posts after they have been viewed by a substantial number of users (Mustafaraj and Metaxas, 2017) allows them to evade accountability for the spread of false information.

2.5. The Threats of Twitter

Since its launch in 2006, Twitter has undergone significant changes and evolution. Initially, many were sceptical of the platform and its potential, with the limited 140-character updates appearing to be of little value (Johnson, 2009). However, as the platform has grown and evolved, it has become an increasingly useful tool for communication, news dissemination, and social engagement. Today, Twitter is widely recognized as a powerful tool for connecting people, sharing information, and fostering discussions on a wide range of topics. Currently, Twitter receives at least 500 million tweets a day (DeMilt, 2017), making it one of the most widely used social media platforms available today. Twitter's ability to help real-time communication and the ability to follow diverse groups of people as well as its role in breaking news and fostering public discourse have made it an essential tool for individuals, organisations, and media companies.

2.5.1. Twitter User Policies

Twitter states that its policy is to ‘serve the public conversation’ and its user guide includes rules related to violence, terrorism, child sexual exploitation, abuse, harassment, and hateful content (Twitter, 2019). The question of how these rules are monitored is an important one. Twitter uses a combination of automatic and human review processes to enforce its rules. Automated systems are used to detect and flag potentially violating content, while a team of human reviewers then evaluates the flagged content to determine if it violates the platform's rules.

However, it is important to note that enforcement of these rules is not always straightforward and may not always be consistent. Personal beliefs and biases can play a role in the enforcement of these rules. Additionally, individuals with a large following and significant amount of power, such as Elon Musk, may have more influence on the platform and may be subject to different enforcement standards. Twitter has been criticised for not consistently enforcing its rules and for not acting against tweets from high-profile individuals that violate its policies.

2.5.2. Elon Musk and His Influence

Elon Musk is a highly influential figure on Twitter, as the CEO of Tesla and SpaceX, Musk has a large following on the platform and his tweets often generate significant engagement and media attention.

However, his influence on Twitter is not limited to his personal following. As of October 27th, 2022, he is also the owner of the platform. Musk now has the power to shape the company's policies and direction, which could have a significant impact on the platform's overall operation and function. This includes the ability to make changes to the user interface, the type of content that is allowed on the platform, and how rules are enforced.

An example of this is the spike in hate speech after Musk took over. According to a recent study (Benton et al., 2022), specific hateful terms that were never tweeted more than 84 times an hour before Musk took control, jumped to 4,778 tweets per hour within the first twelve hours of Musk's takeover. This suggests that the change in ownership may have led to a loosening of restrictions on hate speech and other forms of harmful content. This trend raises concerns about the potential consequences of Musk's ownership of Twitter, particularly regarding the platform's ability to enforce its rules and protect its users from harmful content.

Additionally, Musk's ownership of Twitter could also lead to conflicts of interest, where his personal and business interests may come into play when making decisions about the platform. There is also a concern that his ownership of the platform could lead to censorship of certain content or groups, or that it could be used to promote his own interests and agenda.

2.5.3. The 2022 Blue Check Scandal

The blue check on Twitter is a way to indicate that an account has been verified by Twitter as belonging to a notable public figure, celebrity, brand, or organisation. The verification process is intended to help users quickly identify authentic accounts and to prevent impersonation. Accounts that have been verified are marked with a blue checkmark next to their name. The process of verification is done by Twitter, and it involves applying and providing evidence of identity, such as a government-issued ID or other forms of documentation (Twitter, 2022). The blue checkmark serves as an indication that the account is authentic and that the person or organisation behind the account is who they claim to be.

However, after Musk's takeover, he launched Twitter's premium service, which granted blue check verification labels to anyone willing to pay $8 a month (Milmo, 2022a). This change had significant impacts on the platform, as it made it easier for individuals and organisations to obtain verified status and the credibility that comes

with it. This led to an increase in the number of verified accounts on the platform, including parody accounts that exploited the blue check to spread misinformation for purposes such as satire, awareness of corruption or to deceive the readers (Milmo, 2022b).

The launch of Twitter's premium service raises concerns about the potential for increased manipulation on the platform. This change ushered in a wave of bizarre and potentially harmful impersonation of public figures and brands. For example, Eli Lilly and Company, an American Pharmaceutical company, dropped over 4.37% in the stock market after a parody account, campaigning as the company, tweeted that insulin would now be free (Lee, 2022). This change in the verification process resulted in a loss of trust in the authenticity of verified accounts and a rise in misinformation, deception, and harm.

Figure 2.2 :Eli Lilly and Company Parody Account Tweet (Twitter, 2022)

Since the blue check scandal, Twitter has deleted every blue-checked parody account and their tweets, which is a form of censorship on the platform. This move by Twitter is intended to restore trust in the authenticity of verified accounts and to prevent the spread of misinformation, deception, and harm. However, it also raises concerns about the potential for censorship and the suppression of free speech.

Musk's decision to delete all these accounts can be seen to protect his own integrity and brand. By removing these parody accounts, he may be trying to prevent the public from seeing the mistakes of his idea to offer blue check verification for a fee. This highlights the importance of tracking manipulation on social media platforms like

Twitter, as well as the need for transparency and accountability in the actions of platform owners and operators.

It is important for social media platforms to balance the need to protect users from harmful content with the right to free speech and expression. There is a need for a consistent and fair approach to moderating content, and for transparent policies and procedures for handling issues related to misinformation and impersonation.

2.6. Evaluation Methods

Evaluating manipulation on social media necessitates a comprehensive methodology that incorporates both human and technical evaluations. It is imperative to comprehend these forms of evaluations to effectively mitigate the corruption of data on social media platforms.

2.6.1. Human Evaluation

Human evaluation is important when looking for manipulation on social media because it provides a level of analysis and understanding that automated tools and algorithms cannot match. Human analysts can evaluate the context and nuances of the information shared on social media, which is crucial for identifying deception and manipulation (Hudson et al., 2016). They can review the source of the information, analyse the language used, and identify patterns of deception or manipulation. They can also evaluate the intent behind the information and assess the potential impact of the misinformation (Hocevar, Flanagin and Metzger, 2014). In addition, human analysts can evaluate the credibility of sources and identify potential biases. All these aspects are essential for understanding the scope and nature of manipulation on social media, and for taking appropriate actions to address it.

Human evaluation methods, such as content analysis, media forensics, network analysis, and behavioural analysis, are critical in identifying patterns of deception or manipulation. Content analysis involves reviewing the text, images, and videos shared on social media to identify patterns of disinformation, false information, and potential biases (Morris, 1994). Media forensics involves the analysis of visual and audio information, such as images, videos, and audio recordings, to assess their authenticity and identify any signs of manipulation (Robertson et al., 2019). Network analysis involves the analysis of the relationships and interactions between different accounts on social media to identify patterns of deception or manipulation (Weng

and Lin, 2022). Behavioural analysis involves the analysis of the behaviour of social media users to identify patterns of deception or manipulation (Tagliabue, Squatrito and Presti, 2019).

All these methods are essential for understanding the scope and nature of manipulation on social media and for taking appropriate actions to address it. They can be used in combination with technical evaluation methods to provide a comprehensive approach to identifying manipulation on social media.

2.6.2. Technical Evaluation

Technical evaluation methods are an integral aspect of identifying manipulation on social media, as they provide a means of analysing copious quantities of data in an efficient and precise manner. These methods often employ algorithms and machine learning techniques to detect patterns and anomalies within the data.

One example of a technical evaluation method is the utilisation of social media scraping, which involves the collection of data from social media platforms through automated tools (Dewi, Meiliana and Chandra, 2019). This method can be utilised to gather extensive amounts of data on a specific topic or individual, which can then be analysed to identify patterns of manipulation. Another example is network analysis (Del Fresno García, Daly and Segado Sánchez-Cabezudo, 2016), which involves utilising algorithms to analyse the interactions and relationships between different accounts on social media. This method can be used to detect clusters of coordinated activity and automated accounts.

Moreover, technical evaluation methods play a crucial role in detecting deepfake videos and images (Westerlund, 2019), which are becoming increasingly prevalent on social media platforms. These methods can detect subtle changes in the video or image, such as changes in facial expression, which may not be discernible to the human eye.

In conclusion, technical evaluation methods are a necessary component of identifying manipulation on social media, as they provide a means of quickly and accurately analysing copious quantities of data. They can be used with human evaluation methods to provide a comprehensive approach to identifying manipulation.

2.7. Summary of Review

This literature review has explored the various aspects of Fake News, censorship, parody, and manipulation on social media, with a specific focus on Twitter. It has explored the differences between misinformation and disinformation and the various forms of censorship, both intentional and unintentional. This review has also delved into the purpose and types of parody, and the concept of manipulation, including its targeted audience and the potential for harm or harmlessness. This review has also examined the current state of Twitter, including the policies and user guidelines, the influence of Elon Musk on the platform, and the 2022 blue check scandal. Finally, this review has discussed the importance of evaluating manipulation on social media and the different methods used, including human evaluation and technical evaluation. Overall, this literature review highlights the complex and dynamic nature of Fake News and its associated issues on social media and the importance of understanding these issues to reduce their negative impact on society. The information presented in this chapter provides the fundamental ideas which influence the process and design of building the proposed monitor.

3. Design

3.1. Chapter Introduction

This project's main aim was to assess the likelihood of purposeful deletion of tweets on Twitter. The previous section described the potential hazards and consequences of censorship and manipulation on social media, emphasising the significance of effective monitoring tools. In this section, we will examine the fundamental principles and concepts that guided the development and design of the 'Doubtfire' tweet monitoring and sentiment analysis system. This section will analyse the various classifications, tasks, and approaches employed in creating a system design that meets the project's design goals. A four-step strategy was used in designing the system for this project, as depicted in Figure 3.1.

Step 1: This stage involved identifying the necessary requirements for the project, the design's objective, and the prerequisites for the sentiment analysis library and tweet scraper employed. This stage provided the foundation for the project and enabled us to comprehend its scope.

Step 2: This step in the design process involved assessing various libraries and selecting the ones that were most appropriate for fulfilling the identified system requirements. The libraries were selected based on their ability to perform effectively, whether they met the system's requirements, and their compatibility with the other techniques selected. The chosen libraries were drawn from the same classification as defined in Step 1.

Step 3: This phase allowed us to determine the tasks of the monitor and sentiment analysis required to meet the project's design specifications. This stage facilitated the development of a roadmap for the system's processes and provided a framework for selecting the techniques that would be utilised in the system's ultimate design. After assigning the tasks a successful monitor and sentiment analyser was constructed that met the design objective. This section outlines the design process we followed and the decisions we made.

Step 4: This stage entailed implementing the design techniques we had selected above.

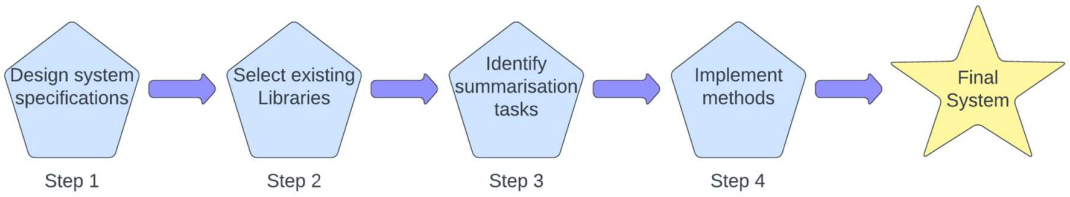

Figure 3.1 :Summarised Design Process used for this System.

3.2. Requirements of The System

To provide a comprehensive explanation of this system, the requirements for this tool were determined based on the project's motivation and design objectives. The objective addressed in this section is to develop a tweet monitoring system that tracks deleted tweets by utilising keyword searches and a sentiment analysis system that assesses the sentiment values of both deleted and non-deleted tweets and compares them. The primary motivation for this system is to investigate how freedom of speech on social media may be compromised. The system's aim is to analyse the relationship between sentiment and tweet deletion. Therefore, based on the project's motivation and design objectives, the following requirements were established for the required tweet monitoring and sentiment evaluation system.

R1: The system functions independently without relying on Twitter's API.

As a result of Twitter's continuously evolving nature, characterised by Elon Musk's frequent introduction of new rules and regulations, it was imperative to employ a scraper that did not depend on Twitter's API, given the uncertainty surrounding its continued availability to the public. This need proved to be crucial, as Twitter announced on February 1st, 2023, that it would no longer offer free access to its API, resulting in significant implications for its internal tools and public-facing APIs (Newton, 2023). The objective of this requirement was to guarantee that any modifications to Twitter made by Musk would not impede our research.

R2: The system exclusively searches for the keywords specified within a particular period.

With the vast number of tweets being generated on Twitter each day, estimated to be at least 500 million tweets, it can be challenging to locate and retrieve specific tweets, particularly those that have been deleted. To address this issue, it is necessary to narrow down the search criteria by focusing on specific keywords and

within a particular time. This requirement is particularly crucial when searching for deleted tweets, as it allows for the timely retrieval of these tweets before they are permanently removed from the system. By limiting the search time to one hour before adding to the value of n = time, this requirement enables us to find and analyse the deleted tweets more efficiently, while also reducing the number of irrelevant tweets that are retrieved.

R3: The system incorporates the ability to recognise sarcasm, irony, and casual human conversational language.

Twitter is a widely used platform where users frequently employ casual language and engage in informal communication. As such, it is common to encounter tweets that utilise sarcasm and irony, which can present significant challenges for sentiment analysis. Distinguishing between sarcastic tweets and those intended to be taken seriously can be a complex and nuanced task, particularly in an open setting (Kunneman et al., 2015). Thus, it is crucial that any sentiment analysis library employed is capable of accurately identifying and accounting for sarcasm, to ensure that the sentiment of a tweet is correctly interpreted and analysed. By incorporating the ability to recognise sarcasm and other forms of conversational language into the system, we can enhance the accuracy and reliability of the sentiment analysis results and obtain a more comprehensive understanding of the attitudes and opinions expressed on Twitter.

R4: The system includes the consideration of emoji sentiment in its sentiment analysis.

Emojis have become an integral and pervasive aspect of online communication and are widely used on social media platforms such as Twitter to express emotions and sentiments. In fact, a study by Kralj Novak et al. (2015) found that over 10 billion emojis were used on Twitter between 2013 and 2015. The inclusion of emojis in tweets can provide valuable additional context, allowing for the conveyance of nuanced meanings that may not be readily apparent through text alone. The sentiment conveyed by emojis can be highly informative and influential in understanding the overall sentiment of a tweet, as they often reflect the emotions and attitudes of the author in a more direct and immediate way than text alone. By considering the sentiment of emojis, the system can obtain a more precise, accurate,

and comprehensive understanding of the sentiment expressed in a tweet, enhancing the quality and reliability of the sentiment analysis results.

3.3. Classifying The System

This section outlines the initial stage in developing a comprehensive system for monitoring and analysing deleted tweets. The first step involves identifying a set of categories based on the system's specific requirements, which serve to define its functionalities and working principles broadly. By establishing these categories early on, the number of potential methods and design options can be narrowed down, streamlining the design process, and ensuring that the final product is aligned with the project's goals and objectives.

In this section, the selected categories for the system are presented, which can be visualised in Figure 3.2.

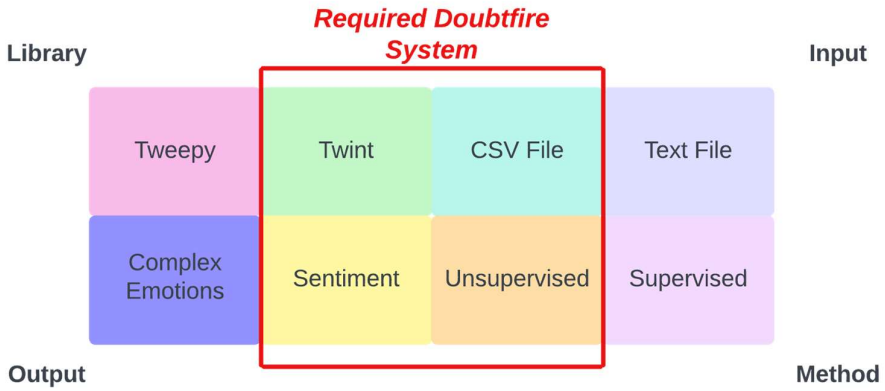

Figure 3.2 :Classifications identified for the Proposed System.

Library: Twint

In compliance with the requirements of ***R1*** and ***R2***, the chosen library must satisfy the condition of operating independently without relying on Twitter's API, and exclusively searching for specific keywords within a particular period. Thus, the Twint library was identified as the suitable choice for the task at hand.

Unlike Tweepy, another tweet scraping library, which heavily relies on Twitter's API to access and retrieve data, Twint employs the use of bots to scrape Twitter data (Hwang, 2020). Twint's utilisation of non-API-based data retrieval strategies makes it an independent system that can execute the task at hand with utmost efficiency, while avoiding potential limitations and restrictions imposed by Twitter's API.

Moreover, Twint's functionality permits the system to exclusively search for the keywords specified within a designated period, thereby meeting the requirements of ***R2***. This approach provides a more straightforward and direct means of identifying relevant data while filtering out irrelevant information. Consequently, the library's effectiveness and reliability in retrieving accurate and precise data, coupled with its non-API-based data retrieval mechanism, makes it an excellent choice for this task.

Input: CSV File

To fulfil the requirements of ***R4*** and ***R3***, it is necessary to convert the tweets scraped by Twint into CSV format. This is due to the layout of the tweets, which require individual analysis for the sentiment analyser to accurately recognize and classify sarcasm, irony, and casual human conversational language, as specified by ***R3***. Additionally, ***R4*** calls for the consideration of emoji sentiment in the sentiment analysis process.

Converting the tweets into CSV format allows for each tweet to be individually parsed and analysed, providing a more accurate and comprehensive understanding of the sentiment conveyed in the tweet. This format also enables the sentiment analyser to consider the sentiment conveyed by emojis, which can often provide additional context and nuance to the sentiment of the tweet.

Output: Sentiment

In accordance with ***R3*** and ***R4***, the output generated by the 'Doubtfire' monitor pertains solely to the sentiment conveyed in each tweet, rather than attempting to capture the nuanced and multifaceted nature of human emotions. This is achieved through three broad categories, namely negative, positive, and neutral, as opposed to more granular descriptors such as anger, sadness, or happiness. By eschewing these finer distinctions, the monitor enables easy access to data and facilitates the identification of common trends among tweets. Also, the system can recognise subtle aspects of human language such as sarcasm, irony, and colloquialisms and considers the sentiment conveyed by any emojis present in the tweet.

Method: Unsupervised

The most vital classification of the 'Doubtfire' system is that the method of sentiment analysis is unsupervised. As described by ***R3*** of this system, the sentiment analysis must be able to recognise sarcasm, irony, and casual human conversation.

Unsupervised methods such as Vader (Valence Aware Dictionary and sEntiment Reasoner) are used for sentiment analysis because they do not require labelled data for training. This means that the algorithm can analyse text data without prior knowledge of what the sentiments should be, making it more versatile and useful in situations where labelled data is not available or is difficult to obtain (Calderon, 2018).

Vader specifically is designed to manage social media text, which can be challenging for traditional supervised methods because it often contains non-standard language, sarcasm, and other forms of irony. Vader uses a lexicon of words and rules to calculate a sentiment score for each text input, which is based on the intensity of positive and negative words and punctuation within the text.

While unsupervised methods like Vader may not be as accurate as supervised methods that are trained on labelled data, they are still useful for quickly and efficiently analysing large volumes of text data and providing a general understanding of the sentiment present in the text.

System Classification

The library, input, output, and method classifications established for the 'Doubtfire' system determine the system's required components as Twint as the library, a CSV file for input, sentiment analysis for the output and supervised learning for the method. These classifications will serve as a framework for subsequent stages of the design process, enabling the identification of essential system tasks and methods.

3.4. Library Selection

This section of the chapter will discuss how the libraries chosen were decided upon and why.

Tweet Scraping Library

The selection of the appropriate tweet scraping library for a project can be a difficult decision, as each library has its own strengths and weaknesses. In this project, four libraries were considered: Tweepy, Snscrape, Selenium, and Twint.

Tweepy is a widely used Python library that offers an easy-to-use interface for accessing the Twitter API (Rigden, 2018). It provides various functionalities, including getting tweets, user information, follower information, retweets, etc.

Although Tweepy offers seamless integration with the Twitter API, it has some limitations as the Twitter API imposes rate limits that can slow down the data scraping process. Furthermore, with the recent change of public API rules by Twitter, Tweepy's functionality has been restricted.

Snscrape is a command-line interface Python library that uses Twitter's web scraping API to retrieve data (Desai, 2021). It has the advantage of being able to retrieve more data than Tweepy because it uses web scraping techniques to retrieve the data. However, it requires a bit more technical knowledge to use than Tweepy and does not offer as many features.

Selenium is a web testing library that can be used for scraping data from Twitter. It automates the process of opening a web browser, navigating to the Twitter website, and retrieving data from the website (Silman, 2019). Selenium is highly customisable, and you can use it to scrape data that is not easily retrievable using other libraries. However, it requires more technical knowledge and has a steeper learning curve than the other libraries.

Twint is a Python library that provides an easy-to-use interface for scraping data from Twitter without using the Twitter API. It can retrieve data such as tweets, user information, and follower information. Twint is faster than Tweepy because it does not use Twitter's API, but it has some limitations in terms of the amount of data that can be retrieved, and the search capabilities compared to the Twitter API. Twint is a suitable choice for this project as it is not dependent on the Twitter API, which is subject to frequent changes and restrictions.

In conclusion, after careful consideration of the strengths and limitations of each library, Twint was chosen as the most suitable library for this project, given its speed and independence from the Twitter API which helps it to fulfil ***R1*** and ***R2***.

Sentiment Analysis Library

The selection of the appropriate sentiment analysis library for this project was also a difficult decision. In this project, four libraries were considered: NLTK (Natural Language Toolkit), TextBlob, AFinn and Vader.

NLTK is a widely used Python library for natural language processing (Reid, 2021). It includes several functions for sentiment analysis, such as the NaiveBayesClassifier

and the SentimentIntensityAnalyzer. However, NLTK may not perform well on tweets with informal language and slang.

TextBlob is a Python library that provides a simple and intuitive interface for performing common natural language processing tasks, including sentiment analysis (Shah, 2020). It uses a machine learning algorithm to classify the sentiment of the text as positive, negative, or neutral. TextBlob can handle informal language and slang better than NLTK. However, it may require additional training data to improve the accuracy of the sentiment analysis.

Afinn is a Python library that uses a list of pre-computed polarity scores for simple words in English to calculate the sentiment of the text (Lohiya, 2018). It can be used for simple sentiment analysis tasks but may not be suitable for complex analysis. Afinn has a smaller dictionary of words compared to other libraries and may not perform well on tweets with unique or uncommon words.

Vader is a rule-based sentiment analysis tool that is specifically designed for analysing social media texts, such as tweets. It uses a combination of lexicon-based and rule-based approaches to classify the sentiment of the text as positive, negative, or neutral. Vader can handle informal language, slang, and emojis, which are commonly used in tweets. It also provides a sentiment score that can be used to quantify the sentiment's intensity.

All these libraries can be used for sentiment analysis of tweets, but their performance and accuracy may vary depending on the specific requirements and use case of the project. NLTK and TextBlob are more suitable for complex analysis, while Afinn and Vader are better for simple analysis. Vader is specifically designed for social media texts and can handle informal language, slang, and emojis better than other libraries hence why it was the right choice for this project hence fulfilling ***R3*** and ***R4***.

3.5. Identifying Tasks

This section will present the outcomes of *Step 3* of the design process outlined above. The tasks of the 'Doubtfire' system are critical actions that the system must perform to effectively detect manipulation on social media platforms. While these tasks may not explicitly specify the functionality of the 'Doubtfire' system, identifying the necessary tasks provides a framework for applying appropriate methods to

perform the tasks. Constructing methods based on the identified tasks leads to the design of the 'Doubtfire' system. The aim of this section is to determine the essential tasks that the system must perform to fulfil the design objectives' requirements. The tasks identified in this section provide a comprehensive overview of the proposed 'Doubtfire' system's operation. These tasks are presented in Figure 3.3.

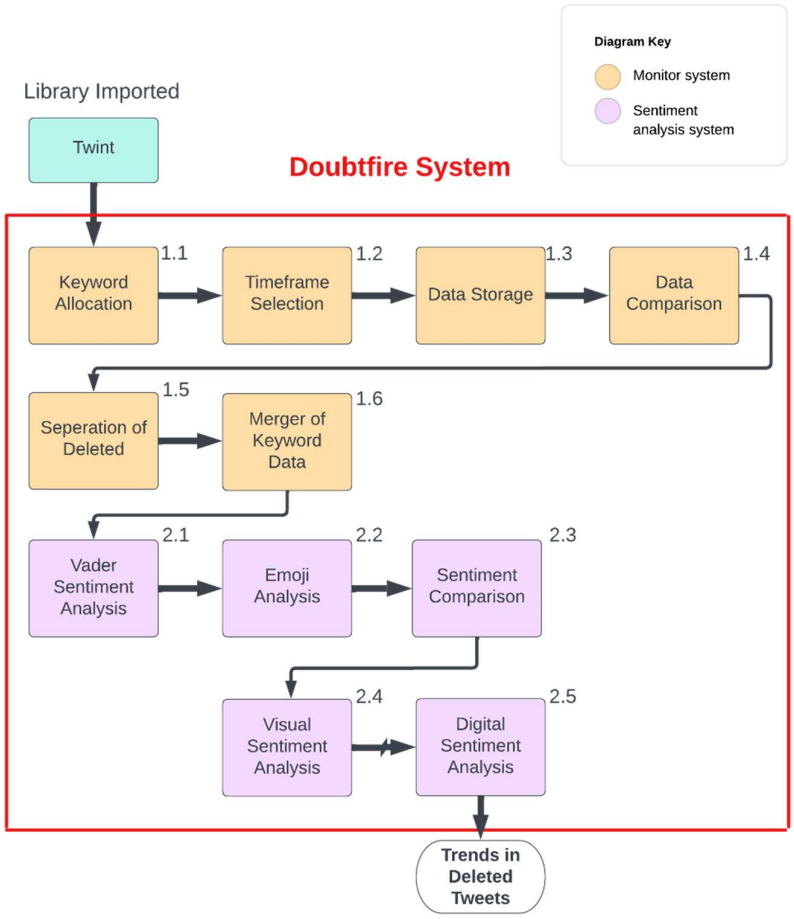

Figure 3.3 :Tasks to be considered for this System.

3.5.1. Task Group 1: Monitoring System

The objectives of this group are tailored towards the monitoring system. The primary goal of this task group is to identify deleted tweets on the Twitter platform for the given keywords. These tasks conduct the fundamental functions required to identify trends in deleted tweets. Based on the classification and specifications outlined for the desired system, the activities conducted by this task group must identify deleted tweets based on the designated keywords within the specified period. The assignments within this group are allocated as follows, with the following outputs:

Task 1.1 - Keyword Allocation: The primary objective of this task is to identify the essential keywords required for the system. The identification of these keywords enables the monitor to conduct more targeted searches on the Twitter platform and locate tweets relevant to issues that may have a significant impact on Twitter. This, in turn, facilitates the achievement of ***R2***, which aims to enhance the system's ability to identify specific deleted tweets. By pinpointing the relevant keywords, the system can streamline its search process, reducing the search time and improving the accuracy of the search results. As a result, this task plays a critical role in enabling the system to achieve its monitoring objectives effectively.

Task 1.2 – Timeframe Selection: This task involves the critical step of selecting an appropriate timeframe to locate deleted tweets in real-time. It is essential to identify an optimal timeframe as tweets are constantly being deleted. Selecting a period that is too extensive may cause the system to miss a considerable number of deleted tweets. Conversely, a period that is too short could lead to difficulties and delays in locating the necessary data. The selection of an appropriate time frame is crucial for the system to meet ***R2***, which aims to facilitate the identification of specific deleted tweets efficiently. By selecting an optimal period, the system can enhance its ability to locate relevant data in real-time and ensure accurate monitoring of Twitter activities. This task, therefore, plays a crucial role in enabling the system to achieve its objectives effectively.

Task 1.3 – Data Storage: This task is critical to the system's overall functioning, particularly in subsequent steps. Given the large amount of data collected each day for each keyword, the proper storage of data is essential to ensure the system's ability to identify deleted tweets correctly and effectively. The accuracy and efficiency of the system's operations are heavily reliant on the effectiveness of this task. By ensuring that the data is stored correctly, the system can streamline the process of locating deleted tweets, enhancing the achievement of its objectives. Proper data storage is also essential for enabling efficient data retrieval and analysis, which is crucial to the system's overall functioning.

Task 1.4 – Data Comparison: This task is one of the critical activities for this task group. Data comparison is essential for locating all the deleted tweets accurately. However, the formatting of each file, the diverse periods, and the extensive volume of data collected can make it challenging to perform this task effectively. Despite

these challenges, the accurate and efficient completion of this task is crucial for the system to achieve its monitoring objectives. By comparing data effectively, the system can streamline its search process, improve the accuracy of its search results, and enhance its ability to identify specific deleted tweets.

Task 1.5 – Separation of Deleted: This task is essential to achieving one of the primary objectives of the monitoring system, which is to identify manipulation on social media by detecting deleted tweets. The separation of deleted tweets from the tweet data is a crucial step in achieving this objective. To ensure that the system can identify as many deleted tweets as possible, it is necessary to check every file and compare them accurately. Any duplication or omission of data could significantly impact the system's ability to identify manipulations on social media accurately. Therefore, this task requires a prominent level of accuracy and attention to detail to avoid any errors in the separation of deleted tweets from the tweet data. The successful completion of this task is essential to meeting the monitoring system's objectives and facilitating effective social media monitoring.

Task 1.6 – Merger of Keyword Data: Each keyword, which is essential for the subsequent task group on sentiment analysis. The integration of data serves as a key reference point for comparing deleted tweets data, which is crucial to achieving accurate results. This task requires meticulous attention to detail and accuracy to ensure that all data is correctly integrated and that there are no omissions or duplications. The successful completion of this task enhances the accuracy and effectiveness of the system in detecting manipulations on social media, which is essential to achieving the overall monitoring system's objectives. By integrating data accurately, the system can improve its ability to perform sentiment analysis effectively, which is essential to achieving ***R3***.

3.5.2. Task Group 2: Sentiment Analysis System

The tasks in this group are specific to the 'Doubtfire' system's sentiment analysis and data analysis capabilities. These tasks are critical in achieving the system's objectives, as they perform the primary operations of sentiment analysis and data analysis. The sentiment analysis helps to identify the tone of the tweets, while the data analysis helps to identify the trends in the data, which are crucial to detecting manipulations on social media accurately. To achieve these objectives, the tasks in this group have been designed to perform the following operations:

Task 2.1 – Vader Sentiment Analysis: This task plays a critical role in achieving ***R3*** of the 'Doubtfire' system, which requires the system to perform sentiment analysis effectively. By using Vader sentiment analysis, this task specifically helps to analyse the tweets, both deleted and non-deleted, while considering sarcasm, irony, and casual language. Unlike other sentiment analysis tools, such as Textblob and Afinn, Vader accounts for this type of language, providing a more accurate sentiment value for this type of data.

The output of this task is a sentiment score for each tweet, which indicates whether the tweet is positive, negative, or neutral. The sentiment score is an essential component of the 'Doubtfire' system as it helps to identify the tone of the tweets, which is critical in detecting manipulations on social media.

Task 2.2 – Emoji Analysis: This task is focused on achieving ***R4*** of the 'Doubtfire' system, which requires the system to consider the impact of emojis on sentiment analysis. This task involves the development of an algorithm that can accurately analyse the sentiment of a tweet while also considering the impact of emojis. The algorithm needs to be trained to recognize the distinct types of emojis and how they can affect the sentiment of the tweet. The output of this task is a sentiment score for each tweet that considers the impact of emojis. This sentiment score is critical in achieving the objectives of the 'Doubtfire' system as it provides a more accurate understanding of the sentiment of the tweets.

Task 2.3 – Sentiment Comparison: This task involves comparing the sentiment scores of the emojis' sentiment analysis and the Vader's sentiment analysis. The aim is to combine both scores to obtain a more accurate sentiment result and achieve ***R3*** and ***R4*** of the 'Doubtfire' system. The sentiment scores obtained from both analyses in the tasks above are combined to produce a final sentiment score. This process helps increase the sentiment analysis's accuracy by considering both the text and emojis used in the tweet. The output of this task is a final sentiment score for each tweet that considers the sentiment of both the text and the emojis.

Task 2.4 – Visual Sentiment Analysis: This task involves using various visualisation techniques such as pie charts and bar charts to present the sentiment analysis results of both the deleted and non-deleted tweets for each keyword. The primary objective of this task is to identify trends in the sentiment of tweets and detect significant differences between deleted and non-deleted tweets. The output of

this task is a set of visualisations that provide a clear and concise presentation of the sentiment analysis results. These visualisations aid in the interpretation of the sentiment analysis data and help to identify patterns and trends in the data.

Task 2.5 – Digital Sentiment Analysis: This task is crucial in identifying trends in deleted tweets. By using digital sentiment analysis and representing the results through percentages and figures, we can easily identify any significant patterns in the sentiment of deleted tweets. This allows us to determine the frequency and timing of such tweets, which is important in identifying potential instances of censorship or manipulation.

3.6. Summary of Design

This section outlines the process for importing the Twint library and setting parameters to identify deleted tweets containing specific keywords over a defined period. The process involved defining a set of requirements for the system based on the project's design objective and motivations. The first three steps of the four-step design process were then completed to develop a system design that met the identified requirements and satisfied the project's design objective.

The system's classification was initially defined, providing a high-level overview of its intended functionality. Libraries were then carefully selected based on their performance and ability to satisfy the identified requirements. The selected libraries were incorporated into the system design, which was subsequently evaluated against the previously outlined requirements.

The design produced in this chapter successfully satisfied all the requirements established at the beginning of the project, thus meeting the overall design objective of creating a monitoring and sentiment analysis system capable of identifying trends in deleted tweets. This design will be used in the subsequent chapter to implement the system.

4. Implementation

4.1. Chapter Introduction

In this chapter, we will delve into the implementation of the proposed 'Doubtfire' system designed for monitoring and analysing the trends in deleted tweets. This implementation is based on the design requirements outlined in the previous chapter. The implementation discussed here was instrumental in evaluating the design's effectiveness and limitations. Since the 'Twint' library, used in this project, is developed in Python, the 'Doubtfire' system was implemented in Python as well. The goal of this chapter is to provide the reader with a comprehensive understanding of the development process of the 'Doubtfire' system and the technologies employed to achieve this objective.

4.2. Technologies Used

The following technologies were used to build this system.

4.2.1. Python

Python is a high-level, interpreted, and object-oriented programming language with a focus on simplicity and ease of learning, making it an excellent choice for ‘Rapid Application Development’ and scripting (Python Software Foundation, 2019). The built-in data structures, combined with dynamic binding and dynamic typing, make it a powerful tool for connecting various components together. Python was chosen to develop the 'Doubtfire' system due to its user-friendly syntax and the fact that the 'Twint' library used in the system is built with Python, making it a natural choice for implementing the system. The development process is streamlined by using Python, resulting in faster development.

4.2.2. Twint

As stated above Twint is a python library used for scraping tweets from Twitter without making use of Twitter’s API.

4.2.3. Codecs Module

Codecs is a Python module that is used in the monitoring system to manage file I/O and encoding/decoding of files. The codecs module is important to ensure that the code can manage non-ASCII characters in the input and output data.

4.2.4. Utf-8 Encoding

UTF-8 encoding is a variable-length character encoding that can represent any character in the Unicode standard, including ASCII characters. The UTF-8 encoding is used in the monitor system's code.

4.2.5. Vader

As discussed in the design chapter, Vader is a sentiment analysis tool that can consider sarcasm, irony, and casual human language.

4.2.6. Matplotlib

Matplotlib is a Python library used in the sentiment analysis system to create a pie chart that shows the counts of each sentiment type (positive, negative, and neutral) and a bar chart that compares the negative/positive sentiment of both deleted and all the tweets.

4.2.7. Emojis

The "import emoji" module was incorporated into the sentiment analysis system to perform sentiment analysis on emojis found in tweets. This module allowed for the recognition of a wide range of emojis and their associated sentiment, which was then used to calculate the overall sentiment score of a tweet.

4.2.8. Os

Os is a Python module used in the sentiment_analysis() function, of the sentiment analysis system, to extract the filename of the input data file without its extension. The os.path.splitext() function is called on the input data_file to separate the filename from its extension. The resulting filename is then used to construct the name of the output file in which the results of the sentiment analysis scores will be stored.

4.2.9. Visual Studio Code

Visual Studio Code is a code editor with a large library of extensions available for download that allow developers to add new features and functionality to the editor (Microsoft, 2016). It supports a wide range of programming languages, including JavaScript, Python, C++, and many others, and was used when coding this project.

4.2.10. GitHub

GitHub is a web-based platform that provides hosting for software development version control using Git. It allows developers to store and manage their code repositories in a collaborative way (Jackson, 2018). GitHub was used in this project to store the data and share it with this project's supervisors.

4.3. 'Doubtfire' System Architecture

The 'Doubtfire' system was created to identify patterns in deleted tweets by analysing their sentiment. It consists of two stages, which are illustrated in Figure 4.1. Stage one is the monitoring system stage, which monitors deleted tweets, and stage two, the sentiment analysis system stage, which evaluates the sentiment of the deleted tweets.

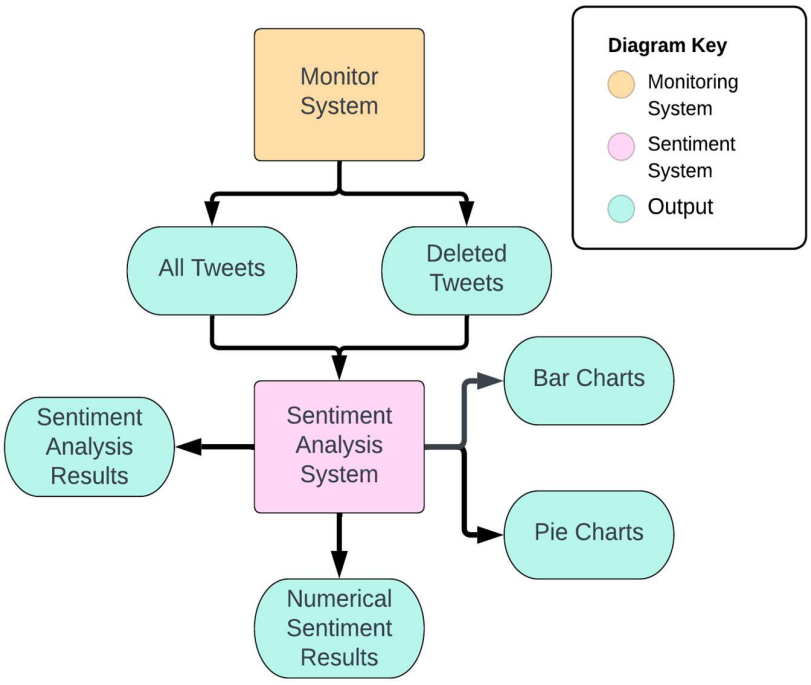

Figure 4.1 :'Doubtfire' System Architecture

4.3.1. Monitoring System

The monitor system serves as the primary component of the Doubtfire system, which uses the Twint library to perform Twitter scraping without using the Twitter API as classified by requirement 1 (***R1)*** of the design requirements. This system is based on task group 1 (the monitoring system tasks) of the design process. This system executes a Python script that imports Twint and performs Twitter searches on

selected keywords over a specific period. For this system, the period covered five days in early February 2023 (the 6th to the 10th), with each day beginning at 10 am and running for an hour. The hour was then set as i, and each subsequent rerun added ten minutes, with the equation i+n used to add ten minutes throughout the day as it reran. The resulting tweets are then saved in CSV files, with file names reflecting the date and time of the search.

The script identifies deleted tweets by comparing the files generated from successive searches and cross-referencing the corresponding data for each day. Each tweet is given a unique 10-character ID and if this ID is absent from both files, the tweet is assumed to be deleted and is saved in a new separate file. All data is then consolidated into a single file. This file is verified against the full file of the day, scraped 24 hours later, for the given timeframe. The deleted tweets file is regularly updated, and duplicates are removed similarly to that of the full dataset file.

4.3.1.1. Keyword Selection for The Monitoring System

The Keyword selection needed to be in line with requirement 2 (***R2***) of the design of this system hence the selection of keywords, which include ‘Musk’, ‘Trump’, ‘Tesla’, ‘SpaceX’, ‘Kanye’, and ‘Biden’, was deliberate and were each chosen for specific reasons. ‘Musk’ was included as Elon Musk is the owner of Twitter, while SpaceX and Tesla were selected due to Musk’s ownership of them. Musk's prior support for Donald Trump made it intriguing to investigate whether ‘Trump’ would receive preferential treatment on Twitter, although Donald Trump currently uses his own social media platform called 'Truth Social' (Milmo, 2023). ‘Biden’ was chosen as a baseline since Musk is not a Democrat and has expressed a dislike for Joe Biden, the current US president, making negative tweets against him unlikely to be removed. ‘Kanye’ was chosen due to Kanye West’s past postings of highly anti-Semitic and hateful content on Twitter, although he was suspended only after posting hateful content and mocking Musk on the same day, which resulted in controversy over his suspension. However, the Doubtfire system was unable to accumulate a significant database on the keyword ‘Kanye’ as Kanye West currently goes by the name 'Ye'. Since ‘ye’ is commonly used as you (plural) and as slang for ‘yes,’ there are too many tweets with the word 'ye' to create an accurate database.

4.3.1.2. Runtime and Scale

It should be noted that the current implementation of the monitoring system exhibits varying runtimes for data collection, ranging from a few seconds to as much as half an hour. This variability can be attributed to differences in the sizes of the datasets being collected, with the number of tweets in each time zone playing a significant role. Additionally, less frequently used keywords may require less time for collection. It is important to acknowledge that the low level of machinery utilised in the project may also contribute to the observed variability in runtime.

4.3.2. Sentiment Analysis System

This sentiment analysis system is a comprehensive tool that excels in analysing both deleted and non-deleted tweets. This system is based on task group 2 (the sentiment analysis system tasks) of the design section. It uses the highly accurate Vader library, renowned for its ability to identify and analyse positive and negative sentiment in texts, including sarcasm and irony. The system's impressive ability to identify and decipher mixed sentiment is one of its unique features, as it can accurately identify and score the sentiment of each text element, even in the presence of mixed sentiment, a step that was needed to fulfil requirement 3 (***R3***) of the design of this system.

One of the unique features of the tool is its ability to analyse emojis. Emojis have a huge sentimental value and can be used to describe various emotions such as happiness, sadness, anger, and others. In fact, emojis can completely change the sentiment of a tweet. For example, consider the tweets "Wow that's amazing 😊" and "Wow that's amazing 😡". Both tweets have the same text, but the emojis change the sentiment of the tweet entirely. Over 10 billion emojis were used on Twitter between 2013-2015, and new emojis are constantly being added. However, emojis can mean different things to different people, making it challenging to analyse them accurately. The emojis used for this system were based on research conducted by Kralj Novak et al. in 2015. This was in accordance with Requirement 4 (***R4***) of the system's design.

The system was then tested with other sentiment analysis tools, such as Textblob and Afinn, and Vader was found to perform better on social media platforms like

Twitter. This is because social media platforms often have their unique language and slang that may not be present in articles and reviews.

The sentiment analysis results are presented in a pie chart that shows the percentage of sentiment distribution in the collected and deleted tweets. This helps to give an overall idea of the sentiment of the tweets. An example of this output is shown in Figure 4.2. Additionally, the system produces a sentiment analysis file that appends the calculated sentiment score to the end of each tweet and labels each tweet, positive, negative, or neutral. This facilitates testing and analysis, making it easier to identify trends in sentiment over time.

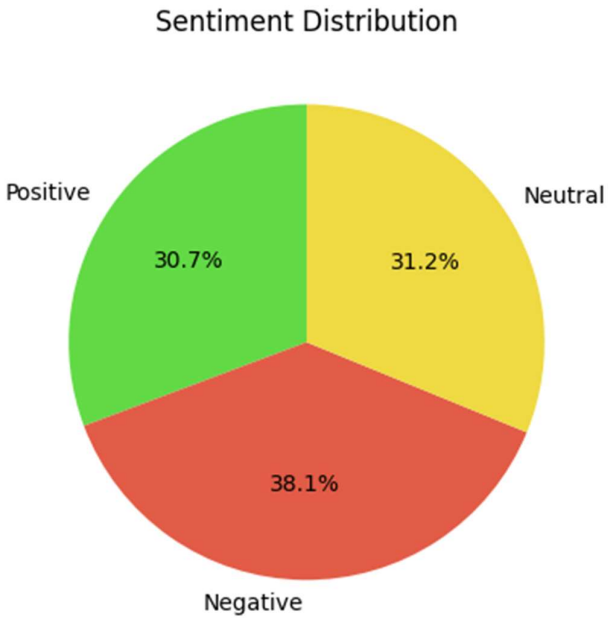

Figure 4.2 :Pie chart depicting the Sentiment Distribution of Deleted Tweets for the Keyword 'Biden'.

The system generates bar charts that highlight the data in a percentage format, comparing the number of tweets to the number of deleted tweets. The bar charts and pie charts generated by this system offer a clear representation of the data, allowing users to interpret the results of the sentiment analysis quickly and easily. The charts are particularly useful when comparing the sentiment of tweets versus deleted tweets, as they provide a quick overview of the overall sentiment of the data set. This facilitates the identification of any underlying trends and patterns in the data, which may not be immediately obvious from the raw sentiment analysis results. It also

allows us to create a digital analysis of the results, making it easier to identify trends and patterns, and draw actionable insights from the data.

4.4. Chapter Summary

This chapter centres on the implementation of the proposed 'Doubtfire' system, intended for monitoring and analysing deleted tweets. The chapters aim was to provide a comprehensive understanding of the development process of the 'Doubtfire' system and the technologies utilised to accomplish this goal. It provides an overview of the technologies employed in building the system and the methodology behind the 'Doubtfire' system's construction. The subsequent chapter will delve into the results and evaluations of the trends discovered.

5. Evaluation

5.1. Introduction

To identify instances of manipulation on social media, a comprehensive examination of the sentiment charts and numerical analyses produced by this ‘Doubtfire’ system was undertaken. For this study, it is assumed that all the tweets are being deleted by Twitter. The ensuing section will explain the discoveries made and the constraints associated with this study.

5.2. Evaluation Results - Keyword Sentiment Distribution Analysis

The primary objective of this study was to identify instances of manipulation on social media, with a focus on Twitter. To achieve this goal, we conducted a thorough analysis of the sentiment analysis results for each of the selected keywords. This approach enabled us to examine potential patterns of manipulation on social media. In this section of the report, each keyword is analysed individually. It is important to note this study was conducted over 5 days from 10 am to 4 pm (IST) from the 6th to the 10th of February 2023.

5.2.1. Kanye

As outlined in the implementation section, it was necessary to remove the keyword ‘Kanye’ from our analysis due to Kanye West now being referred to as ‘Ye’.

5.2.2. Trump

The keyword ‘Trump’ is frequently used on Twitter, this owes to its association with a former President of the United States, Donald Trump. During this study, the keyword ‘Trump’ received an average of 27,888 tweets per hour, resulting in a total of 836,623 tweets over the 30-hour period. Out of these tweets, only a small proportion of 103,720 (12.4%) were deleted. This observation may be attributed to the fact that the former President, Donald Trump, is no longer active on Twitter and has created his own social media platform, "Truth Social" (Forman-Katz and Stocking, 2022). The sentiment analysis of the tweets, both deleted and collected, is presented in Figure 5.1 and 5.2.

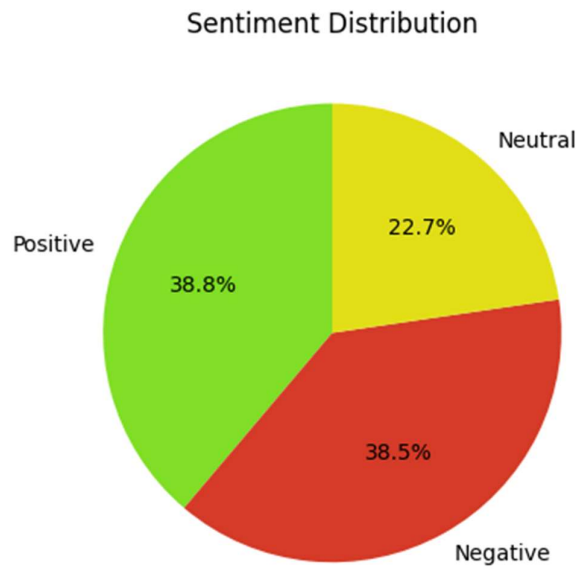

Figure 5.1 :Pie chart depicting the Sentiment Distribution for all Deleted Tweets containing the Keyword 'Trump'.

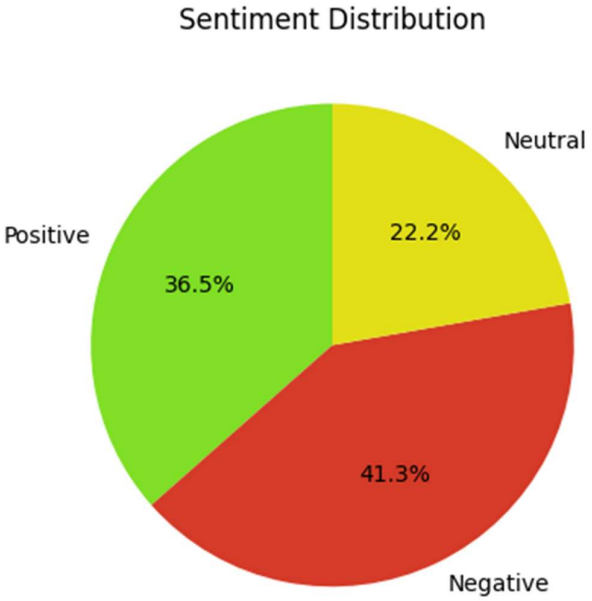

Figure 5.2 :Pie chart depicting the Sentiment Distribution for all Tweets containing the Keyword 'Trump'.

5.2.2.1. Negative Sentiment

Within the period of our study, the keyword ‘Trump’ received a total of 345,525 tweets with a negative sentiment. Out of these, 39,932 were deleted. Although the keyword ‘Trump’ received a high percentage of negative tweets (41.3%), only a small proportion of these negative tweets (11.56%) were deleted. This high percentage observed could be attributed to the recent decision by the Koch network to oppose Donald Trump's bid for the US presidency in 2024 (Williams, 2023) as this happened during the period in which the study was being conducted. This event may have resulted in a surge of negative tweets directed towards the keyword ‘Trump’, as individuals expressed their disapproval of Donald Trump.

5.2.2.2. Positive Sentiment

The keyword ‘Trump’ received a significantly smaller percentage of positive tweets (36.5%) in comparison to negative sentiment tweets (41.3%) during the study period. Despite receiving 305,367 positive tweets during the study, only a small proportion (13.79%), albeit higher than that of the negative tweets, were deleted.

5.2.2.3. Trends Analysis

There is no notable disparity in the sentiment distribution of collected and deleted positive and negative tweets containing the word ‘Trump’. The percentage distribution of positive sentiment tweets between collected (36.5%) and deleted (38.8%) increases by 2.3%. Furthermore, the marginal difference in negative sentiment distribution of 2.8% between deleted (38.5%) and collected (41.3%) negative tweets is insignificant. As a result, it can be inferred that there is not a disparity in the handling of negative and positive tweet deletion, and there is no indication of any manipulation in the deletion of tweets related to the keyword 'Trump.’

5.2.3. Tesla

Compared to the frequently used keyword ‘Trump,’ the keyword ‘Tesla’ is not as commonly found on Twitter. This is likely since ‘Tesla’ is associated with a company and not an individual. The usage of ‘Tesla’ on Twitter was found to be at a rate of approximately 1,855 tweets per hour during the study period. A comparatively smaller number of tweets were collected for this analysis, with only 55,652 tweets being recorded, and of those, only 4,958 were deleted. This represents a remarkably small proportion of 0.089%. The results of the sentiment analysis for both collected and deleted tweets are presented in Figures 5.3 and 5.4, respectively.

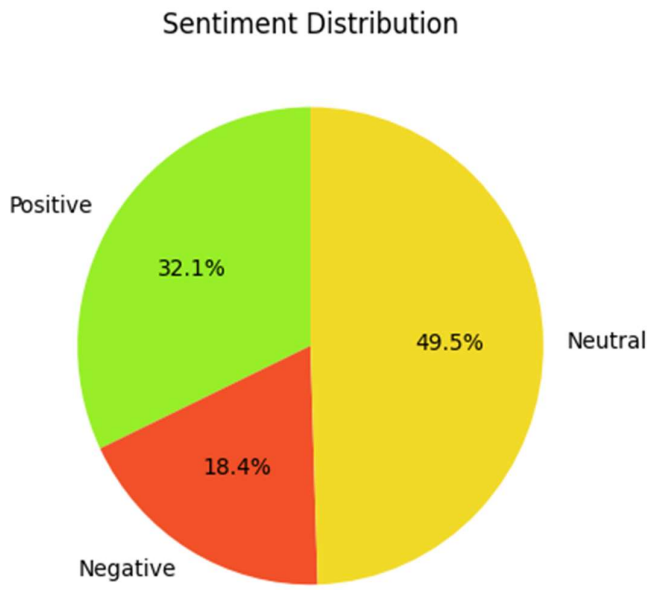

Figure 5.3 :Pie chart depicting the Sentiment Distribution results for all Deleted Tweets containing the Keyword 'Tesla'.

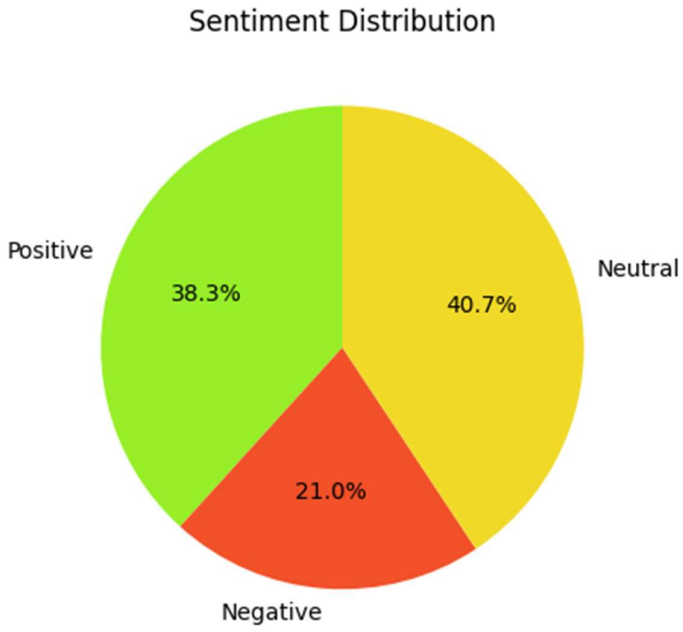

Figure 5.4 :Pie chart depicting the Sentiment Distribution for all Tweets containing the Keyword 'Tesla'.

5.2.3.1. Negative Sentiment

Throughout the duration of this study, tweets containing the keyword ‘Tesla’ demonstrated a negative sentiment percentage of 21%, indicating that 1 in 5 tweets associated with this keyword were negative. Out of the collected tweets, 11,687 were negative, representing many critical views. However, out of those negative tweets, only 912 were deleted, which amounts to a mere 0.078%.

5.2.3.2. Positive Sentiment

The keyword 'Tesla' yielded a higher percentage of positive sentiment tweets, with 38.3% or 21,315 of the collected tweets reflecting a positive sentiment. This surge in positive tweets could be attributed to the increase in Tesla's stock market price throughout February, which climbed by 18.8% (Yang, 2023). This event could have influenced the results obtained. It is worth noting that out of the 21,315 tweets, only 1,592 were deleted, indicating a lower deletion rate in the positive sentiment tweets. Specifically, only 0.075% of positive sentiment tweets were deleted.

5.2.3.3. Trend Analysis

Once again, we observe an insignificant difference between the sentiment distribution percentage of negative sentiment tweets collected and those deleted. Specifically, negative tweets deletion had only a sentiment distribution decrease of 2.6% (21% and 18.4% respectively). However, there was a significant sentiment distribution decrease of 6.2% in the deletion of positive tweets (32.1%) compared to the positive tweets collected (38.3%). Interestingly, the keyword 'Tesla' exhibited a significant sentiment distribution increase in the deletion of neutral tweets, with an 8.8% rise (40.7% and 49.5% respectively). While manipulation cannot be ruled out, it is worth noting that 81.6% of the deleted tweets and 79% of the collected tweets were either positive or neutral. Due to this fact and the small sample size, it is difficult to ascertain if any manipulation occurred.

5.2.4. SpaceX

Like 'Tesla', 'SpaceX' is also associated with a company name and, therefore, received significantly fewer tweets during the experiment period. In fact, it received the least amount of attention, with an average of only 742 tweets per hour. Over the 5-day period, a total of 22,273 tweets mentioned 'SpaceX', of which only 0.0113% (250 tweets) were deleted. The sentiment analysis results for both collected and deleted tweets are presented in Figures 5.5 and 5.6.

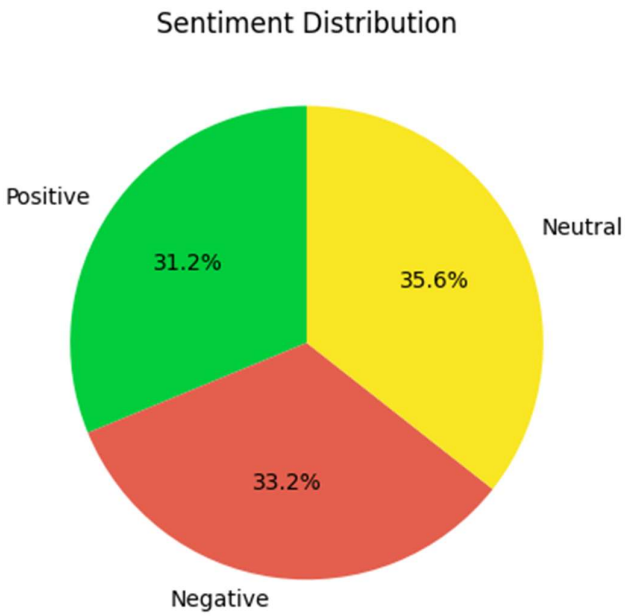

Figure 5.5 :Pie chart depicting the Sentiment Distribution for all Deleted Tweets containing the Keyword 'SpaceX'.

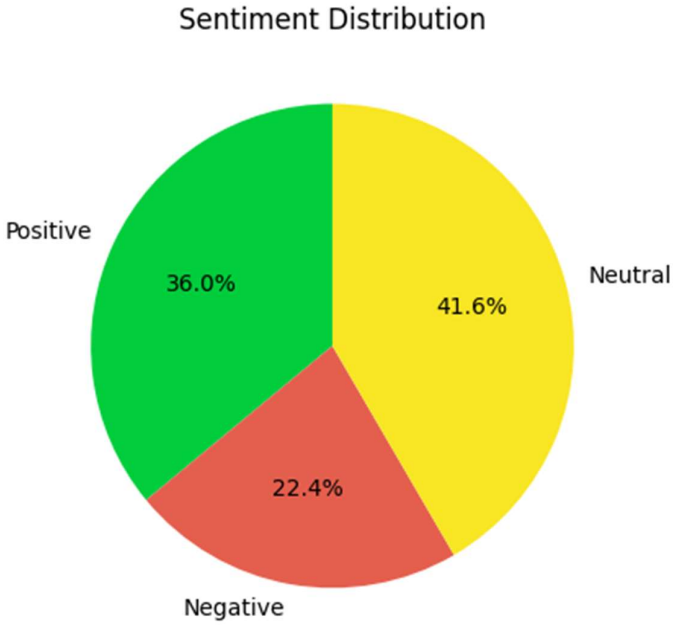

Figure 5.6 :Pie chart depicting the Sentiment Distribution for all Tweets containing the Keyword 'SpaceX'.

5.2.4.1. Negative Sentiment

The term 'SpaceX' was identified in a total of 4,989 negative sentiment tweets, of which only 83 were deleted. While the absolute number and percentage of deletions may seem small (0.016% of all negative tweets), it is worth noting a significant disparity between the percentage of sentiment distribution of the deleted and collected negative sentiment tweets. Specifically, while only 22.4% of collected tweets were negative, 33.2% of deleted tweets had negative sentiment. This represents a notable sentiment distribution increase of 10.8%.

5.2.4.2. Positive Sentiment

The keyword 'SpaceX' exhibited a significantly higher percentage of positive tweets, with a total of 8,018 tweets (36%) displaying a positive sentiment. Out of this group of 8,018 tweets, only 78 were deleted. Although, again, the absolute number and percentage of deleted positive tweets (0.0097% of all positive tweets) may appear small, the difference of 4.8% between the sentiment distribution percentage of deleted (31.2%) and collected (36%) positive tweets could be seen as noteworthy.

5.2.4.3. Trend Analysis

The keyword 'SpaceX' demonstrated a noteworthy increase in the sentiment distribution of deleted negative sentiment tweets and a significant decrease in the distribution of deleted positive sentiment tweets compared to that of the collected tweets. This trend raises the possibility of potential manipulation in the deletion of tweets associated with 'SpaceX'. It is also important to note there was a significant decrease in the sentiment distribution of neutral sentiment for the collected and deleted tweets (6%). However, due to the limited data set available for analysis, we cannot draw definitive conclusions. Nevertheless, it is important to monitor this trend in the future to determine if further action is necessary.

5.2.5. Biden

The keyword 'Biden' served as the baseline for this project. Joe Biden presently holds the position of President of the United States of America, and therefore, it is unsurprising that the term 'Biden' had the largest data set, with an average of 38,455 tweets per hour. The dataset for the keyword 'Biden' comprised 1,153,679 tweets, of which 26.3% (303,462) were deleted. The sentiment analysis results for both the collected and deleted tweets are presented in Figures 5.7 and 5.8, respectively.

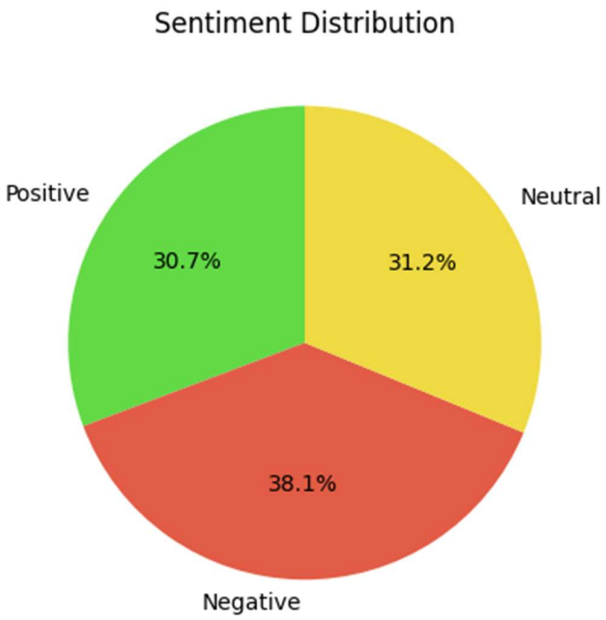

Figure 5.7 :Pie chart depicting the Sentiment Distribution for all Deleted Tweets containing the Keyword 'Biden'.

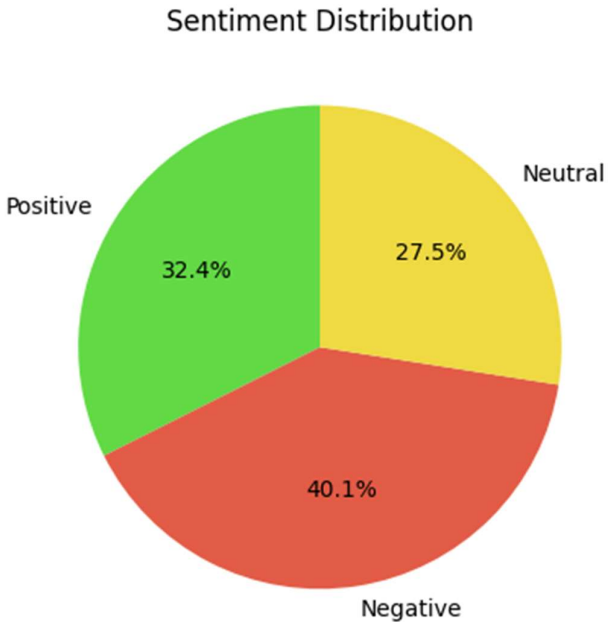

Figure 5.8 :Pie chart depicting the Sentiment Distribution for all Tweets containing the Keyword 'Biden'.

5.2.5.1. Negative Sentiment

The data collected for the keyword 'Biden' has revealed a considerable number of negative tweets, with 46,147 such tweets being identified. Out of these, 115,619 tweets were deleted, resulting in a deletion percentage of 24.99%, which indicates that one in four negative tweets containing the word 'Biden' were deleted. Given Joe Biden's high public profile as the current President of the United States, it is not surprising to observe a higher percentage of negative tweets directed towards him. Additionally, his delivery of the second State of the Union address on February 7th, 2023 (Biden, 2023) may have contributed to the increase in negative tweets related to him.

5.2.5.2. Positive Sentiment

The data also reveals a sizable number of positive tweets for the keyword 'Biden'. In total, there were 373,792 positive tweets, out of which 93,163 tweets were deleted, indicating a deletion percentage of 24.92%. This observation, much like the negative tweet deletions, suggests that one in four positive tweets containing the word 'Biden' were deleted. It is possible that Joe Biden's second State of the Union address contributed to the high number of positive tweets as well as negative tweets.

5.2.5.3. Trend Analysis

Based on the analysis of the data, it can be concluded that there was no significant manipulation with respect to the deletion of tweets associated with the keyword 'Biden'. This conclusion is drawn from the fact that there is no substantial difference between the sentiment distribution and deletion percentages of negative and positive sentiment tweets, with both having a small decrease in the percentage sentiment distribution from that of the collected tweets (2% and 1.7%, respectively).

5.2.6. Musk

The selection of the keyword ‘Musk’ for this study was based on Elon Musk's acquisition of Twitter in October 2022. Being a well-known public figure and the CEO of Twitter, the term ‘Musk’ is commonly used on the platform. During the study, ‘Musk’ was tweeted 32,634 times per hour, with a total of 979,011 tweets collected. However, it is noteworthy that 32.61% (319,211) of these tweets were deleted, the highest percentage of deletion among all keywords studied. The sentiment analysis results for both the collected and deleted tweets are presented in Figures 5.9 and 5.10, respectively.

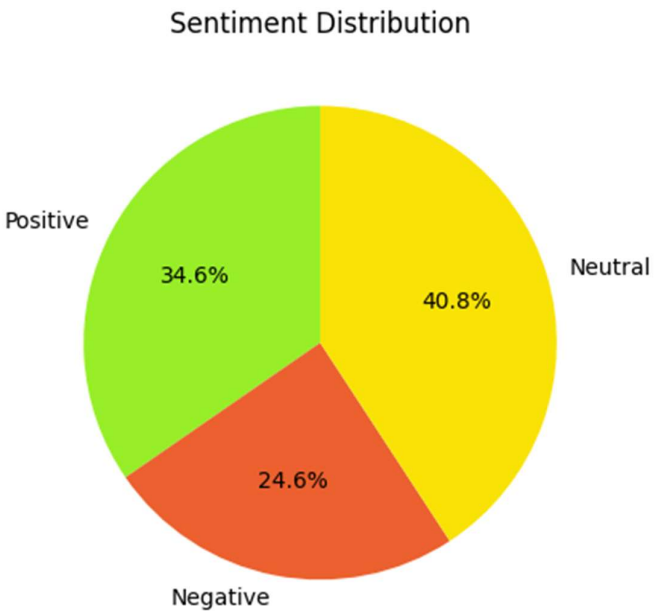

Figure 5.9 :Pie chart depicting the Sentiment Distribution for all Deleted Tweets containing the Keyword 'Musk'.

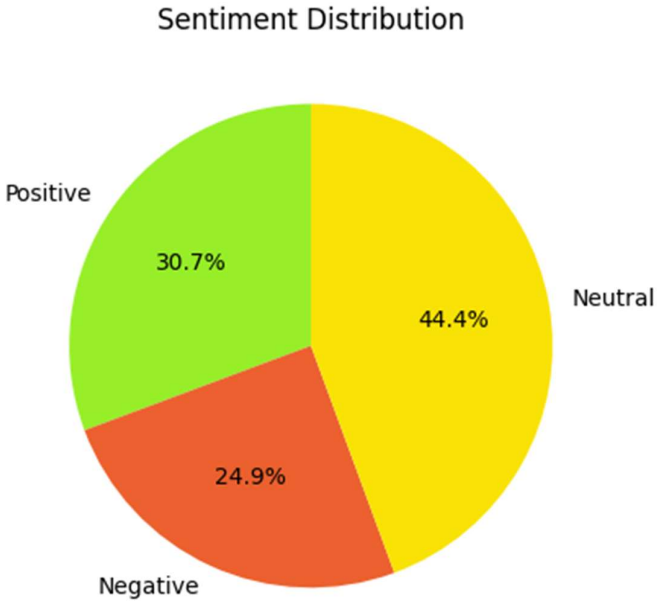

Figure 5.10 :Pie chart depicting the Sentiment Distribution for all Tweets containing the Keyword 'Musk'.

5.2.6.1. Negative Sentiment

The data collected for the keyword 'Musk' revealed a total of 243,774 negative tweets, out of which 78,526 tweets were deleted, representing a significant deletion percentage of 32.21%. While the 0.3% decrease in negative sentiment distribution between deleted (24.6%) and collected (24.9%%) tweets may not be significant, the considerable number of deleted tweets is notable. It is worth noting that the private Twitter API implemented by Elon Musk and the Twitter outage on February 9th, 2023 (Association, 2023) could have contributed to the high number of negative tweets containing the word 'Musk' during that week.

5.2.6.2. Positive Sentiment

The data collected showed that the word ‘Musk’ received a total of 300,556 positive tweets during the experiment period, out of which 110,447 were deleted. This represents a considerable percentage of 36.75% and highlights the high rate of deletion for positive tweets related to Elon Musk. However, the increase in positive sentiment distribution between collected (30.7%) and deleted (34.6%) tweets was only 3.9%, indicating a small difference.

5.2.6.3. Trend Analysis

Based on the data collected, there is a high percentage of positive and negative tweets containing the word 'Musk' being deleted, which is higher than the average for other keywords analysed. However, it is important to note that the sentiment analysis did not show a significant difference in the distribution between the sentiment of collected and deleted tweets. Therefore, while it is possible that there has been manipulation of tweets containing the word 'Musk', this cannot be conclusively determined based on this data alone.

5.3. Evaluation Results - Collective Analysis

In this section of the report, a thorough analysis will be conducted on all the data collected, with particular focus on the sentiment of deleted tweets through a comparative analysis.

5.3.1. Graphical Analysis of Sentiment Distribution

Figure 5.11 presents a visual representation of the comparison between the negative sentiment analysis of collected and deleted tweets for each keyword. The corresponding colour for each keyword is shown in the key on the side of the figure.

The left-side bar represents the negative sentiment analysis of deleted tweets, while the right-side bar represents the negative sentiment analysis of collected tweets for that keyword.

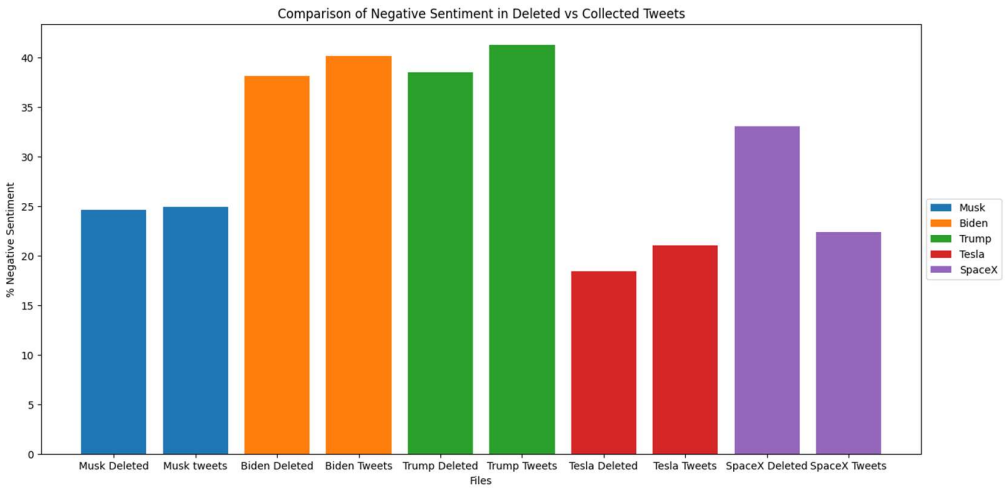

Figure 5.11 :Graphical Comparison of the Negative Sentiment in the Deleted vs. Collected Tweets.

Figure 5.11 reveals that for all keywords, the reduction in the negative sentiment distribution of deleted tweets compared to collected tweets is insignificant, except for SpaceX, which is an outlier with a 10.8% increase in negative sentiment distribution. The data suggests that manipulation may not have occurred in terms of sentiment distribution between deleted and collected tweets.

Similarly Figure 5.12 presents a visual representation of the comparison between the positive sentiment analysis of collected and deleted tweets for each keyword. Similarly, to Figure 5.11, the corresponding colour for each keyword is shown in the key on the side of the figure. The left-side bar represents the positive sentiment analysis of deleted tweets, while the right-side bar represents the positive sentiment analysis of collected tweets for that keyword.

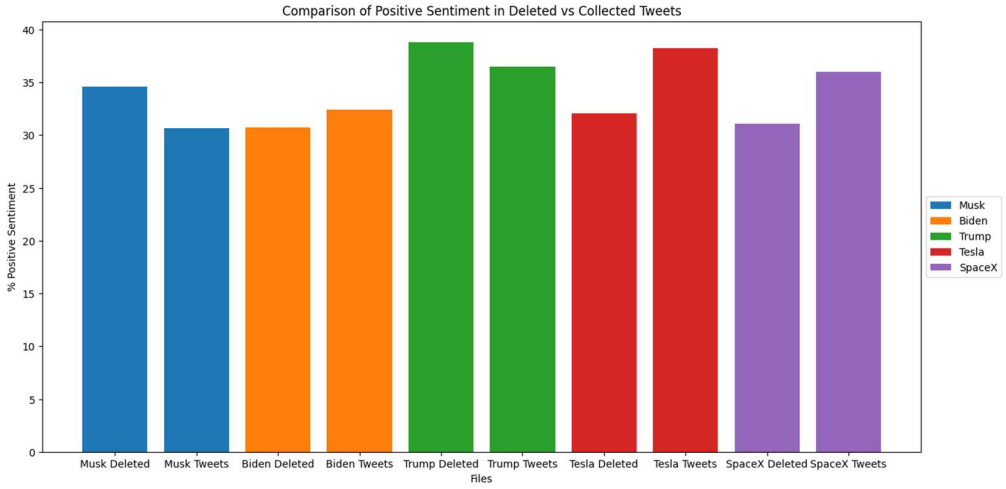

Figure 5.12 :Graphical Comparison of Positive Sentiment in the Deleted vs. Collected Tweets.

The results presented in Figure 5.12 demonstrate that there is no significant variance in sentiment distribution between deleted and collected tweets for the keywords 'Trump' and 'Musk' (with marginal changes of 2.3% and 3.9% respectively). In addition, there is a slight decrease in sentiment distribution for the keyword 'Biden' (1.7%). However, the analysis revealed an unusual pattern for the keyword 'Tesla', with a significant increase of 6.2% in the sentiment distribution of deleted tweets compared to collected tweets. Conversely, the keyword 'SpaceX' exhibited a significant decrease of 4.8% in the sentiment distribution of deleted vs. collected tweets. This could suggest the presence of manipulation for 'SpaceX' but not for 'Tesla'.

5.3.2. Graphical Analysis of Deletions

Figure 5.13 provides a visual representation of the collected tweets over time for each keyword.

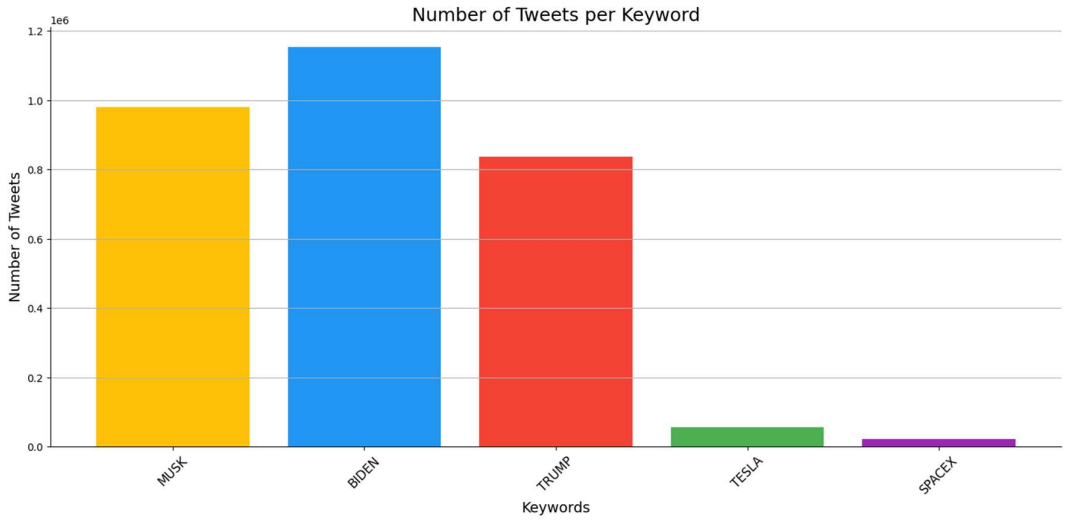

Figure 5.13 :Graphical Representation of The Number of Tweets per Keyword

The data presented in Figure 5.13 clearly indicates that the keywords ‘Tesla’ and ‘SpaceX’ have significantly fewer tweets in the database when compared to ‘Musk,’ ‘Biden’ and ‘Trump’. Therefore, for the purpose of this analysis, we will exclude them and focus solely on the keywords ‘Musk,’ ‘Biden’ and ‘Trump’. Interestingly, during the same period, the keyword ‘Biden’ received 15.2% more tweets than ‘Musk’ and 27.5% more tweets than ‘Trump.’ This observation suggests a higher level of online activity and interest surrounding the keyword ‘Biden,’ which could be attributed to his second State of Union taking place during this project’s period, as well as his position as the current US president. It is also worth noting that the keyword ‘Musk’ was tweeted 14.5% more than ‘Trump.’ This could be due to a variety of factors, including Twitter crashing during this period, Donald Trump not using Twitter as his main social media platform, and the introduction of Twitter’s private API during this period.

Upon analysing the deletions of tweets during the specified period for the keywords 'Trump', 'Biden', and 'Musk', noteworthy trends can be observed, as illustrated in Figure 5.14.

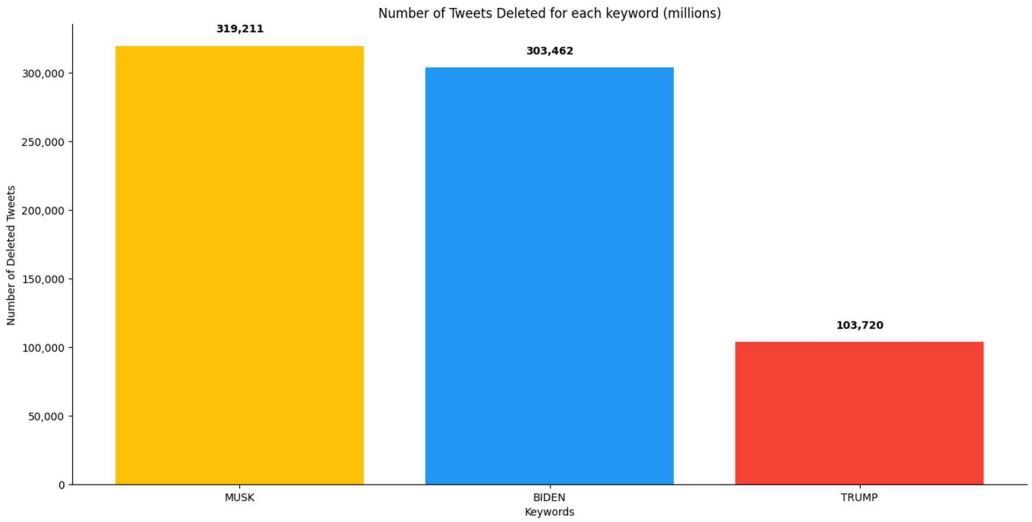

Figure 5.14 :Number of Tweets deleted for Each Keyword

Despite receiving significantly fewer tweets than 'Biden,' the keyword 'Musk' exhibits some interesting trends in tweet deletions when compared to 'Biden' and 'Trump.' Figure 5.14 illustrates that 'Musk' experienced a 5% higher deletion rate compared to 'Biden', despite having fewer total tweets (979,011 versus 1,153,679). Moreover, a negative tweet containing the word 'Musk' had a 23% higher chance of being deleted compared to one containing the word 'Biden', despite 'Biden' receiving 47% more negative tweets than 'Musk'. Interestingly, although 'Musk' only had a 14.5% increase in total tweets compared to 'Trump,' it experienced a staggering 67.5% more deletions. Additionally, a negative tweet containing the word 'Musk' had a 49.2% higher chance of being deleted than one containing the word 'Trump,' despite 'Trump' receiving 29.5% more negative tweets than 'Musk.' Although 'Musk' received slightly fewer positive tweets than 'Trump' (4,811 tweets or 1.6% less) and 'Biden' (73,226 or 19.5%), it also had the most deleted positive tweets. Specifically, a positive tweet containing the word 'Musk' had a 15.6% higher chance of being deleted compared to one containing 'Biden' and a 63.5% higher chance compared to one containing 'Trump.'

5.4. Technical Evaluation

The 'Doubtfire' system, designed for monitoring the deletion and sentiment of tweets, has been found to be an exceptionally effective tool for social media manipulation analysis. This system uses advanced programming systems and languages such as Python and has proven to be efficient in collecting and analysing tweets. It has

effectively demonstrated the capability to discern and account for sarcasm and emojis in the sentiment analysis of tweets, adding an additional layer of accuracy and complexity to its analysis.

Despite the low-level technology on which the system is built, it has effectively met its goals and proven to be a valuable tool in understanding tweet deletion trends and manipulations. It provides researchers and analysts with a powerful means of monitoring social media behaviour, offering a comprehensive view of the changing landscape of online discourse.

Furthermore, the ‘Doubtfire’ system's success can serve as a solid foundation for future studies of social media manipulation, as well as an essential tool for social media platforms and policymakers in combating the spread of disinformation and harmful online behaviour. Its effectiveness in analysing social media content and identifying key trends can provide valuable insights to improve online safety and promote responsible online behaviour.

5.5. Evaluation Summary

The presented evidence suggests the possibility of manipulation on social media when it comes to the deletion of tweets containing the keyword 'Musk'. Despite receiving fewer tweets than 'Biden', 'Musk' had a higher percentage of tweets deleted, indicating targeted efforts to remove negative sentiment tweets related to 'Musk'. The fact that negative tweets containing the keyword 'Musk' were more likely to be deleted than negative tweets containing the keywords 'Biden' and 'Trump' also raises suspicions. Additionally, the higher likelihood of positive tweets containing the keyword 'Musk' being deleted compared to positive tweets containing 'Biden' and 'Trump' suggests a lack of impartiality in handling tweet deletion for 'Musk'.

Moreover, some suspicious trends can be observed in the distribution of sentiment in the deleted versus collected tweets when looking at the keywords 'SpaceX' and 'Tesla'. The significant jump in negative sentiment deletions and the significant drop in positive sentiment deletions for 'SpaceX' can be seen as biassed, while the substantial increase in positive sentiment deletions for 'Tesla' cannot be considered as biassed.

Overall, these trends in the deletion of tweets suggest a potential bias towards 'Musk', and it is possible that some form of manipulation participated in the deletion of tweets related to this keyword e.g., ‘SpaceX’. Further investigation may be necessary to determine the cause of these patterns and to ensure the impartiality of social media platforms in handling content related to public figures.

5.6. Chapter Summary

This chapter provides an overview of the main findings of this study on deleted tweets on Twitter. It highlights the significance of the study's results and the need for further research to better understand the phenomenon of tweet deletion on social media platforms. This study provides valuable insights into the potential manipulation of online discourse and the importance of transparency and impartiality in social media moderation.

6. Future Work

6.1. Chapter Introduction

This chapter aims to explore the future work that can be done with this 'Doubtfire' system. It will evaluate the limitations of the system and propose practical solutions to overcome these issues. Also, it will address this study's importance in the context of the ongoing ‘War on Information’ and how technical advancements can contribute to combating this issue. This chapter highlights the significance of this study for future generations, as it can be used as a foundation for further research in the field of social media analysis and information manipulation.

6.2. System Limitations

This project is subject to several limitations, which will be discussed in this section.

6.2.1. Time Zone

This study was conducted during Ireland Standard Time (IST), which is 7 hours ahead of the centre of America. As of 2023, Twitter has approximately 450 million monthly active users, with the United States having the highest number of Twitter users at 79.6 million (Ruby, 2022). Recent studies suggest that the period between 6 am to 12 pm (UTC-4) is the least optimal time to post on Twitter (Kolowich, 2019). Given that this study was conducted in the UTC+1 time zone, the databases could contain fewer tweets due to the time difference with the United States. Additionally, many of the keywords used in this study are associated with American public figures or American companies, which further reduces the number of available tweets to collect given the time zone.

6.2.2. Tweet Deletions

Tweet deletions in this study are subject to certain limitations. Although in this study, it is assumed that Twitter removed all the deleted tweets, this cannot be definitively confirmed. Users may delete tweets for several reasons, including correcting spelling or grammatical errors, changing opinions, responding to backlash, or improving public image (Almuhimedi et al., 2013). To determine whether a tweet was deleted by the user or by Twitter, manual checking of each tweet in the database against Twitter's user policies would be the most appropriate approach. However, even this

approach is limited, as it can only assume that the user deleted the tweet if it does not violate any of Twitter's policies, which is not always the case.

6.2.3. Bots

An important limitation of the present study is its lack of consideration for the potential presence of bots on Twitter. In recent years, the presence of bots on social media platforms, including Twitter, has raised significant concerns about the reliability and accuracy of information dissemination and public opinion formation. Bots are automated accounts that can create and distribute content, interact with other users, and influence public discourse. Estimates suggest that a substantial portion of Twitter users may be bots. For instance, it has been reported that there are around 16.5 million bots on Twitter, accounting for 5% of the total Twitter users, although some scholars argue that this figure may be underestimated (Maxwell, 2022). This absence of consideration for Bots may impact the reliability and accuracy of the study's findings.

6.2.4. Limited Timeframe and Keywords

Another limitation of this study is the focus on a specific set of keywords. The chosen keywords, while relevant to the public figures and companies analysed in this study, may not provide a comprehensive picture of social media manipulation. Additionally, the study was conducted over a limited period, and therefore, may not capture long-term trends in tweet deletion behaviour.

6.3. War on Information

This study holds significant importance due to the ongoing conflict known as the "War on Information". The term refers to the struggle between entities aiming to prevent the dissemination of false information and those seeking to manipulate and control the flow of information. This conflict is one of power and influence in the digital age. False news has emerged as a growing challenge, with research indicating that it is more likely to be retweeted on social media than true news. A study by Vosoughi, Roy, and Aral (2018) found that, on average, false news stories were retweeted approximately 1.1 times more often than true news stories. Such false narratives are rampant on social media, necessitating caution. Various techniques, such as machine learning, natural language processing, and network analysis, have been developed to detect and combat Fake News (Zhou and Zafarani, 2020). However, detecting censorship or manipulation on social media,

which also constitutes a form of Fake News, remains a challenge. Therefore, further research is required to identify the impact of manipulated deletions on public opinion and to develop strategies to combat such manipulation.

6.4. Future Research

There are many opportunities for future research for this project. All of which will be outlined in this section.

6.4.1. Comprehensive Data

To enhance the scope and validity of this study, a more comprehensive dataset can be incorporated in future research. It may be useful to examine additional keywords, especially using ‘left-winged’ or ‘right-winged’ celebrities and politician’s names as keywords to investigate any potential biases towards a specific political party. Further investigation into keywords and hashtags related to public figures or topics susceptible to social media manipulation is also necessary. Additionally, extending this study to other social media platforms, such as Facebook and Instagram, can help determine whether deletion trends are observed consistently across different platforms.

6.4.2. Bot Analysis

To improve the accuracy and reliability of future studies, it is advisable to adopt techniques that can identify and filter out Bot accounts. One approach is to use machine learning algorithms to examine user behaviour and content patterns, allowing for differentiation between bot and human accounts. Additionally, third-party tools and services that offer insights into bot activity can aid in detecting and eliminating bot accounts from the dataset.

6.4.3. Sentiment Analysis

To enhance the comprehensiveness and accuracy of this study, it would be beneficial to adopt a more comprehensive sentiment analysis. Given that sentiment is constantly evolving and new emojis are regularly introduced, it is critical to ensure that the sentiment analysis tool is continually updated to provide accurate assessments of the tweet’s sentiment. This would enable the study to produce more comprehensive and reliable results.

6.4.4. 24 Hour Monitor

To enhance the quality and size of the dataset used in this study, future research could involve the development of a 24-hour tweet monitoring system. Such a system would provide a more extensive dataset by accounting for the most active times for tweeting in the USA and other regions beyond the IST time zone. This would result in a more comprehensive and representative sample of tweets, potentially leading to more accurate and generalizable findings.

6.4.5. Future Analysis

The ‘Doubtfire’ system is a powerful tool that can be used for future analysis and research in the field of social media manipulation. This system provides a solid foundation for developers to analyse the most used words in different sentiment tweets using word clouds or to examine the most frequently deleted user tweets. Additionally, it can be used to analyse the content of deleted tweets, providing valuable insights into the type of content that is most frequently removed from social media platforms.

The findings of this study are critical for raising questions about the manipulation of social media and its impact on online discourse. Furthermore, the ‘Doubtfire’ system serves as an essential basis for conducting various sentiment evaluations, thereby enabling researchers to obtain a comprehensive understanding of the sentiment dynamics on social media platforms.

By providing an effective means of analysing social media content, the ‘Doubtfire’ system can assist policymakers in addressing issues of online safety and promoting responsible online behaviour. The system's capabilities can also be used to combat disinformation campaigns and enhance the overall transparency of social media platforms. Overall, the ‘Doubtfire’ system represents a significant step forward in the analysis of social media trends and behaviors and is poised to become an essential tool in the fight against harmful online practices.

6.5. Chapter Summary

This chapter delves into the limitations of the 'Doubtfire' system and explores potential solutions to overcome these limitations. The 'War on Information' is also discussed and how this war makes the 'Doubtfire' system extremely relevant.

Furthermore, this chapter highlights potential future work that can be undertaken to improve the effectiveness of the 'Doubtfire' system.

7. Conclusion

The primary objective of this study was to investigate whether it was possible to detect manipulation on Twitter by analysing trends in deleted tweets. To achieve this objective, a monitoring and sentiment analysis system, the ‘Doubtfire’ system, was developed that was independent of Twitter's API and internal systems, based on a review of current events. The system's performance was then evaluated to determine its ability to detect Twitter manipulation through the trends in deleted tweets.

7.1. Objectives Summary

To address the research question of this paper, six objectives were established, ensuring the system's relevance and effectiveness. The objectives and their outcomes are summarised below, as outlined in Chapter 1:

O1- Examining the Phenomenon of Fake News and Parody.

To optimise the effectiveness of the ‘Doubtfire’ system, it was imperative to first have a clear understanding of the concepts of Fake News and parody. Chapter 2 provided a comprehensive review of the impact of Fake News and parody, as well as their definitions and motivations. This review provided a deeper understanding of the detrimental consequences that Fake News and parody can have on various scenarios, including elections. This knowledge was then applied to the development of the ‘Doubtfire’ system, ensuring that it was designed to effectively address the issue of manipulation on Twitter.

O2 - Investigating Censorship and Manipulation on Social Media.

To address the research question posed in this project, it was crucial to attain a thorough understanding of censorship and manipulation on social media. Chapter 2 presents a comprehensive review of these topics, wherein the risks and harms associated with censorship and manipulation were deliberated upon. Furthermore, the specific threat posed by Twitter was expounded upon in detail. The insights garnered from this review were then used to inform the development, execution, and evaluation of the ‘Doubtfire’ system.

O3 - Developing a Monitoring System to Identify Manipulation.

To adequately tackle the research question of this project, a system needed to be constructed for monitoring deleted tweets. However, due to the constant changes and privatisation of Twitter's external and internal systems, it was imperative that the system be built without relying on the Twitter API. Additionally, the system needed to have the capability to exclusively search for specific keywords within a given period. In Chapter 3, the project's motivations and design objectives were used to define a set of requirements, which were then categorised based on their relevance to the project's objectives. To perform the identified tasks, appropriate libraries were selected based on their performance, ability to fulfil the requirements, and compatibility with other libraries that would be used in the system. Lastly, the tasks that were essential to the ‘Doubtfire’ system and those that helped to fulfil the project's requirements and classifications were identified, providing a high-level description of the system's design and functionality. In Chapter 4, the implementation of the designed system was discussed, which enabled the monitoring of deleted tweets by using the identified libraries and fulfilling the requirements and classifications defined in Chapter 3.

O4 - Analysing Deleted Tweets for Trends and Patterns of Manipulation.

Similarly, to O3, a system needed to be developed for sentiment analysis of both collected and deleted tweets to effectively address the research question of this project. The casual nature of Twitter required the sentiment analysis system to account for sarcasm and irony. Additionally, the utilisation of over 10 billion emojis on Twitter during 2013-2015 highlights the need for the system to account for emoji sentiment to ensure accurate results. Chapter 3, as explained above, dived into the requirements, tasks and libraries used for this system. Chapter 4 delves into the implementation of this step of the system, in a comparable manner to that of O3, where the appropriate libraries were used to achieve the sentiment analysis of the collected and deleted tweets, accounting for sarcasm, irony, and emoji sentiment.

O5 - Evaluating the Extent of Manipulation on Twitter.

To comprehensively evaluate the level of manipulation on Twitter, the outcomes generated by the 'Doubtfire' system were considered. In Chapter 5, an in-depth examination of the sentiment analysis results was conducted, considering the number of tweets, the percentage of deletions, and the distribution of sentiment. The

sentiment distribution of each keyword was scrutinised in detail, and a comparative analysis of all the keywords was performed to gain a better understanding of the results and assess if manipulation was indeed transpiring.

O6 - Identifying Opportunities for Further Research.

Chapter 6 of this study focused on identifying opportunities for further research. Firstly, the limitations of the 'Doubtfire' system were addressed, acknowledging areas that require improvement to enhance its overall effectiveness. Additionally, the significance of this study was examined considering the ongoing 'War on Information,' and the potential impact it may have on future generations. Subsequently, opportunities for future research were discussed, including potential avenues for expanding upon the current study, and addressing any identified limitations of the 'Doubtfire' system. Moreover, potential solutions to these limitations were explored, highlighting areas that require further investigation and research to develop more effective solutions.

7.2. Research Question Summary

The comprehensive analysis conducted in this study leads to the conclusion that the research question, "Can manipulation on Twitter be detected by examining trends in deleted tweets?" can be answered affirmatively. The findings demonstrate that manipulation on Twitter can be detected through the examination of trends in deleted tweets, and the 'Doubtfire' system is a valuable tool for this purpose.

The placement of influential individuals in charge of regulating free speech has been proven to hinder efforts to combat Fake News. This study demonstrates the necessity of implementing a system for analysing deleted tweets on social media, which can aid in identifying the individuals responsible for deletion and their motives. The risk of censorship poses a serious threat to the preservation of free speech on Twitter.

Moreover, the results of the study highlight the significance of addressing the problem of manipulation on social media, especially in the current context of the 'War on Information.' This issue will continue to be relevant and require ongoing attention in the future.

8. Bibliography

ACLU (2019). What Is Censorship? [online] American Civil Liberties Union. Available at: https://www.aclu.org/other/what-censorship.

Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, [online] 31(2), pp.211–236. doi:https://doi.org/10.1257/jep.31.2.211.

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N. and Acquisti, A. (2013). Tweets are forever: a large-scale quantitative analysis of deleted tweets. Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13. doi:https://doi.org/10.1145/2441776.2441878.

Aral, S. and Eckles, D. (2019). Protecting elections from social media manipulation. Science, 365(6456), pp.858–861. doi:https://doi.org/10.1126/science.aaw8243.

Association, P. (2023). Twitter outage results in many users being unable to tweet or follow accounts. [online] TheJournal.ie. Available at: https://www.thejournal.ie/twitter-outage-results-in-many-users-being-unable-to-tweet-or-follow-accounts-5990790-Feb2023/ [Accessed 1 Apr. 2023].

Barrabi, T. (2022). Twitter employee seen sleeping on office floor as Elon Musk pushes tight deadlines. [online] New York Post. Available at: https://nypost.com/2022/11/02/twitter-employee-sleeps-on-office-floor-amid-elon-musk-deadlines/.

Benton, B., Choi, Jin-A., Luo, Y. and Green, K. (2022). Hate Speech Spikes on Twitter After Elon Musk Acquires the Platform. School of Communication and Media, Montclair State University. [online] Available at: https://digitalcommons.montclair.edu/scom-facpubs/33/.

Biden, J. (2023). State of the Union 2023. [online] The White House. Available at: https://www.whitehouse.gov/state-of-the-union-2023/.

Bossaller, J., Bernier, A., McQueen, S. and Peterson, M. (2019). The hoax and the president: Historical perspectives on politics, truth, and academia. Critical

librarianship. [online] Available at: https://www.ideals.illinois.edu/items/112473 [Accessed 2 Feb. 2023].

Bradshaw, S. and Howard, P. (2018). Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation. [online] Available at: https://holbrook.no/share/papers/computational_social_media_fake.pdf.

Buchanan, T. and Benson, V. (2019). Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of 'Fake News'?. Social Media + Society, 5(4), p.205630511988865. doi:https://doi.org/10.1177/2056305119888654.

Calderon, P. (2018). VADER Sentiment Analysis Explained. [online] Medium. Available at: https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9.

Condren, C., Davis, J.M., McCausland, S. and Phiddian, R. (2008). Defining parody and satire: Australian copyright law and its new exception: Part 2 - Advancing ordinary definitions. Media and Arts Law Review, [online] 13(4), pp.401–421. Available at: https://researchnow.flinders.edu.au/en/publications/defining-parody-and-satire-australian-copyright-law-and-its-new-e [Accessed 2 Feb. 2023].

Deibert, R.J. (2003). Black Code: Censorship, Surveillance, and the Militarisation of Cyberspace. Millennium: Journal of International Studies, 32(3), pp.501–530. doi:https://doi.org/10.1177/03058298030320030801.

Del Fresno García, M., Daly, A.J. and Segado Sánchez-Cabezudo, S. (2016). Identificando a los nuevos influyentes en tiempos de Internet: medios sociales y análisis de redes sociales / Identifying the new Influencers in the Internet Era: Social Media and Social Network Analysis. Revista Española de Investigaciones Sociológicas. doi:https://doi.org/10.5477/cis/reis.153.23.

Demilt, J. (2017). The Origins of Twitter. [online] Pennington Creative. Available at: https://penningtoncreative.com/the-origins-of-twitter/.

Desai, R. (2021). How to scrape millions of tweets using snscrape. [online] DataSeries. Available at: https://medium.com/dataseries/how-to-scrape-millions-of-tweets-using-snscrape-195ee3594721.

Dewi, L.C., Meiliana and Chandra, A. (2019). Social Media Web Scraping using Social Media Developers API and Regex. Procedia Computer Science, 157, pp.444–449. doi:https://doi.org/10.1016/j.procs.2019.08.237.

Dizikes, P. (2018). Study: On Twitter, false news travels faster than true stories. [online] MIT News. Available at: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308.

Dušan V., P. (2021). Freedom of Expression on Social Networks: An International Perspective. The Impact of Digital Platforms and Social Media on the Freedom of Expression and Pluralism : Analysis on Certain Central European Countries, pp.277–310. doi:https://doi.org/10.54237/profnet.2021.mwsm_9.

Fisher, M., John Woodrow Cox and Hermann, P. (2016). Pizzagate: From rumor, to hashtag, to gunfire in D.C. The Washington Post. [online] 7 Dec. Available at: https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html.

Fitzpatrick, N. (2018). Media Manipulation 2.0: The Impact of Social Media on News, Competition, and Accuracy. Athens Journal of Mass Media and Communications, 4(1), pp.45–62. doi:https://doi.org/10.30958/ajmmc.4.1.3.

Forman-Katz, N. and Stocking, G. (2022). Key facts about Truth Social. [online] Pew Research Center. Available at: https://www.pewresearch.org/fact-tank/2022/11/18/key-facts-about-truth-social-as-donald-trump-runs-for-u-s-president-again/.

Funke, D. and Flamini, D. (2014). A guide to anti-misinformation actions around the world - Poynter. [online] Poynter. Available at: https://www.poynter.org/ifcn/anti-misinformation-actions/.

Germani, F. and Biller-Andorno, N. (2021). The anti-vaccination infodemic on social media: A behavioral analysis. PLOS ONE, [online] 16(3). doi:https://doi.org/10.1371/journal.pone.0247642.

Ghanem, B., Rosso, P. and Rangel, F. (2020). An Emotional Analysis of False Information in Social Media and News Articles. ACM Transactions on Internet Technology, 20(2), pp.1–18. doi:https://doi.org/10.1145/3381750.

Good, K.D. (2012). From scrapbook to Facebook: A history of personal media assemblage and archives. New Media & Society, 15(4), pp.557–573. doi:https://doi.org/10.1177/1461444812458432.

Guess, A. and Lyons, B. (2020). Misinformation, disinformation, and online propaganda, Chapter Two. Cambridge University Press. [online] doi:https://doi.org/10.1017/9781108890960.

Haman, M. (2020). The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. Heliyon, 6(11), p.e05540. doi:https://doi.org/10.1016/j.heliyon.2020.e05540.

Herrero-Diz, P., Conde-Jiménez, J. and Reyes de Cózar, S. (2020). Teens' Motivations to Spread Fake News on WhatsApp. Social Media + Society, 6(3), p.205630512094287. doi:https://doi.org/10.1177/2056305120942879.

Hocevar, K.P., Flanagin, A.J. and Metzger, M.J. (2014). Social media self-efficacy and information evaluation online. Computers in Human Behavior, 39, pp.254–262. doi:https://doi.org/10.1016/j.chb.2014.07.020.

Höller, M. (2021). The human component in social media and fake news: the performance of UK opinion leaders on Twitter during the Brexit campaign. European Journal of English Studies, [online] 25(1), pp.80–95. doi:https://doi.org/10.1080/13825577.2021.1918842.

Hudson, S., Huang, L., Roth, M.S. and Madden, T.J. (2016). The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors. International Journal of Research in Marketing, 33(1), pp.27–41. doi:https://doi.org/10.1016/j.ijresmar.2015.06.004.

Hwang, J.P. (2020). What Python package is best for getting data from Twitter? Comparing Tweepy and Twint. [online] Medium. Available at: https://towardsdatascience.com/what-python-package-is-best-for-getting-data-from-twitter-comparing-tweepy-and-twint-f481005eccc9.

Jackson, B. (2018). What Is GitHub? A Beginner's Introduction to Github. [online] Kinsta Managed WordPress Hosting. Available at: https://kinsta.com/knowledgebase/what-is-github/.

Johnson, S. (2009). Back to Article Click to Print How Twitter Will Change the Way We Live. [online] Available at: http://individual.utoronto.ca/kreemy/proposal/04.pdf.

Kolowich, L. (2019). The Best Times to Post on Social Media in 2023 [New Data]. [online] Hubspot.com. Available at: https://blog.hubspot.com/marketing/best-times-post-pin-tweet-social-media-infographic.

Kralj Novak, P., Smailović, J., Sluban, B. and Mozetič, I. (2015). Sentiment of Emojis. PLOS ONE, [online] 10(12), p.e0144296. doi:https://doi.org/10.1371/journal.pone.0144296.

Kunneman, F., Liebrecht, C., van Mulken, M. and van den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. Information Processing & Management, 51(4), pp.500–509. doi:https://doi.org/10.1016/j.ipm.2014.07.006.

Lee, B.Y. (2022). Fake Eli Lilly Twitter Account Claims Insulin Is Free, Stock Falls 4.37%. [online] Forbes. Available at: https://www.forbes.com/sites/brucelee/2022/11/12/fake-eli-lilly-twitter-account-claims-insulin-is-free-stock-falls-43/.

Lohiya, H. (2018). Sentiment Analysis with AFINN Lexicon. [online] Medium. Available at: https://himanshulohiya.medium.com/sentiment-analysis-with-afinn-lexicon-930533dfe75b.

Maxwell, T. (2022). How Many Bots Are on Twitter and Does It Matter? [online] MUO. Available at: https://www.makeuseof.com/how-many-bots-on-twitter/.

McDermott, R. and Hatemi, P.K. (2020). Ethics in field experimentation: A call to establish new standards to protect the public from unwanted manipulation and real harms. Proceedings of the National Academy of Sciences, [online] 117(48), pp.30014–30021. doi:https://doi.org/10.1073/pnas.2012021117.

Microsoft (2016). Visual Studio Code. [online] Visualstudio.com. Available at: https://code.visualstudio.com/docs/editor/whyvscode.

Milmo, D. (2022a). Musk proposes charging $8 for verified Twitter account despite user backlash. [online] the Guardian. Available at: https://www.theguardian.com/technology/2022/nov/01/musk-charging-twitter-verified-accounts [Accessed 2 Feb. 2023].

Milmo, D. (2022b). Twitter blue check unavailable after impostor accounts erupt on platform. [online] the Guardian. Available at: https://www.theguardian.com/technology/2022/nov/11/twitter-blue-check-verification-impostor-accounts.

Milmo, D. (2023). Donald Trump's Truth Social posts bode ill for his return to Facebook. The Guardian. [online] 26 Jan. Available at: https://www.theguardian.com/us-news/2023/jan/26/donald-trump-truth-social-posts-bode-ill-return-facebook [Accessed 30 Mar. 2023].

Morris, R. (1994). Computerized Content Analysis in Management Research: A Demonstration of Advantages & Limitations. Journal of Management, 20(4), pp.903–931. doi:https://doi.org/10.1177/014920639402000410.

Mustafaraj, E. and Metaxas, P.T. (2017). The Fake News Spreading Plague. Proceedings of the 2017 ACM on Web Science Conference - WebSci '17. [online] doi:https://doi.org/10.1145/3091478.3091523.

Newton, C. (2023). How a single engineer brought down Twitter. [online] The Verge. Available at: https://www.theverge.com/2023/3/6/23627875/twitter-outage-how-it-happened-engineer-api-shut-down [Accessed 20 Mar. 2023].

Oberlo (2023). Find out How Many People Use Social Media in 2019 | Oberlo. [online] Oberlo. Available at: https://www.oberlo.com/statistics/how-many-people-use-social-media.

Python Software Foundation (2019). What is Python? Executive Summary. [online] Python.org. Available at: https://www.python.org/doc/essays/blurb/.

Reid, J.M. (2021). Building a Sentiment Analysis report using NLTK and Altair. [online] Medium. Available at: https://towardsdatascience.com/building-a-sentiment-analysis-interactive-report-using-nltk-and-altair-83cb9fcb36fe [Accessed 26 Mar. 2023].

Rigden, J. (2018). Tweepy: a Python Library for the Twitter API. [online] Medium. Available at: https://medium.com/@jasonrigden/tweept-a-python-library-for-the-twitter-api-9d0537dcebd4 [Accessed 26 Mar. 2023].

Ringel, S. and Davidson, R. (2020). Proactive ephemerality: How journalists use automated and manual tweet deletion to minimize risk and its consequences for social media as a public archive. New Media & Society, p.146144482097238. doi:https://doi.org/10.1177/1461444820972389.

Robertson, E., Guan, H., Kozak, M., Lee, Y., Yates, A.N., Delgado, A., Zhou, D., Kheyrkhah, T., Smith, J. and Fiscus, J. (2019). Manipulation Data Collection and Annotation Tool for Media Forensics. [online] openaccess.thecvf.com. Available at: https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Robertson_Manipulation_Data_Collection_and_Annotation_Tool_for_Media_Forensics_CVPRW_2019_paper.html [Accessed 2 Feb. 2023].

Rojas, H., Shah, D.V. and Faber, R.J. (1996). FOR THE GOOD OF OTHERS: CENSORSHIP AND THE THIRD-PERSON EFFECT. International Journal of Public Opinion Research, [online] 8(2), pp.163–186. doi:https://doi.org/10.1093/ijpor/8.2.163.

Ruan, L., Knockel, J., Ng, J. and Crete-Nishihata, M. (2016). ONE APP, TWO SYSTEMS How WeChat uses one censorship policy in China and another internationally. [online] Available at: https://tspace.library.utoronto.ca/bitstream/1807/96729/1/Report%2384--oneapp-twosystems.pdf.

Ruby, D. (2022). Twitter Statistics: Facts and Figures After Elon Musk Takeover (2022). [online] demandsage. Available at: https://www.demandsage.com/twitter-statistics/.

Setia, M.S. (2016). Methodology Series Module 3: Cross-sectional Studies. Indian Journal of Dermatology, [online] 61(3), pp.261–264. doi:https://doi.org/10.4103/0019-5154.182410.

Shah, P. (2020). Sentiment Analysis using TextBlob. [online] Medium. Available at: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524.

Shahi, G.K., Dirkson, A. and Majchrzak, T.A. (2021). An exploratory study of COVID-19 misinformation on Twitter. Online Social Networks and Media, 22, p.100104. doi:https://doi.org/10.1016/j.osnem.2020.100104.

Shepherd, J. (2022). 22 Essential Twitter Statistics You Need to Know in 2022. [online] The Social Shepherd. Available at: https://thesocialshepherd.com/blog/twitter-statistics.

Silman, J. (2019). Scrape Tweets using Selenium. [online] Medium. Available at: https://medium.com/@jamessilman/scrape-tweets-using-selenium-3f713873439 [Accessed 26 Mar. 2023].

Southern, M.G. (2022). Elon Musk's Twitter Takeover: A Timeline Of Events. [online] Search Engine Journal. Available at: https://www.searchenginejournal.com/elon-musks-twitter-takeover-a-timeline-of-events/470927/.

Tagliabue, M., Squatrito, V. and Presti, G. (2019). Models of Cognition and Their Applications in Behavioral Economics: A Conceptual Framework for Nudging Derived From Behavior Analysis and Relational Frame Theory. Frontiers in Psychology, [online] 10. doi:https://doi.org/10.3389/fpsyg.2019.02418.

Twitter (2019). The Twitter Rules. [online] Twitter.com. Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules.

Twitter (2022). About verified accounts. [online] help.twitter.com. Available at: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

Vese, D. (2021). Governing Fake News: The Regulation of Social Media and the Right to Freedom of Expression in the Era of Emergency. European Journal of Risk Regulation, [online] 13(3), pp.1–41. doi:https://doi.org/10.1017/err.2021.48.

Vosoughi, S., Roy, D. and Aral, S. (2018). The spread of true and false news online. Science, [online] 359(6380), pp.1146–1151. doi:https://doi.org/10.1126/science.aap9559.

Vraga, E.K. and Bode, L. (2020). Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation. Political Communication, 37(1), pp.136–144. doi:https://doi.org/10.1080/10584609.2020.1716500.

Weng, Z. and Lin, A. (2022). Public Opinion Manipulation on Social Media: Social Network Analysis of Twitter Bots during the COVID-19 Pandemic. International

Journal of Environmental Research and Public Health, 19(24), p.16376. doi:https://doi.org/10.3390/ijerph192416376.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, [online] 9(11). Available at: https://timreview.ca/article/1282.

Williams, A. (2023). Billionaire Koch's donor network says it opposes Trump's re-election. Financial Times. [online] 5 Feb. Available at: https://www.ft.com/content/72d8a8b3-6ac4-45a9-ab5a-53dc1c6c4998 [Accessed 1 Apr. 2023].

Wu, L., Morstatter, F., Carley, K.M. and Liu, H. (2019). Misinformation in Social Media. ACM SIGKDD Explorations Newsletter, 21(2), pp.80–90. doi:https://doi.org/10.1145/3373464.3373475.

Yang, R. (2023). Why Tesla Climbed 18.8% in February. [online] The Motley Fool. Available at: https://www.fool.com/investing/2023/03/05/why-tesla-climbed-188-in-february/ [Accessed 1 Apr. 2023].

Zhou, X. and Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys, 53(5). doi:https://doi.org/10.1145/3395046.