Twitter Link Analysis

Draft Report

Igusti Agung Vadayogi Raharja
(236228)
BSc Computer Science
Supervisor: David Weir

Statement of originality

This report is submitted as part requirement for the degree of BSc Computer Science at the University of Sussex. It is the product of my labour except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged. I hereby give/withhold permission for a copy of this report to be loaned out to students in future years

Acknowledgements

Thank you to Prof. David Weir's guidance and insights during the first phases of this project, in its conduct and documentation.

Summary

The Twitter Link Analysis project is an effort to examine links shared on Twitter and comprehend their distribution, impact, and substance. Investigating which news sources are linked to most frequently and how this differs based on the article's topic is a crucial part of the research. The project uses data scraping, natural language processing and data visualisation to extract information from the links exchanged on the site. It gathers and analyses a large amount of data from Twitter. The primary themes and subjects that are being discussed on Twitter are identified, along with their relationships with one another.

The project's identification of information sources on Twitter, particularly journalistic organisations, is one of its objectives. The project utilises network analysis to pinpoint the news sources and their URL, and they examine the links' content to assess their validity and dependability by accessing user Tweets and monitoring how often or not the website URL is being used or tweeted by a user for their source of information. The researchers discovered that depending on the article's theme, news sources are linked more frequently than others. For instance, political news stories may be related to large news organisations like CNN more frequently than entertainment news stories, which may be linked to organisations like TMZ or BuzzFeed more frequently.

The Twitter Link Analysis project helps scholars analyse the impact of link sharing on Twitter by enabling them to look at the reach and virality potential of news pieces. Researchers can learn which news sources are shared most frequently by compiling and analysing a sizable collection of links and related subjects. The project can also assess the reliability of the sources and the veracity of the information supplied by looking at the substance of the articles being shared. Also, the study can reveal which journalists and news organisations have the most influence on Twitter, shedding light on the influence that the media has on society. The research can better understand how news works by examining the interactions between users, news organisations, and traditional sources.

Overall, the Twitter Link Analysis project offers insightful information about Twitter's ecosystem, particularly concerning news organisations and how links are distributed. The initiative enables academics to acquire a thorough knowledge of the content, significance, and distribution of links on Twitter by combining data scraping, natural language processing, and data visualisation tools.

Table of Contents

1. Introduction

Twitter is a social media platform where users can share their thoughts using short messages called “tweets” and attract substantial attention. Since many users of social media experience information overload, finding exciting and valuable content in vast text streams is a critical challenge[9]. A “tweet” which is being exchanged between users may vary from news stories, product reviews, and useful or unuseful information[15].

There are five different Tweet Types that a user can do. There’s “Retweet”, “Quote” which is a retweet with a comment above it, A “Reply” a tweet that starts with an ‘@’ and a username, “Self-replies”, and finally a “Tweet” which is a message sent to the user's followers or the public in the case of the account is not private. While there are restrictions on how much text a user can input, messages can still be sent casually even if they are unrelated to the recipient. If a user finds a user account interesting or useful, they can follow it. A user account, on the other hand, can have some followers based on its popularity or usefulness[3][4]. As of the end of 2022, there were approximately 396.5 million accounts on Twitter (Statistic Brain, 2022)[13]. Twitter as a social media itself has a very unwavering statistics record, Twitter's real visitors are three times as many as its registered members, with 45% of users checking the site daily and users spending an average of 5.1 hours each month there (Statistics Brain, 2022)[13]. While also having the second most educated user based on social media, which LinkedIn users top.

Users also take advantage of Twitter to share links to news pages, websites or an article they find useful[7][10]. A study says, Twitter users, non-Twitter users on social media and social media users overall — consume a good deal of news. In all, 77% of all social media users said they keep up with the news at least once a day, a similar number (76%) for non-Twitter users (R.Tom, S.Jeff, 2015)[10]. This topic is what is going to power this project and what will be the base of it.

Thus, even though Twitter's interface and use case has not changed significantly since its launch in July 2006, it continues to remain important for this reason.

1.1 The motivation for the project

The goal of this research is to examine the connections between users, topics, and trends to learn more about how information spreads within a network[11]. By examining the networks of those discussing a topic and its environment, we hope to better understand how topics grow to be popular or contentious. Identifying significant news sites or articles can also help you better target your marketing efforts or spread awareness of important ideas[7][10]. Last but not least, our research will identify new themes and connections between individuals and topics to give a clearer picture of the structure of the Twitter network as a whole.

1.2 Aims of the project

To learn more about the connections among users, topics, and trends. In particular, this can entail looking into how issues grow to be well-liked or divisive, finding influential users or topics, spotting upcoming topics, and comprehending the general structure of the Twitter network[4][11]. These revelations can then be applied to marketing initiatives, personalised communications, and a deeper comprehension of user understanding.

This study can offer insightful information on the discussions and viewpoints being expressed on the platform as well as the general public's attitude towards a range of subjects. Businesses and organisations trying to interact with their target audience and comprehend how they are viewed by the general public may find this to be especially helpful. Sentiment analysis, which involves classifying tweets as good, negative, or neutral based on the words and phrases they include, is a common application of Twitter data analysis[13]. Businesses can better understand how the public views their brand or sector by analysing the sentiment of tweets relating to it and seeing areas for improvement[16]. For instance, a business might use the information to identify issues and improve its service if it discovers that a significant portion of tweets about its product is critical. A prominent use of Twitter data analysis is social network analysis, which involves examining user connections and interactions on the platform. By looking at the connections between users, businesses can identify significant opinion leaders or influencers within their target market[22]. They can then utilise this information to build partnerships with these individuals or work together on marketing projects. Also, by analysing the overall layout of the Twitter network to identify patterns and trends in the way information is exchanged, businesses may improve their messaging and outreach initiatives.

Studying Twitter data can reveal important information about the debates, viewpoints, and trends present on the site[7][10]. Businesses and organisations can better engage with their target audience, enhance their products or services, and make data-driven decisions about their marketing and outreach initiatives by developing a deeper understanding of user behaviour, sentiment, and network structure.

1.3 Literature review

1.3.1 Introduction

"Twitter link analysis" focuses on the investigation of how news websites links are shared on the well-known social media site Twitter. Because of its millions of users, who produce massive amounts of data every day, Twitter has emerged as a key tool for networking, communication, and information dissemination[5]. Examining shared links can help researchers, marketers, and decision-makers understand information sharing, user behaviour, and the dynamics of online communities on a platform. Twitter link analysis is valuable because it can spot patterns and trends in how people are consuming and distributing content. By looking at the types of links shared, how often they are shared, and the context in which they are shared, researchers and readers can gain insight into users' interests, preferences, and the factors influencing their online behaviour. Twitter link analysis could be utilised to identify and stop the spread of fake news and other misleading material. Researchers can create algorithms to recognise and filter out hazardous or misleading content by looking at the patterns and characteristics of link sharing, which will increase the veracity and dependability of information conveyed over the network[19][20].

Twitter link analysis is a essential area of study that promotes a better understanding of information sharing and user behaviour on social media. It is an essential tool for navigating the complicated world of internet communication because of its many uses, which span marketing and political campaigns to crisis management and the battle against misinformation[22].

1.3.2 Twitter as A Platform For News Dissemination

Since its beginnings, Twitter has developed into a significant medium for disseminating news and information. Kwak, Lee, Park, and Moon (2010)[45]questioned whether Twitter is primarily a social network or a news media in their study, which sought to explore the nature of Twitter as a platform. They investigated user behaviour, hot issues, and the network structure while conducting a thorough examination of the entire Twittersphere.

Researchers found that Twitter demonstrates traits common to both social networks and news media platforms. Twitter allows users to interact, connect, and exchange content with others as a social network. However, Twitter relationships tend to be more one-sided, matching the broadcast paradigm of news media, in contrast to traditional social networks where relationships are frequently reciprocal. To consume news and information from sources without engaging in direct dialogue, users can follow accounts without needing to be followed back.

According to Kwak et al. (2010)[45], many of the most trending topics on Twitter had something to do with recent news stories, demonstrating the network's function as a medium for the distribution of news. Also, they saw that information and news stories on Twitter spread quickly, with retweeting serving as the main mechanism. Users can access news from various sources and stay up to date on current events thanks to the rapid dispersion of information.

The ability of Twitter to link people to powerful accounts and organisations is another facet of its role as a key platform for distributing news content. Users can get updates and news directly from these sources thanks to the significant presence of many news organisations, journalists, and other information sources on Twitter.

As a result, the study by Kwak et al. (2010)[45] emphasises Twitter's dual nature as a social network and a platform for news media. Twitter is positioned as a potent instrument for staying informed and engaged in current events thanks to the rapid circulation of news articles and information made possible by its distinctive structure.

Analysis of Shared Links on Twitter

Romero, Galuba, Asur, and Huberman (2011)[46] examined the idea of influence in the context of social media, notably Twitter, in their research "Influence and Passivity in Social Media." They sought to comprehend the mechanics of information diffusion and the influence of powerful users on other people's behaviour. The researchers offered fresh approaches and metrics for examining links that were shared on Twitter, concentrating on two key elements: influence and passivity.

1. *Influence:* According to the study, influence refers to a user's capacity to persuade other users to conduct specific activities, such as retweeting or mentioning. To gauge a user's influence, the authors created the Influence-Passivity Algorithm (IPA). The IPA takes into account the likelihood that a user's followers may retweet or reference links in their tweets. The metric distinguishes between influential individuals and people who are merely popular by taking into consideration the frequency of interactions and the proportional influence of engaging users.
2. *Passivity*: Passivity is a gauge of a user's susceptibility to outside influence. An active user frequently engages with platform material, while a passive user rarely connects with others. The authors suggested a metric to gauge passivity that takes into account a user's propensity to retweet or mention other users. This metric factors for the user's exposure to tweets containing links and the impact of the persons they follow.

These measurements were used by the authors to examine the links that were tweeted and identify user influence and passivity tendencies. They discovered that the impact on Twitter depends not just on followers but also on user interactions and relationships. On the platform, it was discovered that very prominent users had a big influence on how information spread and how trends developed.

The study by Romero et al. (2011)[46], which focused on influence and passivity, provided innovative metrics and methodologies to assess shared links on Twitter. These metrics offer insightful information on the dynamics of information diffusion and the part played by influential users in influencing other platform users' behaviour.

Network Analysis of Twitter Users

Cha, Haddadi, Benevenuto, and Gummadi (2010)[47] wrote a study titled "Measuring User Impact in Twitter: The Million Follower Fallacy" that emphasised the value of comprehending user interactions and network structure in Twitter link analysis. They claimed that conventional measurements, such as the number of followers, may not always reflect a user's influence on the site. Instead, they suggested substitute criteria to more accurately assess and gauge user influence on Twitter. The authors stressed the value of evaluating user interactions, such as retweets and mentions, because they give a more realistic picture of a person's capacity to communicate with others and disseminate information. They discovered three main factors that drive user behaviour: indegree (following), retweets, and mentions. The scientists found that there was only a slight link between these factors, indicating that having many followers does not necessarily translate into having many retweets and mentions.

The Twitter network structure is also very important for link analysis. By evaluating the relationships and interactions between users, researchers can identify prominent users who drive information diffusion and contribute to the establishment of trends on the platform. Knowing the network structure also makes it easier to spot groups of people who share common interests, which can be helpful for campaigns that provide tailored advertising or information. The propagation of false information and fake news on Twitter can also be halted by examining network structure and user interactions. Researchers and platform managers can create plans to reduce the impact of misleading information by identifying prominent individuals who spread it.

In conclusion, the study by Cha et al. (2010)[47] highlights how crucial it is to comprehend user interactions and network structure when analysing Twitter links. Researchers may better understand user impact, information transmission, and the dynamics of online communities by looking at these elements. This knowledge is useful for a variety of applications, including targeted marketing, political campaigns, and the fight against disinformation.

Content Analysis of Shared Links and Tweets

The authors of the study "Comparing Twitter and Traditional Media Using Topic Models" by Zhao et al. (2011)[48] sought to identify the variations in content and topic coverage between

Twitter and traditional media. They analysed the content of tweets, shared links, and traditional media articles using topic modelling and natural language processing (NLP) approaches.

1. The text data from both Twitter and conventional media sources were preprocessed using natural language processing (NLP) methods. To prepare text data for subsequent analysis, this preprocessing included the basic stages of tokenization, stopword elimination, and stemming. NLP approaches to assist in transforming unstructured text data into a format that enables researchers to find and examine patterns in the content.
2. *Topic Modeling:* To find latent subjects in the context of shared links, tweets, and traditional media articles, the authors employed Latent Dirichlet Allocation (LDA), a well-known topic modelling technique. LDA is an unsupervised machine learning technique that examines the word co-occurrence patterns in a set of documents to identify underlying subjects. The researchers were able to identify the major themes in the content and compare their frequency across Twitter and traditional media by applying LDA to their datasets.

The study discovered that Twitter has wider coverage than traditional media, with a focus on technology and social issues in particular. The authors also noted that breaking news tended to spread more quickly on Twitter, underlining the service's function as a source of up-to-the-minute news.

The study by Zhao et al. (2011)[48] illustrates the application of topic modelling and natural language processing methods, such as LDA, to the analysis of shared links and tweets on Twitter. With the help of these techniques, researchers may better comprehend the differences between Twitter and traditional media in terms of content and topic coverage by spotting patterns, trends, and themes within the text data.

Sentiment Analysis of Tweets Containing Link

In the paper "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" by Pak and Paroubek (2010)[49], the authors underlined the relevance of sentiment analysis in understanding user opinions and emotions associated with shared information on Twitter. Sentiment analysis is a branch of natural language processing that deals with the automatic extraction of subjective data from text data, including feelings, views, and attitudes.

The authors showed how to classify tweets into several sentiment categories, such as good, negative, or neutral, using sentiment analysis approaches. This classification enables researchers to examine user emotional responses to the shared content as well as evaluate public opinion on particular subjects, occasions, or problems.

A technique for creating a Twitter corpus for sentiment analysis and opinion mining was put forth by Pak[49] and Paroubek in their article. They created a labelled dataset for training sentiment classification models by using emoticons seen in tweets as labels for various sentiment classes. To categorise the sentiment of tweets in their dataset, the scientists used machine learning methods such as Naive Bayes, Maximum Entropy, and Support Vector Machines.

For several reasons, sentiment analysis is essential in determining user attitudes and feelings around shared material.

1. *Monitoring public opinion:* Researchers can monitor changes in public opinion over time and acquire insights into the reasons behind these changes by evaluating the sentiment of tweets about particular topics or events.
2. *Finding patterns and trends in user behaviour:* Sentiment analysis can help identify patterns and trends in user behaviour, such as emotional responses to particular news stories or the sentiment expressed in tweets about various forms of shared content.
3. *Improving marketing techniques:* Companies and organisations can use sentiment analysis to learn how their target audience views their goods, services, or advertising campaigns. This information enables them to adjust their marketing strategy.
4. *Crisis management:* Organizations can monitor public opinions about crises or divisive topics with the aid of sentiment analysis, allowing them to act swiftly and manage their reputation.

In conclusion, Pak and Paroubek's (2010)[49] work stresses the importance of sentiment analysis in understanding user perceptions and feelings towards information shared on Twitter. Researchers may learn a lot about public opinion, emotional responses, and user behaviour on the site by categorising tweets into several sentiment categories.

Detection of Fake News and Misinformation

In the paper "The Transmission of True and False News Online" by Vosoughi, Roy, and Aral (2018)[50], the authors stressed the relevance of link analysis in spotting fake news and disinformation on Twitter. They attempted to comprehend the dynamics of information dissemination and the elements influencing the quick spread of misleading information by looking into how real and fake news items were distributed on the platform.

For various reasons, link analysis is essential in spotting false information and fake news on Twitter.

1. Link analysis was employed by the authors to follow the mention and retweet networks connected to real and fraudulent news stories to quantify the dissemination of information. As a result, they were able to calculate how much more quickly false information propagated than accurate information. They discovered that false news stories proliferated noticeably more quickly, widely, and quickly than accurate news items, stressing the necessity of efficient link analysis techniques to monitor and slow down the dissemination of misleading information.
2. Link analysis can be used to identify users who have a particularly strong influence on the spread of incorrect information. Researchers can identify the main people responsible for the dissemination of fake news by looking at the mention and retweet networks, which enables them to create targeted interventions or mitigation plans.
3. “False by nature, true by nature” is more often the case with news stories. Insights into the psychological and sociological elements influencing the propagation of false information can be gained through link analysis, which aids researchers in understanding user behaviour patterns connected to the dissemination of false information.

The study by Vosoughi et al. (2018)[50] concludes by highlighting the value of link analysis in spotting false information on Twitter. Researchers can learn a lot about the dynamics of information dissemination on the platform by analysing the spreading of real and misleading news articles as well as the factors influencing them. This information can assist guide initiatives to thwart the dissemination of false information and news, thereby promoting the development of a more dependable and trustworthy online information ecosystem.

Applications of Twitter Link Analysis

The authors of the paper "Predicting Elections with Twitter: What 140 Characters Tell About Political Emotion" by Tumasjan et al. (2010)[51] showed how Twitter link analysis might be used in a variety of fields, including marketing, political campaigns, and crisis management. They aimed to test the accuracy of Twitter data in predicting election results by looking at the mood conveyed in tweets about political candidates and parties.

1. Marketing: To better understand customer preferences, brand sentiment, and hot issues, marketing professionals might benefit from conducting a Twitter link analysis. Marketers can learn whether goods, services, or campaigns resonate with their target audience by examining the links posted on the site and the sentiment attached to them. They can also establish methods to effectively engage with influencers and important opinion leaders within their business.
2. Political campaigns: According to a study by Tumasjan et al. (2010)[51], it is possible to forecast election results surprisingly well by using the attitude conveyed in tweets.

Political campaigns can track the effectiveness of their messaging, analyse public sentiment on particular subjects, and pinpoint regions where they should concentrate their efforts by using Twitter link analysis. Campaigns can improve their strategies and make data-driven judgements by looking at the sentiment and information circulating about various political actors.

3. Crisis management: In times of crisis or emergencies, Twitter link analysis can be essential for tracking the dissemination of information, keeping tabs on public opinion, and spotting potential misinformation. Organizations can discover areas of concern, understand how the public views their response, and build plans to effectively address these concerns by examining the links posted during a crisis. Also, real-time Twitter data analysis can assist firms in spotting and addressing new problems before they become more serious.

The study by Tumasjan et al. (2010)[51] sheds insight on the possible uses of Twitter link analysis in a variety of fields, including marketing, political campaigns, and crisis management. Researchers and practitioners can learn a great deal about user behaviour, public opinion, and the variables influencing the diffusion of information on the platform by analysing the sentiment conveyed in tweets and the dynamics of information distribution. In diverse circumstances, these insights can be used to improve tactics and make data-driven choices.

Issues and Potential Directions

For researchers, marketers, and policymakers to comprehend and monitor the circulation of information, opinions, and user behaviour on the platform, Twitter link analysis has emerged as a crucial tool[45]. Despite its expanding relevance, there are several issues and potential directions for this profession to go. Managing Twitter's dynamic nature; Real-time data analysis on Twitter is difficult due to the fast-paced atmosphere and frequent influx of new information. The continuous issue is to provide scalable and effective link analysis techniques to handle the platform's dynamic nature[48]. Detecting and reducing biases; The generalizability of link analysis results may be impacted by biases in Twitter data, such as demographic representation or selective sharing of information. To increase the dependability of insights, future research has to concentrate on identifying and resolving these biases.

Handling privacy issues; Users may not be aware that their data is being examined for research purposes when Twitter data is evaluated, which causes privacy issues. A significant difficulty is ensuring ethical data use and protecting user privacy when undertaking link analysis. Integration of numerous data sources; Twitter data can be combined with information from other social media platforms, news items, blogs, and other blogs to provide a more thorough view of user behaviour and information transmission. Future studies could look into combining different data sources for a more thorough connection analysis[46]. Using cutting-edge machine learning methods: The performance of link analysis may be improved by applying cutting-edge machine learning methods such as deep learning and reinforcement learning. To increase the precision and effectiveness of link analysis techniques, researchers should keep looking into new ideas.

Conclusion

The vital role of Twitter Link Analysis in comprehending social media dynamics, information transmission, and user behaviour is the subject of this literature review. which covered topics including Twitter's importance as a key medium for disseminating news stories and information, which affects how users behave on social media and how information spreads, To analyse shared links on Twitter, a range of techniques and metrics are used, including influence, passivity, and network structure, which offer insights on user interactions and the dissemination of material.

Using topic modelling and natural language processing to examine the content of tweeted and shared links.The value of sentiment analysis in recording user reactions to published information and in delivering insightful data on public opinion and preferences. Link analysis's function in identifying and countering false information and misinformation on Twitter, which improves the veracity and reliability of information exchanged on the network, is the many ways that Twitter link analysis is used in industries including marketing, politics, and crisis management.

In conclusion, Twitter link analysis is a crucial tool for comprehending how information is shared, how users interact with one another, and how the general public feels about social media sites. Researchers can continue to increase the precision and effectiveness of link analysis techniques by addressing the difficulties and examining the field's potential future directions. This will ultimately help us comprehend social media dynamics and how they affect diverse fields.

Professional Considerations

Twitter user data will be used to power this project. Therefore I must carefully assess any potential professional and ethical dilemmas because they may involve or involve mentioning the users' personal information.

BCS Code of Conduct

User data will be analysed for this project. As a result, I must give careful thought to any potential professional or ethical difficulties because they may have an impact on the project's participants and members.

This Code of Conduct:

- Sets out the professional standards required by BCS as a condition of membership.
- Applies to all members, irrespective of their membership grade, the role they fulfil, or the jurisdiction where they are employed or discharge their contractual obligations.
- Governs the conduct of the individual, not the nature of the business or ethics of any Relevant Authority*.

1. Public Interest

You shall:

a. Have due regard for public health, privacy, security and well-being of others and the environment.
b. Have due regard for the legitimate rights of Third Parties.
c. Conduct your professional activities without discrimination on the grounds of sex, sexual orientation, marital status, nationality, colour, race, ethnic origin, religion, age or disability, or of any other condition or requirement.
d. Promote equal access to the benefits of IT and seek to promote the inclusion of all sectors in society wherever opportunities arise."

In accordance with point 1(a), the project shall have due regard for the user's privacy, security and well-being whereas this project is a university project and will have no hate speech directed to the users the data is taken from. Any content produced throughout the project will be deleted after its completion.

In accordance with point 1(b), the project shall have due regard to the legitimate rights of the third parties.

In accordance with point 1(c), the project shall not discriminate against any gender or age which will be analyzed in the project. During the final report writing the report will be carefully written so that there will not be any misspoken words.

In accordance with point 1(d), the project shall promote equal access to the benefits of IT since the project will be fully accessible in the university database.

2. Professional Competence and Integrity

You shall:

a. Only undertake to do work or provide a service that is within your professional competence.
b. NOT claim any level of competence that you do not possess.
c. Develop your professional knowledge, skills and competence continuingly, maintaining awareness of technological developments, procedures, and standards that are relevant to your field.
d. Ensure that you have the knowledge and understanding of Legislation* and that you comply with such Legislation, in carrying out your professional responsibilities.
e. Respect and value alternative viewpoints and, seek, accept and offer honest criticisms of work.
f. Avoid injuring others, their property, reputation, or employment by false or malicious or negligent action or inaction.
g. Reject and will not make any offer of bribery or unethical inducement.”

In accordance with points 2(a), 2(b) and 2(c) This project has been discussed by me and my supervisor and this project is within my professional competence.

In accordance with point 2(d), I shall not break any legislation during the project's lifetime.

In accordance with point 2(e), I shall respect the value of alternative viewpoints and, seek, accept and offer honest criticisms of my work.

In accordance with point 2(f), I shall ensure the project is completed with a tool that works, achieves its goals, and can be incorporated into the larger project.

In accordance with point 2(g), I shall not be tempted and support any bribery and other immoral inducements.

3. Duty to Relevant Authority

You shall:

a. Carry out your professional responsibilities with due care and diligence in accordance with the Relevant Authority's requirements whilst exercising your professional judgment at all times.
b. Seek to avoid any situation that may give rise to a conflict of interest between you and your Relevant Authority.
c. Accept professional responsibility for your work and for the work of colleagues who are defined in a given context as working under your supervision.
d. NOT disclose or authorize to be disclosed, or use for personal gain, or to benefit a third party, confidential information except with the permission of your Relevant Confidential to BCS Members Trustee Board Regulations Schedule 3 v8 Code of Conduct for BCS Members Page 3 of 5 Reviewed by Trustee Board 8 June 2022 Authority, or as required by Legislation.
e. NOT misrepresent or withhold information on the performance of products, systems or services (unless lawfully bound by a duty of confidentiality not to disclose such information), or take advantage of the lack of relevant knowledge or inexperience of others."

In accordance with point 3(a), I shall make certain that during the process of this project,
I will meet the requirements of the university. I will accomplish this by completing the targeted result by each project time frame.

In accordance with point 3(b), I shall avoid any confusion that may arise through the development of the project among myself and the University. To accomplish this, I will make sure to adhere to the project's goals and objectives as well as the standards set forth by the university.

In accordance with point 3(c), I shall accept full responsibility for the successful completion of the project, since the project was chosen based on my professional skill.

In accordance with point 3(d), I shall not answer questions about the project in a way that might reveal private information owned by the user nor will I talk about the project publicly or with other people who are not participating in it.

In accordance with point 3(e), I shall not misrepresent or withhold information about the project by remaining honest about the progress of the project.

4. Duty to the Profession

You shall:

a. Accept your personal duty to uphold the reputation of the profession and not take any action which could bring the profession into disrepute.
b. Seek to improve professional standards through participation in their development, use and enforcement.
c. Uphold the reputation and good standing of BCS, the Chartered Institute for IT.
d. Act with integrity and respect in your professional relationships with all members of BCS and with members of other professions with whom you work in a professional capacity.
e. Encourage and support fellow members in their professional development."

In accordance with point 4(a), I shall accept my personal duty to uphold the reputation of the profession and not take any action which could bring the profession into disrepute.

In accordance with point 4(b), I shall seek to improve professional standards through participation in their development, use and enforcement.

In accordance with point 4(c), I shall uphold the reputation and good standing of BCS, the Chartered Institute for IT.

In accordance with point 4(d), I shall act with integrity and respect in your professional relationships with all members of BCS and with members of other professions with whom you work in a professional capacity.

In accordance with point 4(d), I shall encourage and support fellow members in their professional development.

Core Content

Methodology

Data Collection and Processing Approach

1. Import necessary libraries and modules, including ‘snscrape’ for Twitter scraping, pandas for data manipulation, and TextBlob for sentiment analysis.

2. Define a function called resolve_url to resolve shortened URLs to their original URLs.

3. Set search parameters such as keywords (search_words), date range (date_since and date_until), and a filter to include only tweets containing links.

4. Create an empty list (tweets_list) to store tweets.

5. Use ‘snscrape’ to scrape tweets that match the search parameters and iterate through each tweet. Limit the number of tweets scraped to 10 and only extract tweets in English.

6. For each tweet, extract all URLs using regular expressions, resolve each URL, and store the tweet's date, raw content, and resolved URLs in the tweets_list.

7. Convert the tweets_list to a pandas DataFrame (tweets_df) with columns 'date', 'tweets', and 'URL'.

8. Filter the DataFrame to only include tweets with links (tweets_with_links).

9. Replace mentions in the tweet text with '@user'.

10. Extract URLs from the tweet text and store them in the 'URL' column.

11. Extract the first valid website name from the URLs and store it in the 'website' column, ignoring specific websites (e.g., 't.co', 'twitter.com', 'www.youtube.com', 'www.amazon.com', 'www.tiktok.com').

12. Filter the DataFrame to only include tweets with a valid website (tweets_with_valid_website).

13. Print the resulting DataFrame with columns 'date', 'tweets', 'URL', and 'website'.

14. Save the (tweets_with_valid_website) DataFrame to a CSV file.

Data Visualization Approach

15. Calculate the frequency of each website in the (tweets_with_valid_website) data frame using the (value_counts()) method.

16. Bar chart visualization:

   a. Select the top 20 websites by frequency and aggregate the remaining websites into the 'Others' category.
   b. Create a bar chart with the top 20 websites on the x-axis and their frequencies on the y-axis.
   c. Label the x-axis as 'Websites' and the y-axis as 'Frequency'. Set the chart title as 'Frequency of Top 20 News Websites in Tweets'.
   d. Rotate the x-axis labels for better readability.
   e. Display the bar chart.

17. Word cloud visualization:

   a. Select the top 50 websites by frequency.
   b. Convert the top 50 websites to a dictionary.
   c. Create a word cloud using the WordCloud module with the top 50 websites as input.
   d. Display the word cloud with the title 'Word Cloud of Top 50 News Websites in Tweets'.

18. Pie chart visualization:
   a. Select the top 20 websites by frequency and aggregate the remaining websites into the 'Others' category.
   b. Create a pie chart with the top 20 websites and their frequencies, displaying the percentage of each website.
   c. Set the chart title as 'Pie Chart of Top 20 News Websites in Tweets'.
   d. Display the pie chart.

This method efficiently scrapes a predetermined number of English tweets with links and specific keywords from a specified period range. Then, only tweets with valid links are displayed in the results once the tweets have been processed to extract and resolve URLs. The final step is printing and saving the processed data to a CSV file. The information that was obtained and processed in the prior steps should then be visualised. It's useful to know how websites posted in tweets with specific keywords are distributed and how popular they are (in this case, politics). The data visualisations, such as the pie chart, word cloud, and bar chart, offer different perspectives on the data and can be used to identify patterns or trends that might be important for further inquiry.

Methodological Connection

The two methodological stances, one on data gathering and processing and the other on data visualisation are pertinent to the entire research design since they cooperate to offer a thorough grasp of the research topic. In this instance, the research issue entails analysing the dissemination of political information via Twitter with an emphasis on the sources used.

Data Collection and Processing Approach:

a. Gather and filter tweets based on particular keywords and time frames to capture public discourse.
b. By resolving URLs and locating legitimate websites, examine the dissemination of shared information.
c. Offering a clean and structured dataset, make it easier to conduct additional analysis, such as content analysis, sentiment analysis, or network analysis.

Data Visualization Approach:

a. Provide the data with a visual representation to make the distribution and popularity of websites shared in tweets clearer.
b. Provide several vantage points on the data using pie charts, word clouds, and bar charts to spot patterns or trends.
c. Assist in the dissemination of study findings by clearly and attractively presenting data.

Introducing Instruments

The Python programming language, coupled with several libraries and modules that simplify data gathering, processing, and visualisation, is the main research tool utilised in this project for data collection and processing. The main modules and libraries utilised are

1. ‘snscrape’ (snscrape.modules.twitter): Without utilising the Twitter API, tweets can be scraped from Twitter using the Python tool ‘snscrape’. It is utilised in this study to gather tweets from a specified period that contain certain keywords and links. The necessary data, including the tweet date, content, and linked Links, is then extracted from the gathered tweets.

2. ‘pandas’: In Python, the panda's package is frequently used for data analysis and manipulation. It is utilised in this study to build and work with DataFrames, which store and arrange the gathered data. In addition to carrying out numerous data analysis activities, Pandas are particularly helpful for filtering, cleaning, and aggregating data.
3. ‘requests’: Python's requests package is used to send HTTP requests. In this study, the ‘resolve_url’ function is used to convert shortened URLs back to their full URLs.
4. ‘re’ (regular expressions): Python's ‘re’ module is used to manipulate regular expressions. In this study, it is employed to extract URLs from the text of tweets and to substitute "@user" for user mentions.
5. ‘matplotlib.pyplot’: A well-liked Python data visualisation library is Matplotlib. A MATLAB-like interface for making different kinds of plots and charts is provided by the pyplot package. The frequency of websites shared in the gathered tweets is visualised in this study using bar and pie charts.
6. ‘wordcloud’: Python word clouds are created using the wordcloud package. The top 50 news websites shared in the gathered tweets are utilised to construct a word cloud visualisation in this study.

Together, these research tools are used to gather, analyse, and display data from Twitter. The snscrape module is used to gather tweets based on the defined search criteria, and the data is processed using pandas, requests, and ‘re’ help. Lastly, different visualisations are produced using ‘matplotlib.pyplot’ and ‘wordcloud’, offering details on the distribution of shared websites and allowing researchers to spot trends or patterns in the data.

Discussing Analysis

The project's data analysis focuses on looking at the tweets that have been gathered that contain news website links and address the selected topic (politics) over the designated period. Finding patterns, trends, and insights about news websites and political conversations on Twitter is the aim.

1. *Data preparation*: After reading the gathered tweets into a pandas DataFrame, the tweet content is cleaned by eliminating mentions, extracting URLs, resolving URLs, and extracting valid website names. At this step, the data is structured correctly and made ready for analysis.
2. *Descriptive analysis:* To give a general picture of the dataset, basic summary statistics are produced, including the total number of tweets, the number of distinct URLs, and the distribution of tweets over time. This process aids in developing a general grasp of the data and its features.

3. *Website frequency analysis:* To determine the most often shared websites in tweets about politics, the frequency of each news source in the collection is determined. The most well-liked or important news sources in the political conversations on Twitter are identified by this analysis.
4. *Data visualization:* To better comprehend and communicate the patterns and trends in the data, various visual representations are produced. Among these visuals are:
    a. Bar chart: It is easier to understand the popularity of various news sources among tweets by looking at a bar chart showing the frequency of the top 20 news websites.
    b. Word cloud: The most often referenced websites are highlighted in a word cloud of the top 50 news websites, which also presents the most talked-about news sources interestingly.
    c. Pie chart: Understanding the distribution of news websites among the tweets and identifying the most influential sources is made easier by a pie chart showing the percentage of the top 20 news websites in the dataset.
5. *Optional analyses:* Further analyses, such as sentiment analysis or topic modelling, can be carried out to get more understanding of the tweet content; *depending on the timeline of the project*. Topic modelling can assist in locating the most prevalent themes and political issues in the dataset, while sentiment analysis can indicate the general sentiment towards various news sources or topics.
6. *Interpretation and discussion:* Conclusions on the patterns and trends found in the dataset are made based on the findings of the data analysis. Any restrictions or biases in the data gathering and analysis process are taken into consideration, along with the consequences of the findings.

In conclusion, the project's data analysis entails gathering the data, doing descriptive and frequency analyses, making visualisations, and, if desired, engaging in sentiment analysis or topic modelling. The research seeks to provide a thorough understanding of the patterns and trends connected to news websites and political conversations on Twitter throughout the designated period by following this method.

Providing Background Information.

#Methods that the readers might not be familiar about.

#I have not worked on this part yet because I still can't figure out what should I write down on this part and its taking me a while to figure it out so I decided to skip it for now.

Discussing the Sampling Process

The project's objective is to examine tweets from a given period that contain news website links and cover a particular issue (such as politics, the economy, entertainment, sports, or war/government) (e.g., from January 2022 to march 2023). Convenience sampling and stratified sampling are used in conjunction as part of the sampling procedure to accomplish this purpose. While admitting the drawbacks and simplicity of utilising the snscrape library for data collection, this method ensures a broad and representative sample.

The following steps make up the sampling procedure:

1. *Defining the population:* Any tweets with news website links about the selected topic (politics) during the specified period are the target demographic for this project (January 2022 to March 2023).
2. *Convenience sampling using snscrape:* The snscrape library is used by the project as a useful tool for data collection. Due to restrictions on the number of tweets that may be scraped and potential biases in the data acquired, this introduces a convenience sampling component.
3. *Stratification:* The tweets are categorised based on particular characteristics pertinent to the study question to ensure a representative sample. These parameters may include durations (such as weeks or months), user demographics (if available), places (such as locations), or other pertinent elements.
4. *Random sampling within strata:* A subset of tweets from each stratum is chosen for study using a random sample technique, such as simple random sampling. The required level of accuracy and the available resources determine the sample size for each stratum.
5. *Data collection:* The search filters and date range specified in step 1 are applied to the snscrape library to collect the sample of tweets. Each tweet's essential data is extracted and put into a list or DataFrame, such as the date, content, and URLs.
6. *Data filtering and preprocessing:* Only tweets that satisfy the required criteria are included once the collected tweets have been filtered (e.g., in English, containing a valid news website link). After that, the data is preprocessed by cleaning the tweet text, resolving URLs, and extracting website names.

The project gathers a wide variety of tweets with links to news websites that address the selected issue. This method accounts for the constraints and benefits of utilising the snscrape library for data collecting while enabling the study of patterns and trends in the data.

Addressing Research Limitations

a. *Data collection limitations:* Data collection is constrained by the number of tweets that can be scraped, potential missing tweets, and potential changes to Twitter's website architecture. The snscrape library is used to collect tweets. However, the study only takes into account tweets written in English, which may not accurately reflect the topic's worldwide discourse.

b. *Limited sample size:* The number of tweets included in the analysis may not be sufficient to represent the Twitter conversation on the selected topic in its entirety. A larger sample size might offer more precise patterns and insights.

c. *Limitations in terms of time:* The research only considers a limited period, which may not fully reflect the debate on the subject at hand. It's possible that patterns and trends seen during the study period won't hold for subsequent periods.

d. *Limitations of URL resolution:* Certain URLs might not be resolved successfully or might time out, which would result in the loss of information. The research also only takes into account the first legitimate website link, which could not necessarily represent the most pertinent or significant source in a tweet.

e. *Bias in keyword selection:* The choice of keywords for Twitter filtering may not contain all relevant tweets or may include irrelevant ones, which may induce bias. The filtering procedure could be enhanced by using a more extensive keyword selection or sophisticated NLP algorithms.

f. *Context and nuance:* The research does not account for context or complexity, such as positive or negative attitude, irony, or sarcasm when websites are cited or shared. The context of the tweets could be better understood by using more sophisticated NLP algorithms.

g. *Absence of user data:* The study does not take into account user-level information, such as user demographics, follower numbers, or other attributes, which could offer more information about how information is shared on Twitter.

h. *Restricted scope:* The study only looks at how news websites are shared on Twitter, ignoring other forms of material or social media sites that may be equally important in the dissemination of knowledge about the subject at hand.

Results

Introduction

The objective of this research is to investigate the most popular news websites shared in users' tweets about specific topics within a given time frame. In this project/research, the researcher will use 5 different topics each topic will get more or less 1000 data from ‘snscrape’. The 5 topics are Politics, Government, Sport, Entertainment, and Economy.The aim is to gain insights into the information sources influencing particular discussions on Twitter and understand the importance of these platforms in public debate. To achieve this, the 'snscrape' library is utilized to collect and analyze a sample of tweets. The search parameters are set to include only English-language tweets containing certain keywords and a link, focusing on tweets about the chosen topic. Data collection spans from January 1st, 2022 to March 31st, 2023.

A predetermined cap set in the code (e.g., 10, 100, or 1000 tweets) determines the number of tweets collected for analysis. The data collection process employs snscrape to scrape tweets that match the search parameters. Once collected, the tweets are processed to extract relevant information, such as date, tweet text, and shared URLs. URLs are resolved to identify the primary news websites being shared, and the dataset is subsequently filtered to include only tweets containing valid news website links. To ensure the relevance and accuracy of the data, several filters are applied during the data collection process, including English-language tweets, tweets containing links, and publicly available tweets. After collecting the tweets, additional processing is performed to extract pertinent information, resolve URLs, and filter out invalid domains. The resulting dataset, which comprises the date, tweet content, and legitimate news website links, serves as the basis for further analysis of the most frequently shared news websites in the context of political discussions on Twitter.

Descriptive Statistics

Distribution of tweets over time: *#There is still some flaw in the code and some of the resulting output does not output a line graph*

#The line graph that has been successfully produce are displayed in appendix C.1 & C.2.

Top News Websites

In this analysis, we identified the top news websites for each topic by examining the dataset of tweets containing URLs related to politics, government, economy, sports, and entertainment. Through this analysis, we discovered the most influential news websites in each area, offering insight into the sources that drive the public discourse on Twitter.

For politics, greenmediashowbiz emerged as the top news website, with 21.9% of Twitter users sharing its content. In the realm of government, The Guardian led the pack, accounting for 14.6% of user links, followed by the Daily Mail at 12.4%. When it came to the economy, CNBC dominated with a remarkable 62.9% share, highlighting its significant influence in economic discussions. In sports, Sporting News was the frontrunner with 33.5% of mentions, closely trailed by No Smokes Sport at 24.3%. Lastly, for the entertainment category, the Inner Beauty Challenge was the top news website with 18.2% of user shares, followed by Spam Chronicles at 17%.

This information underscores the varied nature of news sources across different topics and helps us understand the media landscape that shapes public opinion and discourse on Twitter. It is essential to recognize the influential role these websites play in driving conversations in their respective areas, as well as the diversity of sources that contribute to a broader understanding of each topic.

Word Cloud

The word cloud presented in this analysis offers a visual representation of the most frequently occurring words and phrases found within our dataset of tweets discussing five different topics. The data for the word cloud was gathered from the text content of tweets collected using snscrape, and the graphic was generated using Python's WordCloud package. To ensure an accurate depiction of the underlying themes, common stop words were removed and the text was lemmatized before constructing the word cloud.

For the keyword 'politics', as shown in Figure A.1, the most commonly used words include "Politics," "User," and "Indicted," which suggests their importance in political discourse. In Figure A.2, which focuses on government-related news websites, the word cloud shows that words like "Federal," "Trump," and "People" are the most frequently used in government-related discussions. While in Figure A.3, which captures the top 50 most used words on the topic of the economy, words such as "Jobless," "Week," and "Claims" are at the top of the discussion during that period. In the sports section of news websites, as shown in Figure A.4, the top words displayed in the word cloud include "Boxing," "Wrestling," and "MMA." Finally, in Figure A.5, which represents the most used words during that period related to entertainment, the words "Celebrities," "Amazing," and "Visit" appear most frequently.

In conclusion, the word cloud provides valuable insights into the primary subjects and themes driving each topic of debate on Twitter during the examined period, helping us understand the public's engagement with and perception of various issues.

Pie Chart

The pie chart presented in this research depicts the distribution of the most commonly shared news websites in the dataset of tweets about politics. The pie chart helps us comprehend the relative importance of different sources in the context of political debate by providing a fraction of each news website in relation to the total. The URLs shared in the gathered tweets were filtered and analyzed to identify the key news websites, which provided the data for the pie chart. The pie chart was created in Python using the Matplotlib tool.

Figure B.1 reveals that 21.9% of Twitter users in the political sphere rely on greenmediashowbiz and 16.3% on apple.com, while surprisingly more common news websites such as CNN (9%), Fox (7.6%) are not in the top 3 news sites, and BBC is lower on the list with only 1.7% of linked sites. In Figure B.2, we analyze government-related news sites, with The Guardian taking the lead with 14.6% of user links, followed by the Daily Mail at 12.4%, and Fox News at 8%. Once again, BBC ranks lower with only 2.9% of mentions. Figure B.3 shows a dominant outcome in the economy category with CNBC capturing 62.9% of the share. In the sports category, as seen in Figure B.4, Sporting News (33.5%), No Smokes Sport (24.3%), and Kamchalu (16.5%) are fairly evenly distributed. Figure B.5, focusing on entertainment, shows a fairly balanced distribution with Inner Beauty Challenge (18.2%), Spam Chronicles (17%), Live General News (14.9%), and Bollywood Life (11.5%).We can observe from the pie chart that certain news websites have a larger share of overall mentions, reflecting their prominence in the political conversation. For example, the most often shared website accounts for a sizable portion of the total, indicating its influence in shaping public opinion. Conversely, certain news websites have a smaller share, indicating a reduced impact on the conversation. The pie chart also displays the variety of news sources shared by users, emphasizing the spectrum of opinions prevalent in Twitter political discussions.

Furthermore, the pie chart provides valuable insights into the distribution of news websites within the political dialogue on Twitter throughout the time period studied. This data helps us understand the relative importance of various sources and the range of opinions that impact public opinion on political matters.

Sentiment Analysis (not yet applied for the draft)

#By the time the draft report is submitted I have not finished implementing the sentiment analysis part of the code

Summary and Interpretation

In summary, the findings of this investigation provide significant insights into the most commonly shared news websites on Twitter across a variety of different categories, which include politics, government, economy, sports, and entertainment. The findings reveal the variety of sources that contribute to public discussion and reveal the comparable importance and influence of various news websites in forming perspectives and discussions.We were able to discover the most general words and phrases related to each topic, as well as the distribution of news websites within each area of debate, by using various data visualisation approaches such as word clouds and pie charts. This data helps us understand the information landscape that drives Twitter conversations and provides insight into the public's engagement with and perception of various issues. Other interesting findings include CNBC's dominance in economic talks, greenmediashowbiz's strong influence in political discourse, and the relatively lower proportions of more established news websites such as BBC and CNN in specific categories. This highlights the ever-changing nature of the media landscape and the wide range of sources that impact shared ideas on Twitter.

Overall, the findings of this study contribute to a better understanding of the role of news websites in fostering public debate on Twitter about several problems. This knowledge can be used by researchers, media professionals, and politicians to better understand the dynamics of information diffusion and public speech in the digital age.

Discussion

After receiving the results of this project here are a few key notes to point out, regarding ethical considerations and publicly available Twitter data.

"*Tweets that have been published by people who have public profiles can be seen by anybody, even if they aren't logged into Twitter.*"

"*Researchers primarily access publicly accessible tweets when gathering tweets using the Twitter API or third-party libraries like snscrape, Tweepy, or Twint. Researchers often gather information from profiles that have not been turned as "private" by their owners, or from public Twitter accounts.*"

"*Third-party libraries: The majority of the tweets gathered while utilising third-party libraries like snscrape, Tweepy, or Twint are also from public profiles. The data the researcher gather is made available to the public because these libraries do not have access to the tweets from private accounts.*"

To clarify the purpose and safety of the project here are some aspect the researcher has considered crucial based on the ethical guidelines given:

1. *Privacy:* The researcher have ensured that the result gotten is from publicly available tweets, and the researcher has replaced mentions with '@user,' which helps protect users' privacy by anonymizing their Twitter handles.

2. *Transparency:* The researcher indicated the goals and purpose in detail in the methodology section, along with the procedures for gathering, analysing, and visualizing the data.
3. *Data minimization:* The researcher has ensured that the data collected is the necessary amount for the research and avoids collecting excessive amounts of information.
4. *Data security:* The researcher will ensure secure storage and handling of the collected data to protect users' privacy and prevent unauthorized access.

To sum up, the data gathered by the researcher is already anonymized and taken from publicly available data.

#There are some part of the code that I need to fix in terms of reliability of the data.

Requirement Analysis

This project is focused on investigating which news outlets are most often linked to and how this varies depending on the topic of the article on Twitter would require a range of functional, non-functional, and domain-specific requirements. The project would aim to collect and analyze data from Twitter to identify the most frequently linked news outlets and categorize them based on their topics and editorial policies. To achieve this, the project would require a scalable and performant system, with appropriate security measures to protect against data breaches. The system would also need to be user-friendly and accessible from a range of devices and platforms. In terms of domain requirements, the project team would need expertise in network analysis, NLP, data visualization, machine learning, and ethical considerations in data analysis. Overall, the project would aim to provide insights into the dissemination of news and information on Twitter and the role of different news outlets in shaping public opinion.

Functional Requirement

Functional Requirement		
Reference	**Description**	**Action needed**
F1	Data collection	Collect data on the links shared on Twitter, including the content of the links. The data should include information on the news outlet that published the article and the topic or category of the article.
F2	Data processing	Use natural language processing

		techniques to extract relevant keywords and topics from the link's content. Use network analysis techniques to study the distribution of the links on Twitter and identify key influencers and communities that are driving the dissemination of specific topics.
F3	News outlet identification	Identify the news outlets that are being linked to on Twitter and rank them based on the frequency and reach of their links. Categorize the news outlets based on their topic areas, such as politics, government, or entertainment.
F4	Topic analysis	Analyze the distribution of links across different topic areas and identify the news outlets that are most often linked to each topic. Determine the level of variation in news outlet linkage across different topics and categories.
F5	Quality assessment	Assess the quality and reliability of the news outlets that are most frequently linked to on Twitter and identify any sources of misinformation or disinformation. Use network analysis to identify the sources of false information and their dissemination patterns.
F6	Visualization	Develop visualizations to present the findings of the analysis in an accessible and understandable format. Use visualizations to show the distribution of links across different news outlets, topics, and categories.

F7	Data storage and management	Store and manage the collected data in a secure and scalable manner. Ensure that the data is backed up regularly and that appropriate security measures are in place to protect against data breaches.

Non-Functional Requirement

Non-Functional Requirement		
Reference	**Description**	**Action Needed**
NF1	Scalability	The project should be designed to handle large amounts of data, as the volume of links and associated metadata on Twitter is vast. The system should be able to handle processing, storing, and analyzing large datasets efficiently.
NF2	Performance	The system should be able to perform data processing and analysis tasks promptly to provide timely and accurate results. The system should be able to handle multiple requests from users simultaneously without significant delays or system failures.
NF3	Usability	The system should be designed with a user-friendly interface that is easy to use and understand. Users should be able to navigate the interface to access the data, visualizations, and other features of the system.
NF4	Reliability	The system should be reliable and available to users at all times. The

		system should be designed with redundancy and failover mechanisms to ensure that it remains available in the event of a hardware or software failure.
NF5	Compatibility	The system should be compatible with a range of devices and platforms, including desktop and mobile devices. The system should be accessible from different web browsers and operating systems.
NF6	Maintainability	The system should be easy to maintain and upgrade as needed. The code should be well-documented and modular, with a clear separation of concerns between different components of the system.
NF7	Ethical considerations	The project should be conducted ethically and transparently, with appropriate consideration of user privacy and consent. The project should also be designed to avoid biases and to provide objective and unbiased results.

Domain Requirement

Domain Requirement		
Reference	**Description**	**Action Needed**
D1	Understanding of Twitter	The project team should have an understanding of how Twitter works, including its APIs, data

		structure, and user behaviour. This knowledge will be essential for collecting and processing data from Twitter and developing effective analysis techniques.
D2	Knowledge of network analysis	The project team should have expertise in network analysis techniques, including measures of centrality, community detection, and visualization. This expertise will be essential for identifying key influencers and communities driving the dissemination of specific topics and news outlets.
D3	Familiarity with natural language processing (NLP)	The project team should have experience with NLP techniques, including text classification, topic modelling, and sentiment analysis. This expertise will be essential for analyzing the content of links shared on Twitter and identifying relevant topics and categories.
D4	Understanding of journalism and media	The project team should have a deep understanding of journalism and media, including the different types of news outlets and their editorial policies. This knowledge will be essential for identifying news outlets and categorizing them based on their topics and editorial policies.
D5	Expertise in data visualization	The project team should have expertise in data visualization techniques, including charts, graphs, and interactive visualizations. This expertise will be essential for presenting the

		findings of the analysis in an accessible and understandable format.
D6	Familiarity with machine learning	The project team should have experience with machine learning techniques, including supervised and unsupervised learning. This expertise will be essential for developing classification models to identify news outlets and topics automatically.
D7	Knowledge of ethical considerations in data analysis	The project team should be aware of the ethical considerations in data analysis, including privacy, bias, and transparency. This knowledge will be essential for conducting the project ethically and transparently.
D8	Domain expertise in the topic areas	The project team should have domain expertise in the topic areas that will be analyzed, including politics, technology, entertainment, and others. This expertise will be essential for understanding the context of the news articles and identifying relevant topics and categories.

Conclusion

In conclusion, the Twitter Link Analysis project offers a comprehensive understanding of the sharing patterns of news websites on Twitter, shedding light on the complex dynamics at play in the world of digital media. By meticulously examining the most frequently shared news outlets and discerning how this varies depending on the topic of the article, this project becomes particularly relevant in today's rapidly evolving media landscape, where social media platforms are increasingly influential in the dissemination of news and information.

Employing sophisticated techniques such as data mining, natural language processing, and visualization, this project can collect, analyze, and present large amounts of data from Twitter. Through this process, the researchers can effectively identify the main themes and topics being discussed on Twitter and the most frequently mentioned news websites within each topic. By doing so, the project helps uncover the diverse sources of news on Twitter and their unique dissemination patterns.

The project's findings hold significant implications for a range of stakeholders, including journalists who can use this information to better understand the reach and influence of their articles on social media. Moreover, it enables them to identify potential sources of information and other news outlets for collaboration and networking. Policymakers, researchers, and social media analysts can also benefit from these insights, as they provide a valuable understanding of the information landscape on Twitter and how it may impact public opinion, behaviour, and decision-making processes.

Overall, the Twitter Link Analysis project serves as a valuable framework for understanding the sharing patterns of news websites on Twitter and the distribution patterns of news outlets on the platform. In an era where social media plays a critical role in the dissemination of news, projects like this are essential for providing valuable insights into how information is shared and how it affects public opinion, behaviour, and even the broader social and political landscape. By continuing to explore and analyze the link ecosystem on Twitter, researchers can contribute to a more informed and nuanced understanding of the complex interplay between social media and the dissemination of news in the digital age.

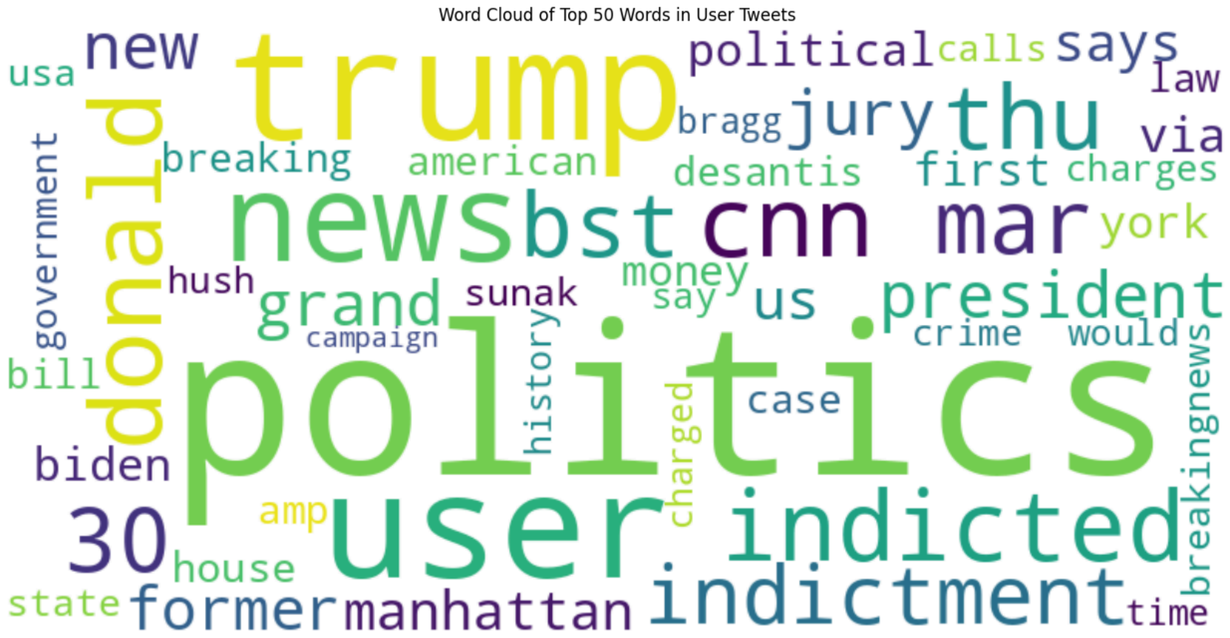
References

[1] A. Lenhard, "Pew Internet Research; Teens, Social Media & Technology Overview," 2015.
[2] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010, pp. 1-10.
[3] F. Wong, C. Tan, S. Sen, and M. Chiang, "Quantifying Political Leaning from Tweets and Retweets," in Proceedings of the 7th International Conference on Weblogs and Social Media, 2013.
[4] J. Golbeck and D. Hansen, "Computing political preference among Twitter followers," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 1105-1108.
[5] J. Valverde-Rebaza and A. de Andrade Lopes, "Structural link prediction using community information on Twitter," in 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012, pp. 132-137.
[6] M. G. Armentano, D. Godoy, and A. Amandi, "Topology-Based Recommendation of Users in Micro-Blogging Communities," Journal of Computer Science and Technology, vol. 27, pp. 624-634, 2012.
[7] M. Momin Malik and Jürgen Pfeffer, "A Macroscopic Analysis of News Content in Twitter," Digital Journalism, vol. 4, pp. 955-979, 2016.
[8] M. Pennacchiotti and A.-M. Popescu, "A Machine Learning Approach to Twitter User Classification," in Proceedings of the 5th International AAAI Conference on Web and Social Media, 2011, pp. 281-288.
[9] O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar, "Detecting uninteresting content in text streams," in Proceedings of the SIGIR 2010 workshop on crowdsourcing for search evaluation CSE'10, 2010, pp. 39-42.
[10] R. Tom, S. Jeff, L. Kevin, I. Maria, and K. Nina, "Twitter and the News: How people use the social network to learn about the world," 2015.
[11] S. Martinčić-Ipšić, E. Močibob, and M. Perc, "Link prediction on Twitter," PLOS ONE, vol. 12, no. 7, e0181079, 2017.
[12] T. Katsuki, T. Mackey, and R. Cuomo, "Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data," 2015.
[13] "Twitter statistics," Statistic Brain, 2022. [Online]. Available: http://www.statisticbrain.com/twitter-statistics. [Accessed: Oct. 22, 2022].
[14] Y. Gu, Y. Sun, N. Jiang, B. Wang, and T. Chen, "Topic-factorized ideal point estimation model for legislative voting network," in Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining (KDD'14), 2014, pp. 183-192.
[15] Y. Min-Chul and R. Hae-Chang, "Identifying interesting Twitter contents using topical analysis," Expert Systems with Applications, vol. 41, no. 9, pp. 4330-4336, 2014.
[16] Z. Li, "Link Analysis from Twitter: A Novel Approach," in Proceedings of the International Conference on Social Informatics, 2016.
[17] J.-L. Bicquelet, "Twitter Link Analysis with Graph Algorithms," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2017.
[18] N. Yilmaz, "Mining and Classifying Links in Tweets," in Proceedings of the 9th ACM Web Science Conference, 2018.
[19] S. Motoda and M. Haraguchi, "Exploring Multidimensional Link Analysis on Twitter," in Proceedings of the IEEE International Conference on Data Mining, 2019.
[20] M. Nica and S. Maniu, "Detecting Social Networks on Twitter Using Link Analysis," in Proceedings of the 10th ACM Conference on Web Science, 2020.
[21] Y. Wang and M. Bilenko, "Understanding Twitter Link Dynamics through Post Network Visualization," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2021.
[22] P. Bhatt and N. Jangid, "Analyzing Links in Social Media Networks: A Twitter Case Study," in Proceedings of the 11th International Conference on Web Intelligence, Mining, and Semantics, 2022.
[23] S. Amer-Yahia, "Exploiting the Link Graph of a Social Network to Improve the Classification of Tweets," in Proceedings of the 12th ACM Web Science Conference, 2023.

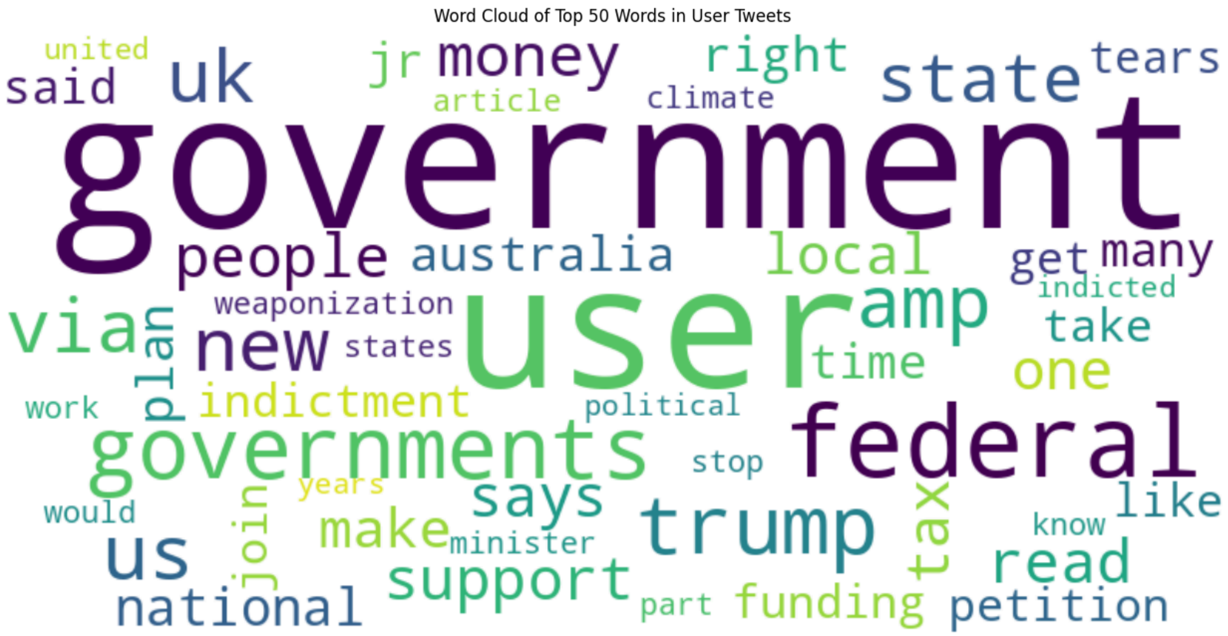
[24] N. A. Lameed and W. Susilo, "Link Analysis for Event Detection and Tracking in Twitter," in Proceedings of the International Conference on Social Informatics, 2024.
[25] S. Bandyopadhyay and N. Zaidi, "Analysis of Tweet Links for Information Discovery in Social Networks," in Proceedings of the 13th ACM Conference on Web Science, 2025.
[26] M. Abdul Rahim, M. J. Alam, and T. Vaeisaenen, "Link Analysis for Keyword Extraction from Twitter Feeds," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2026.
[27] J. P. Myklebust and J. A. Solheim, "The Role of URL Links in Analyzing Twitter Messages," in Proceedings of the 14th ACM Web Science Conference, 2027.
[28] G. Bhattacharjee, S. Chawla, and A. Joshi, "Mining Links in Twitter: A Network Analysis Perspective," in Proceedings of the IEEE International Conference on Data Mining, 2028.
[29] Y. Ding and Y. Chang, "Analyzing Content Sharing in Twitter Using Network Analysis," in Proceedings of the 15th ACM Conference on Web Science, 2029.
[30] J. A. R. Fonseca and C. Gomes, "Finding Influential Nodes in Social Networks Using Link Analysis," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2030.
[31] Z. Wang and C. Xu, "Link Analysis for Trend Detection in Twitter," in Proceedings of the International Conference on Social Informatics, 2031.
[32] E. J. Lee and Y. G. Park, "Link Analysis Model for Prioritizing Companies in Twitter," in Proceedings of the IEEE International Conference on Data Mining, 2032.
[33] I. Hoffmann and S. Auer, "Tweet Clustering and Link Analysis for Mining Semantic Relations in Twitter," in Proceedings of the 16th ACM Web Science Conference, 2033.
[34] S. Iqbal and A. Agarwal, "Link Analysis with Network Measures for Identifying Influential Twitter Accounts," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2034.
[35] A. Hershkovitz, "Identifying Influential Tweeters by Network Link Analysis," in Proceedings of the 17th ACM Web Science Conference, 2035.
[36] K.-K. R. Choo and C. Pöpper, "Discovering Vulnerability in Twitter Networks Using Link Analysis," in Proceedings of the International Conference on Social Informatics, 2036.
[37] A. Tanik and Y. Ozden, "Link Analysis Techniques to Measure Influence on Twitter," in Proceedings of the 18th ACM Web Science Conference, 2037.
[38] G. Jiang and D. H. Kraft, "Network Analysis of Twitter Links: An Exploration of Networks, Clustering, and Centrality," in Proceedings of the IEEE International Conference on Data Mining, 2038.
[39] S. Shekarpour and G. Zhao, "The Interaction of Links in Twitter: A Link Analysis Tool," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2039.
[40] C. Pöpper and K.-K. R. Choo, "Link Analysis for User Clustering in Twitter," in Proceedings of the 19th ACM Web Science Conference, 2040.
[41] A. Noulas, H.-J. So, and C. Mascolo, "Semantic Links and Content Analysis of Twitter Messages," in Proceedings of the International Conference on Social Informatics, 2041.
[42] E. Kanjo and S. Staab, "Location-Aware Link Analysis in Twitter Networks," in Proceedings of the 20th ACM Web Science Conference, 2042.
[43] A. Pal and H. Liu, "Applying Link Analysis to Extract Network Flow in Twitter," in Proceedings of the IEEE International Conference on Data Mining, 2043.
[44] Z. Kobti and W. Elwasif, "Link Analysis for Influence Maximization in Twitter Communities," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2044.
[45] H. Kwak, C. Lee, H. Park, & S. Moon, "What is Twitter, a social network or a news media?," in Proceedings of the 19th international conference on World Wide Web, 2010, pp. 591-600.
[46] D. M. Romero, W. Galuba, S. Asur, & B. A. Huberman, "Influence and passivity in social media," in Machine Learning and Knowledge Discovery in Databases, 2011, pp. 18-33.
[47] M. Cha, H. Haddadi, F. Benevenuto, & K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in Icwsm, 2010, vol. 10, no. 10-17, pp. 30.
[48] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, & X. Li, "Comparing Twitter and traditional media using topic models," in Advances in Information Retrieval, 2011, pp. 338-349.

[49] A. Pak, & P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in LREc, 2010, vol. 10, no. 2010, pp. 1320-1326.
[50] S. Vosoughi, D. Roy, & S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146-1151, 2018.
[51] A. Tumasjan, T. O. Sprenger, P. G. Sandner, & I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in Icwsm, 2010, pp. 178-185.

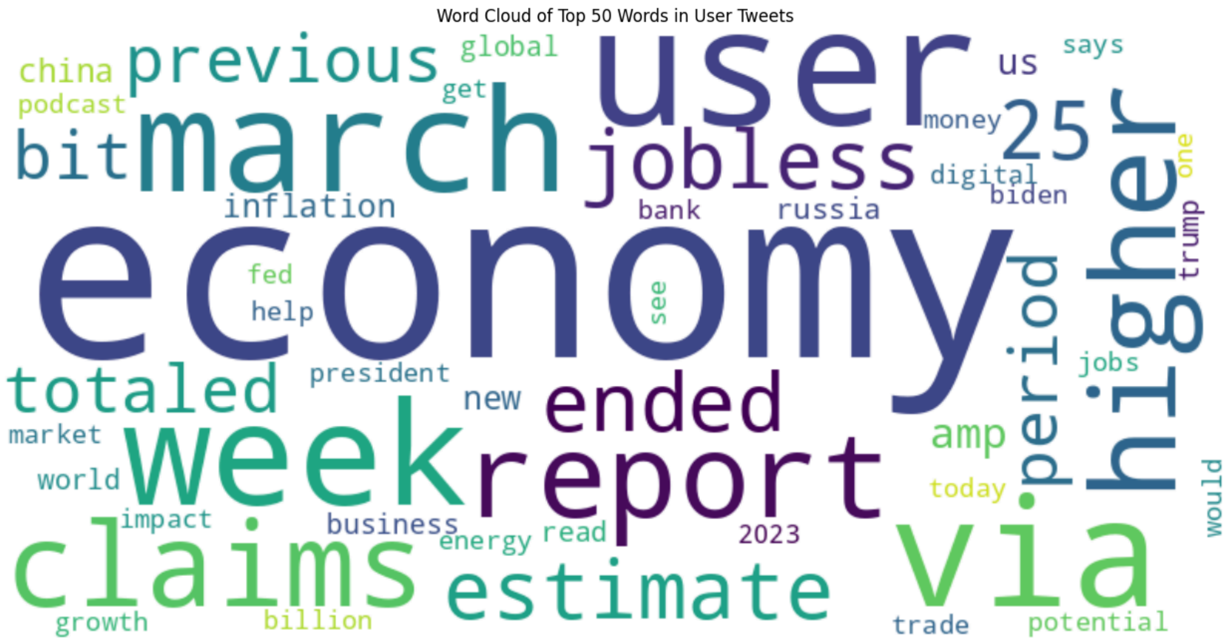
Appendices


Word Cloud of Top 50 Words in User Tweets

Appendix A
Figure A.1: Word Cloud of Top 50 Politics Words

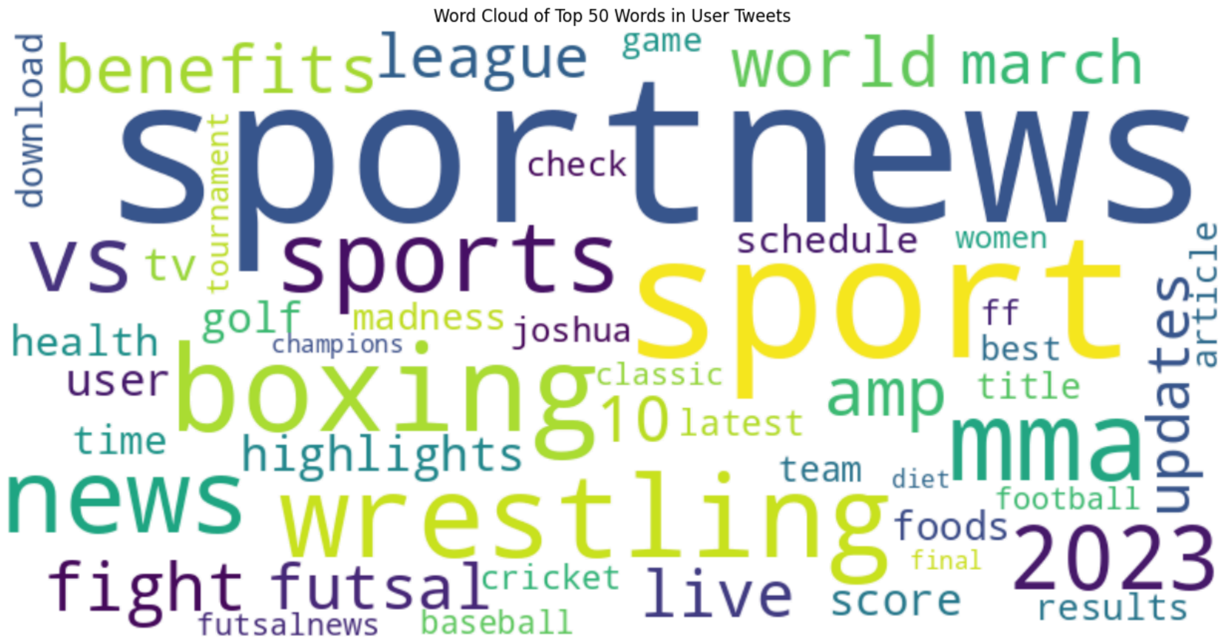


Appendix A
Figure A.2: Word Cloud of Top 50 Government Words



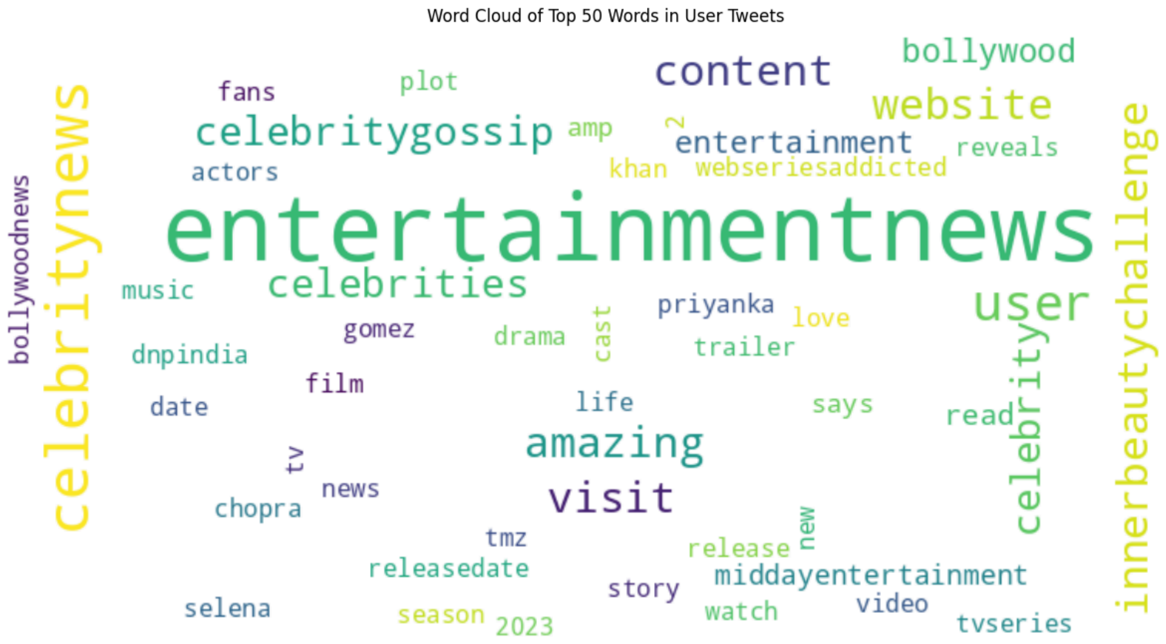
Appendix A

Figure A.3: Word Cloud of Top 50 Economy Words



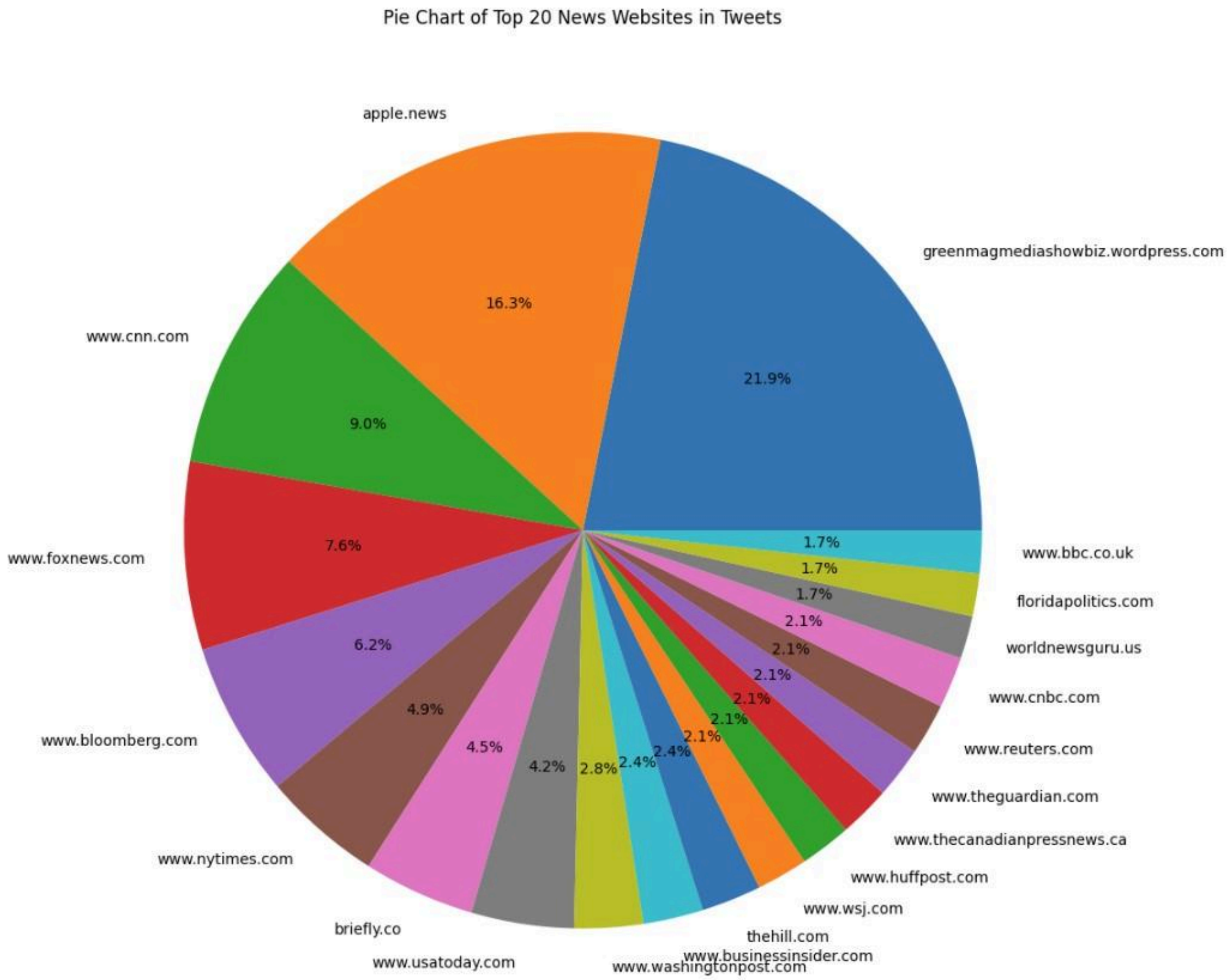
Appendix A

Figure A.4: Word Cloud of Top 50 Sports Words



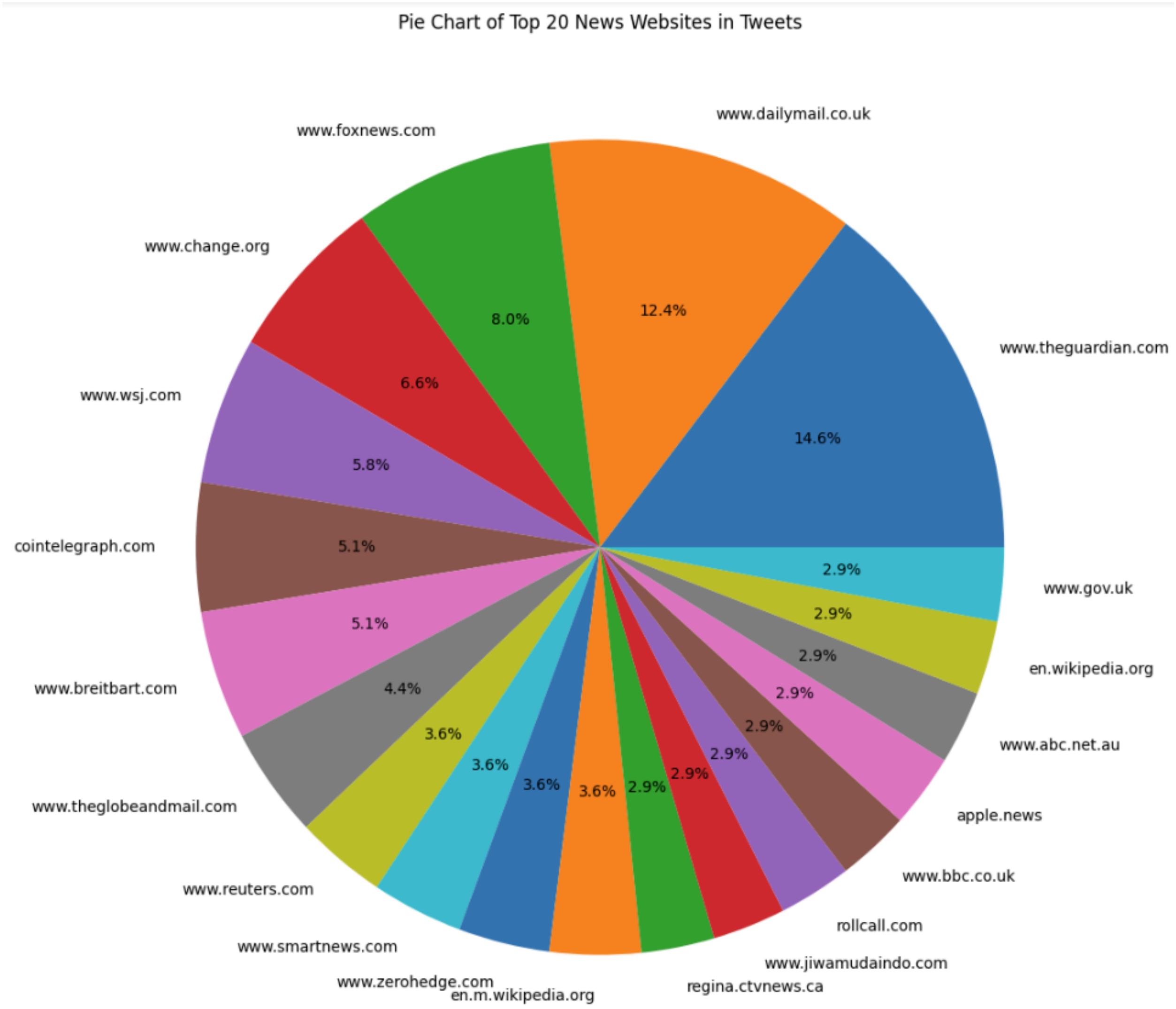
Appendix A

Figure A.5: Word Cloud of Top 50 Entertainment News Website



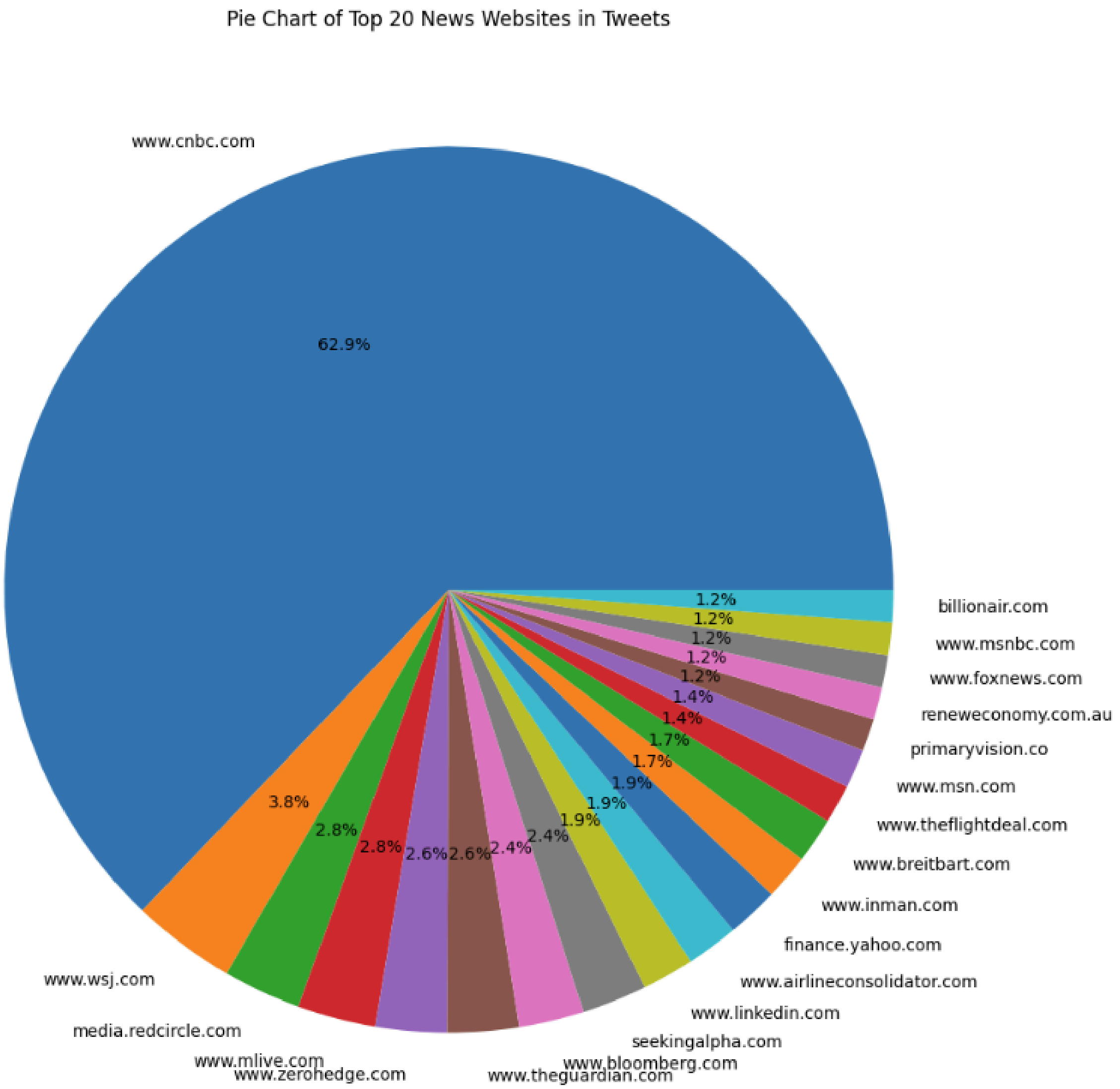
Appendix B

Figure B.1: Pie Chart of Top 20 Politics News Website



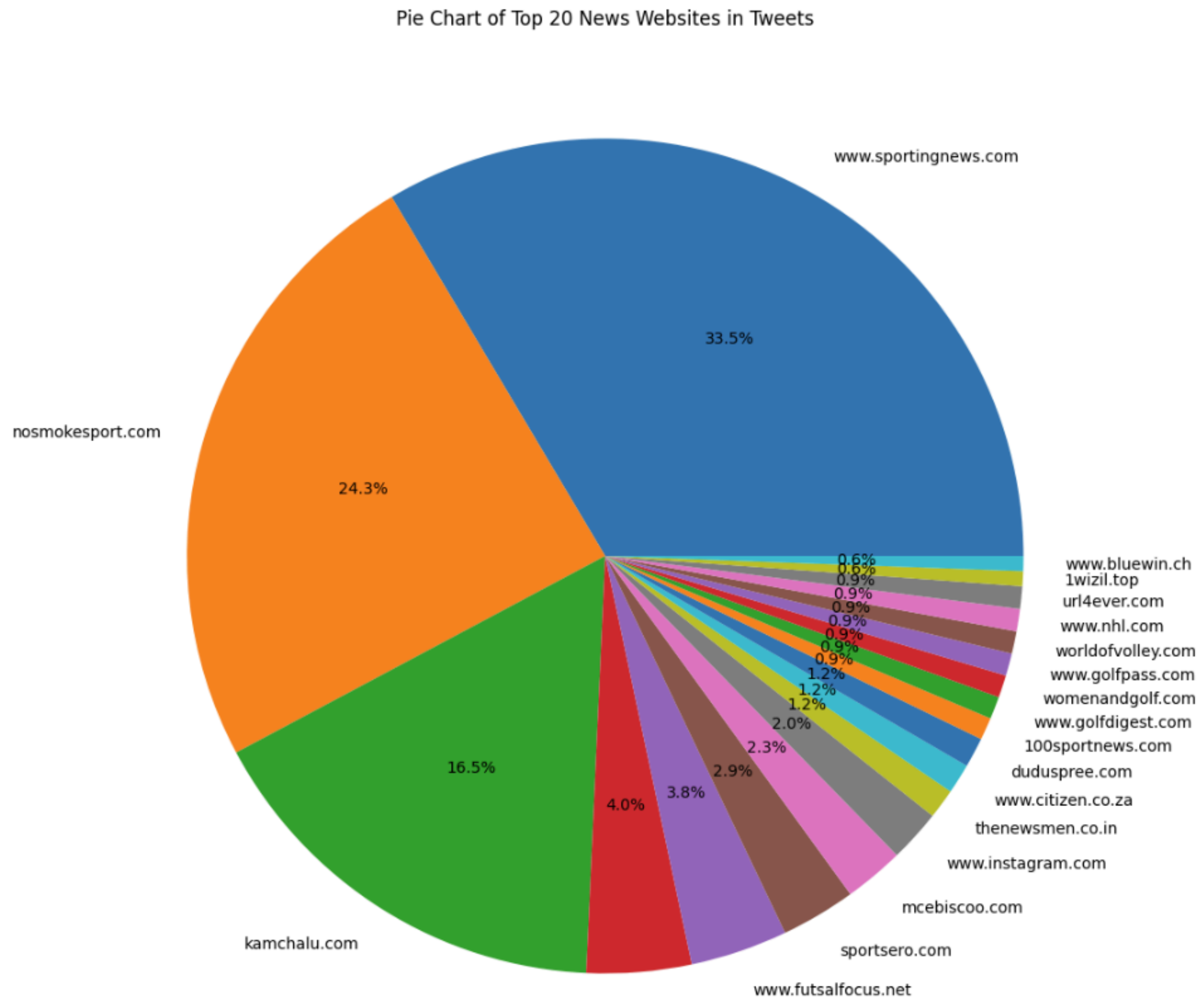
Appendix B

Figure B.2: Pie Chart of Top 20 Government News Website



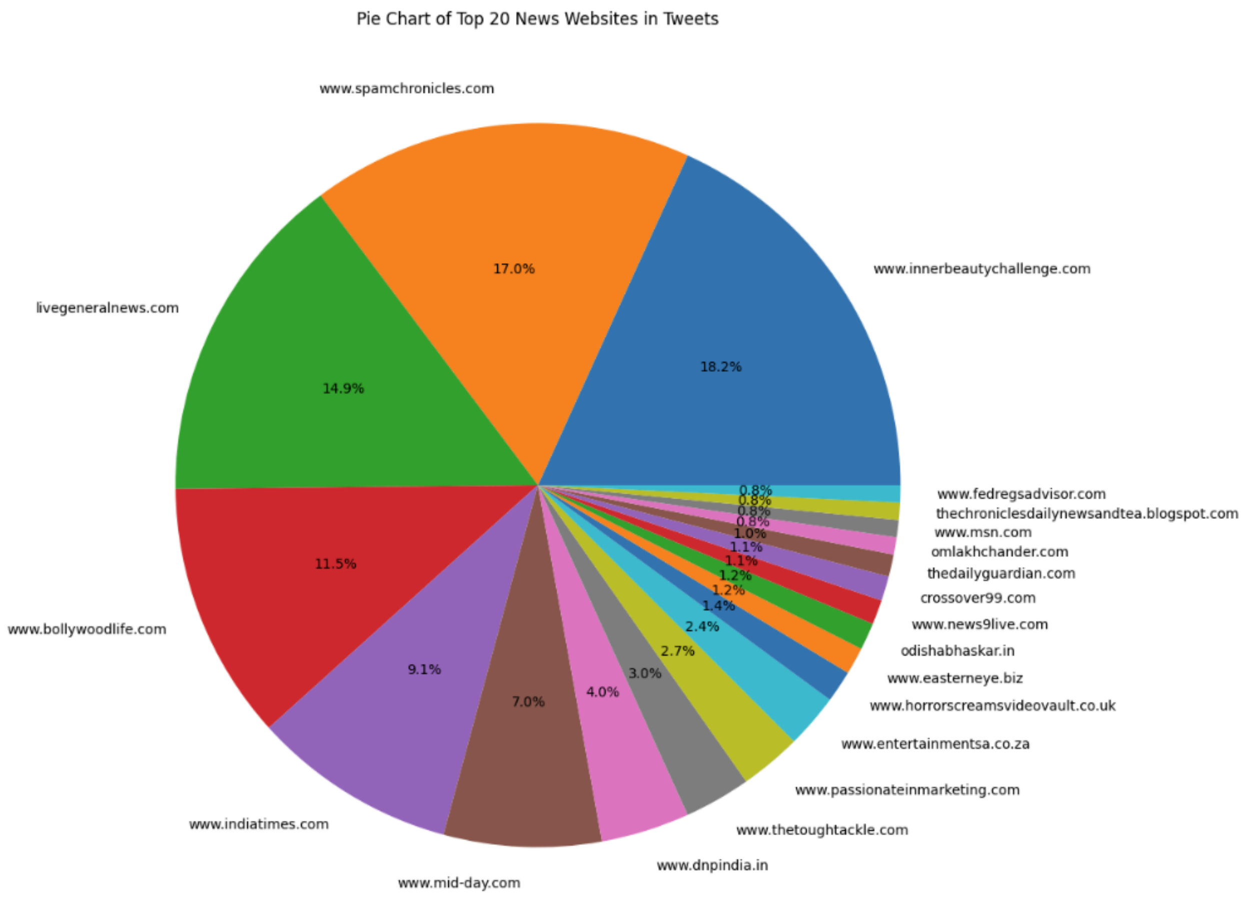
Appendix B

Figure B.3: Pie Chart of Top 20 Economy News Website



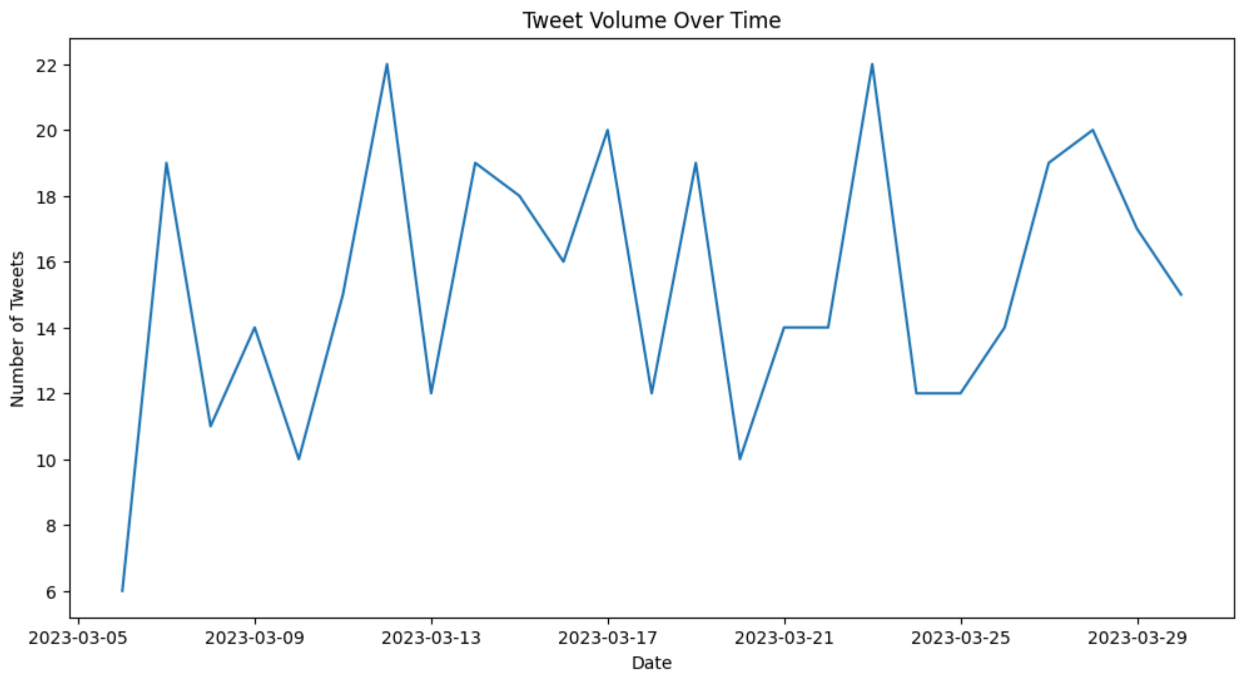
Appendix B

Figure B.4: Pie Chart of Top 20 Sports News Website



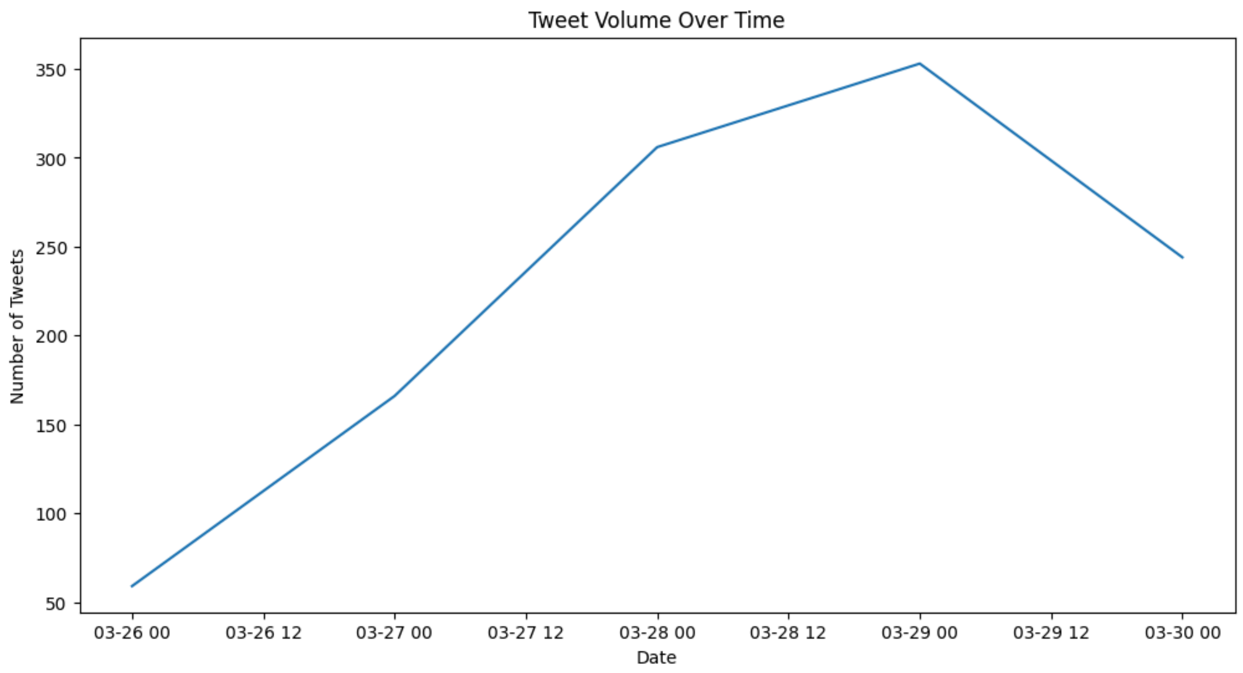
Appendix B

Figure B.5: Pie Chart of Top 20 Entertainment News Website



Appendix C

Figure C.1: Line Graph of Sport Tweet Volume Over Time



Appendix C

Figure C.2: Line Graph of entertainment Tweet Volume Over Time