Twitter Data Mining and Sentiment Analysis: Uncovering Insights from the Tweet Dataset

Silmara Coelho

CCT College Dublin

Computing.IT.Business

MSc in Data Analytics SB+/FT

Lecture(s)

Muhammad Iqbal

David McQuaid

<https://github.com/SilmaraCCoelho/SilmaraCCoelho>

Date 10/11/2023

**Author Note**

Contact information: [2023140@student.cct.ie](mailto:2023140@student.cct.ie) or silmaraccoelho@gmail.com

#### Acknowledgements:

I would like to extend my sincere thanks to the following people, without whom this project would not have been possible: my professors, my classmates and my flatmate, for their support. In particular, I would like to thank my mentor Luis Maldonado for being so patient, encouraging me to learn, taking the time to guide me, and sharing your knowledge with me. Thank you very much.

**Disclosure:** The author in this paper have no conflicts of interest to disclosures.

# Abstract

This project focuses on analysing a substantial dataset obtained from the Twitter API called "ProjectTweets.csv" for understanding public sentiment, tracking trends and gaining insights on various topics. The project develops in several stages, starting with Data Pre-processing to ensure data integrity, data quality, cleaning the data and exploring the data to reveal the structure, content and possible irregularities of the data set.

Once the data is refined, we employ the Text Blob and NLTK libraries, to label the dataset using count-based vectorization that is an estimation of sentiments indicated by individual polarity of words. Subsequent data visualization reveals our findings through sentiment distribution histograms and time series charts. These visualizations offer insights into prevalent topics over time and help answer research questions.

Then, we incorporate different neural network machine learning techniques to predict sentiments, categorize tweets, and unveil hidden patterns within the data. The purpose is to select the most appropriate algorithms to process and show the results as fast as possible.

Finally, in order to speed up the process, data is migrated to a database architecture in order to perform transformation and summarization processes there. A comparison of the database architectures and their interaction with a Big data development environment is made.

*Keywords:* **Twitter; Sentiment Analysis; Neural Networks; Big Data Architecture;**

# Introduction

In an age marked by digital communication and social media's ubiquitous presence, platforms such as Twitter have emerged as rich sources of unfiltered, real-time data offering unique insights into public sentiment, trending topics and the pace of global conversations. The wealth of information contained in the Twitterverse presents an important opportunity for data analysis and exploration. Twitter has wide-ranging applications, including monitoring public opinion on political issues, tracking consumer sentiment toward brands, or simply understanding the ever-evolving landscape of global discussions.

This project aims to analyse a substantial set of Twitter data, aptly named "ProjectTweets.csv", acquired through the Twitter API, with the aim of discovering valuable insights, trends and sentiments hidden within. To achieve these objectives, this project follows a structured approach that encompasses data pre-processing, sentiment analysis, data visualization and application of machine learning techniques to predict sentiment behaviour over time.

Data pre-processing is the most fundamental step to ensure data quality and consistency. The key process here is achieving an acceptable labelling of the dataset (marking each sentence as positive, negative or neutral). a task that is even difficult to a real person. I tested two simple count-based unsupervised methods: Text Blob and Vader[[1]](#footnote-1) that are based on rules. Other methods may be better suited (trained and tuned to twitter data, as SentiWordNet or LIWC) and we want to compare them to feature based methods like logistic regression, support vector machines, or embedding based methods like lbl2vec, Word2Vec, FastText or Flair. This preparatory phase lays the foundation for further analysis. After cleaning the data, we employ Vader sentiment analysis to assess the emotional tone and polarity of tweets in the dataset. However, unsupervised methods can impact in unpredicted ways the performance of the rest of the algorithms and lead to wrong conclusions if the labelling is ambiguous or whether tweets convey real positive, negative, or neutral sentiment.

By exploring hashtags, mentions, and keywords, I identify topics that dominated the Twitter landscape during specific periods. This information, combined with the sentiment analysis can shed light on the interests, concerns, and discussions that have caught the attention of Twitter users.

Some Machine Learning algorithms were tested to predict sentiments, classify tweets into categories, or discover hidden patterns in data. In the same way, some different database architectures were tested in the big data environment in order to have faster processing, storing and retrieval of data.

## Objectives

Our objective is to uncover concealed insights, trends, and sentiments contained within this extensive data repository. I intend to provide a complete approach to my project, covering sentiment analysis, trending topics, visualization and architecture.

**Objective 1:** The main objective is to perform sentiment analysis on the Twitter dataset to evaluate the emotional tone of tweets. This includes classifying tweets as positive, negative, or neutral. The goal is to understand sentiment trends over time and how they relate to specific events or topics.

**Objective 2:** Identify and analyse trending topics and hashtags in the dataset. This objective aims to discover the most discussed and influential topics on Twitter during the dataset period.

**Objective 3:** In this part I will analyse how tweet volume and sentiment change over time. This objective will help to understand how events, seasons, or external factors impact Twitter activity and sentiment.

**Objective 4:** Create informative and visually coherent data visualizations to clearly present findings. These include sentiment distribution charts, time series charts, and word clouds to make insights accessible to a broad audience.

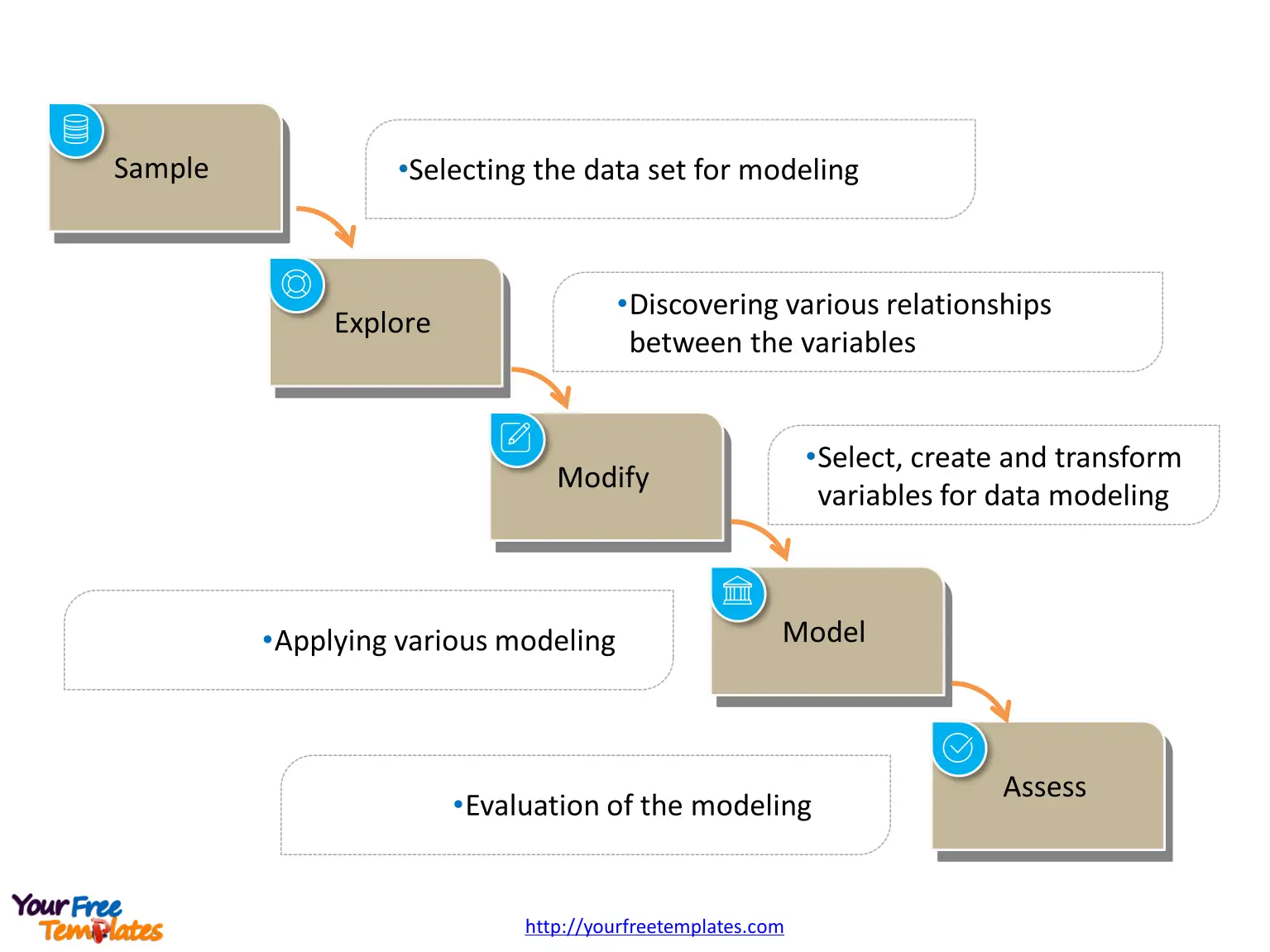
Project Management

### Participants

Academic student: The study was conducted by a healthcare professional, studying MSc in Data Analytics SB+/FT. The main objectives are sentiment analysis and identifying trends, with the option to apply machine learning to gain deeper insights.

### Data framework selected

The choice to employ the SEMMA framework for project management in the analysis of my Twitter dataset (Against more comprehensive approaches as CRISP-DM or KDD) is based on its simple and empiric approach, which precisely suits the distinctive needs of my data-driven project, because the business knowledge and the expected insights are not that clear from the beginning, they have to be discovered, and they are not as critical as testing the technology stack. SEMMA presents a well-organized process that harmonizes seamlessly with my project objectives and the dataset's characteristics, simplifying efficient and effective project management. Phases of SEMMA are:



##### Figure 1. The SEMMA stages is a straightforward empirical process aimed to discover unforeseen information

##### Source: <https://www.datascience-pm.com/semma/>

**Sample:** The project revolves around the analysis of an extensive Twitter dataset obtained via the Twitter API. This dataset encompasses tweets from a specified timeframe and covers a wide array of content. The analysis stems from the necessity to derive insights into public sentiment and trending topics on the Twitter platform.

**Explore:** The dataset was provided to me through my lecture, and its origin traces back to data collected from the Twitter API. Within the SEMMA framework, it entails a comprehensive examination of the Twitter dataset structure, content, and quality. This involves description, exploration, and initial data analysis to uncover insights and ascertain any pre-processing requirements. It involves examining the structure, content and quality of the dataset through description, exploration and initial analysis of the data to gain insights and identify any necessary pre-processing steps

**Modify:** This phase ensures that data is transformed into a clean, structured format conducive to subsequent analysis or modelling, aligning with SEMMA's systematic data preparation requirements.

**Model:** The sentiment analysis methodology will be implemented in this phase, employing specific tools, libraries, and programming languages consistent with SEMMA's modelling guidelines.

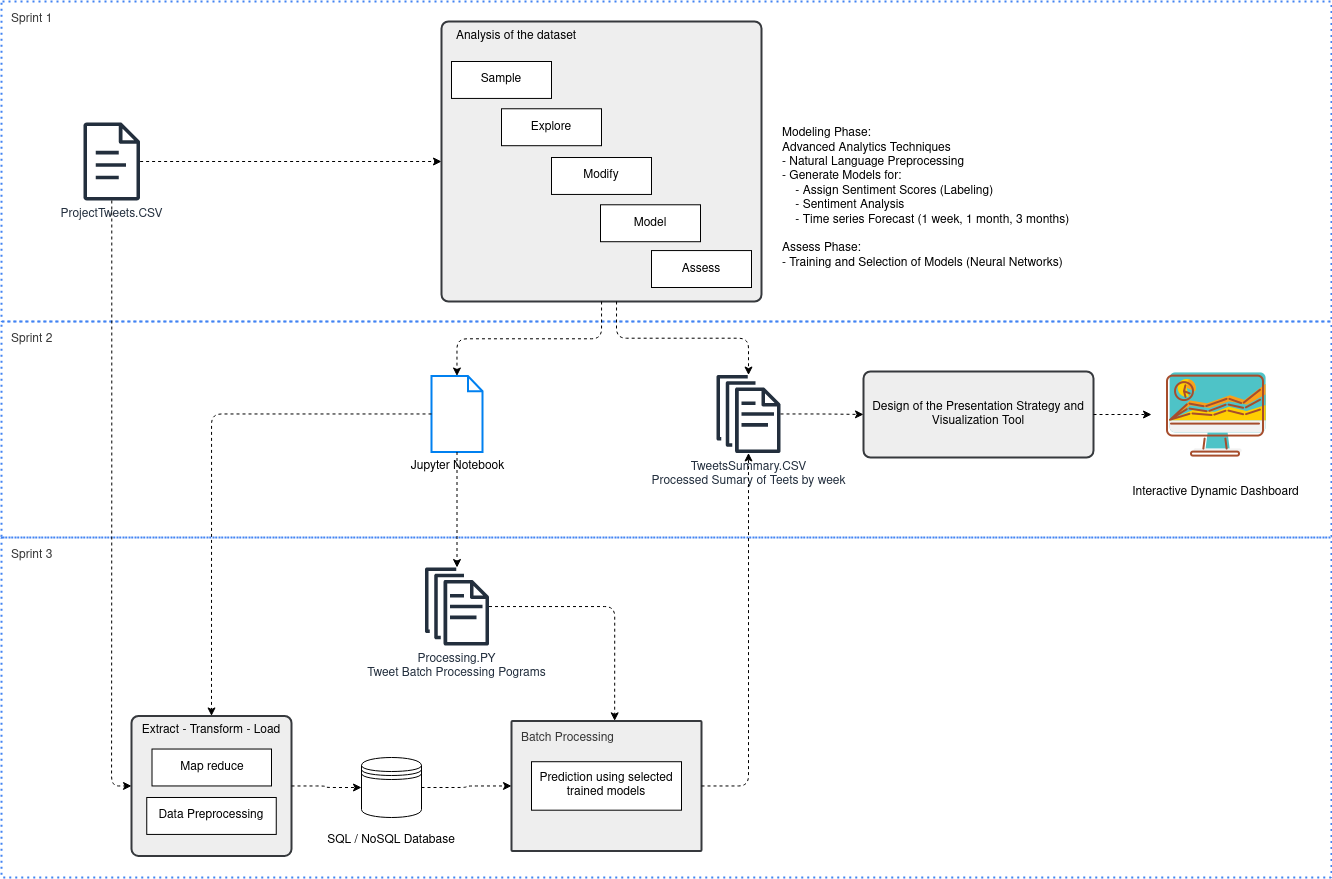
Assess: The SEMMA framework assesses the quality and effectiveness of the analysis or models, utilizing specific metrics and criteria to confirm alignment with project objectives and practical applicability. In this case, identifying actionable insights within the analysis, encompassing sentiment trends, trending topics, or user engagement insights.

### Project Management framework selected

Due to the nature of the project, it started with an exploration stage, after which a proper planning and organization happened. As a consequence, the report has an initial data exploration and analysis and then an in depth advanced analytical stage, after analysing the complete dataset. An Agile approach was selected, because the project aligned better with stages of incremental development following the exploration stage. See Figure 2.

### Developing Environment - Hadoop

When selecting the development environment, I opted for th**e Hadoop** development environment because it is a specialized setup where software development and data processing tasks are performed within the Hadoop ecosystem. It includes configuring tools, libraries, and resources that enable developers to write, test, and run code for analysing, processing, and managing data using the Hadoop distributed computing framework. In the context of a Jupyter notebook, it serves as an interface for defining and executing tasks in the Hadoop environment, allowing developers to work on big

data and distributed computing projects.

##### Figure 2. Simplified Process map of the project, showing a timeline of 3 Sprints aligned with corresponding increments in the complexity and technology stack of the project: Phase 1: Exploration using SEMMA framework and training of the Machine Learning Algorithms. Phase 2: consolidation of processing and transforming algorithms, and visualize the dataset insight summary. Phase 3: Transformation optimization by using a distributed processing environment and testing different database architectures.

#### Versioning

The project was developed in ***Hadoop,*** *which provides its own version control system*. In terms of Git versioning, the decision was to upload only finished versions of datasets and notebooks. The primary justification for this choice was that the datasets are very heavy and the risk associated with mastering the long learning curve of Git efficiently. Given the project's resource constraints, I decided to minimize the learning curve and potential errors that could arise from extensive version control management and instead focus on the algorithms and design decisions.

# 1. Data Analysis

## Exploratory Data Analysis

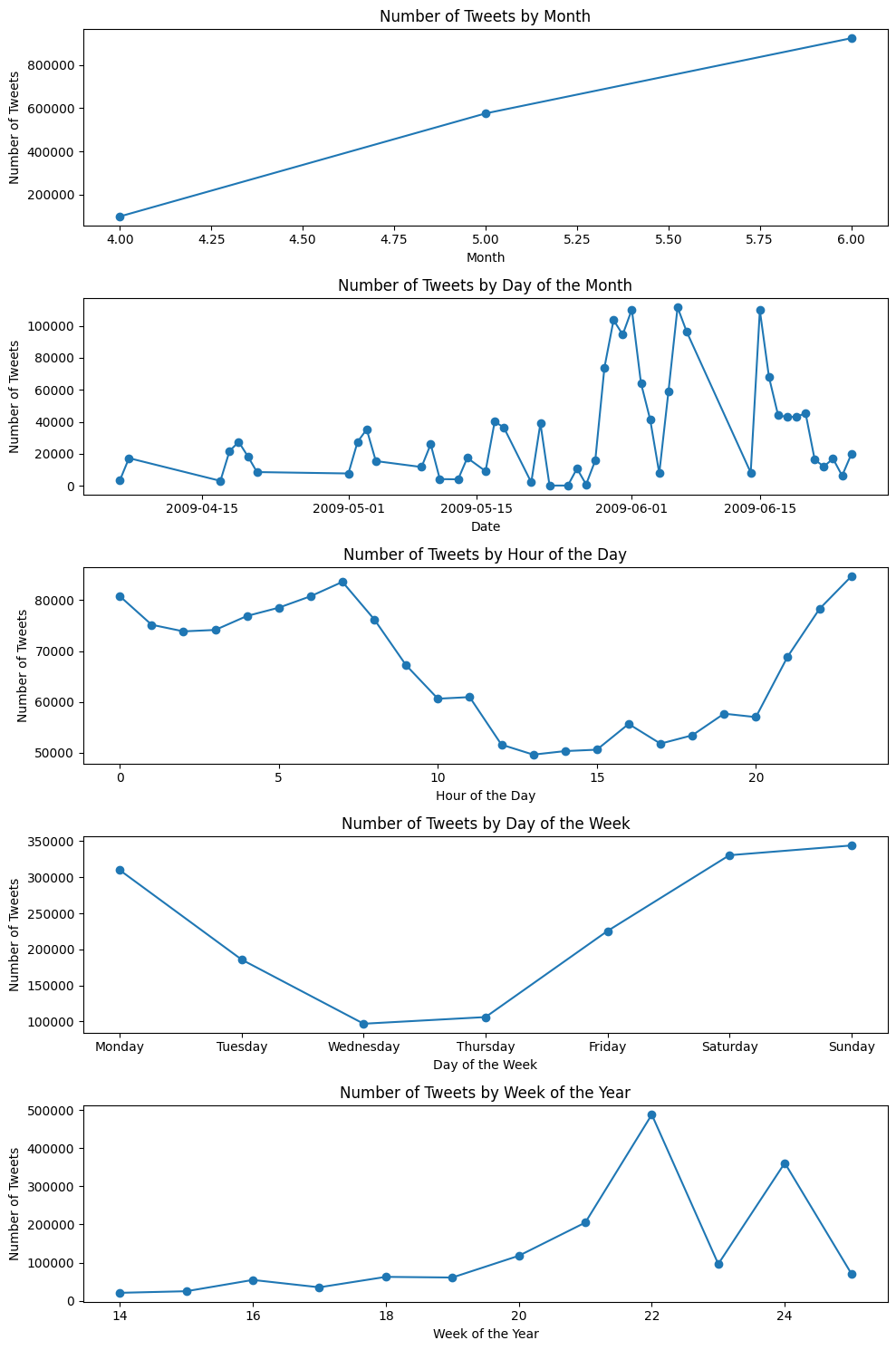
This process involved a series of techniques, both statistical and computational, to thoroughly explore the data. First the dataset were cleaned, transformed, and structured from the raw Twitter dataset. That includes the handling of missing values, identification and treatment of outliers, elimination of duplicated rows, conversion of data types, and the creation of engineered features (in this case, most of the new features were time related variables) . The ultimate objective is to ensure data accuracy, completeness, and suitability for analysis and prepare the dataset in a manner that allows for effective modelling. Some exploration of the dataset gives us a lot of preliminary information:

#### Assumptions:

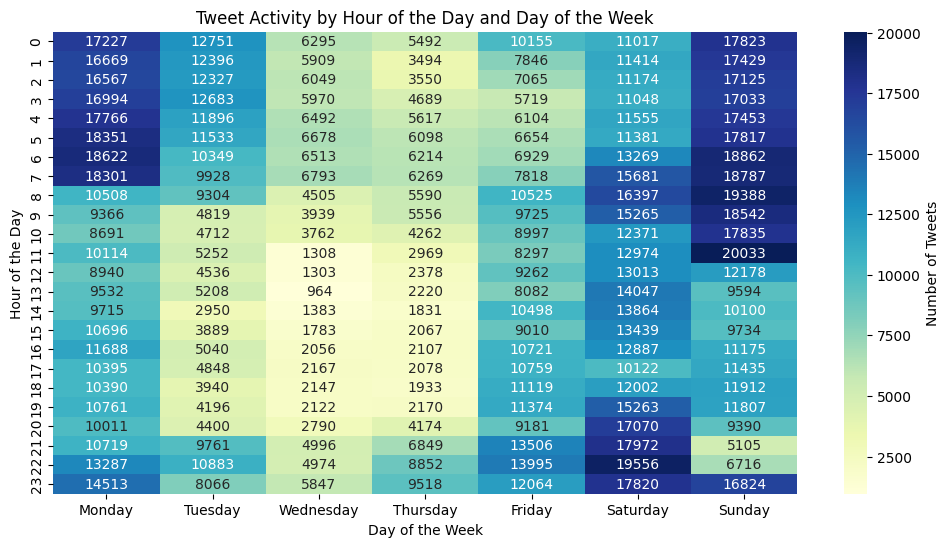
* The data has been collected without any known filter. It corresponds to all tweets made in a limited geographical area.

#### Some interesting findings about the dataset are:

* Data has been collected from April 6th 2009 to June 16th 2009, it is 12 weeks
* Number of tweets is increasing, especially after the 20th week of the year
* Majority of tweets are posted between 20 hours in the night and 10 hours in the morning with peaks at 0 Hrs and 6 hrs. All tweets are in Pacific Daytime. This suggests that tweets have been filtered by a geographical zone that is different from Pacific Time. See figure 3.
* Majority of tweets are posted between Friday evening to Sunday noon, and on Monday early morning. This also suggests that the geographical zone is different from pacific day time. See figure 4.
* Majority of tweets are concentrated in a couple weekends. Showing the strong seasonality of the number of tweets, See figures 5 and 6.
* There are many repeated rows, which correspond to generic copy-pasted messages

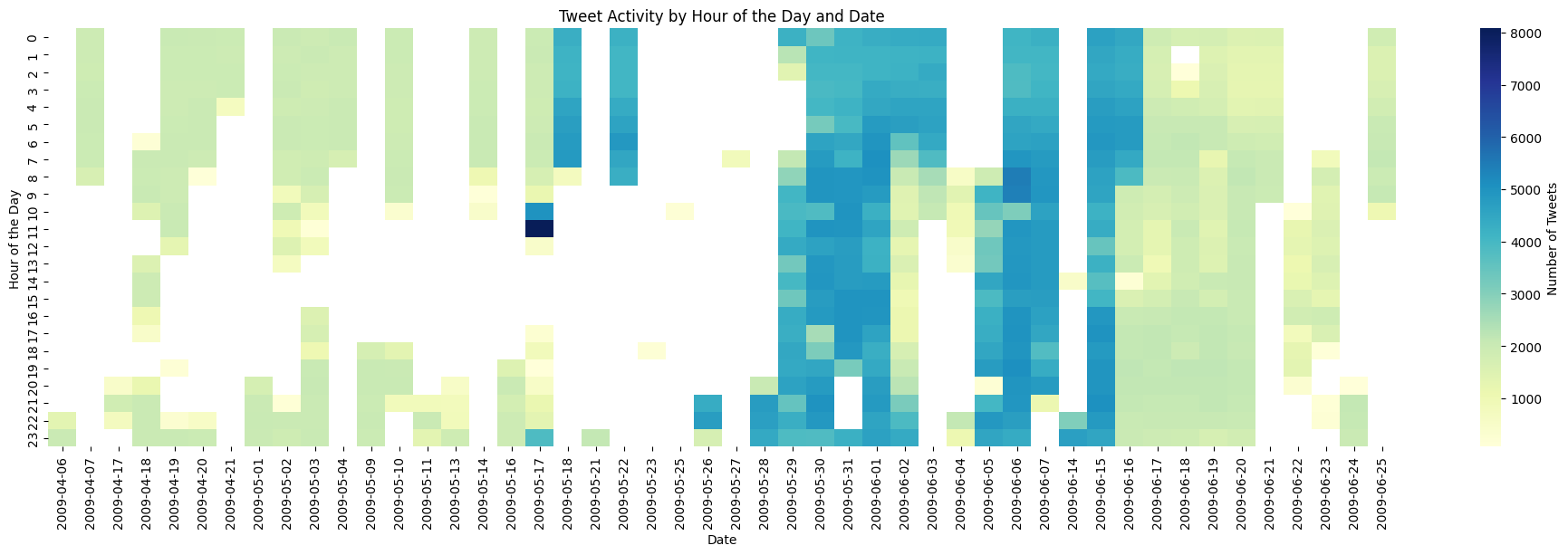


##### Figure 3. Quantity of Tweets compared with different ways to measure time: by month, day, hour, Day of the week and week.



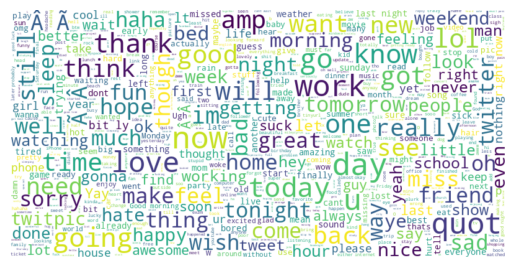
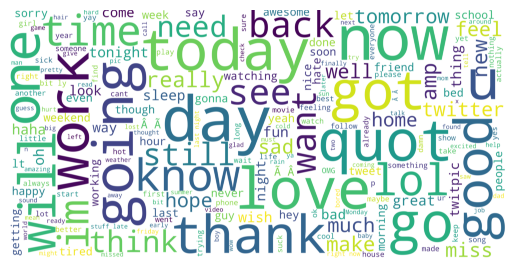
##### Figure 4. This heatmap reveals when most tweet activity occurs during the day and which days of the week.

##### 



##### Figure 5. This heatmap also shows when most tweet activity happens on different dates and times of the day.

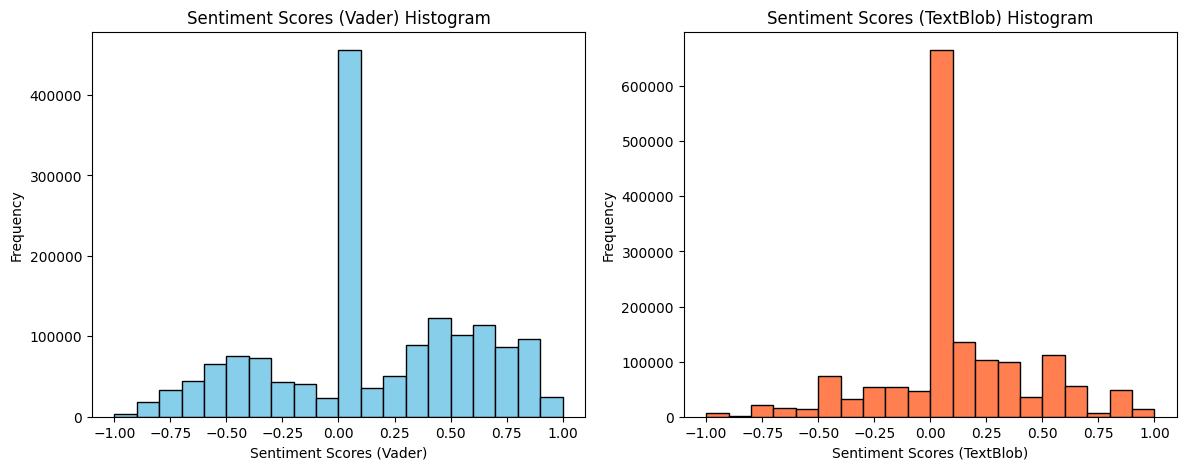
##### Figure 6. This heatmap provides insights into tweet activity patterns on the weekends



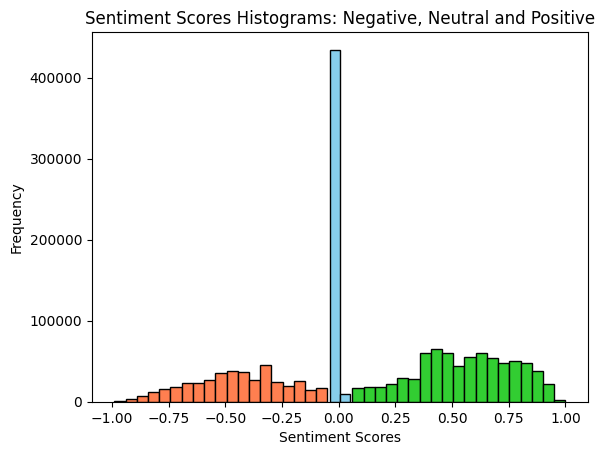
##### Figure 7. This word clouds visually represents the most frequently used words in the dataset for different samples of 10000 tweets. The size of each word corresponds to its frequency, larger words indicate greater occurrence. It provides insights into prevalent keywords and topics in the data, helping you identify common themes and discussions. A weakness of this visualization is that it can misguide the human eye, that is why the second and third plot has less variability in font scale and can make many more words visible. I

## Data Pre-processing

This is the critical phase of preparing the dataset to use in Natural Language Processing and to train Machine Learning Algorithms. First I labelled the entire dataset using a count-based method known as Vader (Valence Aware and sentiment Reasoner, a Parsimonious rule-based model tuned for social media). Vader is a human-validated sentiment analysis method developed for Twitter and social media contexts. VADER was created from a generalizable, valence-based, human-curated gold standard sentiment lexicon. It seemed to work better than TextBlob.



##### Figure 8. Histograms of sentiment scores assigned by Vader (left) and TextBlob (right) methods



##### Figure 9. Assigning categories to the Vader sentiment scores: the majority of tweets are neutral

The behaviour of the output variable (target variable, the sentiment score) will change a lot depending on the labelling method, This is the point of the analysis where most bias can be

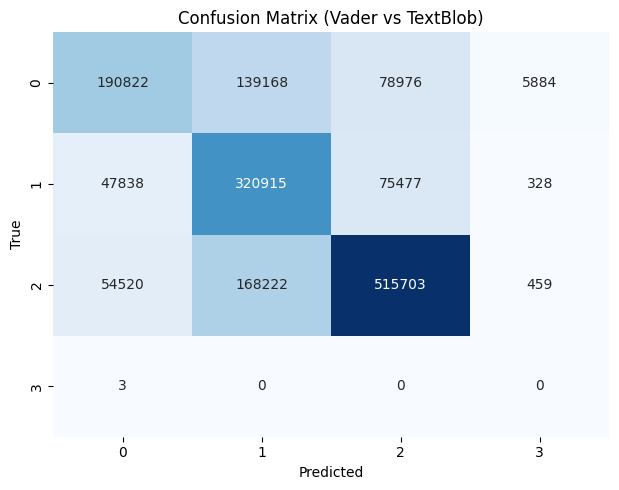
##### 

##### 

##### 

##### Figure 10. Comparison of the number of tweets classified as Positive, Neutral or Negative over time.

Introduced, thus, the importance of the selection of a well-tuned labelling algorithm and the difficulties of using a naive method for unsupervised learning.



##### Figure 11. Confusion matrix comparing the labelling methods Vader (True) and TextBlob (Predicted) Ideally, it should be a diagonal matrix. Also, ideally, we should have a better benchmark reference than Vader.

Then, the next phase of pre-processing was to create a vector representation of the words and word-embedding and used that representation to train the models.

# 

# 2. Analysis of Sentiment Change Over Time

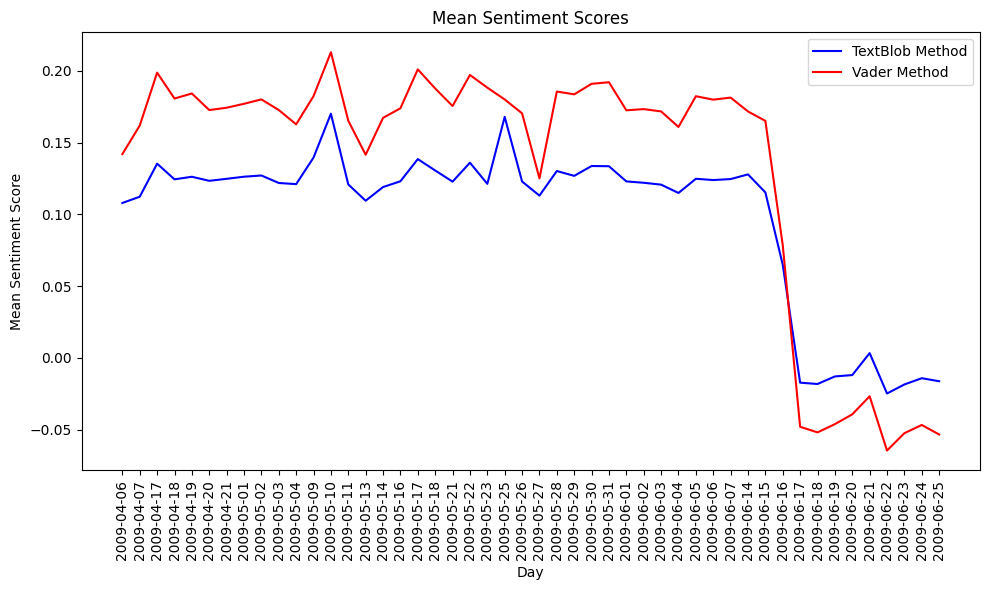
### 

### Research Questions

Four Research Questions were proposed at the beginning of the project:

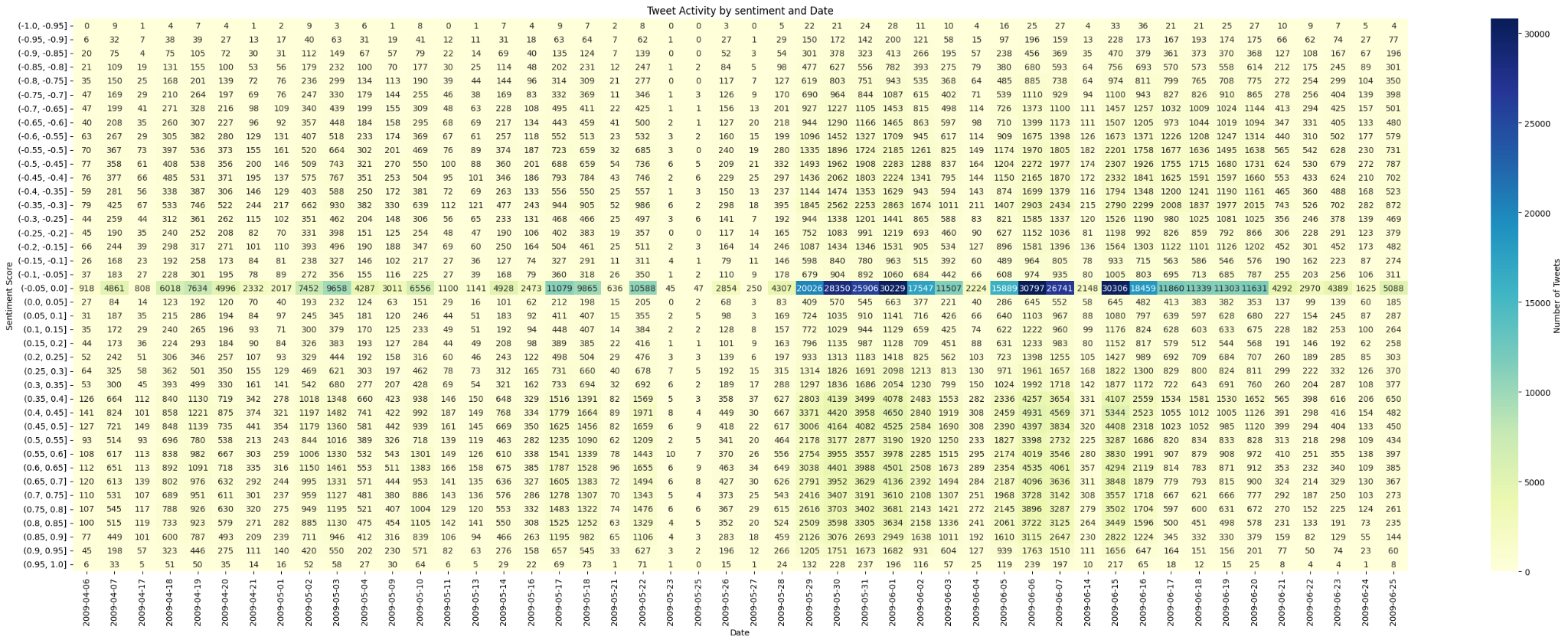
#### 1. What are the prevailing sentiment trends in tweets related to a specific topic or event on Twitter, and how have these sentiments evolved over time?

Average sentiment scores in tweets exhibit fluctuations over time. As in Figure 12, that shows varying levels of positive, negative, and neutral sentiment. Figure 14 shows the main trend: the majority of tweets are neutral. Also there is a tendency of tweets to decrease the average sentiment score over time to 0 (Neutral).



##### Figure 12. Behaviour of mean of sentiment scores for all tweets comparing the two labelling methods.

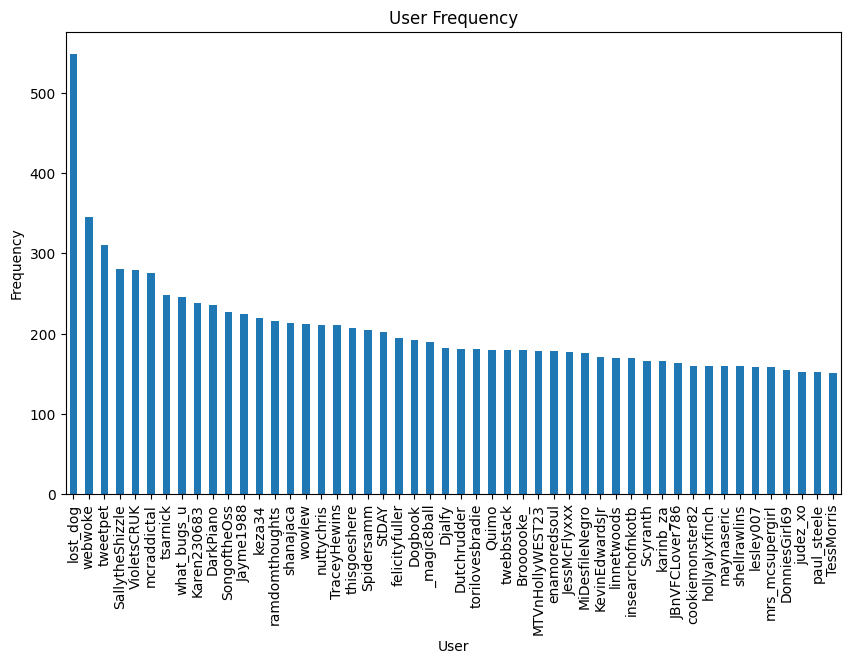
##### Figure 13. Boxplots of the distribution of sentiment scores over time for all tweets using only Vader method.



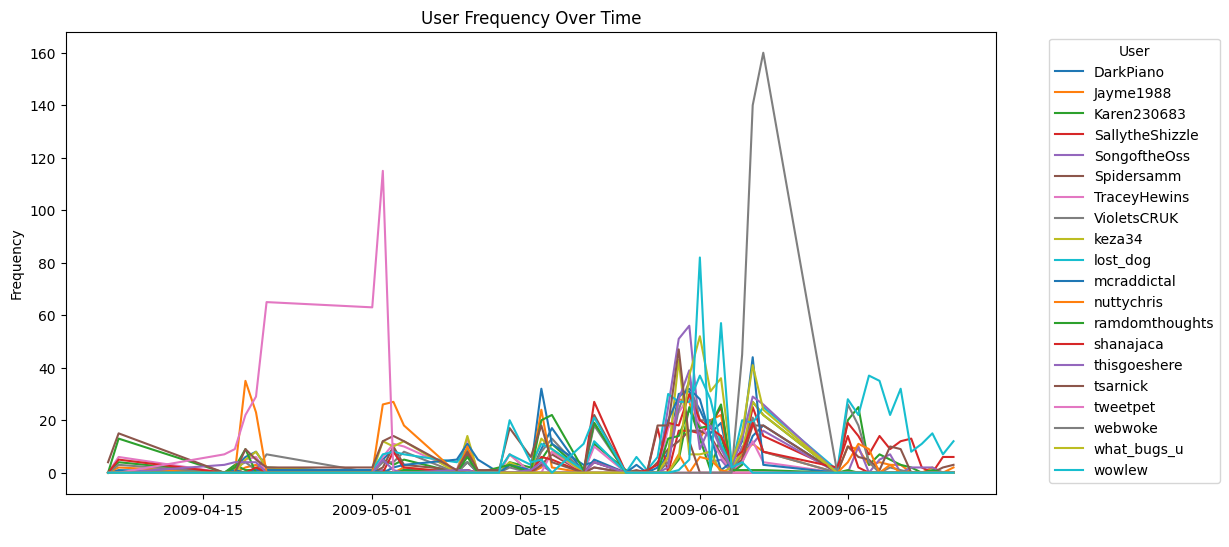
##### Figure 14. Heatmap of the volume of tweets posted that are classified in each sentiment score scale over time

#### 2. Can you identify the most influential users in the dataset by analysing their tweet engagement metrics and what factors contribute to their influence?

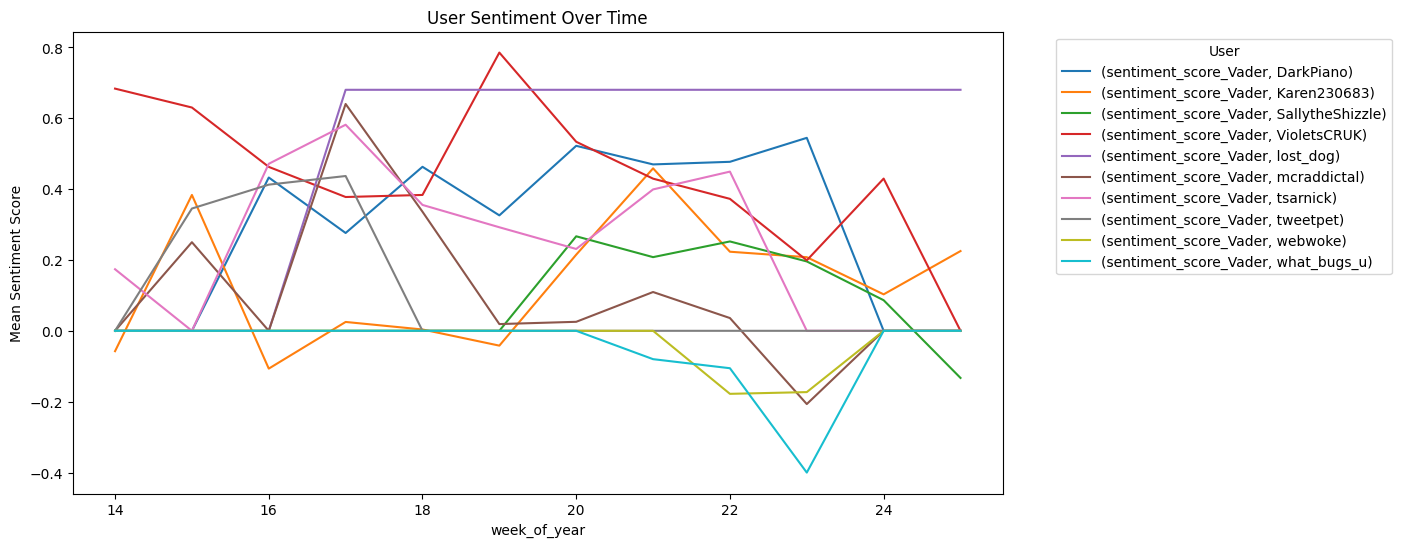
By analysing tweet engagement metrics, I identified the most influential users in the dataset. Visual representations like bar plots and line plots emphasize key contributors.



##### Figure 15. Histogram of the number of tweets by each of the 40 most prolific users in the platform.



##### Figure 16. Number of posts for the 20 most prolific users over time. They might be the most influential users, but not every day.



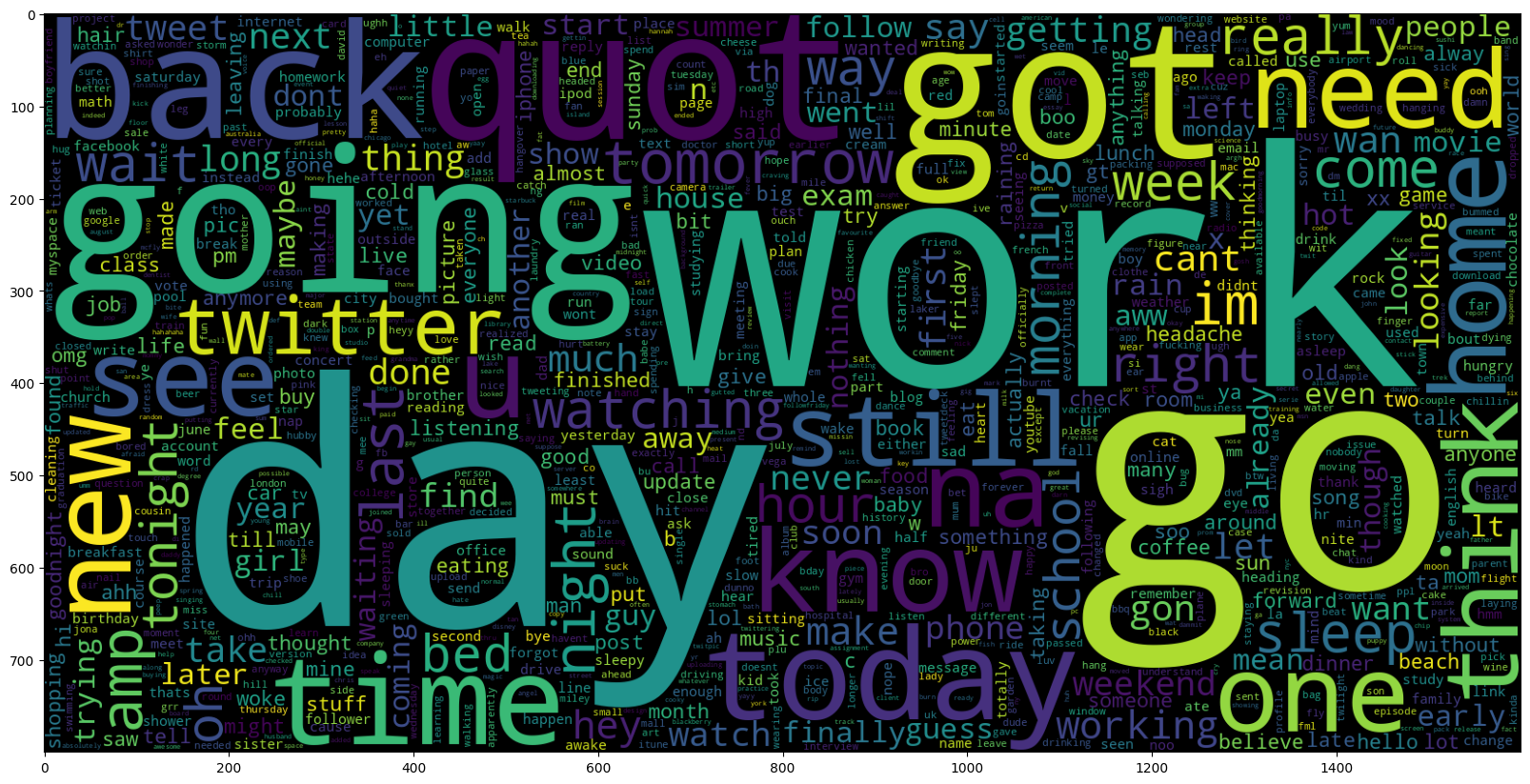
##### Figure 17. Mean Sentiment score variation over time for the 10 most prolific users in the platform.

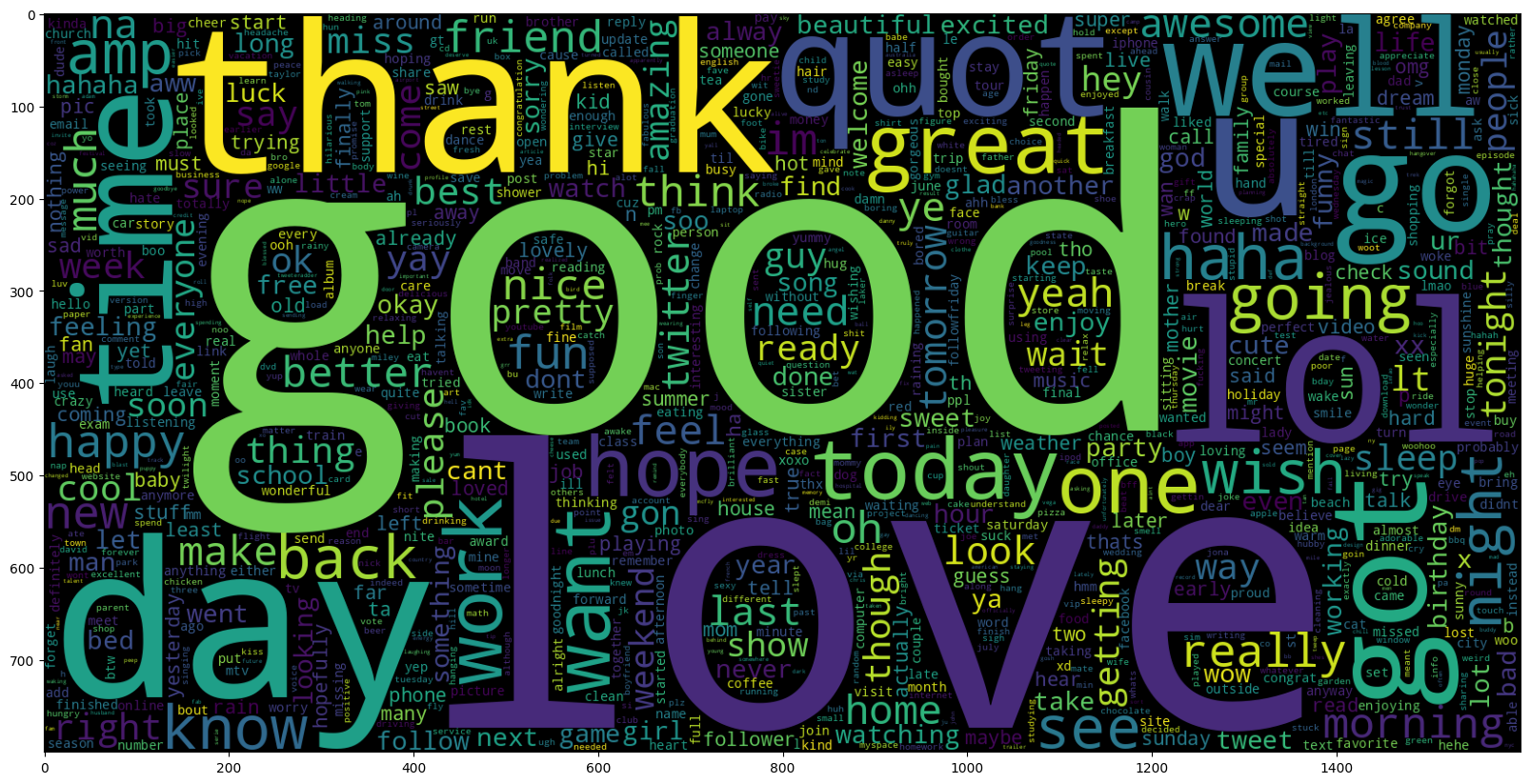
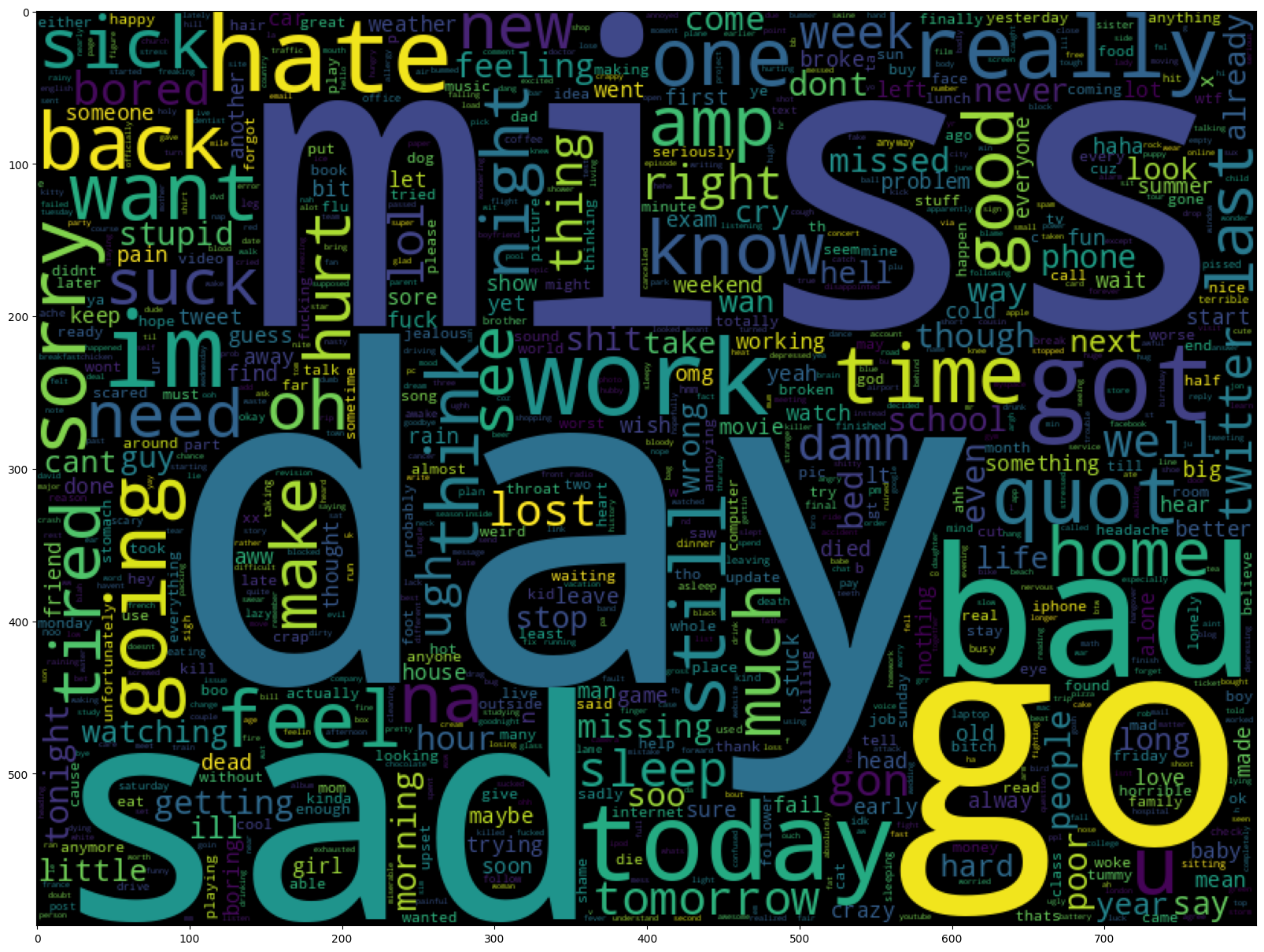
##### 

#### 

#### 3. Are there distinct patterns or clusters of topics in the Twitter dataset, and can you determine the factors that lead to the emergence of these patterns?

Visual representations like word clouds show interesting patterns: many words are present in negative, neutral and positive posts to a great extent, for example the words *“day”, “going”, “go’, “got”*. By examining tweet characteristics and correlation analysis, we identified that factors such as keyword frequency, user engagement, and trending events strongly contribute to the emergence of these patterns. This analysis provides insights into the underlying structure and factors that influence conversations on Twitter.





##### Figure 18. Word Clouds of Negative (left) Neutral (Upper right) and Positive (Lower right) Tweets.

#### 4. What are the prevailing sentiment trends in the tweets in the dataset, and how have these sentiments evolved over time?

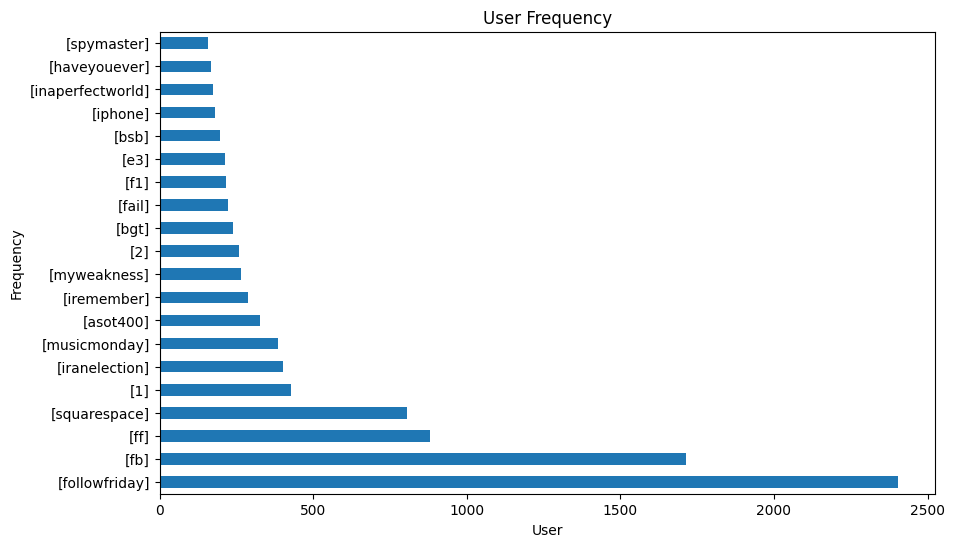
Analysis of my project's tweet dataset revealed dynamic changes in sentiment over time. I observed that sentiment is influenced by external events and factors, with trends varying between positive, negative and neutral sentiment. This knowledge provides information about the mood of Twitter users and the potential impact of events on sentiment.

##### 

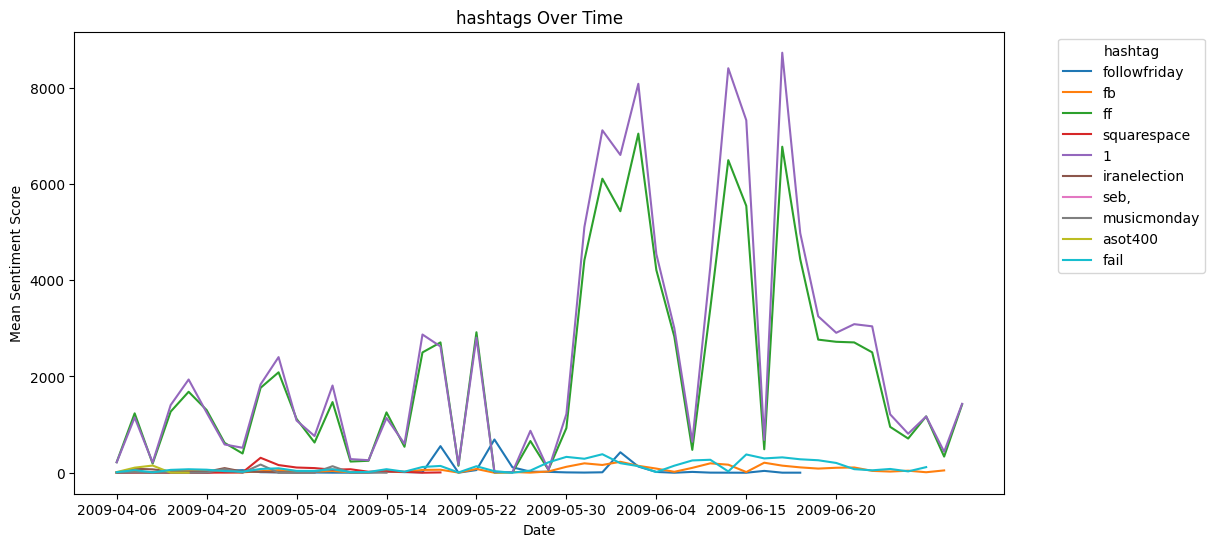
##### Figure 19. Number of posts that include the 10 most mentioned tags.

##### 

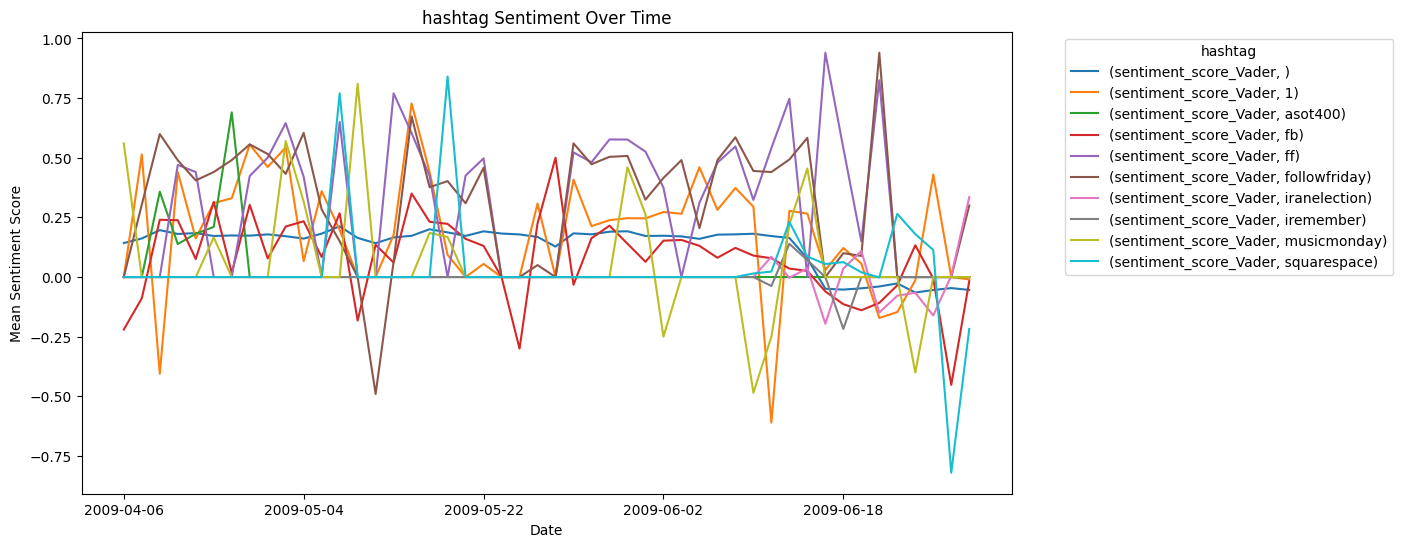
##### Figure 20. Average sentiment of posts that include the 10 most mentioned tags



##### Figure 21. Most common hashtags. Some of them are not very informative.



##### Figure 22. Volume of tweets of the 10 most common hashtags.



##### Figure 23. Sentiment variation of the 10 most common hashtags.

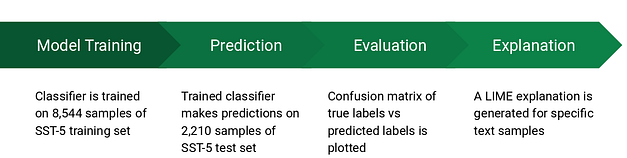
Hashtags can be interpreted as trending topics. As is visible in Figures 22 and 23, the volume and sentiment of hashtags is very variable, and it follows the main trends of change in volume and the trend to neutral sentiment.

# 3. Forecast of Sentiment

First, some classifiers were tested. I used five different models, respectively:

* **Bernoulli Naive Bayes Classifier**
* **SVM (Support Vector Machine)**
* **Logistic Regression**
* **Linear Support Vector Classification (LinearSVC)**
* **Classification using Ensemble (RandomForestClassifier)**

The idea behind choosing these models is to test all the classifiers in the dataset, from the simplest to the complex models, and then I will try to find the one that offers the best performance among them. A general workflow for model training and evaluation is shown below.

Figure 24. Sentiment classification: Training & Evaluation pipeline:

The process of classification involved the following steps:

1. Classifier is trained on a Train sample of the first 2 weeks of the dataset[[2]](#footnote-2).
2. Trained Classifier makes predictions on a Test Sample of the first weeks of the dataset
3. A confusion matrix comparing Vader Labelling with predicted labels is plotted
4. A ROC curve is plotted to evaluate how much information is used and lost

##### 

## Assessments and Measures

After training the model, I will apply the evaluation measures to check how the model is performing. Accordingly, I use the following evaluation parameters to check the performance of the models respectively:

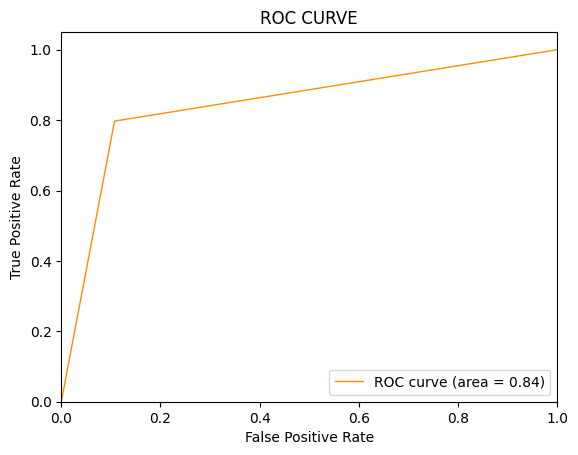
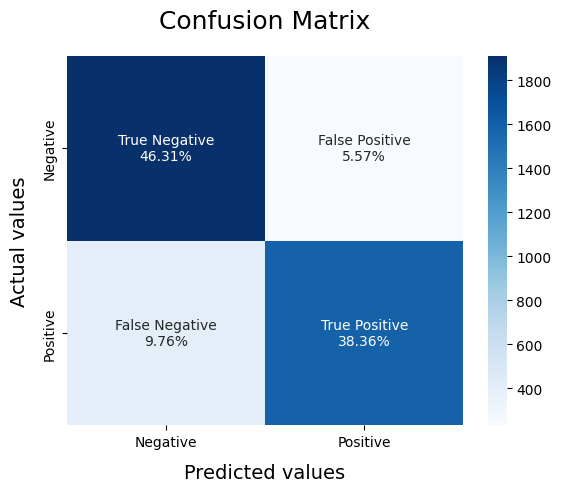
* Accuracy Score: Accuracy score is a metric that quantifies the overall accuracy of a model's predictions, expressed as the proportion of correctly predicted instances out of the total number of instances.
* Confusion Matrix with Plot: A Graphed Confusion Matrix is ​​a visual representation of a model's performance, categorizing predictions into true positives, true negatives, false positives, and false negatives to evaluate classification accuracy.
* ROC-AUC Curve: The ROC-AUC Curve is a graphical measure that illustrates the ability of a binary classification model to differentiate between positive and negative classes. The Area Under the Curve (AUC) quantifies the discrimination performance of the model, with a higher AUC indicating better discrimination ability.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Precision | Recall | F1-Score | Support |
| Bernoulli Naive Bayes | 0 | 0.80 | 0.87 | 0.83 | 2142 |
| 1 | 0.85 | 0.76 | 0.80 | 1987 |
| accuracy |  |  | 0.82 | 4129 |
| Support Vector Machine | 0 | 0.82 | 0.87 | 0.85 | 2142 |
| 1 | 0.86 | 0.80 | 0.83 | 1987 |
| accuracy |  |  | 0.84 | 4129 |
| Logistic Regression | 0 | 0.82 | 0.89 | 0.86 | 2142 |
| 1 | 0.87 | 0.79 | 0.83 | 1987 |
| accuracy |  |  | 0.85 | 4129 |

##### Table 1. Comparison of the algorithms for prediction of sentiment labels against Vader labels training with just the first week of data.

##### Figure 25. Confusion Matrix and Roc Curve of the Bernoulli Naive Bayes Classifier

##### Figure 26. Confusion Matrix and Roc Curve of the Support Vector Machine Classifier

****Figure 27. Confusion Matrix and Roc Curve of the Logistic Regression Classifier

# 6. Presentation

Interactive Dynamic dashboard

This is a visual data presentation tool that provides interactive or real-time access to information, allowing users to explore, analyse and visualize data in a user-friendly and responsive way. It often combines data visualizations, charts, and widgets that can be manipulated and customized to support data-driven decision-making and insights.

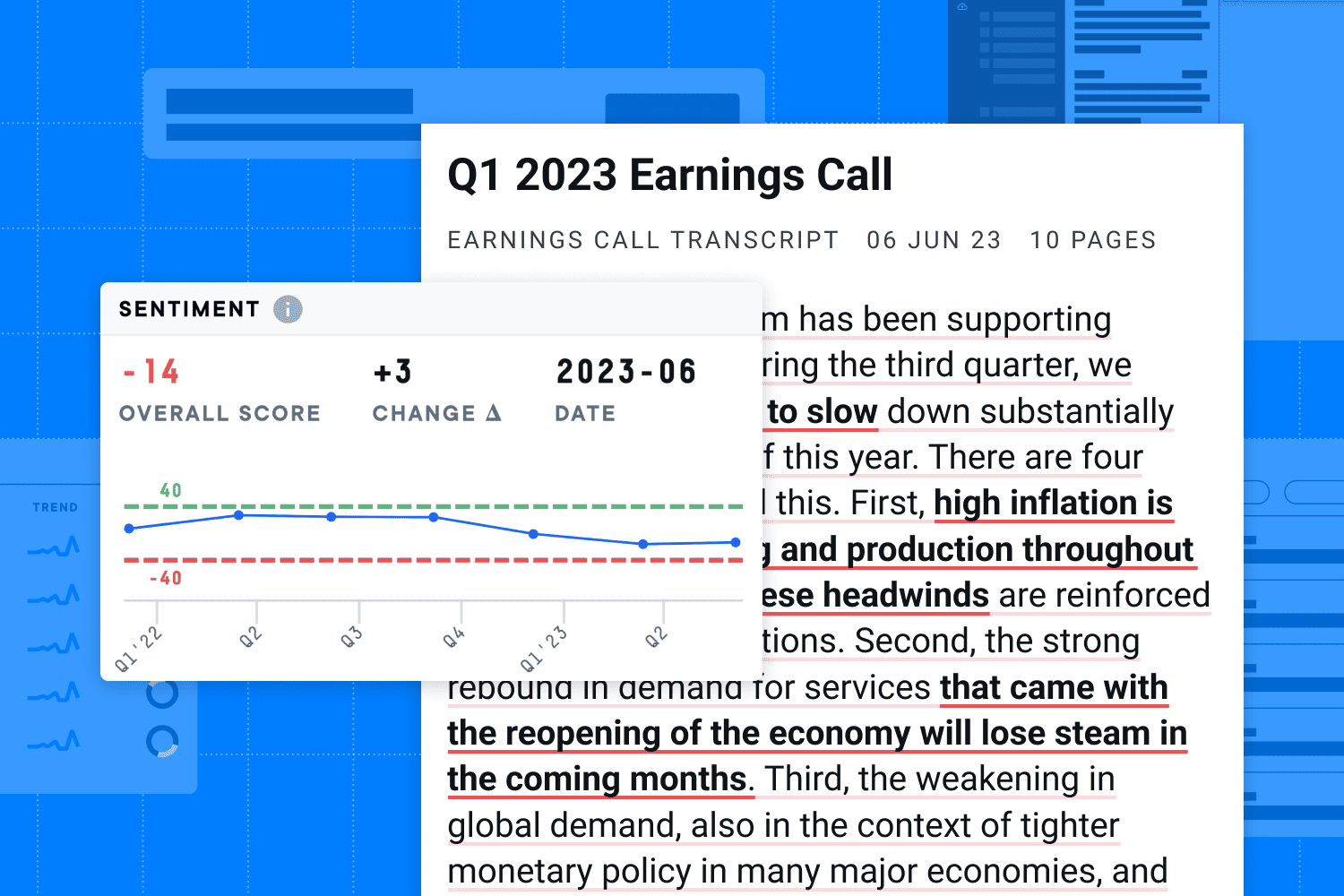


Figure . Proposed Dashboard

# 7. Results

## Methodology Discussion

In my project, "Twitter Data Mining and Sentiment Analysis: Uncovering Insights from the Tweet Dataset", we explore various aspects of sentiment trends, influential users, and topic clusters in the Twitter dataset.

My analysis of sentiment trends revealed valuable insights into the mood and emotions expressed by Twitter users. I have observed that sentiment fluctuates in response to external events such as holidays, product launches, and crisis situations. These observations highlight the dynamic nature of sentiment on social media.

Discovering that identifying influential users has practical implications for marketing and engagement strategies. Content type, hashtags, and sentiment have been recognized as essential factors shaping user influence. This insight can guide businesses and marketers as they seek effective online engagement and partnerships with influential voices in the Twitter community.

Exploration of topic clusters revealed distinct thematic patterns in Twitter conversations. Discovering that keyword frequency, user engagement, and trending events significantly contribute to the emergence of these clusters. This knowledge equips companies and researchers with tools to decode trending topics and create content that resonates with Twitter's diverse user base. In short, my project not only deepens the understanding of Twitter, but also provides practical insights for harnessing the potential of this dynamic social media platform.

## Conclusions

In the course of my project, "Twitter Data Mining and Sentiment Analysis: Uncovering Insights from the Project Tweets Dataset", I diligently follow the SEMMA (Sample, Exploration, Modification, Model, and Evaluation) framework to extract valuable insights from a set of substantial Twitter data. Twitter. The project was driven by clear goals: identify sentiment trends, measure user engagement, recognize influential users, and discover emerging topics in Twitter conversations.

Sample: The beginning of the project was to select a representative sample of Twitter's vast data set, which allows us to work with manageable data, maintaining its essential characteristics.

Explore: The exploration phase allowed us to gain a comprehensive understanding of the dataset's structure, content, and preliminary insights. This set the stage for further analysis.

Modify: Data pre-processing was a key step in which we meticulously cleaned, transformed, and structured the dataset. This process ensures that data is accurate, complete, and ready for analysis.

Model: Our data analysis phase was marked by the application of various techniques, from fundamental statistical analyses to advanced machine learning algorithms. These methods equip us to uncover sentiment trends, recognize influential voices, and dissect emerging discussions.

Assess: Analysis findings illuminated sentiment trends, identified influential voices, and expanded our understanding of dynamic online conversations. We used evaluation metrics to evaluate the quality and performance of our models.

The project had its challenges, including resolving missing data and ensuring data quality. However, the SEMMA framework provided a structured approach to successfully address these challenges.

As the project progressed, it laid the groundwork for future projects, such as deeper analysis of user engagement, real-time sentiment monitoring, and continuous exploration of Twitter's evolving landscapes are exciting possibilities.

In conclusion, "Twitter Data Mining and Sentiment Analysis" showcases the power of the SEMMA framework in handling complex data analysis projects in today's digital landscape. The project met the defined commercial objectives, highlighting the importance of research questions and the SEMMA framework. It exemplifies the value of data in making informed decisions and embracing the dynamic world of social media data.

In conclusion, this project is a journey into the world of Twitter data, where I seek to analyze and understand public sentiment and gain knowledge from sentiment analysis in a digital society. As I navigate the breadth of this dataset, I have the opportunity to discover trends, patterns and sentiments, contributing to a deeper understanding of the dynamic nature of social media discourse.

Our discoveries not only shed light on Twitter user sentiment but also provide valuable insights into trending topics, enhancing our understanding of the platform's influence on public discourse.

# References:

[1] Murray, A. (2013). Big data: How it is transforming business, society, and everyday life. New York: Bloomsbury Publishing USA.

[2] Baylor, D.; Breck, E.; Cheng, H.T.; Fiedel, N.; Foo, C.Y.; Haque, Z. Tfx: A tensorflow-based production-scale machine learning platform. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017.

[3] Setareh, H.; Deger, M.; Petersen, C.C.H.; Gerstner, W. Cortical Dynamics in Presence of Assemblies of Densely Connected Weight-Hub Neurons. *Front. Comput. Neurosci.* **2017**, *11*, 52.

[4] Curry, B.; Morgan, P.; Beynon, M. Neural networks and flexible approximations. IMA J. Manag. Math. 2000, 11, 19–35.

[5] Sarma, S.; Brock, D.; Ashton, K. The Networked Physical World; TR MIT-AUTOID-WH-001; MIT: Cambridge, MA, USA, 2000.

[6] Kelleher, J. D., & Tierney, B. (2018). Data science: An introduction. Boca Raton: CRC Press.

[7] Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. Sebastopol: O'Reilly Media, Inc.

[8] Peltola, M.J.; Forssman, L.; Puura, K.; Van Ijzendoorn, M.H.; Leppänen, J.M. Attention to Faces Expressing Negative Emotion at 7 Months Predicts Attachment Security at 14 Months. *Child Dev.* **2015**, *86*, 1321–1332.

[9] Anderson, J. R. (2016). Modern data science with R. Boca Raton: CRC Press.

[10] Efron, B., & Hastie, T. (2016). Computer age statistical inference: Algorithms, evidence, and data science. Cambridge: Cambridge University Press.

[11] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge. Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA,

[12] Bing Liu. "Sentiment Analysis and Subjectivity." A chapter in Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

[13] Ribeiro, F.N. et al. (2016) Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods - EPJ data science, SpringerOpen.

[14]Ortega, R.; Fonseca, A.; Montoyo, A. SSA-UO: Unsupervised Twitter sentiment analysis. In Proceedings of the Second Joint Conference on Lexical and Computational, Atlanta, GA, USA, 13–14 June 2013; pp. 501–507

[15] Laney, Douglas. 2001. “3-D Data Management: Controlling Data Volume, Velocity and Variety.” META Group Research Note, February, 6. Retrieved January 28, 2015 (http://gtnr.it/1bKflKH).

[16]Simeon, C.; Hamilton, H.J.; Hilderman, R.J. Word segmentation algorithms with lexical resources for hashtag classification. In Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17 October 2016; pp. 743–751.

[17] Artificial intelligence applications and innovations: 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006, June 7-9, 2006, Athens, Greece

[18]Javed, A.; Lee, B.S. Sense-level semantic clustering of hashtags in social media. In Proceedings of the 3rd Annual International Symposium on Information Management and Big Data (SIMBig), Lima, Perú, 1–3 September 2016; pp. 140–149. [[**Google Scholar**](https://scholar.google.com/scholar_lookup?title=Sense-level+semantic+clustering+of+hashtags+in+social+media&conference=Proceedings+of+the+3rd+Annual+International+Symposium+on+Information+Management+and+Big+Data+(SIMBig)&author=Javed,+A.&author=Lee,+B.S.&publication_year=2016&pages=140%E2%80%93149)]

**Websites referred in the notebook:**

* *https://www.cio.com/article/189218/what-is-sentiment-analysis-using-nlp-and-ml-to-extract-meaning.html* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* *https://www.linkedin.com/advice/0/what-some-best-practices-adapting-general-sentiment* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* *https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* W&oacute;jcik, R. (2020) *https://huggingface.co/blog/sentiment-analysis-python*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* W&oacute;jcik, R. (2020) *https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* W&oacute;jcik, R. (2020) *https://www.analyticsvidhya.com/blog/2021/12/different-methods-for-calculating-sentiment-score-of-text/*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* *https://medium.com/analytics-vidhya/sentiment-analysis-with-vader-label-the-unlabeled-data-8dd785225166* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* *https://stackoverflow.com/questions/61185290/is-it-possible-to-do-sentiment-analysis-of-unlabelled-text-using-word2vec-model* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* *https://stackoverflow.com/questions/63024842/how-to-assign-labels-score-to-data-using-machine-learning* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483 (Accessed: 01 November 2023).
* *https://doi.org/10.1140/epjds/s13688-016-0085-1* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* *https://plainenglish.io/blog/twitter-sentiment-analysis-using-vader-tweepy-b2a62fba151e* (2020) *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* W&oacute;jcik, R. (2020) *https://textblob.readthedocs.io/en/dev/quickstart.html*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* W&oacute;jcik, R. (2020) *https://medium.com/analytics-vidhya/sentiment-analysis-with-vader-label-the-unlabeled-data-8dd785225166*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* W&oacute;jcik, R. (2020) *https://stackabuse.com/python-for-nlp-introduction-to-the-textblob-library/*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483 (Accessed: 01 November 2023).
* W&oacute;jcik, R. (2020) *https://towardsdatascience.com/unsupervised-text-classification-with-lbl2vec-6c5e040354de*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* W&oacute;jcik, R. (2020) *Unsupervised sentiment analysis*, *Medium*. Available at: https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483
* Wilame (2023) *How to label text for sentiment analysis - good practises*, *Medium*. Available at: https://towardsdatascience.com/how-to-label-text-for-sentiment-analysis-good-practises-2dce9e470708
* *What is sentiment analysis? Using NLP and ML to extract meaning* (2021) *Maria Korolov*. Available at: https://www.mariakorolov.com/2021/what-is-sentiment-analysis-using-nlp-and-ml-to-extract-meaning/

# 

# 

# Annex 1. Hadoop Commands

Command to upload CSV to HDFS: First, upload the data in CSV file to HDFS. I use the hdfs dfs command to copy the file to HDFS. to an HDFS directory named input\_data.

hdfs dfs -mkdir -p /user/your\_username/input\_data

hdfs dfs -put input.csv /user/your\_username/input\_data/

Command to Run MapReduce Job: Use Hadoop Streaming to run a MapReduce job. It is needed to specify the mapper script, reducer script (if needed), input and output directories, and any other required configurations. This command runs the MapReduce job on the CSV data and writes the results to the HDFS output directory

mapreduce.py is the preprocessing program and mapper.py and reducer.py produce the sentiment prediction

hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-\*.jar \

-files mapreduce.py,mapper.py,reducer.py \

-mapper mapreduce.py \

-mapper mapper.py \

-reducer reducer.py \

-input /user/your\_username/input\_data/input.csv \

-output /user/your\_username/output\_data

Retrieve and View Results: After the job is completed, retrieve the results from HDFS This command will display the word counts on the console.

hdfs dfs -cat /user/your\_username/output\_data/part-00000

# Annex 2. Preprocessing program for MapReduce

MapReduce Python Program: A MapReduce program to process the data in Python with Hadoop Streaming. it takes all the preprocessing and then assigns a sentiment score for each text that is input.

import re

import string

import sys

import nltk

from nltk.corpus import stopwords

from nltk.stem import WordNetLemmatizer

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

# Initialize NLTK resources

lemmatizer = WordNetLemmatizer()

stop\_words = set(stopwords.words("english"))

# Initialize VADER sentiment analyzer

analyzer = SentimentIntensityAnalyzer()

def calculate\_sentiment\_score(text):

# Calculate sentiment score using VADER

sentiment = analyzer.polarity\_scores(text)

return sentiment['compound']

def preprocess\_text(text):

# Make the text lowercase

text = text.lower()

# Remove URLs from the text

text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)

# Remove hashtags and mentions

text = re.sub(r'#\w+|\@\w+', '', text)

# Remove repeating characters (e.g., loooove -> love)

text = re.sub(r'(.)\1{2,}', r'\1\1', text)

# Remove punctuation

text = text.translate(str.maketrans('', '', string.punctuation))

# Remove numeric characters

text = re.sub(r'\d', '', text)

# Tokenize the text

tokens = text.split()

# Remove stopwords and lemmatize

tokens = [lemmatizer.lemmatize(token) for token in tokens if token not in stop\_words]

return " ".join(tokens)

for line in sys.stdin:

line = line.strip()

# Assuming the text is in a CSV format with text in the first column

parts = line.split(",")

if len(parts) > 0:

text = parts[0]

preprocessed\_text = preprocess\_text(text)

print(preprocessed\_text)

sentiment\_score = calculate\_sentiment\_score(preprocessed\_text)

print(sentiment\_score)

# Annex 3. A program for Sentiment Prediction

**Mapper Phase:** In the mapper phase, you can read in each tweet (or its associated sentiment score) from your input file, and emit key-value pairs based on a key (e.g., user ID or timestamp). This allows you to group tweets by the same user or time period. Your key could contain the user or time information, and the value would contain the sentiment score.

# Read each tweet from input

for line in sys.stdin:

tweet\_data = line.strip().split(',')

user\_or\_time\_key = tweet\_data[0] # User ID or timestamp

sentiment\_score = tweet\_data[1] # Sentiment score

print(user\_or\_time\_key, sentiment\_score)

**Reducer Phase**: In the reducer phase, you can collect all the sentiment scores for a specific user or time period and make predictions using your pre-trained models.

from sklearn.externals import joblib # Or use your model library

# Load your pre-trained models

next\_week\_model = joblib.load("next\_week\_model.pkl")

next\_month\_model = joblib.load("next\_month\_model.pkl")

next\_quarter\_model = joblib.load("next\_quarter\_model.pkl")

for line in sys.stdin:

user\_or\_time\_key, sentiment\_scores = line.strip().split('\t')

sentiment\_scores = [float(score) for score in sentiment\_scores.split(',')]

# Use your models to make predictions

prediction\_next\_week = next\_week\_model.predict(sentiment\_scores)

prediction\_next\_month = next\_month\_model.predict(sentiment\_scores)

prediction\_next\_quarter = next\_quarter\_model.predict(sentiment\_scores)

print(user\_or\_time\_key, prediction\_next\_week, prediction\_next\_month, prediction\_next\_quarter)

1. Ribeiro, F.N., Araújo, M., Gonçalves, P. *et al.* SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* **5**, 23 (2016). https://doi.org/10.1140/epjds/s13688-016-0085-1 [↑](#footnote-ref-1)
2. Train set size: 36366 and Test set size: 9092. This was the maximum number of tweets that the environment in jupyter (Colab) was able to take to train and test. The TF-IDF matrix size was 45458 rows by 10000 features [↑](#footnote-ref-2)