WEB SOCIAL MEDIA ANALYTICS AND VISUALISATION (CMP7202) REPORT

Module Coordinator
Ogerta Elezaj

Submitted By
Meera Mohan Mullamkuzhy (22185279)

MSc Big Data Analytics

Faculty of Computing, Engineering and the Built Environment

5 May 2023

Contents

1 INTRODUCTION

Social media analytics is a way to collect and interpret data from social media channels to support business choices and evaluate the effectiveness of activities taken because of those decisions. The scope of social media analytics goes beyond channel-specific data like following, likes, retweets, previews, clicks, and impressions. Social listening is an element of social media analytics (IBM, 2022).

This report discusses about the statistical tests and text mining done on Social Media data. Trend analysis and sentiment analysis were carried out on scrapped Twitter datasets and graph analysis was done on Facebook ego network. Topic modelling was performed scrapped news articles.

2 STATISTICAL ANALYSIS

This section focuses on the statistical analysis of a trend from Twitter data and a graph dataset from Facebook.

2.1 TREND ANALYSIS USING TWITTER DATA

ChatGPT is a natural language processing tool powered by AI. With the chat-bot, you can communicate in a way that is similar to that of a real being. The language model can provide answers to your questions (Ortiz, 2023). In one month, there were 91 new postings on ChatGPT from dark web forums, up from 37. It is possible to use ChatGPT as a criminal tool (Read, 2023). Both pro and con opinions can be found on ChatGPT and it is on of prominent topics of discussion in the world right now. Statistical analysis will be performed on the tweets regarding the topic 'Chat GPT'.

2.1.1 DATA COLLECTION

'snscrape' library was used scrape data from Twitter. The dataset contains 4001 instances of the selected topic.

2.1.2 EXAMINATION OF TWITTER PLATFORMS

The proportion of tweets from each platform is displayed in Figure 1. Web applications were used by most users (34.4%). According to Table 1, the top 4 sources account for 88.5 percent of all tweet volume.

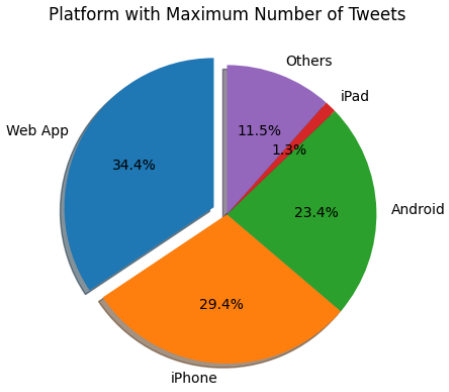

Figure 1 : Percentage of tweets from each platform

Table 1 : Tweet counts from each platform

Sources	Tweet Count
Web App	1376
iPhone	1175
Android	937
iPad	51

2.1.3 SOURCE RELIABILITY CHECK

With millions of users worldwide, Twitter has emerged as a major source for online news consumption. However, it has evolved into a key medium for disinformation dissemination with serious societal repercussions. Users of social media have been seen to spread misinformation more quickly than reliable information (Vosoughi, Roy and Aral, 2018). So, the reliability of the tweets can never be guaranteed.

A donut representation of percentages of tweets and retweets are given in the Figure 2. Obviously, retweets are more than the tweets in the dataset.

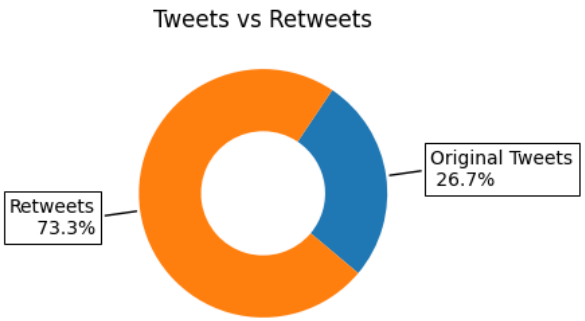

Figure 2: Percentage of tweets and retweets

On Twitter, only 20% of the content was actual information. Furthermore, a significant proportion of users with solid online reputations and verified accounts were in charge of disseminating false information. The fake tweets grew once high-reach individuals reposted the fake tweets, which made them become viral (Cohen, 2013).

Neither the verified accounts nor the number of retweets can guarantee the accuracy of the information. The comparison of verified and non-verified users could not be performed due to the current limitation of data scrapping tool.

2.1.4 EVALUATING CRUCIAL HASHTAGS

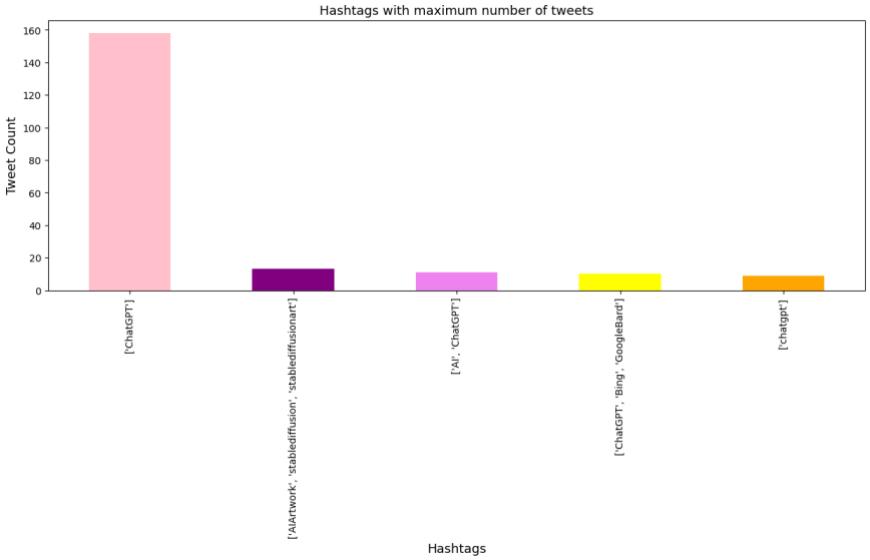

Figure 3 : Hashtags with highest number of tweets

Hashtags allow people to quickly follow the topics that interest them. It expands the audience and visibility.

From Figure 3, the hashtags used for improved reach are ‘ChatGPT’, ‘AIArtwork’, ‘stablediffusion’, and ‘AI’. ‘ChatGPT’ was used by 158 users and became the most widely used hashtag in the dataset.

2.1.5 ASSESSING THE MOST ENGAGED USERS

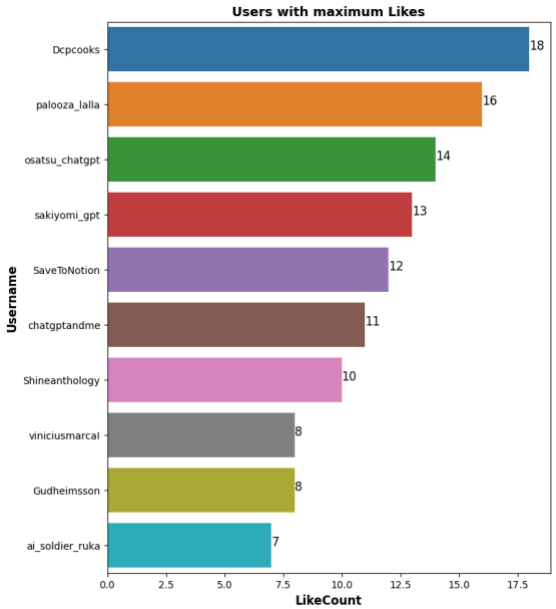

Figure 4 : Users with maximum number of likes

The users with highest number of likes for their tweets were found and visualized in the bar plot Figure 4. The account ‘Dcpcooks’ tweeted the most about ‘Chat GPT’.

2.1.6 FINDING THE MOST COMMONLY USED LANGUAGE

Twitter supports 34 languages (developer.twitter.com, n.d.). Users can tweet in any language supported by Twitter.

Figure 5 shows a Treemap for the languages used in the dataset. 'English' was the most commonly used language, followed by Japanese.

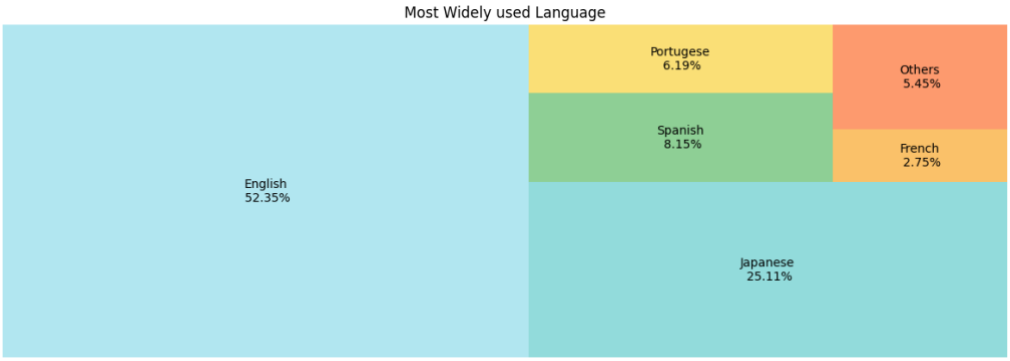

Figure 5 : Languages used by the users.

2.1.7 LOCATION ANALYSIS

Figure 6 is a heat map created after extracting the city names from the location of each tweet using 'Folium' library.

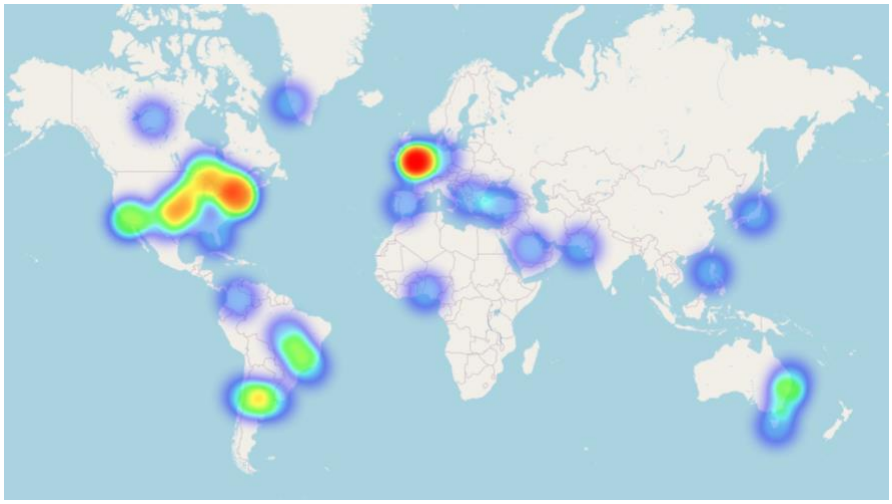

Figure 6 : Location heatmap.

Given that ChatGPT, a corporation owned by OpenAI, is based in San Francisco, a North American city, it seems obvious that more tweets are coming from this continent. The United Kindom is another significant region with a lot of tweets. The majority language in the UK is English, and ChatGPT is an English language model. The likelihood that someone in the UK may utilise ChatGPT for conversations on social media sites like Twitter may therefore be higher.

2.2 GRAPH ANALYSIS

Facebook ego dataset from Stanford Large Network Dataset Collection was used for graph analysis. This dataset is based on the 'friends lists' in Facebook. The nodes represent the users, and the edges represent the connections between them. The statistical analysis of the Facebook graph is given Table 2.

Table 2 : Statistics of Facebook Ego Network Graph

Number of Nodes	4039
Number of Edges	88234
Density	0.01082
Shortest Path	3.693
Diameter	8
Mean	43.691

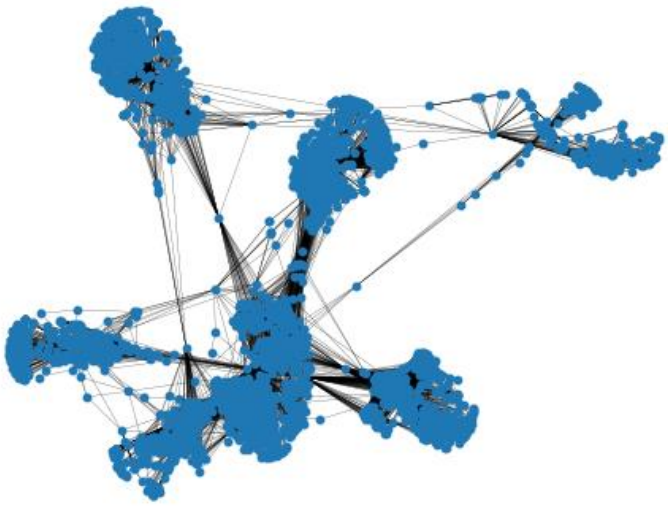

Figure 7 : Visualisation of Facebook Ego Network Graph

Figure 7 shows the visualization of Facebook Ego Network graph.

Considering the computational cost, 107.edges, a subset of Facebook ego-network, was used for further analysis. Table 3 and Figure 3 show the statistics of 107.edges Facebook graph and the visualisation of the graph, respectively.

Table 3 : Statistics of 107.edges subgraph

Number of Nodes	1034
Number of Edges	26749
Density	0.0501
Shortest Path	2.952
Diameter	9
Mean	51.739

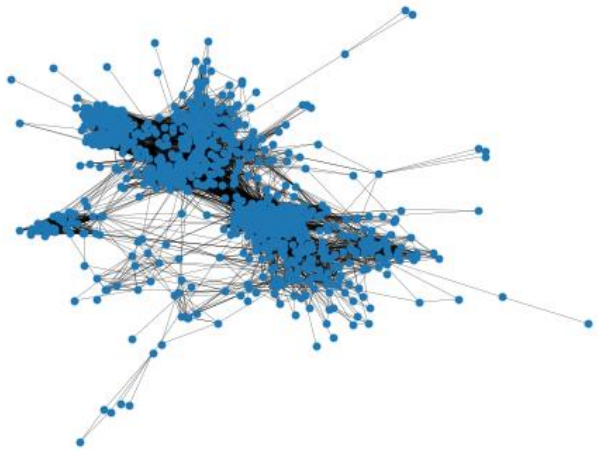

Figure 8 : Visualisation of 107.edges subgraph

2.2.1 BETWEENNESS CENTRALITY

The measure of a node's influence over a graph's information flow is called betweenness centrality. Higher betweenness centrality ratings will be achieved by nodes that regularly lie on the shortest paths between other nodes.

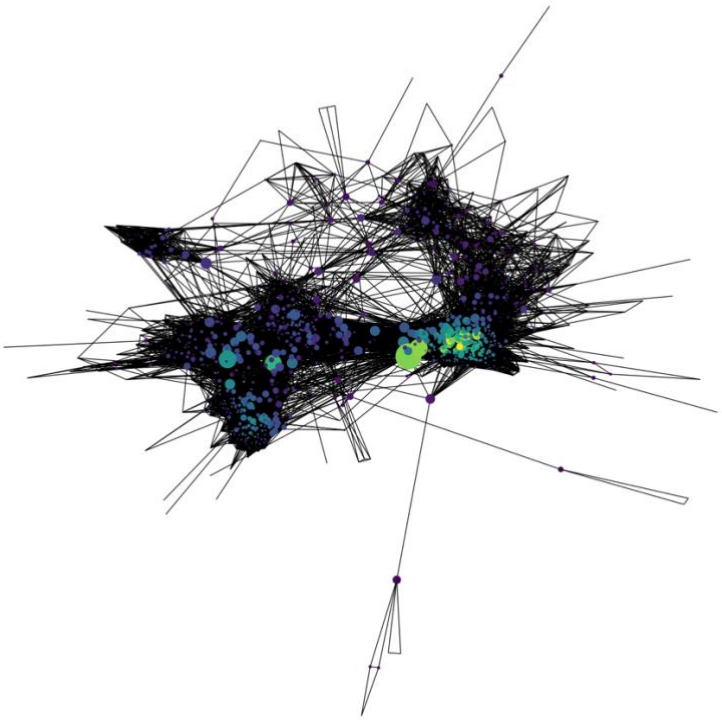

Figure 9 : Visualisation of Betweenness Centrality of 107.edges subgraph

Figure 9 shows the betweenness centrality visualization of 107.edges subgraph. The nodes with

lighter color and larger size represents nodes with higher centrality. Table 4 lists the top five nodes with high betweenness centrality score.

Table 4 : Top five nodes with high betweenness centrality

Rank	Node	Centrality Values
1	1086	0.106557
2	1584	0.050252
3	917	0.048953
4	483	0.042573
5	1334	0.018894

2.2.2 DEGREE CENTRALITY

A node's degree centrality in a graph is just a count of the edges that connect to it. Higher the number of connections, higher will be the centrality score (Powell and Hopkins, 2015).

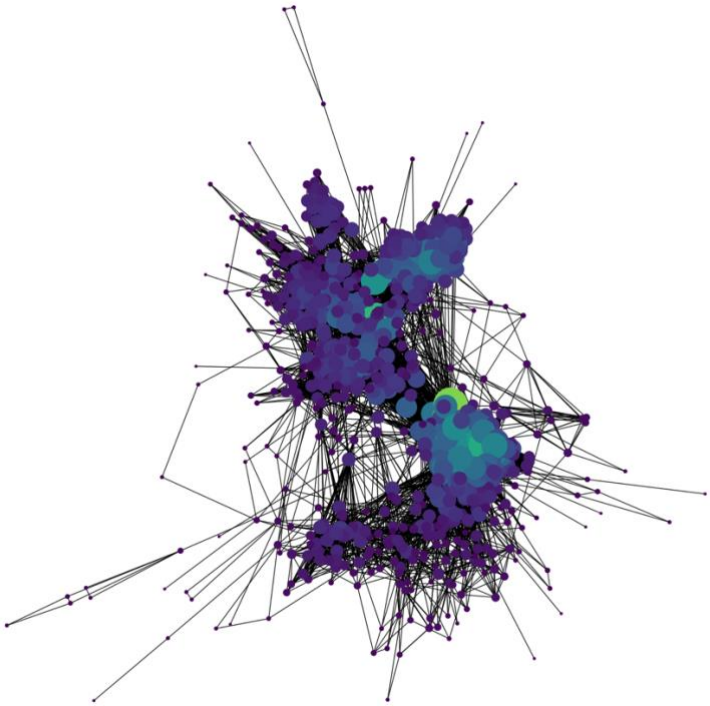

Figure 10 : Visualisation of Degree Centrality of 107.edges subgraph

Figure 10 shows the degree centrality visualization of 107.edges subgraph. The nodes with lighter color and larger size represents nodes with higher centrality. Table 5 lists the top five nodes with high degree centrality score.

Table 5 : Top five nodes with high degree centrality

Rank	Node	Centrality Values
1	1888	0.244918
2	1800	0.236205
3	1663	0.226547
4	1352	0.225557
5	1730	0.217812

2.2.3 EIGENVECTOR CENTRALITY

An approach that measures the transitive effect of nodes is called eigenvector centrality. A node that has a high eigenvector score is linked to numerous other nodes that also have high scores (Neo4j Graph Data Platform, n.d.).

Figure 11 shows the eigenvector centrality visualization of 107.edges subgraph. The nodes with lighter color and larger size represents nodes with higher centrality. Table 6 lists the top five nodes with high eigenvector centrality measure.

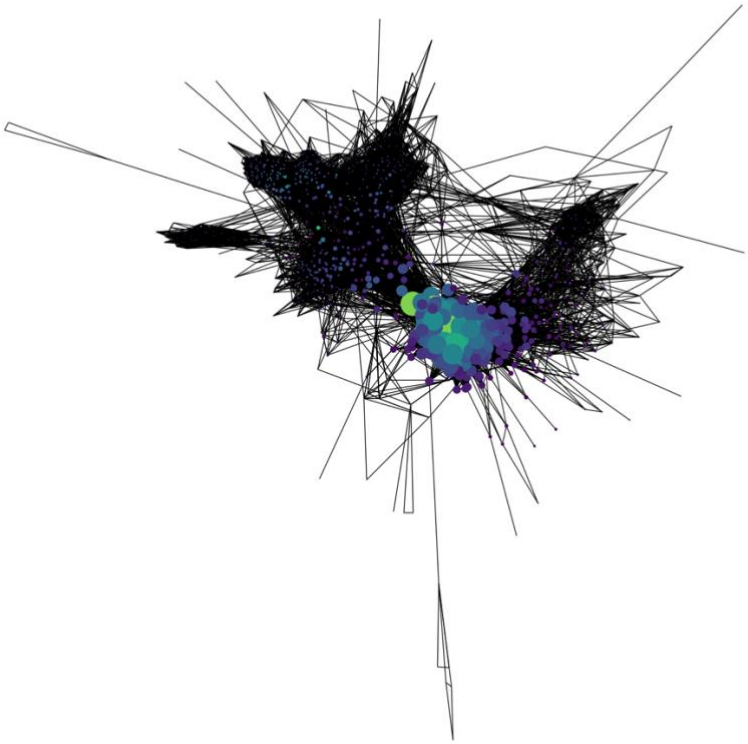

Figure 11 : Visualisation of Eigenvector Centrality of 107.edges subgraph

Table 6 : Top five nodes with high eigenvector centrality

Rank	Node	Centrality Values
1	1888	0.116633
2	1800	0.112373
3	1663	0.109571
4	1352	0.108547
5	1431	0.104803

2.2.4 COMMUNITY DETECTION

Louvain community detection method, a hierarchical clustering algorithm, was used to identify 8 communities with a modularity score of 0.5396. If the modularity is positive, the entities are put in a community where the relationships are denser than they would be in a random network.

Figure 12 and Figure 13 depict the graph's discovered communities in a spiral and spring layout, respectively.

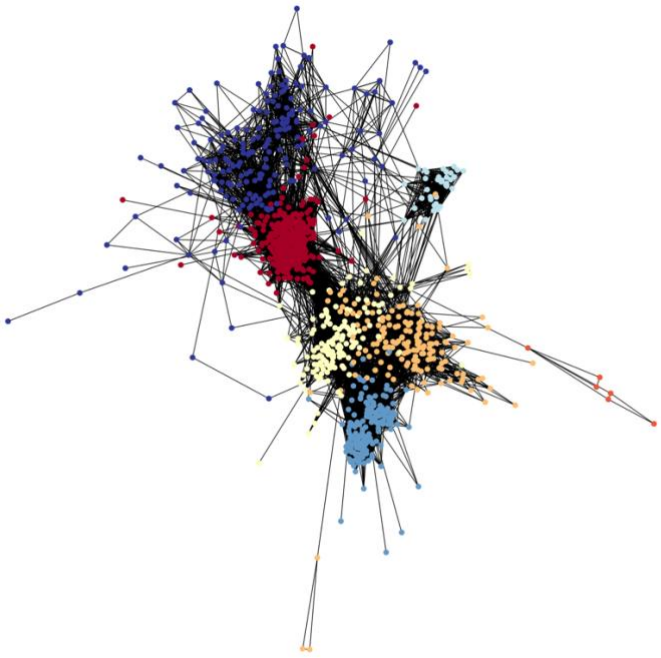

Figure 12 : Spring layout representation of detected communities

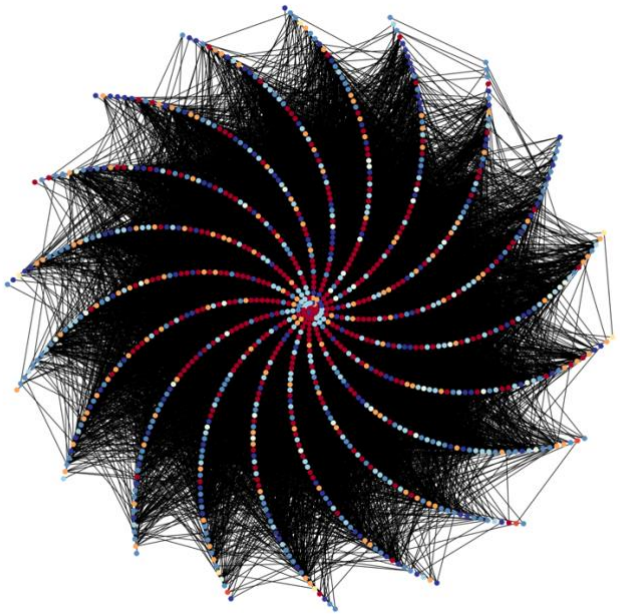

Figure 13 : Spiral layout representation of detected communities

3 TEXT MINING

3.1 SENTIMENT ANALYSIS

Sentiment analysis was performed on the topic '**Pride Community Acceptance'**.

3.1.1 DATA COLLECTION

The data for the analysis was scrapped from Twitter using 'snscrape' library. The dataset contains 5001 instances of the selected topic.

3.1.2 SENTIMENTS ANALYSIS

Sentiments, emotions, and sarcasm of each tweet were identified using different Python libraries.

Figure 14 shows the sentiments counts of the total tweets. It is obvious that majority of the people have 'positive' attitude towards the pride community.

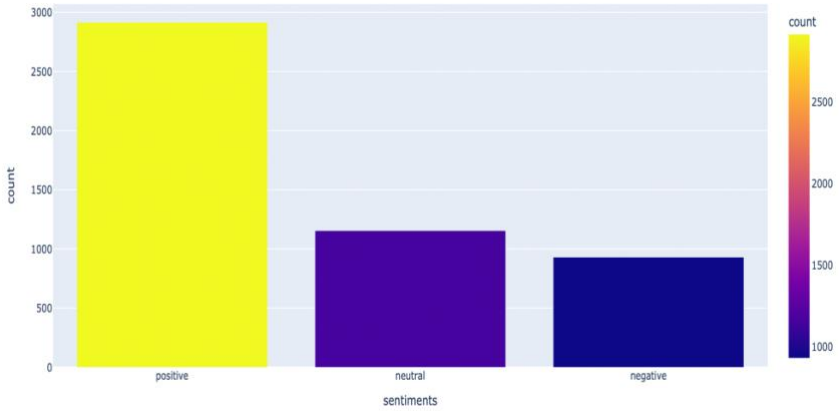

Figure 14 : Sentiments count of total tweets.

A further study on the tweets in Figure 15, shows that the neutral comments are objective, as they are typically factual and informative.

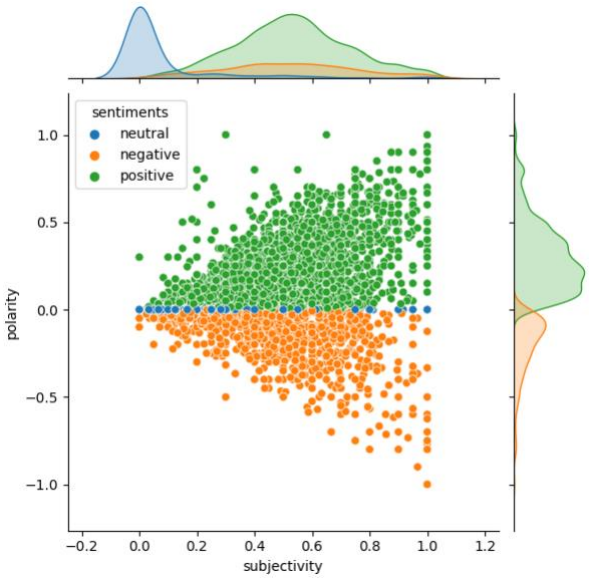

Figure 15 : Polarity vs Subjectivity

To provide an improved comprehension of the sentiment, emotion analysis can be utilised in conjunction with sentiment analysis. The AFINN-165 lexicon is put into a Pandas DataFrame and transformed into an emotional word-to-emotion dictionary. The sentence's overall emotion score is calculated by adding the scores of each word. Finally, based on the overall score, the emotion score is categorised as positive, negative, or neutral.

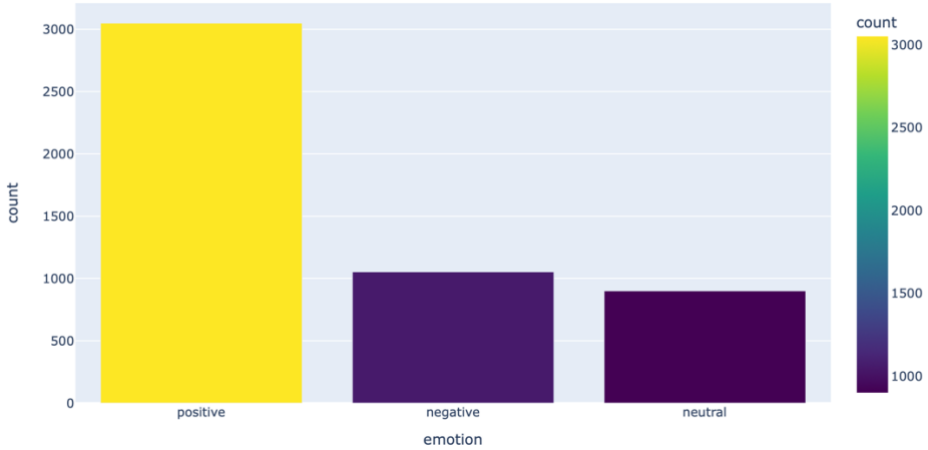

Figure 16 : Emotions count of the total tweets.

From Figure 16, it is evident that there is a positive emotion towards the 'Pride Community'. As mentioned above for sentiment analysis, neutral tweets are also objective in emotional analysis. Figure 16 illustrates the relationship between emotion score and subjectivity.

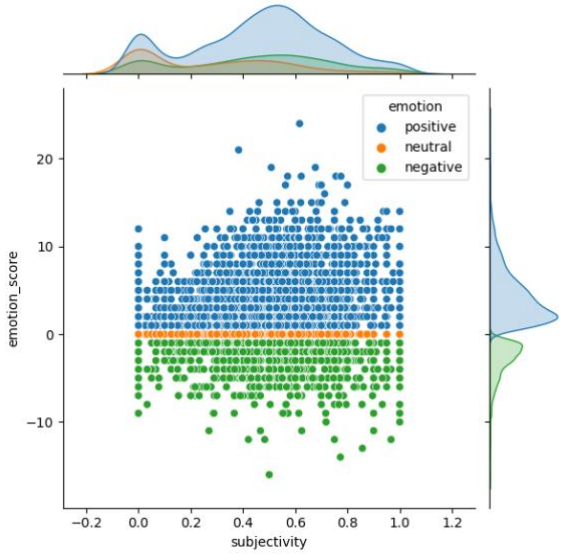

Figure 17: Emotion score vs subjectivity

Sarcastic tweets made up a very small percentage (6.02%) of all tweets. The remaining tweets were straightforward. The proportion of sarcastic tweets among all tweets is in Figure 18.

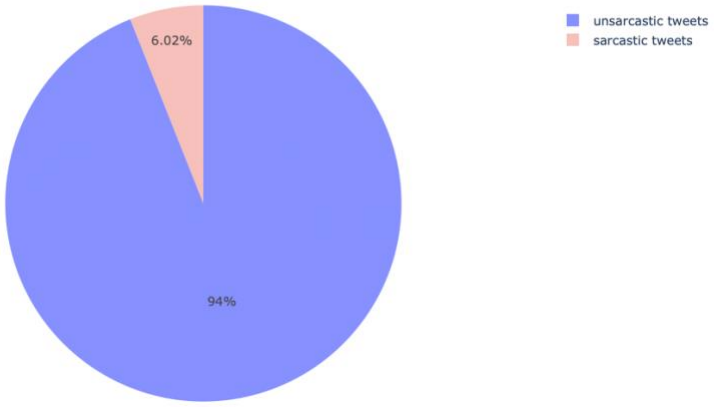

Figure 18 : Percentage of sarcastic tweets.

3.1.3 FREQUENCY DISTRIBUTION AND WORDCLOUD FOR POSITIVE SENTIMENTS

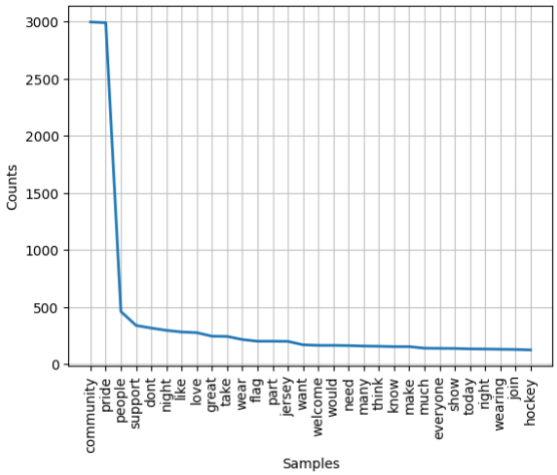

Figure 19 : Word frequency distribution of Positive Tweets

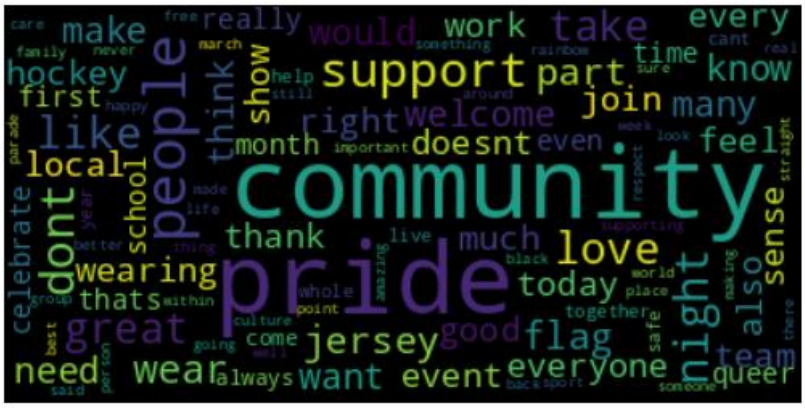

Figure 20 : Word cloud of positive tweets.

Figure 19 and Figure 20 are the frequency distribution and word cloud for the positive tweets. ‘community’, ‘pride’, ‘people’, and ‘support’ are the most used words, and it is evident in the word cloud.

3.1.4 FREQUENCY DISTRIBUTION AND WORDCLOUD FOR NEGATIVE SENTIMENTS

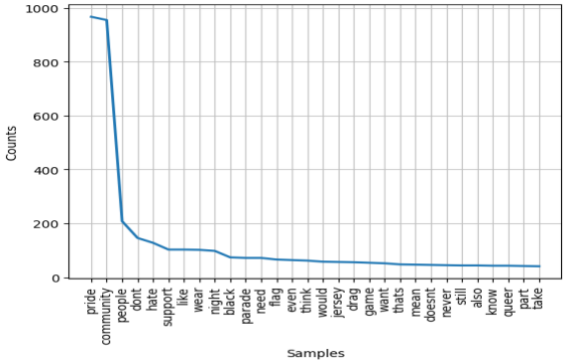

Figure 21: Word frequency distribution of negative tweets.

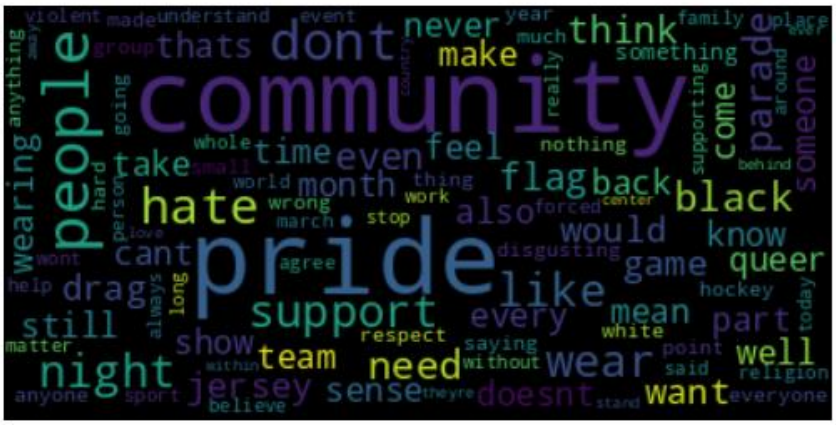

Figure 22 : Word cloud of negative tweets.

Figure 21 and Figure 22 show the frequency distribution and word cloud for the negative tweets. The most used words are ‘pride’, ‘community’, and ‘people’. The word ‘hate’ has negative connotation, so it is evident why the word appeared in the negative word cloud.

3.1.5 FREQUENCY DISTRIBUTION AND WORDCLOUD FOR NEUTRAL SENTIMENTS

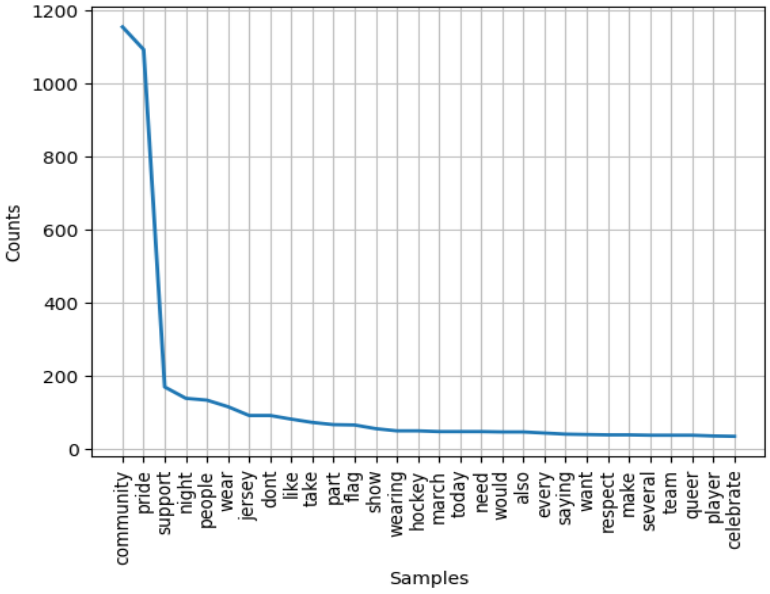

Figure 23 : Word frequency distribution for neutral tweets.

Figure 24 : Word cloud for neutral tweets.

Figure 23 and Figure 24 are the frequency distribution and word cloud for the neutral tweets. Some of the less common but noticeable words are ‘march’, ‘national’, ‘flag’, ‘equality’ and ‘month’. These words have neutral annotation and un-opiniated.

3.1.6 MACHINE LEARNING

Logistic Regression mode was built to predict the sentiments, a dichotomous variable. A new column ‘label’ is created based on the polarity of each tweet. If the tweet is positive, the label is marked as 1, otherwise 0. ‘sklearn’ library was used to build the machine learning model. The vectorized data was the input to the model. The model was trained on the training and test data set.

An accuracy of 84% was achieved for the model. Figure 25 is the Confusion Matrix generated for the classifier.

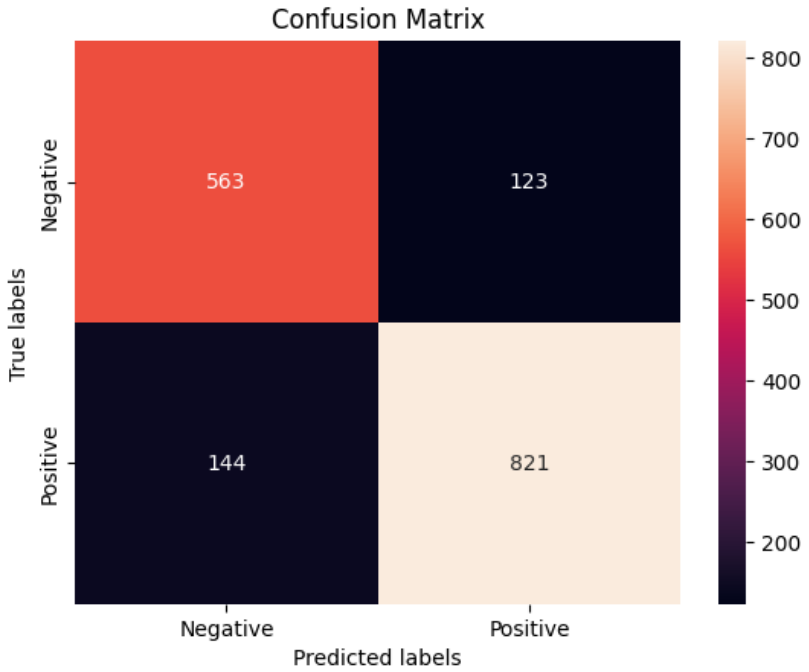

Figure 25 : Confusion matrix for machine learning model.

3.1.7 DEEP LEARNING

Sentiment analysis can be done with deep learning (Zhang, Wang and Liu, 2018). Deep learning has recently gained popularity as a technique for addressing linguistic problems and subjective interpretation that affect word polarity (Rojas-Barahona, 2016).

The dataset was split into training and testing set, and a TensorFlow model was built with Python library support. Performance of the TensorFlow model with three optimizers such as RMSprop, Adadelta, and SGD. From Table 7 it is evident that the TensorFlow model with optimizer as RMSprop has higher accuracy compared to others. The confusion matrix for the model with highest accuracy is given in Figure 26.

Optimizer	Model Accuracy
RMSprop	85%
Adadelta	52%
SGD	59%

Table 7 : Model accuracies.

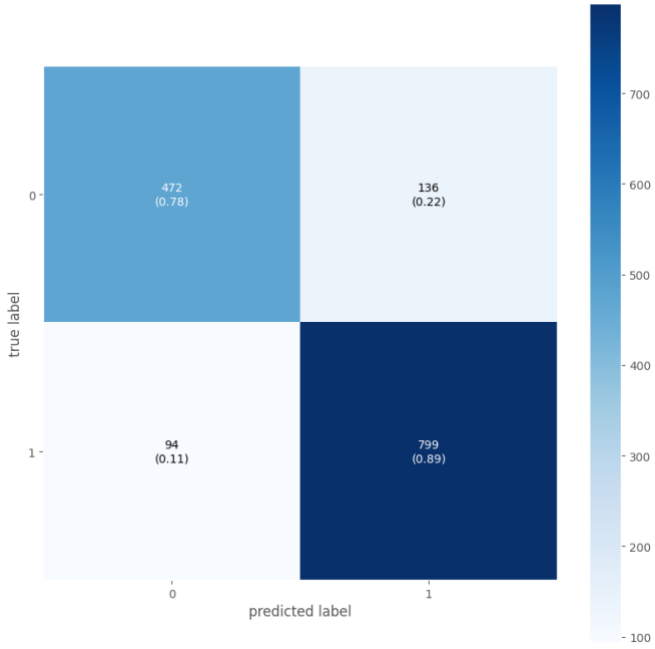

Figure 26 : Confusion matrix for model with RMSprop optimizer

3.2 TOPIC MODELLING

50 news articles related to the topic '**Turkey Syria Earthquake**' were scrapped to perform Topic Modelling.

3.2.1 DATA COLLECTION AND CLEANING

News API in Python was used to scrape the articles (datacareer.ch, 2022).

The following operations were performed to clean the data

1. Lowercase was applied to the text.
2. Square brackets were removed.
3. Links were removed.
4. Punctuations and special characters were removed from the data.

5. Numbers with numbers were removed.
6. Stopwords were removed.
7. Word tokenization was done.
8. Lemmatization was applied to the text.

3.2.2 DESCRIPTIVE ANALYSIS

There are a total of 1952 words in the news article contents. The word Frequency Distribution and the Word Cloud of the news article texts is given in the Figure 27 and Figure 28, respectively. The most used word is ‘Syria’ followed by ‘Turkey’. Given that the news articles were about the earthquake at these places, the words ‘Syria’ and ‘Turkey’ were widely repeated.

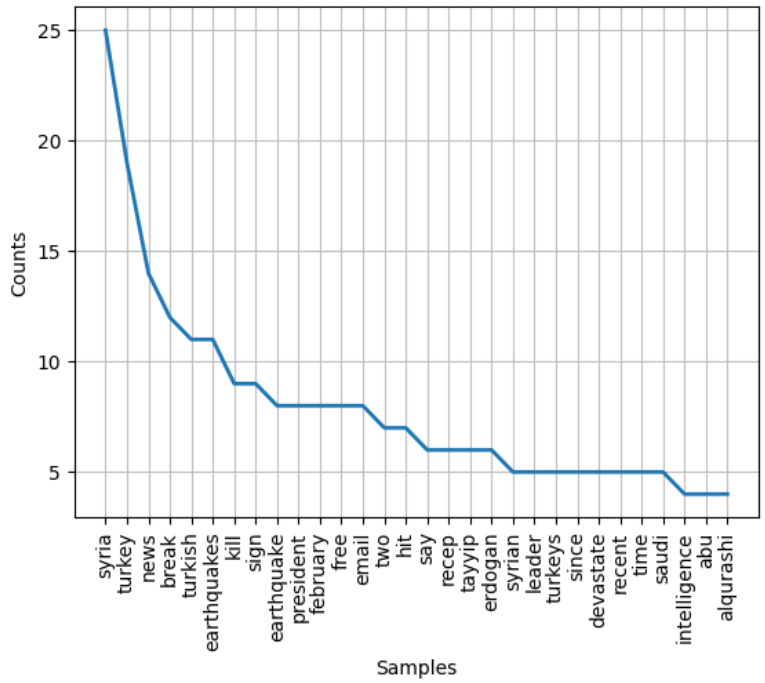

Figure 27 : Word frequency distribution of the news article contents.

Figure 28 : Word Cloud for the news article contents.

Figure 29 shows top eight most used words and their respective counts.

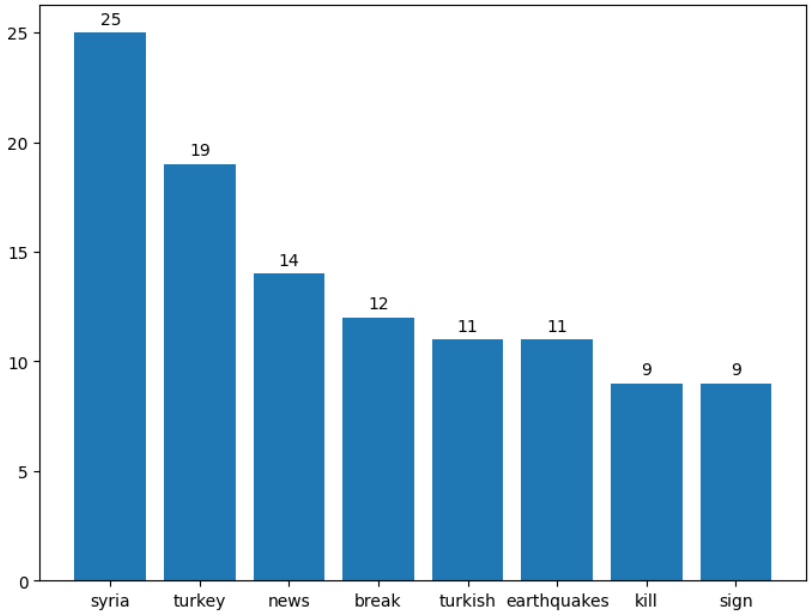

Figure 29 : Most used words and frequency

3.2.3 TOPIC MODELLING USING LDA AND LSA

Latent Dirichlet Allocation (LDA) technique was used to identify topics within the news article contents.

Although choosing the right number of topics for the LDA topic model has not been thoroughly researched, choosing the right number of topics is crucial for text analysis (Gan and Qi, 2021). The optimal number of topics was determined by calculating the coherence score for an LDA model with topic numbers ranging from 2 to 10. The model with the highest coherence was chosen. Table 8 shows the number of topics and the coherence score of the respective LDA model. The number of topics for the LDA model was selected as 4.

Table 8 : Coherence score of LDA model for each number of topics

Number of Topics	Coherence Score
2	0.352292
3	0.353517
4	**0.358899**
5	0.331823
6	0.345890
7	0.337199
8	0.345501
9	0.336077
10	0.341321

Figure 30 is the Word Cloud for the 4 topics selected using LDA model.

Figure 30 : Word Cloud for topics selected using LDA model.

Latent Semantic Analysis (LSA) was also used to find the number of topics. LSA model was able to find only one topic from the news article contents. This might be because the corpus is small, and the content is very similar. The Word Cloud of topic identified using LSA model with a Coherence score of 0.35929 is Figure 31.

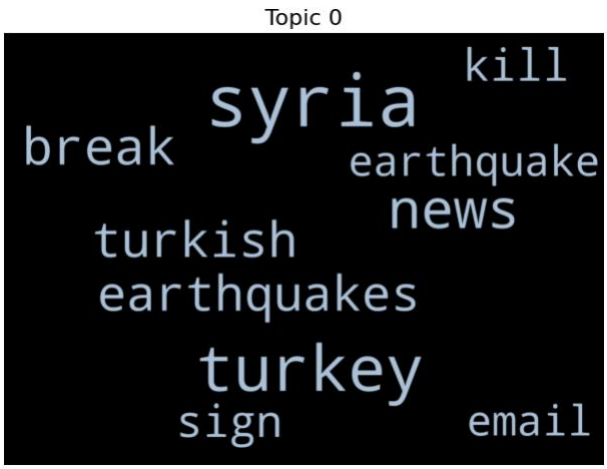

Figure 31 : Word Cloud for the topic identified using LSA model.

3.2.4 TEXT SUMMARISATION

Manual text summarization is laborious, expensive, and almost impossible when there is a large volume of text. To construct the final output as a summary, Extractive Summarisation

selects the pertinent statements from the provided input text or document and convolves those statements (Rathi et al., 2022).

Score of each sentence was calculated and the sentences with score 2 times greater than the average score were extracted and used in the summary. This value can be changed to increase or decrease the summary size. The summarized version of the article on Kılıçdaroğlu, the head of Turkey's opposition Republican People's Party (Luhn, 2023) is as follows.

Erdoğan claims it also "reset" the two-term limit, paving the way for him to seek a third five-year presidential term this May; he previously served as prime minister from 2003 to 2014, then the country's most important office. As chairman, Kılıçdaroğlu has transformed the party from the mouthpiece of the secularist urban elite into a European-style social democratic party, calling for "reconciliation" between Turkey's various political, ethnic, and religious groups.

4 SUMMARY AND CONCLUSION

Statistical analysis and text mining were carried out on Social Media datasets. Various aspects of Twitter data on 'Chat GPT' was analysed using visualization tools. Due to the unavailability of Twitter API and the limitations of snscrape library, it was not possible to analyse the current world trends. The centrality measures and community detection using Louvain Algorithm were performed on a subgraph of Facebook ego network. Use of subgraph was required due to processing costs. Analysis on Twitter data showcased that majority of the tweets had a positive sentiments and emotion towards pride community. Four topics were identified using LDA model and one topic using LSA model on scrapped news article.

5 REFERENCES

Ortiz, S. (2023). What is ChatGPT and why does it matter? Here's everything you need to know. [online] ZDNET. Available at: https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/.

Read, J. (2023). Why ChatGPT Is A Cyber Threat To Businesses. [online] EMS NOW. Available at: https://www.emsnow.com/why-chatgpt-is-a-cyber-threat-to-businesses/ [Accessed 29 Apr. 2023].

Vosoughi, S., Roy, D. and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), pp.1146–1151.

Cohen, H. (2013). *How Reliable Is Twitter? [Research] - Heidi Cohen*. [online] Heidi Cohen. Available at: https://heidicohen.com/reliable-twitter-research/.

developer.twitter.com. (n.d.). *Supported languages and browsers*. [online] Available at: https://developer.twitter.com/en/docs/twitter-for-websites/supported-languages.

Powell, J. and Hopkins, M. (2015). *A Librarian's Guide to Graphs, Data and the Semantic Web*. [online] ScienceDirect. Available at:

https://www.sciencedirect.com/book/9781843347538/a-librarians-guide-to-graphs-data-and-the-semantic-web [Accessed 4 May 2023].

Neo4j Graph Data Platform. (n.d.). *Eigenvector Centrality - Neo4j Graph Data Science*. [online] Available at: https://neo4j.com/docs/graph-data-science/current/algorithms/eigenvector-centrality/.

Rojas-Barahona, L.M. (2016). Deep learning for sentiment analysis. *Language and Linguistics Compass*, 10(12), pp.701–719. doi:https://doi.org/10.1111/lnc3.12228.

datacareer.ch. (2022). *Accessing the News API in Python*. [online] Available at: https://www.datacareer.ch/blog/accessing-the-news-api-in-python/ [Accessed 5 May 2023].

Gan, J. and Qi, Y. (2021). Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. *Entropy*, 23(10), p.1301. doi:https://doi.org/10.3390/e23101301.

Rathi, K., Raj, S., Mohan, S. and Singh, Y.V. (2022). *A Review of State-Of-The-Art Automatic Text Summarisation*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4107774.

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4). doi:https://doi.org/10.1002/widm.1253.

Luhn, A. (2023). *Exclusive: The Man Who Could Beat Erdoğan*. [online] Time. Available at: https://time.com/6274645/exclusive-kemal-kilicdaroglu-turkey-election/.

IBM (2022). *What is social media analytics?* [online] www.ibm.com. Available at: https://www.ibm.com/topics/social-media-analytics.

6 APPENDIX

1. Twitter Trend Analysis : Statistical_Analysis_22185279.ipynb
2. Graph Analysis : Graph_Analysis_22185279.ipynb
3. Sentiments Analysis : Sentiment_Analysis_22185279.ipynb
4. Topic Modelling : News_Articles_22185279.ipynb