Fake News

Github link: https://github.com/FuzelahamedShaik/SocialNetworkAnalysis

Fuzel Ahamed Shaik
ITEE
Uniiversity of Oulu
Oulu, Finland
fuzel.shaik@student.oulu.fi

Nirzor Talukder
ITEE
University of Oulu
Oulu, Finland
nirzor.talukder@student.oulu.fi

Miro Kakkonen
ITEE
University of Oulu
Oulu, Finland
miro.kakkonen@student.oulu.fi

***Abstract*—The goal of this project is to compare the characteristics of fake and real news on social media using various analytical techniques and determine whether these differences can be used to discriminate between them, while also providing insights into the spread of fake news and identifying relevant graph attributes and other factors that could be used for fake news detection, as well as a literature review of relevant research. In the social context, the research takes a network-based approach, employing descriptive statistics, visualization, bot identification, and network analysis as detection methods. Fake news has a huge impact on society, causing confusion, divisiveness, and a loss of faith in the media and institutions. The significance of the project lies in the potential to contribute to the development of effective approaches and tactics for detecting and countering the spread of fake news on social media. This study aims to find the user interaction between fake news and real news on Twitter. From the previous literature, we observe that the impact of fake news is more compared to real news, this intrigued us to identify the behavior of fake news in social media mainly on Twitter. Later this study suggests some future developments in this domain to create a structured pipeline to identify fake news in social media. To achieve this we leverage stweet library in Python and the source code along with datasets are uploaded in the GitHub repo given.**

***Index Terms*—Twitter, stweet, Social network, fake news, and real news**

I. Introduction

In today's society, the propagation of fake news is a major worry. Social media's growth has made it simpler for fake news to spread swiftly to a large audience. Researchers have been able to examine techniques for differentiating false news from legitimate news thanks to the FakeNewsNet dataset created by [1]. However, since humans are inherently bad at it, spotting fake news is not an easy process. False news is spread by two different categories of people: those who do it on purpose and those who don't. Since 68% of Americans acquire their news from social media, it is imperative to stop the propagation of false information.

Real news can be distinguished from fake news by several characteristics. It has a minimal cost of creation when created by bots and is purposefully crafted to deceive. The language employed in false news is biased, intended to elicit an emotional response, and frequently uses clickbait. Fake news has a high user engagement rate and travels quickly and in bursts. Spreading bogus news thrives on echo chambers, confirmation bias, and the bandwagon effect.

Researchers have put out a number of automatic false news detection techniques. These techniques include approaches that are knowledge-based, style-based, and propagation-based. Natural language processing (NLP) combined with external sources and databases is used in knowledge-based approaches to fact-check statements made in news content. Style-based methods can identify clickbait, biased wording, etc. Information like retweets, likes, the number of comments, and temporal evolution are used in propagation-based techniques.

The value of researching bogus news has plenty of literary support. According to a study by [2], the speed and originality of information dissemination might serve as reliable markers of a news item's veracity. According to [3], bogus news frequently receives more user interactions than legitimate news. [4] discovered that people with more followers were more likely to distribute fake news and that a larger percentage of the shares of fake news stories came from Twitter "bots." Finally, [5] study contend that social network analysis measures like degree centrality and clustering coefficient can be employed to differentiate between legitimate and false news.

II. Problem Description

The primary objective of this research project is to investigate the differences between fabricated news and genuine news in social media. We will use the FakeNewsNet dataset, which has a number of attributes that will be thoroughly examined, to achieve this purpose. We will pay close attention to the statistical patterns of key characteristics, the distribution of followers and followers for both types of news, the temporal evolution of user engagement, the verification of user IDs using the botometer program, and the contrast between the graph structures for real and fake news.

To better understand the differences between the two categories, we will also examine the degree distribution of fake and real news. Finally, we will examine how information spreads on social media and explore if it is possible to

distinguish between fake and real news using statistical patterns, user interaction, graph properties, etc.

We will use a variety of pertinent academic studies on social media fake news identification to inform our analysis throughout the investigation. Our final goal is to find out if it is possible to distinguish between fake news and real news on social media using the aforementioned factors. The impact of fake news in the real world is massive. Here are some of the effects of fake news in the real world.

A. The Impact of Fake News

According to [7], advertising is the main driving force behind the popularity of websites with sensational headlines, leading to companies taking advantage of the high traffic to these sites. The creators of fake news and information can profit from this phenomenon through automated advertising that rewards high traffic, as discovered by subsequent research cited in the same reference. The impact of misinformation on the public is a pertinent question, as it can cause confusion and unnecessary stress [16]. Digital disinformation, which is intentionally created to deceive and harm the public, is a form of fake news [17]. Its effects can be far-reaching, potentially impacting millions of people within minutes [16]. Disinformation has been known to disrupt election processes, cause unease, and incite disputes and hostility among the public [17].

B. Fake News and Social Media

Nowadays, the internet plays a crucial role in our daily lives, as mentioned in [21]. Social media platforms have largely replaced traditional methods of acquiring information [21]. As of 2023, Twitter has emerged as one of the largest social media platforms, hosting over 353 million active users worldwide [19]. According to [18], Twitter has played a significant role in the dissemination of fake news, surpassing other social media platforms. [18] further reports that around 48% of Twitter users worldwide obtain their news from the platform. Shockingly, 29% of Twitter users have admitted to sharing false information, either knowingly or unknowingly [20]. This rapid spread of fake news on social media platforms, particularly Twitter, is a growing concern [18].

III. Dataset Description

This study aims to investigate methods to classify social media handles based on common social network metrices focusing especially on activities of Twitter Ids. Twitter data is generated through users posting short messages or ”tweets” on the platform, which can contain text, images, videos, and links. In Twitter, mentions or tags originate when a user includes the ”@” symbol followed by another user’s username (e.g., @username) in a tweet. This creates a link to the mentioned user’s profile. The universe of Twitter mentions is extremely vast, as they can include individuals, organizations, and even fictional characters. As a result, analyzing Twitter mentions can provide valuable insights into a wide range of topics, from political discussions to brand perception. Twitter’s user base is massive, with over 368 million active users as of 2023 [8]. Identifying fake and real accounts from this large pool can be a daunting task, as fake accounts can be difficult to distinguish from real ones and can be created quickly. It is not possible to know with certainty all the real and fake news that were present on Twitter. Our approach towards addressing this issue is based on the observations possible through the public Twitter API by outside researchers without commercial or internal access to data.

The GossipCop and PolitiFact datasets are two publicly available datasets on GitHub. The GossipCop dataset includes news articles from gossip websites, while the PolitiFact dataset includes articles from political news websites. These datasets have been widely used in research on fake news detection and have contributed to the development of techniques for identifying fake news. The GossipCop and PolitiFact datasets were generated by using fact-checking websites to obtain news contents for fake and real news, and by utilizing claims from these websites as ground truths. The source URLs of web pages that published news articles were obtained from PolitiFact’s fact-checking evaluation result, while GossipCop’s rating scores on a scale of 0 to 10 were used to classify news stories as fake or real [1]. We used the Twitter ids already classified and made available through GitHub for our analysis from this dataset. We used these Ids to scrape Twitter.

This study quantifies different network attributes and tries to identify distinguishing characteristics for fake and real news in Twitter. We got our Ids for scraping Twitter from the GitHub repository. The datasets include information about Twitter users and their activity on the platform. All the datasets have some common non-numerical columns, namely - “created_at”, “user_id_str”, “id_str”, “lang”, “location”, “description”, “full_text”, “user_screen_name” and “user_name.” The numerical value containing columns for all the datasets are - followers_count (total number of followers a user has), normal_followers_count , fast_followers_count, favourites_count (total number of tweets a user has marked as favorites), friends_count (total number of accounts a user is following), media_count (total number of images or videos uploaded by the user), statuses_count (total number of tweets a user has posted), retweeted, favorited, and favorite_count (total number of times a tweet has been marked as favorite). As users post and mention users, they create networks which can be extensively analyzed using the tools of social network analysis.

IV. Data collection

Tweet ids are provided in from the FakeNewsNet datasets (GiHub link: https://github.com/KaiDMML/FakeNewsNet). It contains four different files named gossipcop_real.csv, gossipcop_fake.csv, politifact_fake.csv, and politifact_real.csv. These files has all the tweet ids under the column tweet_ids. These tweet ids are used are input to tweet data scraper. The tweet data is extracted using the try_tweet_by_id_scrap method from stweet library in Python. This method takes the tweet id as input and returns the tweet-related data in a JSON file. The

```
created_at                datetime64[ns]
user_id_str                        int64
id_str                             int64
lang                              object
location                          object
description                       object
full_text                         object
user_screen_name                  object
user_name                         object
followers_count                    int64
normal_followers_count             int64
fast_followers_count               int64
favourites_count                   int64
friends_count                      int64
media_count                        int64
statuses_count                     int64
retweeted                           bool
retweet_count                      int64
favorited                           bool
favorite_count                     int64
dtype: object
```

Fig. 1. Dataset features and their datatypes.

JSON files for different datasets are uploaded to GitHub. The raw data in JSON format is transformed into a CSV file by considering only the necessary columns. The structure of the CSV dataset is shown in Figure 1.

V. General Methodology

The primary objective of this project is to analyze and compare the characteristics of fake news and real news on social media platforms using various analytical techniques such as data collection, descriptive statistics, visualization, bot detection, network analysis, degree distribution, and information spread analysis. The project aims to determine whether there are differences in the characteristics of fake and real news and whether these differences can be used to discriminate between them. Additionally, the project aims to provide insights into the spread of fake news on social media and to identify relevant graph attributes and other factors that could be used for fake news detection. Also, the project scope includes a literature review of relevant research on fake news detection in social media.

We utilized analytic approach towards the data for the project, which involves collecting and analyzing data to draw conclusions and insights. Our approach towards fake news detection is based on the Social Context, specifically the Network-based one [9]. The analysis also involves the use of various techniques and tools such as descriptive statistics, visualization, bot detection, network analysis in the following manner –

Descriptive Statistics: Calculating descriptive statistics for fake and real news categories, including tweet messages, distinct user IDs, and various statistical measures of retweets, followers, and followees, to determine whether these attributes can help distinguish between the two categories.

Visualization: Creating plots to compare the follower and followee distribution for fake and real news of both datasets, explore the temporal evolution of user engagement for both categories, and drawing corresponding plots.

Bot Detection using botometer score: Using the botometer program to determine whether fake news originated from bots or humans.

Network Analysis: For extracting the follower relationship graph structure, we utilized NetworkX to calculate global network attributes, compare the attributes for fake and real news categories in gossipcop and politifact datasets, and draw high-level illustrations of each network.

Degree Distribution: Creating plots to compare the degree distribution of fake and real news for both datasets, and determine whether some graph attributes are relevant to distinguish between fake and real news.

Information Spread: Extracting date attributes for retweets and comparing the timely evolutions of retweets in both real and fake news datasets to conclude about the comparison between the two scenarios.

The analysis can help in fake and real news detection by providing insights into the characteristics and patterns of each category, such as the statistical measures, network structure, and information spread, which can aid in developing effective detection algorithms.

VI. Detailed Methodology

A. Data Collection

For the four datasets almost the same procedure was followed to collect data. Here, `politifact_fake_news` data collection is described as an example:

I. We installed the `stweet` package by running "`!pip install -U stweet`" in our Python environment.

II. We imported the `stweet` library using "`import stweet as st`".

III. We loaded the `politifact_fake_news` dataset into a pandas dataframe using "`pd.read_csv`".

IV. We extracted the tweet IDs from the dataset and stored them in a list.

V. We initialized an empty list to store the final tweet IDs.

VI. We iterated through the list of tweet IDs and filtered out any non-numeric IDs, storing the remaining IDs in our final list.

VII. We defined a function called `try_tweet_by_id_scrap` that takes in a tweet ID as input and performs a scraping task using the `stweet` library.

VIII. Inside the `try_tweet_by_id_scrap` function, we created an `id_task` object using the `st.TweetsByIdTask` function, passing in the tweet ID as a parameter.

IX. We created two output objects: one to print the raw `JSON` data and another to save it to a `JSON` line file.

X. We executed the scraping task using the `st.TweetsByIdRunner` function, passing in the `id_task object` and the two output objects.

XI. We initialized a counter variable to keep track of our progress.

XII. We iterated through our final list of tweet IDs, calling the `try_tweet_by_id_scrap` function on each ID and incrementing our counter variable.

XIII. Every 1000 iterations, we printed out a progress message to track our progress.

B. Preprocessing

Conversion of raw data in json to csv files for analysis:

I. We began by importing the necessary libraries - `pandas` and `json`. Then, we created a list of the `JSON` files that we wanted to process. We looped through each file in the list, read it into a Pandas DataFrame using the `pd.read_json()` function, and created an empty DataFrame to store the extracted data.

II. Next, we extracted the required fields from each `JSON` object and appended them to respective lists. These fields included data such as the timestamp of the tweet (`created_at`), the user ID (`user_id_str`), the text of the tweet (`full_text`), and various user profile information (`followers_count`, `description`, etc.). We then added these lists as columns to the empty DataFrame we created earlier.

III. Once we had extracted and compiled all the required data, we wrote the DataFrame to a CSV file using the `to_csv()` method. We also appended the name of this CSV file to another list so we could later check the shape of each processed file.

IV. Finally, we looped through the list of CSV files, read them into a new Pandas DataFrame, and printed their shapes for confirmation.

C. Analysis

1) Task 1: In this analysis, we aimed to compare user engagement metrics such as followers and favourites for users sharing real and fake news articles on two different datasets: `gossipcop` and `politifact`.

I. We started by importing the required Python libraries for data manipulation and visualization: `pandas` and `matplotlib`. We then loaded the `gossipcop` and `politifact` datasets containing real and fake news articles using Pandas' `read_csv()` function.

II. Next, we calculated the number of distinct user IDs for real and fake news articles in each dataset using the unique() method. This gave us an idea of the number of unique users involved in sharing real and fake news articles.

III. To extract user engagement metrics such as followers and favourites, we defined a function called `get_follower_favourite_count()`. This function takes as input the dataset and the user ID and returns the latest follower count and favourites count for that user.

IV. We then defined another function called `get_user_data()` that takes as input the dataset and the list of users and returns a Pandas DataFrame with columns for user ID, followers count, and favourites count.

V. To calculate statistics such as mean, standard deviation, kurtosis, and skewness, we defined a function called `calculate_statistics()`. This function takes as input the user engagement data DataFrame and returns a dictionary containing these statistics.

VI. Finally, we defined a function called `print_statistics()` that takes as input the user engagement data DataFrame and prints the statistics in a table format using the Pandas `DataFrame.to_string()` method. Using these functions, we compared the user engagement metrics for users sharing real and fake news articles in the `gossipcop` and `politifact` datasets separately. We printed the statistics for each dataset separately, allowing us to compare the results and draw conclusions about user behavior when sharing real and fake news articles.

2) Task 2: We analyzed the distribution of follower, followee count for all four datasets. I. we imported the necessary libraries for data manipulation and plotting, including `pandas` and `matplotlib.pyplot`.

II. Next, we loaded the data into a pandas dataframe and separated the fake news and real news into different dataframes for both `gossipcop` and `politifact` datasets.

III. We then plotted the distribution of follower count and followee count for both fake news and real news separately for both datasets. Using `matplotlib`'s subplot function, we created two subplots on the same figure for the fake news and real news categories. The histograms were plotted using the `plt.hist` function with the follower and followee counts as the x-axis and frequency as the y-axis. We also added labels to the x and y-axes using the `plt.xlabel` and `plt.ylabel` functions, and added a legend using the `plt.legend` function.

IV. Next, we calculated the total follower and followee counts for both fake news and real news separately for both datasets. We created a bar chart using the `plt.bar` function to show the total follower counts for each dataset. To add labels to each bar in the chart, we used the `addlabels` function. We also added a title to the chart using the `plt.title` function.

V. Finally, we created a similar bar chart to show the total followee counts instead of the total follower counts. We added labels to each bar in the chart using the `plt.text` function.

3) Task 3: We analyzed the evolution of Twitter interactions (likes, followers, favorites) for fake and real news from two different sources: `gossipcop` and `politifact` following the same steps. Here the steps we followed for the likes -

I. We imported two necessary libraries: `pandas` and `matplotlib.pyplot`.

II. We defined a date format string that will be used for parsing the date and time information in the `'created_at'` column of the data frames.

III. We used the pandas `to_datetime()` function and the defined date format to convert the `'created_at'` column of the `'gossipcop_df_fake'`, `'gossipcop_df_real'`, `'politifact_df_fake'`, and `'politifact_df_real'` data frames to `datetime` format.

IV. We grouped the data frames by the date portion of the `'created_at'` column and summed the `'favorite_count'` column to calculate the number of likes per day for fake and real news for each of the four data frames.

V. Finally, we used the `matplotlib.pyplot` library to plot the timely evolution of the number of likes for fake and real news on two subplots. One subplot was for the `'gossipcop'` data frames, and the other was for the `'politifact'` data frames.

4) Task 4: Taking inspiration from [22] and [23] we would have taken the steps below to fulfill the requirements for the project's 4^{th} task:

I. Downloaded the botometer-python program from the GitHub repository and extracting in the working directory. (GitHub link: https://github.com/saroarjahan/Twitter_Bot_post_detection)

II. This Python script reads in the user tweets from the four datasets and uses the Botometer program to determine the probability that each user ID is a bot or human.

III. The threshold of 0.5 is taken to determine whether a user tweet is classified as a bot or human. If the probability was above 0.5, then the program would've classified the user ID as a bot. If the probability was below or equal to 0.5, then it would've classified the user ID as a human.

IV. Calculating the percentage of user tweets classified as bots in the fake news and real news datasets.

V. Comparing these percentages to distinguish between news.

5) Task 5: We need the follower list for the Twitter Ids in our datasets for carrying out the analysis in task 5. Since we were having trouble with the `'stweet'` package too, we couldn't get the follower list. The following are the steps we would've followed if we had the follower list for Task 5:

I. Importing the required libraries, including pandas, networkx, and matplotlib.pyplot.

II. Copying the contents of the gossipcop_df_real dataframe to a new dataframe called df.

III. Creating an empty dictionary called followers to store the followers of each user. For each unique user_id in the df dataframe, we would've stored their followers in the followers dictionary.

IV. Next, we would've created an empty graph called G using the networkx library and added a node to the graph for each unique user_id in the df dataframe. For each unique pair of user_ids in the df dataframe, we would've checked if they are connected (i.e., if one follows the other), and if they were, we would've added an edge between them in the graph.

V. Calculating various global attributes of the network, including degree centrality, diameter, clustering coefficient, and the size of the largest component. Finally, we would've drawn the network using the spring layout algorithm and the matplotlib library.

6) Task 6: As mentioned in the previous section, we couldn't manage to get the follower list, so here we are describing the steps we would've taken to solve task 6 instead:

I. Two separate graphs would've been created for "fake news" and "real news", and edges were to be added between nodes based on their connection status and the type of news they represent (Using the same steps in task 5 to make graphs).

II. Additionally, the degree distribution of each graph would've been plotted separately and differentiated with a legend.

7) Task 7: The methodology followed to perform task 7 is described below.

I. To check the spread of information in both fake and real news datasets, we extracted the date attributes for the retweets using the Pandas library in Python. We used the `strftime` function to specify the date format and converted the `'created_at'` column to `datetime` format.

II. Next, we grouped the data by date using the `groupby` function and calculated the total number of retweets per day for fake and real news in both datasets. We used the `sum` function to calculate the total number of retweets.

III. After calculating the total number of retweets, we plotted the timely evolution of retweets for both fake and real news using the `matplotlib` library. We created a figure with two subplots, one for each dataset, and plotted the retweet counts against the date using the plot function. We set the color of the lines to red for fake news and blue for real news and added a legend to identify the lines.

VII. Results and discussion

A. Task 1

The tweet data is collected from the stweet library for the tweet ids in gossipcop and politifact real and fake. Fig. 2 describes the number of tweets and the unique user id strings in the datasets. The data collected is approximately balanced for fake and real news. Fig. 3 represents the statistical description of the four datasets. The mean values of the fake news datasets are much higher than the mean values of the real news datasets. We can also observe that followers count and the followee counts have a huge variation in fake news datasets compared to real news datasets. This explains that the population is more into fake news than real news. To explain this phenomenon, we like to explore the distribution of the followers count and the followee count.

B. Task 2

Task 2 summarizes the distribution of followers and followees regarding real and fake news in both datasets. The outcome of this task helps us to understand in what range the follower's count of real and fake news is concentrated and similarly with the followees count of real and fake news. Fig 4 and 5 display the distribution of the followee's count in gossipcop and politifact. Fig 6 and 7 interpret the followers

Dataset	No. Tweets	No. Distinct User Id str
Gossipcop Fake	10770	5869
Gossipcop Real	14394	3465
Politifact Fake	4462	3944
Politifact Real	6312	3513

Fig. 2. Datasets Overview.

```
Politifact Real News Data:
statistics    followers    favourites
      mean 9.680508e+04 35512.049305
       std 2.357307e+06 82873.025558
  kurtosis 1.252652e+03    65.073854
      skew 3.455083e+01     6.029922
Politifact Fake News Data:
statistics    followers    favourites
      mean 4.113219e+05  40123.752693
       std 5.198846e+06 124623.714313
  kurtosis 2.954838e+02    132.169450
      skew 1.656974e+01      9.680933
Gossipcop Real News Data:
statistics    followers    favourites
      mean 1.097900e+05  9312.357718
       std 1.210238e+06 46281.885686
  kurtosis 5.677006e+02   217.058651
      skew 1.811486e+01    12.282866
Gossipcop Fake News Data:
statistics    followers    favourites
      mean 3.272999e+05 35681.170566
       std 3.646276e+06 92413.976951
  kurtosis 5.678891e+02    65.093220
      skew 2.181053e+01     6.913182
```

Fig. 3. Statistical description of the datasets.

count distribution in a histogram for gossipcop and politifact datasets.

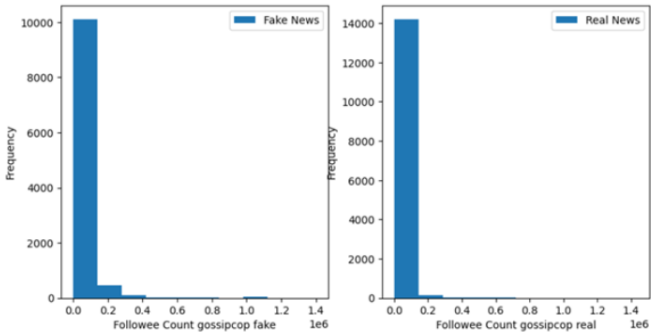

Fig. 4. Followees frequency for Gossipcop.

From these figures we can see that the user counts in fake news are much higher than the real news users. For instance, in Fig 6. the frequency of the follower count is more in between 0 to 20,000,000 for fake news whereas for real news the frequency is more in between 0 to 10,000,000.

C. Task 3

Temporal evolution of the number of likes in fake and real news is considered a valid metric to conclude that fake news will spread quickly and users are more intrigued with fake news compared to real news [2]. Fig 8. describes the temporal evolution of the likes of fake and real news in both gossipcop and politifact datasets. From the graphs, we can see that fake news has the highest number of like counts in both datasets. According to [2], false news stories on Twitter were 70% more likely to be retweeted than true stories, and the spread of false

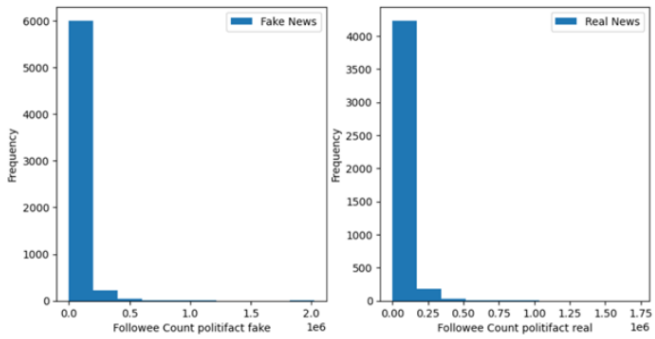

Fig. 5. Followees frequency for Politifact.

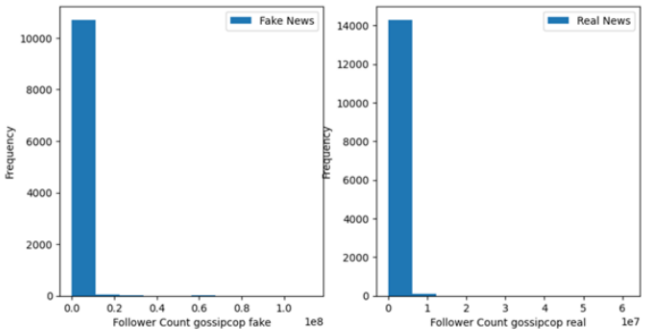

Fig. 6. Followers frequency for Gossipcop.

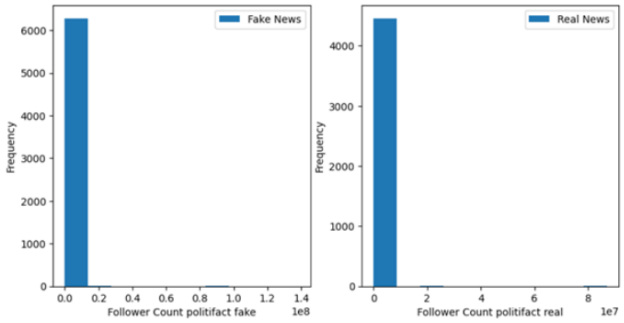

Fig. 7. Followers frequency for Politifact.

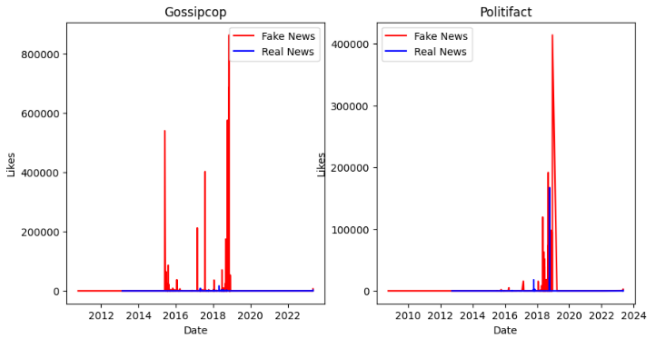

Fig. 8. Evolution of the likes in both datasets.

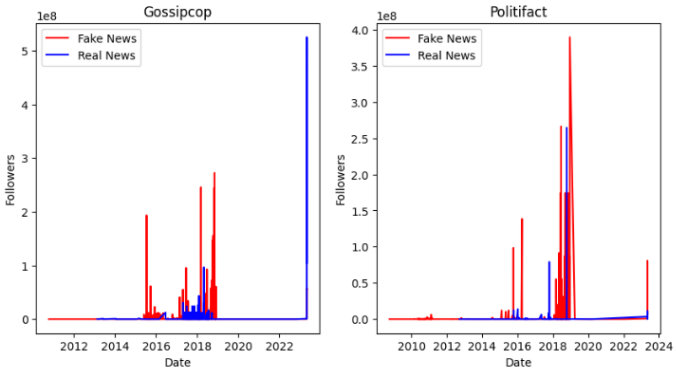

Fig. 9. Evolution of the followers in both datasets.

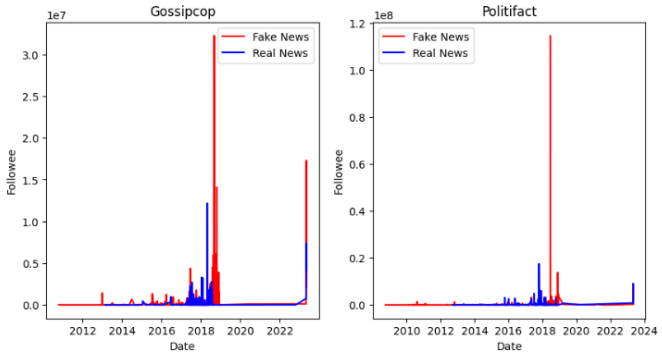

Fig. 10. Evolution of the followees in both datasets.

news tended to peak within the first few hours of its initial propagation. This pattern can be clearly seen in the results. A similar trend has been following since 2010.

Similar trends have been seen in both followers and followee of fake and real news. But in Fig.9 the number of followers of gossipcop real news is higher in recent years compared to fake news. The followers for both the Twitter pages increased from 2015 to 2019, mainly for the politifact page there have observed a peak in 2019. This coincides with the number of likes to the tweets from that page in the year 2019. The frequency of the number of followees over the year was been constant. But during 2018 and 2019, the number of followees increased for a short period, and later in recent years gossipcop page has observed an increase in the followees. Overall, the followers and the followees are more attracted to the fake news rather than the real news.

D. Task 4

To investigate the fake news generation through bots in the Twitter platform we are leveraging the botometer public library mentioned in section VI under subsection 4. This helps us to test the hypothesis of whether fake news globally originated from bots or humans and also to validate the probability of real news generation through humans. In this study, we considered two Twitter pages news namely Gossipcop and Politifact. Each

Dataset	Labelled as Bot	Labelled as Human
Politifact fake	1509	1491
Politifact real	1053	1947
Gossipcop fake	1407	3593
Gossipcop real	954	4046

Fig. 11. Botometer test.

dataset is divided into two different datasets namely, fake news and real news. As these datasets have more than 10000 rows, we have considered 3000 and 5000 rows or tweets from politifact and gossipcop datasets to reduce the computation.

From the Fig 11. we can observe in gossipcop dataset under real news data the model has been classified very less as a bot which can support our hypothesis regarding real news that most of the real news is generated by humans. Whereas under fake news datasets, our hypothesis is contradicting the botometer results.

In politifact dataset under the real news dataset, the botometer model again supports the hypothesis of real news whereas in the fake news dataset, the model has classified an equal number of human and bot-generated tweets. One of the major limitations is, this classification can be made clear if we use more data for the classification.

E. Task 7

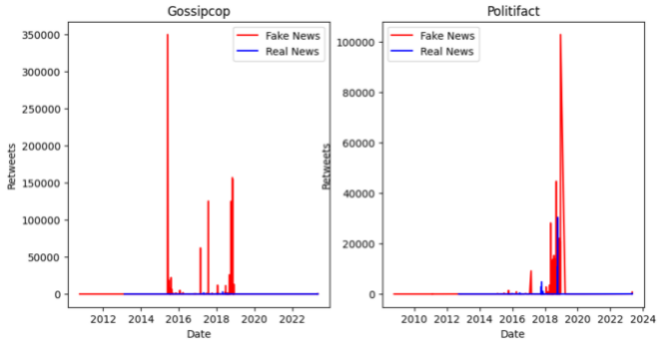

Fig. 12. Evolution of the retweets in both datasets.

Apart from checking the evolution of the likes, number of followers, and number of followees, we are interested to look into the virality of retweets in fake and real news. [24] described virality as the likelihood that a piece of content will be forwarded or shared to a large, yet unknown, audience.From the Fig 12. we can observe that the virality of the tweet in fake news is very frequent and observed for shorter time periods. Whereas the retweet frequency of the real news has been constant over the considered time. This correlates with the fact that the users are more intrigued or interested to follow fake news and have a high probability to share fake news than real news.

VIII. Limitations

The methodologies used in the research have a number of limitations. Descriptive statistics provide simply a brief description of the data and do not consider the underlying distribution or potential outliers. This can result in an oversimplification of the facts, which can lead to inaccurate conclusions and we might need inferential statistics before decision-making based on the statistics only [13]. We use a lot of visualizations in this project and for visualizations, few argue that if not correctly created, visualizations can be deceptive and subjective, and different types of representations might lead to different conclusions [12]. While bot identification is useful, it is not always precise and might result in false positives or negatives. Furthermore, bot behavior might evolve over time and differ depending on the platform [11]. Since our analysis only focuses on Twitter the application has limited scope. The quality and completeness of the data can limit network analysis, and it may not capture all significant links. Furthermore, global network features may not be the best technique to differentiate between fake and legitimate news [10]. Comparing the degree distributions of fake and actual news may not be sufficient to differentiate between the two groups, as the distributions can vary based on network topology and other factors [15]. Lastly, analyzing the temporal history of retweets might help us understand how information spreads, but it may not capture all essential elements, such as tweet content or user behavior [14].

IX. Future Studies

Several of the limitations can be addressed more effectively with alternative approaches. One approach could be to use additional statistical approaches to supplement descriptive statistics. Inferential statistics, such as hypothesis testing and confidence intervals, can be used to estimate the likelihood that observed differences between fake and legitimate news are real and not due to chance. While this is being used as a matrix for decision-making, the usage of many types of clear, labeled, visualizations of proper chart types accurately representing data could eliminate the issues with misleading representations. Furthermore, bot behavior may change between platforms, and a full investigation of social media sites might provide more insights into the problem of fake news. Combining network and content analysis can assist overcome the limitations imposed by using global network measurement as a single metric. The examination of degree distributions can be extended to the implementation of community detection techniques covered in the course. Along with the temporal evolution of user engagements, sentiment of the tweet and other metrics for user behavior studies can be coupled to provide a more comprehensive knowledge of fake news and its spread.

X. Conclusion

Using the `gossipcop` and `politifact` datasets, our examination of the FakeNewsNet dataset shown that it is difficult to identify fake news. But prior studies have indicated that user participation is a typical strategy for spotting bogus news. This assertion is supported by our research, which shows that fake news frequently receives more engagement than legitimate news. Furthermore, our examination of the temporal evolution of retweets for real and false news in the datasets from `gossipcop` and `politifact` showed that false news typically exhibits abrupt and transient spikes in engagement while true news exhibits more gradual and long-lasting periods of activity. These results are in line with earlier studies, like the one by [2], which discovered that false news stories on Twitter were `70%` more likely to be retweeted than true stories and that the spread of false news tended to peak within the first few hours of its initial propagation. Overall, our study contributes to the expanding body of knowledge on the identification of fake news and offers insights that can guide future research and the creation of efficient defenses against the spread of false information.

Acknowledgment

We would like to humbly express our deep appreciation to Mourad Oussalah for placing his trust in us by assigning this project and allowing us to apply our theoretical knowledge. We are also grateful to Saroar Jahan for his exceptional guidance and unwavering support throughout the course, particularly in navigating the challenges we faced while working with the Twitter API. Their contributions have been invaluable to the success of this project and our growth.

References

[1] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

[2] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. *doi: 10.1126/science.aap9559*.

[3] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

[4] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236.

[5] Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we RT?. In Proceedings of the First Workshop on Social Media Analytics (pp. 71-79).

[6] Satya, Prudhvi & Lee, Kyumin & Lee, Dongwon & Tran, Thanh & Zhang, Jason. (2016). Uncovering Fake Likers in Online Social Networks. 2365-2370. *doi: 10.1145/2983323.2983695*.

[7] Burkhardt JM. History of fake news. Libr. Technol. Rep. 2017;53(8):5–9.

[8] J. León-Quismondo, "Social Sensing and Individual Brands in Sports: Lessons Learned from English-Language Reactions on Twitter to Pau Gasol's Retirement Announcement," *International Journal of Environmental Research and Public Health*, vol. 20, no. 2, p. 895, Feb. 2023, doi: 10.3390/ijerph20020895.

[9] I. Ali, M. N. B. Ayub, P. Shivakumara, and N. F. B. Mohd Noor, "Fake News Detection Techniques on Social Media: A Survey," *Journal of Information Processing Systems*, vol. 16, no. 5, pp. 1033-1055, Oct. 2020, doi: 10.3745/JIPS.03.0143.

[10] B. Holthoefer, J., et al., "The spread of misinformation in social media," Proceedings of the National Academy of Sciences, vol. 113, no. 3, pp. 554-559, 2016.

[11] E. Ferrara, et al., "Bot detection in social media," IEEE Intelligent Systems, vol. 31, no. 2, pp. 76-81, 2016.

[12] S. Few, "Data visualization for human perception," in Proceedings of the International Conference on Advanced Visual Interfaces, 2012, pp. 1-6.

[13] A. Field, Discovering statistics using IBM SPSS statistics. Sage, 2013.
[14] J. Liu, et al., "The roles of online community engagement and emotion in the influence of social media recommendations," Computers in Human Behavior, vol. 88, pp. 75-81, 2018.
[15] Z. Wang, et al., "Statistical analysis of the network structure of Chinese online social networks," Physica A: Statistical Mechanics and its Applications, vol. 505, pp. 730-741, 2018.
[16] Figueira Á, Oliveira L. The current state of fake news: challenges and opportunities. Proc. Comput. Sci. 2017;121:817–825. doi: 10.1016/j.procs.2017.11.106.
[17] EC South Africa: Real411. Keeping it real in digital media. Disinformation Destroys Democracy (2019).
[18] Singh, R. (2020). Fake News Detection on Social Media: A Review. In Machine Learning Techniques for Analyzing and Detecting Fake News in Social Media (pp. 15-27). Springer, Cham.
[19] Statista. (2023). Twitter: number of monthly active users worldwide 2010-2023. Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.
[20] Pew Research Center. (2021). Social Media and Misinformation. Retrieved from https://www.pewresearch.org/politics/2021/06/23/social-media-and-misinformation-2021/.
[21] Bondielli A, Marcelloni F. A survey on fake news and rumour detection techniques. Inf. Sci. 2019;497:38–55. *doi: 10.1016/j.ins.2019.05.035*.
[22] Yang, K.-C., Ferrara, E., & Menczer, F. (2022). Botometer 101: social bot practicum for computational social scientists. Journal of Computational Social Science, 5, 1511-1528. https://doi.org/10.1007/s42001-022-00263-x
[23] Alhoori, H., Alnuaimi, O. A., Al-Badi, A. H., Al-Emadi, N. A., Almehrezi, M., & Albadi, N. (2020, December). A Multi-faceted Approach to Identifying Bots on Twitter during Emergencies. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 310-315). IEEE. https://doi.org/10.1109/ICICT51770.2020.00063
[24] Berger, J., Milkman, K. L. (2012). What makes online content viral?. Journal of marketing research, 49(2), 192-205.