Analyzing Tweets regarding the COVID-19 Pandemic and Black Lives Matter Protests

Julian George Agius
University of Malta
Email: julian.agius.14@um.edu.mt

***Abstract*—In today's society, Twitter is used as an outlet for expressing one's opinions, with millions of tweets tweeted per day. Understanding how users are reacting to certain events can be leveraged to find better solutions for current subjects of debate. In this paper, a pipeline used to analyze tweets is presented and described. The two topics analyzed relate to the COVID-19 pandemic and Black Lives Matter protests. Text mining and graph mining approaches were employed to identify and visualize what Twitter users are writing about present trending topics.**

I. Introduction

In today's day and age, people take to multiple social media platforms to voice out their criticisms, preferences and opinions on a variety of topics. Twitter is one such platform that offers this opportunity to its users.

With more than 145 million active daily users[1] tweeting approximately 500 million tweets per day[2], Twitter can provide governments and organisations the opportunity to understand the reaction of their end users to certain news, events or products and improve future policies or updates [1], [2]. However, due to the sheer volume of Tweets being tweeted out every second, the problem of information overload [3] quickly arises.

In this paper, a pipeline used to analyze tweets was developed. The goal of this pipeline is to extract information from a stream of tweets and visualize it in a manner where the reaction of Twitter users can be understood easily and quickly. The two topics chosen to be analyzed for this project, the global COVID-19 pandemic and Black Lives Matter protests are trending in both traditional and social media and hence, are of interest to a wide audience.

A. Aims and Objectives

The main aim of this project is to use data streaming and data mining techniques to analyse tweets relating to the COVID-19 pandemic and Black Lives Matter protests. In order to do so, the following objectives will be carried out:

- **O1**: Analyze tweets to extract information pertaining to the selected topics
- **O2**: Identify whether there is an overlap between tweets regarding the chosen topics

[1]https://blog.hootsuite.com/twitter-statistics/
[2]https://www.dsayce.com/social-media/tweets-day

II. Related Research

This section will outline a number of research initiatives stemming from the analysis of tweets.

The advent of social media applications and its subsequent widespread adoption has provided extensive opportunities and challenges for researchers [4]. Social media platforms are responsible for generating enormous amounts of data at a very rapid pace. Big data obtained from social media can enable organizations and governments to derive meaningful insights [5] by using effective analytic techniques.

Reddick et al. [6] describe how user-generated content from social media can assist governments with policy making by extracting useful insights through text analysis. In their research, they performed entity extraction and linking, keyword extraction, topic modelling and opinion mining on Facebook posts relating to the City of San Antonio Solid Waste Management Department.

Parsons et al. [1] leveraged Twitter data to analyse social network content during natural disasters. Their work researches tweets in relation to the Disaster Management Lifecycle (DML) of flooding in the UK that occurred during 2013 and 2014. Parsons et al. managed to build a timeline of events for UK storms for a particular time frame. The events were classified into six categories described by Faulkner's DML framework [7]. The types of tweets were also categorized as either subjective and critical, informative and supportive or objective and instructive. Li et al. [8] utilized retweets regarding the Fukushima Nuclear Radiation Disaster to highlight the need for reassuring messages via social media during a disaster.

Another popular application of text analysis on social media posts, especially tweets is to figure out the political landscape of a country before an election. Song et al. [9] used topic modeling, Twitter user network analysis, and term co-occurrence retrieval to detect changes in the social issues being discussed in Korea before the 2012 elections. In their research, De Luca et al. [10] highlighted another potential use case for text analysis in politics. They used sentiment analysis to identify the thoughts expressed about the work of United States President Donald Trump.

Besides the use of text analysis to explore user-generated content, there is ongoing research on the use of graph mining on Twitter data. Wang et al. [11] proposed a graph model based on the hashtags found in tweets. After building their

hashtag graph model, they performed sentiment classification and compared the results of their graph-based model against a traditional Support Vector Machine (SVM) model and obtained competitive results.

Talebi et al. [12] developed a pipeline to build a user knowledge graph from tweets. In their approach, they extracted keywords from preprocessed tweets and extracted links using regular expressions.

III. Background

In this section, an overview of the techniques and algorithms used for text analysis and graph mining will be provided.

A. Text Mining

Gandomi and Haider [5] highlight text mining as one of the possible techniques that can be used to transform data generated from social networks into meaningful insights.

Information extraction is the process of extracting important information from text. There are two types of information extraction, rule-based and open. Rule-based systems rely on sets of patterns or regular expressions of part-of-speech tags to extract valuable information from text. Due to the brittleness of this approach, open domain information extraction (open IE) [13] tends to be favoured. Open IE uses a few patterns for simple sentences and extracts self-contained clauses from longer sentences. Natural logic inference is then used to determine the most appropriate arguments for the relation triples.

Jiang [14] states that the two underlying processes for information extraction are relation extraction and named entity recognition. Named entity recognition (NER) is a text analysis process that identifies particular occurrences of words that belong to a particular named entity, such as person, organisation or location [15]. Relation extraction refers to the process of associating these entities using semantic relationships.

B. Graph Mining

PageRank [16] is an algorithm used to measure the importance of website pages. Wang et al. [17] proposed using PageRank to extract keywords from text data. The algorithm to calculate the PageRank of a vertex is as follows:

$$PR(V_i) = (1-d) + d \times \sum_{j \in In(v_i)} \frac{PR(V_j)}{|Out(V_j)|}$$

[16] Where $In(V_i)$ is the set of vertices that point to a given vertex V_i, and $Out(V_i)$ is the of edges going out from vertex V_i. d represents the damping factor which is a number between 0 and 1. The damping factor is generally set to 0.85.

Label Propagation, developed by Raghavan et al. [18], is an algorithm that detects community structures in large-scale networks in near linear time. A community within a graph is defined a set of nodes whose edges between each other is more dense than with the other nodes in the graph. The algorithm for Label Propagation is as follows:

1) Initialize each node in its own community. For a given node x, $C_x(0) = x$
2) Set step t = 1
3) Arrange the nodes in the network in a random order and set it to X
4) For each $x \in X$ chosen in that specific order, find the label occurring with the highest frequency among neighbours and ties are broken randomly
5) If every nodes node has a label that the maximum number of their neighbours have, terminate the algorithm. Else, increment t by 1 and go to (3)

IV. Methodology

This section describes the design and implementation utilized for this project. The architecture used for this research is described in Figure 1.

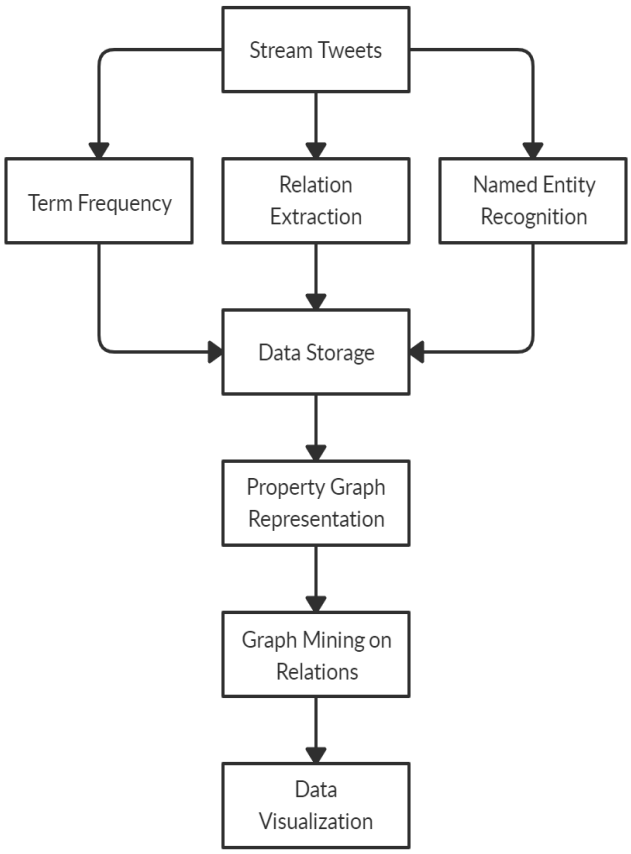

Fig. 1. Pipeline for analyzing tweets

A. Design Architecture

Apache Kafka and the Tweepy Python library were used to read a stream of English tweets with particular hashtags. A Kafka producer was used to send the stream of tweets to topics which were then read by a Kafka consumer to perform text mining, and Spark Streaming to count the frequency of terms in tweets. The results of these processes were stored in separate MongoDB collections.

GraphX was used to represent relation triples extracted earlier within the pipeline as a knowledge graph and perform graph mining.

Finally, the results obtained from the text mining and graph mining approaches were visualized to provide a way for users to understand what Twitter is saying regarding the COVID-19 pandemic and Black Lives Matter protests.

The following sub-sections will provide further detail explaining how these techniques were implemented.

B. Term Frequency

The identification of frequently used terms [6], [10] has been used in several studies to give readers a sense for the most commonly used language within social media posts. This process was implemented by using Spark Streaming while ingesting a stream of tweets from Kafka. The stream of tweets were batched every 10 seconds and preprocessed by removing symbols and numbers from each tweet. A word count process was carried out using MapReduce and the results were stored in a MongoDB collection. The most frequently used terms, excluding stop words and the list of hashtags used, were visualized in a word cloud.

C. Text Mining

The two techniques used for analyzing tweet text were Open Information Extraction and Named Entity Recognition. They were carried out within a Kafka consumer and their results were stored in separate MongoDB collections. For both of these techniques all words were converted to lowercase and all hashtag symbols were removed [10].

A Python wrapper of Stanford OpenIE[3] was used to perform information extraction and produce relation triples from tweet text. The pipeline used by Stanford OpenIE to extract triples is as follows:

- Word Tokenization: Converting a string of text into words.
- Part-Of-Speech Tagging: Labelling words with their POS tag, such as noun, verb, adjective, etc...
- Lemmatization: Converting a word into its dictionary base form.
- Dependency Parsing: Extracting a dependency parse for a sentence representing its grammatical structure and the relations words in the same sentence have with each other.
- Named Entity Recognition: Recognizes named entities in text, in our case diseases.
- Natural Logic [19]: Explaining textual inferences by learning to classify entailment relations
- Open Information Extraction: Extracting open-domain relation triples. Each triple is composed of a subject, relation and object.

The relation triples extracted required further preprocessing as several triples provided a similar semantic meaning. In order to prune some of the unnecessary relations; when multiple triples with the same subject and relation were found, only one triple was kept.

Named Entity Recognition was performed using the Spacy's[4] pre-trained *en_core_web_sm* model. The model was created by using an English multi-task CNN trained on the OntoNotes corpus[5].

D. Graph Mining

GraphX was used to run the PageRank and label propagation algorithms on the knowledge graph that resulted from tweet text. The knowledge graph was created using the relation triples obtained from open information extraction. The subjects and objects are represented as vertices within the graph and the relations are represented as edges.

Similar to the approach adopted by Wang et al. [17], PageRank was used to extract keywords from tweet text.

The label propagation algorithm takes a parameter that defines the maximum number of iterations that will be carried out before the algorithm terminates. In their research, Raghavan et al. [18] found that at least 95% of nodes are correctly classified into communities by the 5th iteration when using randomly generated graphs. In this paper, the maximum number of iterations for label propagation was set to 5.

The communities detected using label propagation were visualized using a Python library called webweb[6] which is built on top of D3.js[7].

E. Challenges

This section will highlight the challenges encountered while implementing this project. The main practical challenges were setting up the required environment and integrating the various components used within the tweet analysis pipeline. Open information extraction was providing a number of relation triples that had similar semantic meaning. In order to limit this, when multiple triples with the same subject and relation were found, only one triple was kept.

V. Results

The following section will describe the results obtained from approximately 480,000 COVID-19 related tweets and 350,000 Black Lives Matter related tweets respectively.

A. Frequent Terms

Some of the most frequent terms used within tweets relating to the corona virus, portrayed in Figure 2 are pandemic, cases, lockdown, people, country and deaths. Popular terms used by the Twitter community with regards to Black Lives Matter protests are people, protest, police and army. The term *btstwt* comes from people tagging *@BTS_twt*. BTS are a K-pop band that made a million dollar donation to Black Lives Matter.

[3]https://nlp.stanford.edu/software/openie.html

[4]https://spacy.io/

[5]https://catalog.ldc.upenn.edu/LDC2013T19

[6]https://webwebpage.github.io/

[7]https://d3js.org/

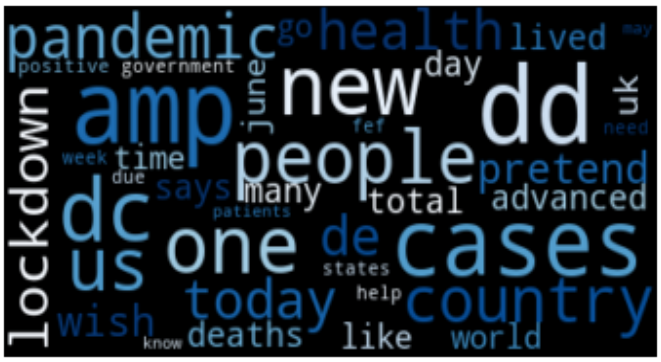

Fig. 2. Frequent words used in tweets regarding the COVID-19 pandemic

Fig. 3. Frequent words used in tweets regarding the Black Lives Matter protests

B. Frequent Named Entities

Table I displays the most frequent persons, organizations and locations extracted from tweets related to the two chosen topics. Neil Ferguson, mentioned in COVID-19 related tweets, was an adviser on the corona virus to the UK government. Breonna Taylor was an African-American woman who was shot by the police in the US.

TABLE I
TABLE DISPLAYING MOST FREQUENT NAMED ENTITIES PER TOPIC

Named Entity	Covid-19	BLM
Person	Neil Ferguson	Breonna Taylor
Organization	govt	army
Location	Earth	the la area

Edward Colston, George Floyd and Sarah Grossman were also featured names extracted from Black Lives Matter related tweets. Edward Colston was an English merchant who was involved in the slave trade in the 1600s and 1700s. George Floyd was an African-American man who was killed by the police and Sarah Grossman was a woman who died after being tear-gassed by the police in Ohio.

C. Results from Graph Mining

Figures 4 and 5 depict the most important nodes in the knowledge graph after carrying out the PageRank algorithm. The knowledge graph was created using all the extracted relations per topic.

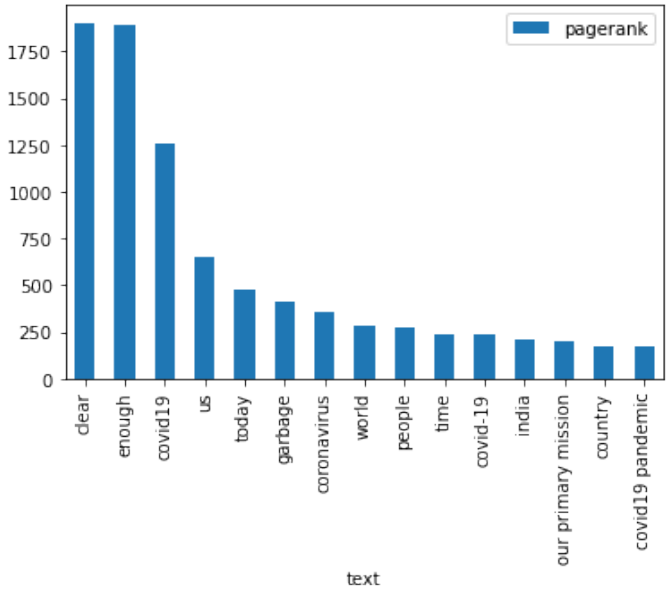

Fig. 4. PageRank scores of entities extracted from tweets regarding the COVID-19 pandemic

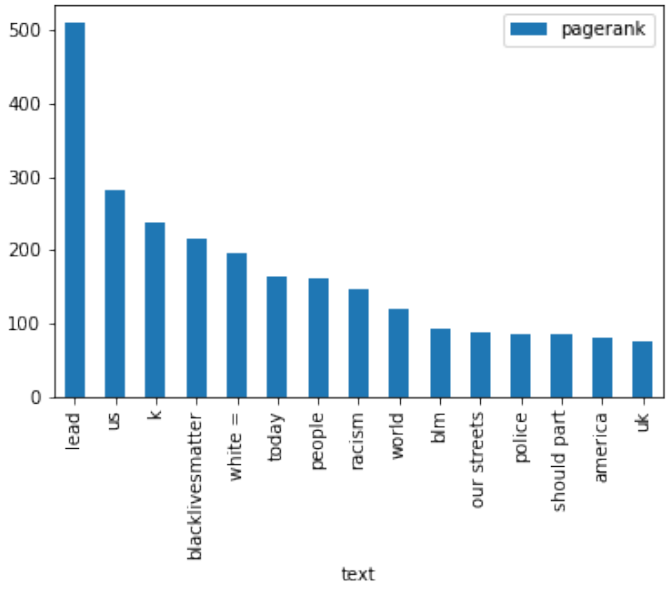

Fig. 5. PageRank scores of entities extracted from tweets regarding the Black Lives Matter protests

Some of the communities detected using label propagation were visualized in a web app using the webweb Python library. These communities provided interesting findings for both topics. Some of the insights obtained from corona virus related tweets are:

- Antifa terrorists in Seattle and Portland. Antifa is an anti-fascist political movement

- IPOB activists protest the killing of biafrans in Nigeria. Interestingly, because the the protests were led by a British Nigerian activist, there is a link between the UK and the IPOB activist. IPOB is a Nigerian political organization.
- Certain people with Nazi ideology wear the swastika with pride

Findings obtained from Black Lives Matter related tweets mention the burden placed on Pakistani women and young girls who require humanitarian action due to gender inequalities. Another community describes a meeting held by the United Nations regarding the COVID-19 Global Humanitiarian response plan. The destruction of Corona virus testing sites during Antifa riots is brought up in a different community.

VI. Conclusion

This paper described a pipeline that ingests a stream of tweets, performs text and graph mining and visualizes the results. The results provided insight into what people are tweeting regarding the COVID-19 pandemic and Black Lives Matter protests. The results of this paper provided:

- A list of the most frequent terms being used in tweets regarding the chosen topics
- A list of the most frequently mentioned named entities, highlighting some of the most tweeted about people
- A number of interesting insights obtained from open information extraction and applying community detection on the resulting relation graph.

Overlap was also found between the two topics. A number of tweets mentioned the protests and riots that were happening due to Black Lives Matter, as well as social distancing when using Corona virus related hashtags. In a community extracted from tweets using Corona virus hashtags, there were mentions of the closing of corona virus testing activities due to antifacist riots. These results signify that both objectives described earlier were reached.

A. Future Work

Canonicalization is a technique that could be implemented to reduce the number of relation triples obtained from open information extraction. Canonicalization is the process of converting data that has multiple possible forms into a single standard form.

Analyzing the sentiment of tweets regarding both COVID-19 and Black Lives Matter could prove to be a worthwhile improvement to the pipeline described in this paper.

One shortcoming of the Named Entity Recognition model used is that it was not trained on social media posts. Implementing a model that was trained on tweets [20] in particular could provide better results.

Rather than analysing the text found in tweets, techniques like community detection could be applied to analyze social network structure and discover sub-groups of users who interact frequently with each other [5].

Although the graph visualization tool allows users to search for particular nodes within a single graph, the ability to search for specific nodes within a list of graphs would allow non-technical users to view particular communities which interest them. Another possible improvement to the graph visualization would be the ability to hide particular nodes to reduce clutter.

References

[1] S. Parsons, P. M. Atkinson, E. Simperl, and M. Weal, "Thematically analysing social network content during disasters through the lens of the disaster management lifecycle," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1221–1226.

[2] A. Katal, M. Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices," in *2013 Sixth international conference on contemporary computing (IC3)*. IEEE, 2013, pp. 404–409.

[3] M. G. Rodriguez, K. Gummadi, and B. Schoelkopf, "Quantifying information overload in social media and its impact on social contagions," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[4] N. A. Ghani, S. Hamid, I. A. T. Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior*, vol. 101, pp. 417–428, 2019.

[5] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International journal of information management*, vol. 35, no. 2, pp. 137–144, 2015.

[6] C. G. Reddick, A. T. Chatfield, and A. Ojo, "A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government facebook use," *Government Information Quarterly*, vol. 34, no. 1, pp. 110–125, 2017.

[7] B. Faulkner, "Towards a framework for tourism disaster management," *Tourism management*, vol. 22, no. 2, pp. 135–147, 2001.

[8] J. Li, A. Vishwanath, and H. R. Rao, "Retweeting the fukushima nuclear radiation disaster," *Communications of the ACM*, vol. 57, no. 1, pp. 78–85, 2014.

[9] M. Song, M. C. Kim, and Y. K. Jeong, "Analyzing the political landscape of 2012 korean presidential election in twitter," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 18–26, 2014.

[10] E. De Luca, F. Fallucchi, R. Giuliano, G. Incarnato, and F. Mazzenga, "Analysing and visualizing tweets for us president popularity," *Int. J. Adv. Sci. Eng. Inf. Technol*, vol. 9, no. 2, pp. 692–699, 2019.

[11] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1031–1040.

[12] S. Talebi, K. Manoj, and G. H. Kumar, "Building knowledge graph based on user tweets," in *Data Analytics and Learning*. Springer, 2019, pp. 433–443.

[13] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 344–354.

[14] J. Jiang, "Information extraction from text," in *Mining text data*. Springer, 2012, pp. 11–41.

[15] A. Mikheev, M. Moens, and C. Grover, "Named entity recognition without gazetteers," in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1999, pp. 1–8.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[17] J. Wang, J. Liu, and C. Wang, "Keyword extraction based on pagerank," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2007, pp. 857–864.

[18] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007.

[19] B. MacCartney and C. D. Manning, "Natural logic for textual inference," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics, 2007, pp. 193–200.

[20] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.