Unraveling the Web of Fake News: An Analysis of Twitter's Information Landscape

Muhammad Hassan
23100199@lums.edu.pk

Aleem Malik
23100179@lums.edu.pk

Jawad Azhar
23100321@lums.edu.pk

Muneeb Munir
23110327@lums.edu.pk

I. ABSTRACT

The primary objective of this research paper is to identify the differences between fake and real news circulating on social media, primarily Twitter. We analyze the dataset by scrutinizing the language used in the tweets, examining engagement metrics, and identifying patterns. This paper clarifies the characteristics of fake news to give us a better idea of how social media influences public discourse.

Our dataset includes two CSV files, one containing Twitter data on real news while the other contains Twitter data on fake news. Each dataset includes 12 attributes that provide a range of information about each tweet and the user who posted it. We use the dataset to gain insight into the behaviour of Twitter users who share news content.

The methodology of the research paper involves several steps for preprocessing and analysis of the dataset, including frequent pattern mining and clustering. We also address the challenges involved in feature selection and data cleaning. We provide insights into the distinctive characteristics of fake news and its spread on Twitter through our analysis, which can be utilized to create strategies to limit the spread of false information and increase the exchange of legitimate data. In ultimately, our research attempts to improve our understanding of how social media influences public discourse and address the challenges caused by the spread of misleading information.

***Index Terms*—news, tweet**

II. INTRODUCTION

In recent years, the rapid expansion of social media platforms has resulted in an unprecedented increase in the circulation of both authentic and fabricated information. The ease with which people can differentiate between information globally has created a complex information landscape, making it challenging to differentiate between facts and falsehoods. This trend has significant ramifications for various domains, including journalism, politics, and public health.

The purpose of this research paper is to investigate the differences between fake and real news in the context of social media, with a specific focus on Twitter as a primary source of data. We will analyze the features of both types of news and examine the sources of fake news to gain a better understanding of the phenomenon.

Our analysis will involve analysing the language utilized in tweets, identifying patterns, and examining engagement metrics. We aim to shed light on the characteristics of fake news and its dissemination on social media to contribute to a better understanding of social media's role in shaping public discourse. We hope that by providing insights into the spread of misinformation and the features of fake news, we can develop strategies for combating the propagation of false information and promoting accurate information dissemination.

III. ABOUT THE DATASET

Our dataset consists of two csv files: "fake news" and "real news", both of which contain news tweets posted by various users. Each dataset includes 12 attributes:

These attributes provide a range of information about each tweet and the user who posted it. The "text" attribute gives the content of the tweet, while "followers" and "following" provide insight into the user's social network. The "verified" attribute indicates whether the user is a trusted source, and "location" can be used to identify where the news originated from. Additionally, the "statuses count" and "favorite count" attributes can be used to gauge the user's activity on Twitter. Overall, this dataset can be used to train machine learning models to distinguish between fake and real news, and to gain insights into the behavior of Twitter users who share news content.

A. Fake and real news

Our data sets include 2 csv files, fake news and real news, both of which contain news tweets posted by various users.

TABLE I
NUMBER OF ROWS IN DATASET

Category	Number of Rows
Fake News	645,528
Real News	995,432

B. Information about the datasets

The data provided includes various attributes related to a tweet posted by a user. The attributes are as follows: user id, which is a unique integer identifier for the user who posted the tweet; tweet id, which is the ID of the tweet sharing the news; text, which is the actual tweet shared by the user; followers, which represents the number of followers the user's account currently has; following, which represents the number of users the user's account follows; location, which is the location specified in the user's account profile; verified, a boolean value indicating whether the user's account is verified

or not; statuses count, which is the number of tweets (including retweets) issued by the user; profile background tile, which is a boolean value indicating whether the user's profile background image is tiled or not; profile use background image, which is a boolean value indicating whether the user's account has a profile background image or not; favorite count, which is the number of tweets the user has liked; and label, which is a binary variable indicating whether the tweet is real news (labeled 1) or fake news (labeled 0).

IV. Preprocessing techniques used

For both the Fake News and Real News datasets, some common preprocessing steps were followed. Firstly, boolean variables such as "verified", "profile background tile", and "profile use background image" were converted into 0 and 1. This was done to standardize the representation of these variables across the datasets.

V. Data Cleaning

Irrelevant variables such as User id and Tweet id were removed from both datasets. These variables did not contain any useful information for the analysis and hence were removed. Next, any row that contained all NULL values was removed from both datasets. This was done to ensure that the analysis was based on valid and complete data only.

A. Graphs

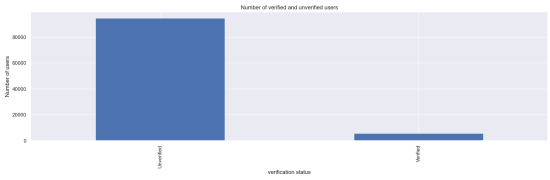

Fig. 1. Number of verified and unverified users (Fake News)

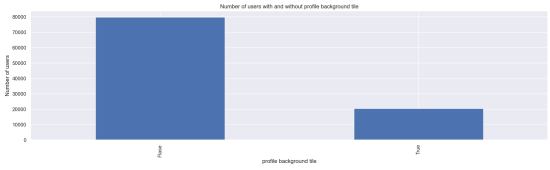

Fig. 2. Number of users with and without background tile (Fake News)

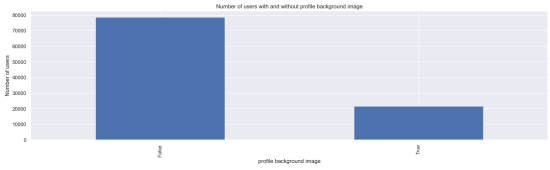

Fig. 3. Number of users with and without background image (Fake News)

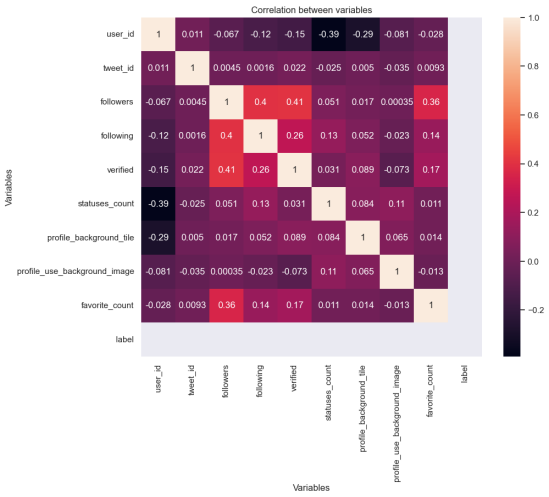

Fig. 4. Correlation between variables (Fake News)

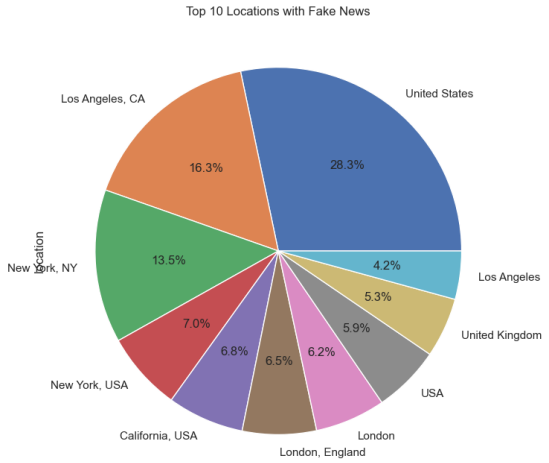

Fig. 5. Top 10 locations with Fake News

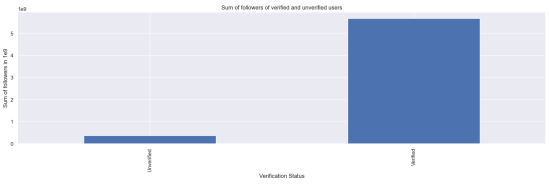

Fig. 6. Sum of followers of verified and unverified users (Fake News)

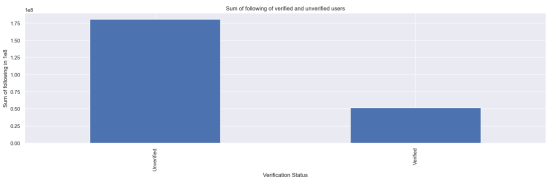

Fig. 7. Sum of following of verified and unverified users (Fake News)

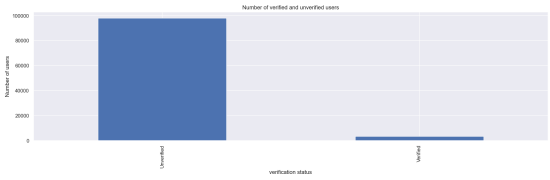

Fig. 8. Number of verified and unverified users (Real News)

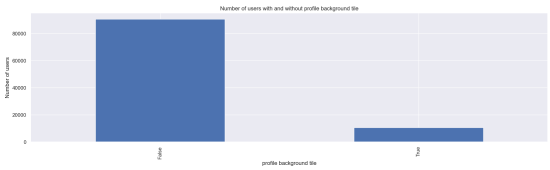

Fig. 9. Number of users with and without background tile (Real News)

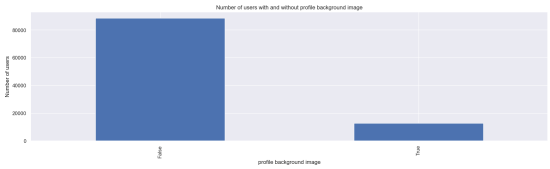

Fig. 10. Number of users with and without background image (Real News)

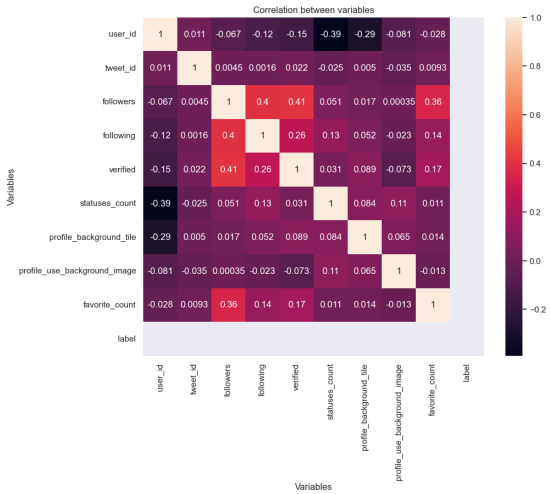

Fig. 11. Correlation between variables (Real News)

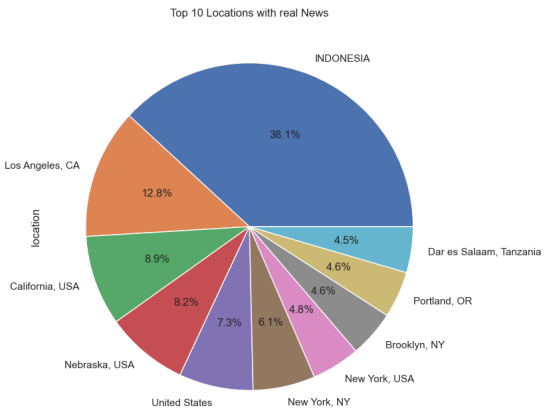

Fig. 12. Top 10 locations with Real News

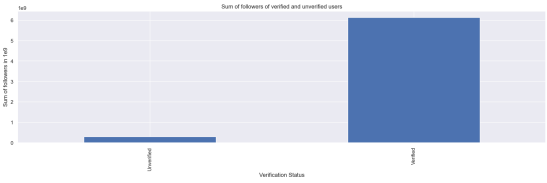

Fig. 13. Sum of followers of verified and unverified users (Real News)

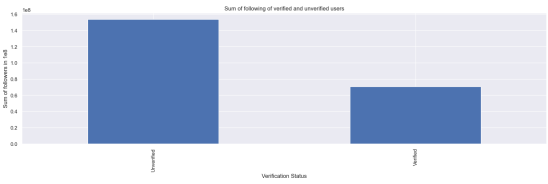

Fig. 14. Sum of following of verified and unverified users (Real News)

B. Analysis

According to the pie charts showing the top countries for Real and Fake news, we can see that the Fake news is mostly originating from USA and UK. Surprisingly, most of the Real news is being originated from Indonesia. USA and Tanzania are also among the top contributors of Real news
Another insight gained from the graphs is that the following of unverified users is more than the following of verified users. However, the sum of followers of verified users is greater than the sum of followers of unverified users.

VI. Methodology

The methodology of the research paper involved several steps for preprocessing and analysis of the dataset including Frequent Pattern Mining and Clustering.

Firstly, a vocabulary was generated from the text data, and a bag of words representation was created using this vocabulary. The text data was tokenized to split it into individual words for further processing. Stop words, which are common words with little significance in text analysis, were removed to reduce noise in the data.

Next, frequent pattern mining was performed using the FP-Growth algorithm, a popular method for mining frequent itemsets. Three variations of FP-Growth were used with different support thresholds of 0.01, 0.03, 0.05, and 0.07.

Association rules were then observed from these three variations, allowing for insights into the relationships between frequent patterns and their implications.

We then applied the Apriori algorithm. Similar to the previous approach, we generated a vocabulary and created a bag of words representation from the text data. Stop words were removed, and the text data was tokenized for further processing.

Next, we performed frequent pattern mining using the Apriori algorithm with various support thresholds. We repeated the tasks of generating association rules and observing patterns for different threshold values to compare the results. However, we encountered similar challenges in terms of time and memory consumption with the Apriori algorithm, especially for lower support thresholds.

As a next step, we explored clustering techniques. We started with the K-means algorithm. We used the elbow method to determine the optimal number of clusters. We ran the K-means algorithm on the number of clusters suggested by the elbow graph, specifically for the variables ”Followers” and ”Following”. We also applied this approach to another variation of the variables ”Followers” and ”Status Count”.

In addition to the above-mentioned approaches, we also explored the use of DBSCAN. The data in our dataset was found to be highly clustered so we used DBSCAN with specific hyperparameter values, including an epsilon value of 5000 and a minimum number of samples set to 100. These values were determined after thorough experimentation and analysis of the dataset.

The DBSCAN approach allowed us to identify dense regions of data points, which could potentially represent meaningful patterns or groups within the data. By carefully tuning the hyperparameter values, we aimed to capture clusters that exhibited significant density, while disregarding noisy or sparse regions.

VII. Issues

During the experimentation process, we encountered an issue with the 0.01 threshold variation of the FP-Growth algorithm. The execution of the algorithm at this threshold took an extensive amount of time and required a significant amount of RAM, resulting in the inability to complete the analysis. This issue may be attributed to the large size of the dataset or the complexity of the patterns being mined. As a result, it was not feasible to obtain meaningful results from this particular threshold variation.

VIII. Frequent Pattern Mining

Finding patterns or correlations within a dataset that occur frequently is known as frequent pattern mining in Data Mining. This is usually accomplished by studying datasets to identify things or groups of objects that recur frequently.

One of the most popular algorithms for frequent pattern mining is the apriori algorithm. It identifies frequently occurring itemsets using a ”bottom-up” method, and then produces association rules from those itemsets. As long as such item sets show in the database frequently enough, it moves forward by detecting the frequent individual items and expanding them to larger and larger item sets.

On the other side, the FP Growth algorithm identifies frequent patterns without candidate generation. Instead of employing Apriori’s generate and test technique, it constructs an FP Tree. The FP Growth algorithm concentrates on fragmenting the item routes and mining common patterns. [2]

A. Graphs

	support	itemsets
0	0.07	(wedding)
1	0.05	(royal)
2	0.03	(x)
3	0.12	(weinstein)
4	0.12	(harvey)
...	...	...
5440	0.05	(assault, allegati, sexual, charges, looking, ...
5441	0.05	(assault, allegati, face, sexual, looking, eve...
5442	0.05	(assault, allegati, face, sexual, charges, loo...
5443	0.05	(assault, face, sexual, charges, looking, even...
5444	0.05	(assault, allegati, face, sexual, charges, loo...

Fig. 15. FP growth with 0.03 support (Fake News)

	antecedents	consequents	confidence
0	(wedding)	(royal)	0.571429
1	(royal)	(wedding)	0.800000
2	(brad)	(weinstein)	0.384615
3	(weinstein)	(brad)	0.416667
4	(pitt)	(weinstein)	0.384615
...	...	...	...
386711	(even)	(assault, allegati, face, sexual, charges, loo...	0.833333
386712	(could)	(assault, allegati, face, sexual, charges, loo...	0.714286
386713	(harvey)	(assault, allegati, face, sexual, charges, loo...	0.416667
386714	(investigations)	(assault, allegati, face, sexual, charges, loo...	0.833333
386715	(weinstein)	(assault, allegati, face, sexual, charges, loo...	0.416667

Fig. 16. FP growth Association rules with threshold 0.25 (Fake News)

	support	itemsets
0	0.07	(wedding)
1	0.05	(royal)
2	0.12	(weinstein)
3	0.12	(harvey)
4	0.05	(think)
...	...	...
2565	0.05	(assault, allegati, face, sexual, looking, eve...
2566	0.05	(assault, allegati, sexual, charges, looking, ...
2567	0.05	(assault, allegati, face, sexual, charges, loo...
2568	0.05	(assault, face, sexual, charges, looking, even...
2569	0.05	(assault, allegati, face, sexual, charges, loo...

Fig. 17. FP growth with 0.05 support (Fake News)

	antecedents	consequents	confidence
0	(brad)	(weinstein)	0.384615
1	(weinstein)	(brad)	0.416667
2	(pitt)	(weinstein)	0.384615
3	(weinstein)	(pitt)	0.416667
4	(brad, pitt)	(weinstein)	0.384615
...	...	...	...
185161	(even)	(assault, allegati, face, sexual, charges, loo...	0.833333
185162	(could)	(assault, allegati, face, sexual, charges, loo...	0.714286
185163	(harvey)	(assault, allegati, face, sexual, charges, loo...	0.416667
185164	(investigations)	(assault, allegati, face, sexual, charges, loo...	0.833333
185165	(weinstein)	(assault, allegati, face, sexual, charges, loo...	0.416667

Fig. 18. FP growth Association rules with threshold 0.20 (Fake News)

	support	itemsets
0	0.07	(wedding)
1	0.12	(harvey)
2	0.12	(weinstein)
3	0.07	(amp)
4	0.13	(brad)
5	0.13	(pitt)
6	0.07	(youtube)
7	0.11	(angelina)
8	0.11	(jolie)
9	0.09	(kids)
10	0.08	(fighting)
11	0.07	(harry)
12	0.08	(sexual)
13	0.07	(could)
14	0.12	(harvey, weinstein)
15	0.13	(brad, pitt)
16	0.08	(pitt, angelina)
17	0.08	(brad, angelina)
18	0.08	(brad, pitt, angelina)
19	0.11	(angelina, jolie)
20	0.08	(pitt, jolie)
21	0.08	(brad, jolie)
22	0.08	(angelina, pitt, jolie)
23	0.08	(brad, angelina, jolie)

Fig. 19. FP growth with 0.07 support (Fake News)

	antecedents	consequents	confidence
0	(harvey)	(weinstein)	1.000000
1	(weinstein)	(harvey)	1.000000
2	(brad)	(pitt)	1.000000
3	(pitt)	(brad)	1.000000
4	(pitt)	(angelina)	0.615385
...	...	...	...
197	(weinstein, sexual)	(harvey)	1.000000
198	(harvey, weinstein)	(sexual)	0.666667
199	(sexual)	(harvey, weinstein)	1.000000
200	(harvey)	(weinstein, sexual)	0.666667
201	(weinstein)	(harvey, sexual)	0.666667

Fig. 20. FP growth Association rules with threshold 0.15 (Fake News)

	support	itemsets
0	0.03	(slams)
1	0.05	(husband)
2	0.03	(matthew)
3	0.07	(harry)
4	0.03	(day)
...	...	...
5440	0.05	(assault, allegati, face, sexual, charges, eve...
5441	0.05	(assault, face, sexual, charges, looking, even...
5442	0.05	(assault, allegati, sexual, charges, looking, ...
5443	0.03	(slams, wants, cop, simpsons, social, husband,...
5444	0.05	(assault, allegati, face, sexual, charges, loo...

Fig. 21. AP growth with 0.03 support (Fake News)

	antecedents	consequents	confidence
0	(slams)	(husband)	1.000000
1	(husband)	(slams)	0.600000
2	(slams)	(following)	1.000000
3	(following)	(slams)	1.000000
4	(slams)	(simpsons)	1.000000
...	...	...	...
386711	(even)	(assault, allegati, face, sexual, charges, loo...	0.833333
386712	(could)	(assault, allegati, face, sexual, charges, loo...	0.714286
386713	(harvey)	(assault, allegati, face, sexual, charges, loo...	0.416667
386714	(investigations)	(assault, allegati, face, sexual, charges, loo...	0.833333
386715	(weinstein)	(assault, allegati, face, sexual, charges, loo...	0.416667

Fig. 22. AP growth Association rules with threshold 0.25 (Fake News)

	support	itemsets
0	0.05	(husband)
1	0.07	(harry)
2	0.06	(face)
3	0.11	(jolie)
4	0.09	(kids)
...	...	...
2565	0.05	(assault, allegati, face, sexual, looking, eve...
2566	0.05	(assault, allegati, face, sexual, charges, eve...
2567	0.05	(assault, face, sexual, charges, looking, even...
2568	0.05	(assault, allegati, sexual, charges, looking, ...
2569	0.05	(assault, allegati, face, sexual, charges, loo...

Fig. 23. AP growth with 0.05 support (Fake News)

	antecedents	consequents	confidence
0	(harry)	(prince)	0.714286
1	(prince)	(harry)	1.000000
2	(harry)	(wedding)	0.714286
3	(wedding)	(harry)	0.714286
4	(harry)	(meghan)	0.714286
...	...	...	...
185161	(even)	(assault, allegati, face, sexual, charges, loo...	0.833333
185162	(could)	(assault, allegati, face, sexual, charges, loo...	0.714286
185163	(harvey)	(assault, allegati, face, sexual, charges, loo...	0.416667
185164	(investigations)	(assault, allegati, face, sexual, charges, loo...	0.833333
185165	(weinstein)	(assault, allegati, face, sexual, charges, loo...	0.416667

Fig. 24. AP growth Association rules with threshold 0.20 (Fake News)

	support	itemsets
0	0.07	(harry)
1	0.11	(jolie)
2	0.09	(kids)
3	0.13	(pitt)
4	0.12	(weinstein)
5	0.08	(sexual)
6	0.11	(angelina)
7	0.07	(youtube)
8	0.13	(brad)
9	0.12	(harvey)
10	0.07	(could)
11	0.08	(fighting)
12	0.07	(wedding)
13	0.07	(amp)
14	0.09	(kids, jolie)
15	0.08	(pitt, jolie)
16	0.11	(angelina, jolie)
17	0.08	(brad, jolie)
18	0.08	(fighting, jolie)
19	0.09	(kids, angelina)
20	0.08	(pitt, angelina)
21	0.13	(brad, pitt)
22	0.08	(fighting, pitt)
23	0.08	(sexual, weinstein)
24	0.12	(harvey, weinstein)
25	0.08	(harvey, sexual)
26	0.08	(brad, angelina)
27	0.08	(fighting, angelina)
28	0.08	(brad, fighting)

Fig. 25. AP growth with 0.07 support (Fake News)

	antecedents	consequents	confidence
0	(kids)	(jolie)	1.000000
1	(jolie)	(kids)	0.818182
2	(pitt)	(jolie)	0.615385
3	(jolie)	(pitt)	0.727273
4	(angelina)	(jolie)	1.000000
...	...	...	...
197	(brad)	(fighting, jolie, pitt, angelina)	0.615385
198	(angelina)	(brad, fighting, pitt, jolie)	0.727273
199	(fighting)	(brad, jolie, pitt, angelina)	1.000000
200	(pitt)	(brad, fighting, jolie, angelina)	0.615385
201	(jolie)	(brad, fighting, pitt, angelina)	0.727273

Fig. 26. AP growth Association rules with threshold 0.15 (Fake News)

	support	itemsets
0	0.10	(scares)
1	0.10	(little)
2	0.10	(help)
3	0.10	(bts)
4	0.10	(ellen)
...	...	...
10342	0.08	(one, video, best, kemper, ellie, making, kali...
10343	0.08	(one, video, best, mindy, kemper, ellie, makin...
10344	0.08	(one, video, best, mindy, kemper, ellie, makin...
10345	0.08	(one, video, best, mindy, kemper, ellie, kalin...
10346	0.08	(one, video, best, mindy, kemper, ellie, makin...

Fig. 27. FP growth with 0.03 support (Real News)

	antecedents	consequents	confidence
0	(slams)	(husband)	1.000000
1	(husband)	(slams)	0.600000
2	(slams)	(following)	1.000000
3	(following)	(slams)	1.000000
4	(slams)	(simpsons)	1.000000
...	...	...	...
386711	(even)	(assault, allegati, face, sexual, charges, loo...	0.833333
386712	(could)	(assault, allegati, face, sexual, charges, loo...	0.714286
386713	(harvey)	(assault, allegati, face, sexual, charges, loo...	0.416667
386714	(investigations)	(assault, allegati, face, sexual, charges, loo...	0.833333
386715	(weinstein)	(assault, allegati, face, sexual, charges, loo...	0.416667

Fig. 28. FP growth Association rules with threshold 0.25 (Real News)

	support	itemsets
0	0.10	(help)
1	0.10	(bts)
2	0.10	(degeneres)
3	0.10	(fangirl)
4	0.10	(ellen)
...	...	...
9220	0.08	(one, video, best, mindy, kemper, ellie, makin...
9221	0.08	(one, video, best, mindy, kemper, ellie, kalin...
9222	0.08	(one, video, best, mindy, ellie, making, kalin...
9223	0.08	(one, best, mindy, video, kemper, ellie, makin...
9224	0.08	(one, video, best, mindy, kemper, ellie, makin...

Fig. 29. FP growth with 0.05 support (real News)

	antecedents	consequents	confidence
0	(brad)	(weinstein)	0.384615
1	(weinstein)	(brad)	0.416667
2	(pitt)	(weinstein)	0.384615
3	(weinstein)	(pitt)	0.416667
4	(brad, pitt)	(weinstein)	0.384615
...	...	...	...
185161	(even)	(assault, allegati, face, sexual, charges, loo...	0.833333
185162	(could)	(assault, allegati, face, sexual, charges, loo...	0.714286
185163	(harvey)	(assault, allegati, face, sexual, charges, loo...	0.416667
185164	(investigations)	(assault, allegati, face, sexual, charges, loo...	0.833333
185165	(weinstein)	(assault, allegati, face, sexual, charges, loo...	0.416667

Fig. 30. FP growth Association rules with threshold 0.20 (Real News)

	support	itemsets
0	0.10	(bts)
1	0.10	(help)
2	0.10	(little)
3	0.10	(fangirl)
4	0.10	(ellen)
...	...	...
4881	0.08	(one, video, best, mindy, kemper, ellie, makin...
4882	0.08	(one, video, best, mindy, kemper, ellie, makin...
4883	0.08	(one, video, best, mindy, kemper, ellie, kalin...
4884	0.08	(one, video, best, mindy, ellie, making, kalin...
4885	0.08	(one, video, best, mindy, kemper, ellie, makin...

Fig. 31. FP growth with 0.07 support (real News)

	antecedents	consequents	confidence
0	(brad)	(weinstein)	0.384615
1	(weinstein)	(brad)	0.416667
2	(pitt)	(weinstein)	0.384615
3	(weinstein)	(pitt)	0.416667
4	(brad, pitt)	(weinstein)	0.384615
...	...	...	...
185161	(even)	(assault, allegati, face, sexual, charges, loo...	0.833333
185162	(could)	(assault, allegati, face, sexual, charges, loo...	0.714286
185163	(harvey)	(assault, allegati, face, sexual, charges, loo...	0.416667
185164	(investigations)	(assault, allegati, face, sexual, charges, loo...	0.833333
185165	(weinstein)	(assault, allegati, face, sexual, charges, loo...	0.416667

Fig. 32. FP growth Association rules with threshold 0.15 (Real News)

	support	itemsets
0	0.07	(trying)
1	0.19	(abuse)
2	0.03	(queen)
3	0.06	(royal)
4	0.13	(day)
...	...	...
10342	0.08	(one, best, mindy, video, kemper, making, kali...
10343	0.08	(one, video, best, mindy, kemper, ellie, makin...
10344	0.08	(one, best, mindy, video, kemper, ellie, makin...
10345	0.05	(something, trying, nikki, bella, excited, tot...
10346	0.08	(one, video, best, mindy, kemper, ellie, makin...

Fig. 33. AP growth with 0.03 support (Real News)

	antecedents	consequents	confidence
0	(trying)	(thats)	1.000000
1	(thats)	(trying)	1.000000
2	(trying)	(total)	1.000000
3	(total)	(trying)	1.000000
4	(trying)	(nikki)	1.000000
...	...	...	...
1113587	(kaling)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
1113588	(life)	(one, video, best, mindy, kemper, ellie, makin...	1.000000
1113589	(said)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
1113590	(office)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
1113591	(days)	(one, video, best, mindy, kemper, ellie, makin...	0.888889

Fig. 34. AP growth Association rules with threshold 0.25 (Real News)

	support	itemsets
0	0.07	(trying)
1	0.19	(abuse)
2	0.06	(royal)
3	0.13	(day)
4	0.06	(like)
...	...	...
9220	0.08	(one, best, mindy, video, kemper, making, kali...
9221	0.08	(one, video, best, mindy, kemper, ellie, makin...
9222	0.08	(one, best, mindy, video, kemper, ellie, makin...
9223	0.05	(something, trying, nikki, bella, excited, tot...
9224	0.08	(one, video, best, mindy, kemper, ellie, makin...

Fig. 35. AP growth with 0.05 support (Real News)

	antecedents	consequents	confidence
0	(trying)	(thats)	1.000000
1	(thats)	(trying)	1.000000
2	(trying)	(total)	1.000000
3	(total)	(trying)	1.000000
4	(trying)	(nikki)	1.000000
...	...	...	...
1075119	(kaling)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
1075120	(life)	(one, video, best, mindy, kemper, ellie, makin...	1.000000
1075121	(said)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
1075122	(office)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
1075123	(days)	(one, video, best, mindy, kemper, ellie, makin...	0.888889

Fig. 36. AP growth Association rules with threshold 0.20 (Real News)

	support	itemsets
0	0.07	(trying)
1	0.19	(abuse)
2	0.13	(day)
3	0.07	(thats)
4	0.09	(kaling)
...	...	...
6917	0.09	(one, video, best, mindy, kemper, ellie, makin...
6918	0.08	(one, best, mindy, video, kemper, making, kali...
6919	0.08	(one, video, best, mindy, kemper, ellie, makin...
6920	0.08	(one, best, mindy, video, kemper, ellie, makin...
6921	0.08	(one, video, best, mindy, kemper, ellie, makin...

Fig. 37. AP growth with 0.07 support (Real News)

	antecedents	consequents	confidence
0	(trying)	(thats)	1.000000
1	(thats)	(trying)	1.000000
2	(trying)	(total)	1.000000
3	(total)	(trying)	1.000000
4	(trying)	(nikki)	1.000000
...	...	...	...
718871	(kaling)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
718872	(life)	(one, video, best, mindy, kemper, ellie, makin...	1.000000
718873	(said)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
718874	(office)	(one, video, best, mindy, kemper, ellie, makin...	0.888889
718875	(days)	(one, video, best, mindy, kemper, ellie, makin...	0.888889

Fig. 38. AP growth Association rules with threshold 0.15 (Real News)

IX. Clustering

In Data Mining Clustering is the process of making a group of abstract objects into classes of similar objects.

When performing cluster analysis, the data set is first divided into groups depending on how similar the data are, and then the groups are given labels. [1]

A. Graphs

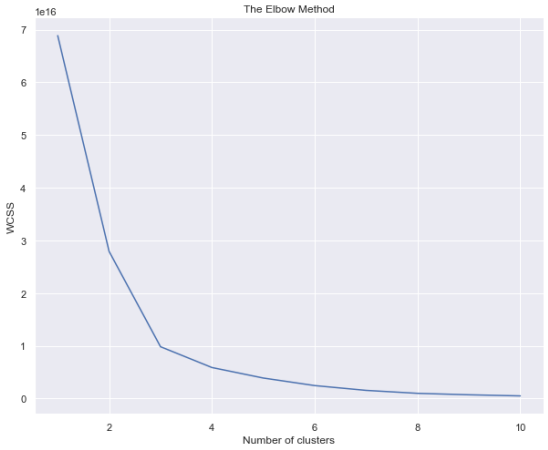

Fig. 39. (Elbow Method - Following (Fake News)

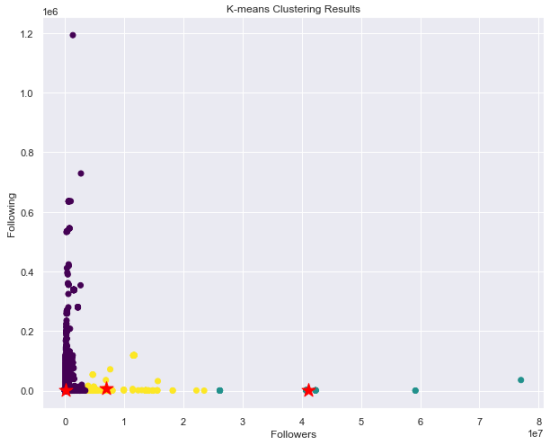

Fig. 40. (Kmeans - Following (Fake News)

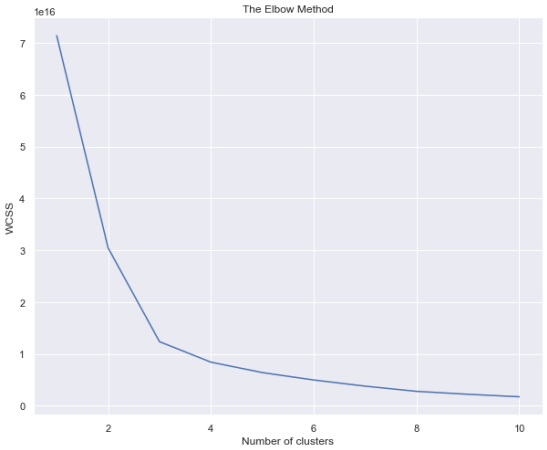

Fig. 41. (Elbow Method - Status (Fake News)

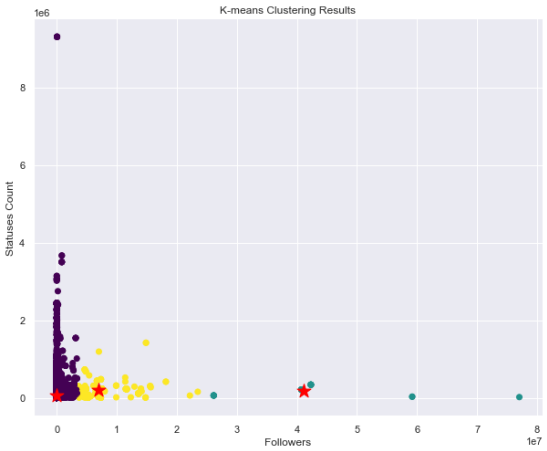

Fig. 42. (Kmeans - Status (Fake News)

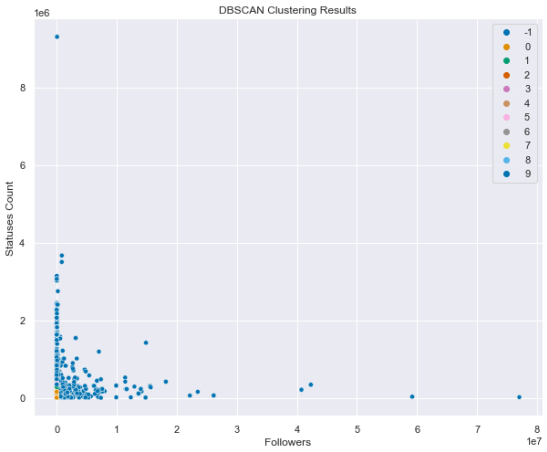

Fig. 43. (DB Scan - Status (Fake News)

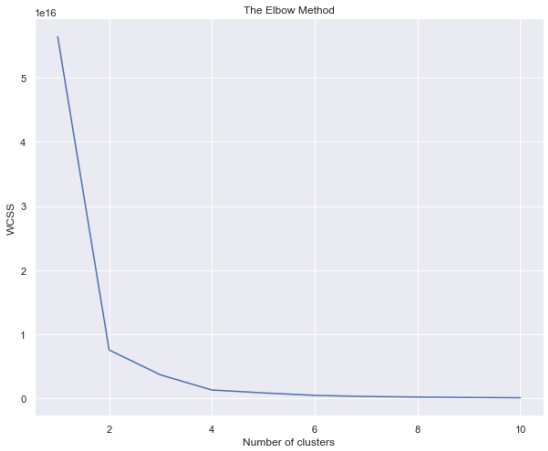

Fig. 44. (Elbow Method - Following (Real News)

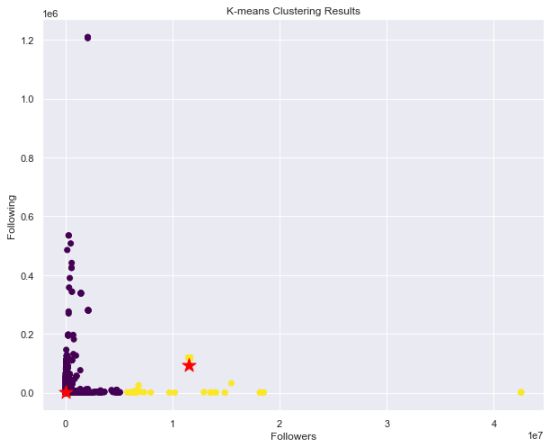

Fig. 45. (Kmeans - Following (Real News)

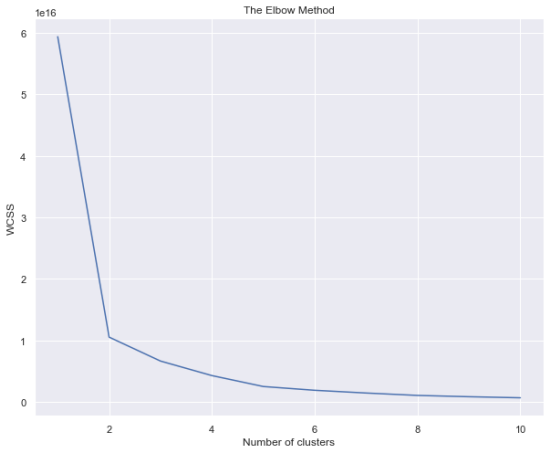

Fig. 46. (Elbow Method - Status (Real News)

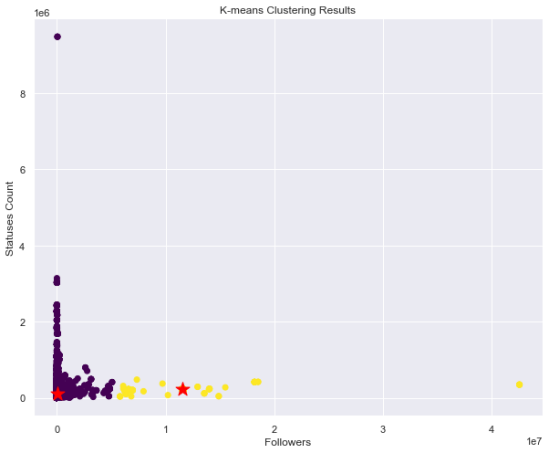

Fig. 47. (Kmeans - Status (Real News)

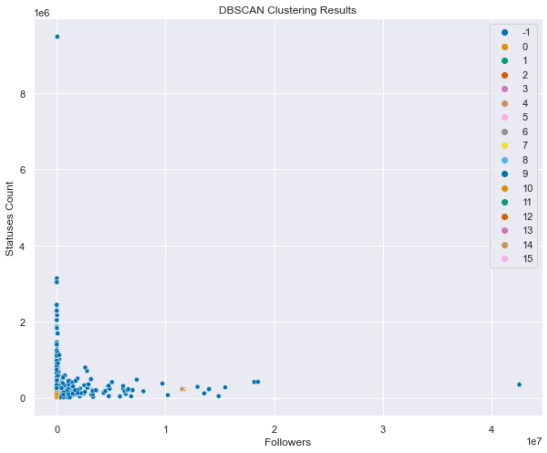

Fig. 48. (DB Scan - Status (Real News)

X. Results

Upon analyzing the obtained results, it was observed that the frequent itemsets formed by fp-growth algorithm suggest that fake news in the dataset contains itemsets such as (assault, allegati, sexual, charges, looking ...), (assault, face, sexual, charges, looking, even...), and (fighting, angelina, pitt, jolie), while the single itemsets include (weinstein), (royal), (wedding), (brad), (angelina), and (kids). This indicates that the fake news primarily revolves around topics such as Weinstein and sexual charges, royal weddings, and celebrities like Brad Pitt and Angelina Jolie. The results were obtained with a minimum support of 0.03, which yielded 5445 itemsets, 2570 itemsets with 0.05 support, and 47 itemsets with 0.07 support.

Similar patterns were also observed in real news, where the frequent itemsets generated by fp-growth algorithm included (one, video, best, mindy, kemper, ellie, makin..), (something, trying, nikki, bella, excited, tot...), and (one, video, best, mindy, kemper, ellie, makin..), and the single itemsets included (trying), (abuse), (day), (that's), (kailing), (queen), and (royal). Again, using a minimum support of 0.03 resulted in 10347 itemsets, 9225 itemsets with 0.05 support, and 4886 itemsets with 0.07 support. Notably, Apriori algorithm also yielded similar results for both real news and fake news.

In addition to using frequent itemsets, k-means clustering was applied to the dataset to analyze the distribution of fake news and real news based on variables such as "followers" and "status count". For fake news, we found the optimal number of clusters to be 3 using the elbow method and thus the dataset was divided into 3 clusters using k-means based on the variables of "followers" and "status count", as well as "followers" and "following". On the other hand, for real news, the elbow method determined the optimal number of clusters to be 2 for both "followers" and "following", and "followers" and "status count", as shown in the graphs. Additionally, DBSCAN was applied, which resulted in 10 clusters and 7570 outliers for fake news, and 16 clusters with 4992 outliers for real news.

Furthermore, the analysis revealed that the majority of real news originated from Indonesia and Tanzania, while most of the fake news came from the United Kingdom and the United States, particularly the state of California. Hence, it is advised that individuals in these countries should exercise caution when encountering news, as fake news can have detrimental effects and contribute to the spread of hate and/or terror.

XI. Suggestions

To deal with similar datasets, We must first remove any nan rows and columns. Similar to this dataset, any new dataset might also have some redundant columns, such as ids which do not provide any meaningful insights and thus should be removed from the dataset. Then we need to convert the categorical variables to numeric representations, as the algorithms used in the paper require data to be in numeric form. In this dataset, we observed that the text and locations needed to be cleaned as special characters, encodings and emojis were present in the dataset which adds noise to the dataset and negatively biases the patterns. Similar datasets might have text in a similar format and thus require similar cleaning. The code used by us can be used as a reference to clean the text. Secondly, for locations, most locations referred to the same area, such as Los Angeles and Los Angeles, CA, are the same location and thus need to be merged. There might also be some redundant locations in the dataset that need to be actively dealt with. The cities can be clustered into Countries to gain a holistic view of the country and fake news. One pitfall that needs to be mentioned is that our algorithm crashed for certain thresholds due to memory constraints. Such might be the case with similar datasets as well. To deal with these problems, we can either sample the datasets and draw patterns from the sampled datasets, or we can choose higher thresholds to avoid running into memory constraints. To evaluate patterns in the dataset, carefully consider the thresholds and draw conclusions. These methods are sufficient to draw out conclusions from the dataset, however, to gain more insights, you can use different algorithms according to the constraints of the dataset. For clustering, you can also implement hierarchical clustering and agglomerative clustering techniques.

This paper does not discuss the classification of real and fake news, but this could also be implemented when dealing with real and fake news. The trained model will take any news as an input and output whether or not this is fake news. Our choice of classifier would be Naive Bayes, which works on assigning a probability to the words. We could also implement a logistic regression model or a neural network-based model as well and chose the best-performing model as our classifier.

XII. conclusion

In conclusion, this paper comprehensively analysed the Twitter fake and real news dataset, employing various techniques such as exploratory data analysis, frequent pattern mining using FP Growth and Apriori algorithms, and clustering using Kmeans and DBscan. The identified common patterns in fake news, including the royal wedding, sexual allegations against Weinstein, and news related to Brad Pitt and Angelina, shed light on the prevalent topics in false information. Additionally, the findings revealed that Los Angeles, California, was a primary source of fake news, while real news was predominantly generated from Indonesia. The insights generated from this paper will help Twitter users and people, in general, differentiate between fake and real news and stay safe from the spread of fake, inaccurate and fabricated information being spread over the internet.

References

[1] Faizan, M., Zuhairi, M. F., Ismail, S., and Sultan, S. Applications of clustering techniques in data mining: A comparative study. *International Journal of Advanced Computer Science and Applications 11*, 12 (2020).

[2] Nasreen, S., Azam, M. A., Shehzad, K., Naeem, U., and Ghazanfar, M. A. Frequent pattern mining algorithms for finding associated frequent patterns for data streams: A survey. *Procedia Computer Science 37* (2014), 109–116. The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)/ The 4th

International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2014)/ Affiliated Workshops.