Chapter 1

Introduction

Social media analytics involves the extraction and analysis of data from social media platforms to derive meaningful insights and patterns. This project specifically focuses on Twitter data, aiming to understand various trends and sentiments expressed by users. Through the utilization of techniques such as text mining, sentiment analysis, probability calculations, time series analysis, and hierarchical clustering, we aim to uncover significant patterns and answer critical business questions.

1.1 Overview

This project utilizes two distinct datasets obtained from Kaggle: a dataset of 1.6 million tweets and a daily website visitors dataset. The Twitter dataset includes fields such as tweet ID, date, user, and text, while the daily visitors dataset contains records of daily traffic measures. Through various analytical methods, we aim to gain insights into user interactions and trends on social media platforms.

- Text Mining: This involves pre-processing the tweet text to remove stop words, usernames, and retweets, allowing us to focus on the original content of the tweets. We then identify the most frequently used words and analyze the sentiment trends using the NRC library.
- Clustering Analysis: By creating a corpus from the text data, we employ hierarchical clustering to group words by their sentiments, visualized through a dendrogram.
- Probability Analysis: We calculate the PMF and CDF of tweet frequencies to understand how the likelihood of tweets changes over time.
- Time Series Analysis: This involves examining the trend of various sentiments over different days of the week and analyzing the number of tweets per day over a specified period.

1.2 Objectives

- **Text Mining:** Identify frequently used words and analyse the sentiment trends in tweets.
- **Clustering Analysis:** Group tweets by sentiments to observe prevalent themes and patterns.
- **Probability Analysis:** Determine the probability distributions of tweet occurrences over time.
- **Time Series Analysis:** Examine trends in tweet volumes and sentiments over different days of the week and months.

Chapter 2

System Requirement

2.1 Software Requirement

- Operating System: Windows 10 or later, macOS 10.15 (Catalina) or later, or a modern Linux distribution.
- R Programming Environment: R version 4.1.0 or later.
- R Packages: tidytext, tidyverse, stringr, lubridate, ggplot2, dplyr, readr, knitr, factoextra, fpc, clValid, cluster, nonlinearTseries, tm.
- Data Analysis Tools: RStudio IDE (version 1.4 or later) for enhanced development and debugging capabilities.
- Data Access and Storage: Access to Kaggle datasets (via URLs provided) and local storage or cloud-based storage with sufficient space for handling large datasets.

2.2 Hardware Requirement

- Processor: Intel Core i5 or equivalent, with at least 2 cores for smooth data processing.
- Memory (RAM): Minimum of 8 GB of RAM, with 16 GB recommended for handling large datasets efficiently.
- Storage: At least 50 GB of free disk space for storing datasets, R packages, and project files.
- Network Connectivity: Stable internet connection for downloading datasets, R packages, and updates, with at least 5 Mbps download speed recommended.
- Display: A monitor with at least 1920x1080 resolution for clear visualization of data and results.

Chapter 3

Methodology

3.1 System Design & Data Flow

The Social Media Analytics system processes Twitter and website visitor data to extract insights. Data is acquired from both sources, cleaned, and stored. Twitter data undergoes text mining, sentiment analysis, and clustering, while website visitor data is analyzed for traffic trends. Data flows from acquisition to cleaning and storage. Twitter data is processed for text mining and sentiment analysis, leading to insights that are visualized. Website visitor data is analyzed for time series trends and also visualized. Both outputs are presented to users for interpretation.

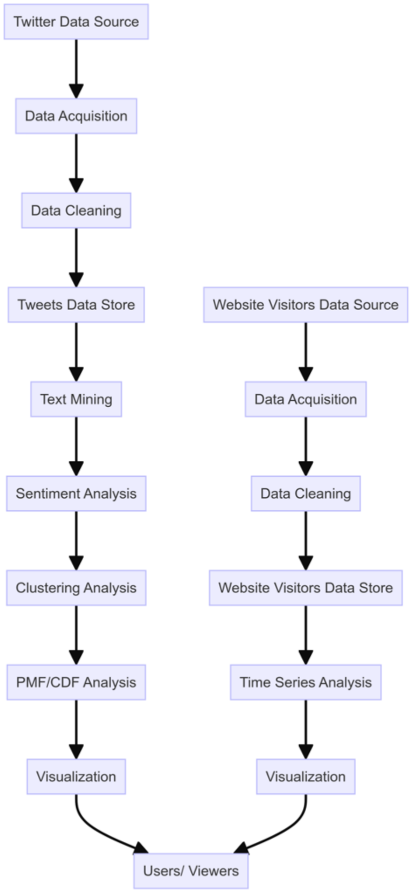


Figure 3.1 Data flow diagram

In this chapter, we outline the comprehensive methodology adopted to analyze social media data, particularly focusing on Twitter data and website visitor data. The analysis encompasses several critical steps: data acquisition, preprocessing, text mining, sentiment analysis, clustering, probability analysis, and time series analysis. Each step is crucial in transforming raw data into meaningful insights.

3.2 Data Acquisition

We acquired two datasets from Data Source (Kaggle):

- Twitter Data: This dataset contains 1.6 million tweets, including fields such as target (polarity of the tweet), ids, date, flag, user, and text. The data was sourced from Sentiment140. [1]
- Website Visitor Data: This dataset contains daily time series data of website traffic over five years, with fields including Row, Day, Day.Of.Week, Date, Page.Loads, Unique.Visits, First.Time.Visits, and Returning.Visits. The data was sourced from daily-website-visitors. [2]

3.3 Data Preprocessing

Twitter Data

- Loading Data: The Twitter data was loaded into R and appropriate column names were assigned.
- Cleaning Text Data: We removed URLs, retweets, mentions, and special characters to focus on the original content of the tweets.
- Tokenization: The tweets were tokenized into individual words, and common stop words were removed.

Website Visitor Data

- Loading Data: The website visitor data was loaded into R and kept ready for time series analysis.

3.4 Text Mining

- **Finding Frequently Used Unique Words**

  We used the ‘tidytext’ package to process the tweet text and identify the most frequently used unique words after removing common stop words and other non-essential text elements.

- **Sentiment Analysis**

  Using the NRC sentiment lexicon, we identified the sentiment associated with each word in the tweets. We then visualized the distribution of different sentiments across the dataset.

3.5 Clustering Analysis

Hierarchical Clustering: We performed hierarchical clustering on the words in the tweets to identify groups of words with similar sentiments. This involved creating a term-document matrix, calculating distances between words, and plotting a dendrogram to visualize the clusters.

3.6 Probability Analysis

Probability Mass Function (PMF) and Cumulative Distribution Function (CDF): We calculated the PMF and CDF for the frequency of tweets over time, providing insights into the probability distribution of tweet activity.

3.7 Time Series Analysis

- **Sentiment Trends by Day of the Week:**

  We analyzed the trends of different sentiments for each day of the week to identify patterns in how sentiments vary over time.

- **Trend Analysis of Number of Tweets:**

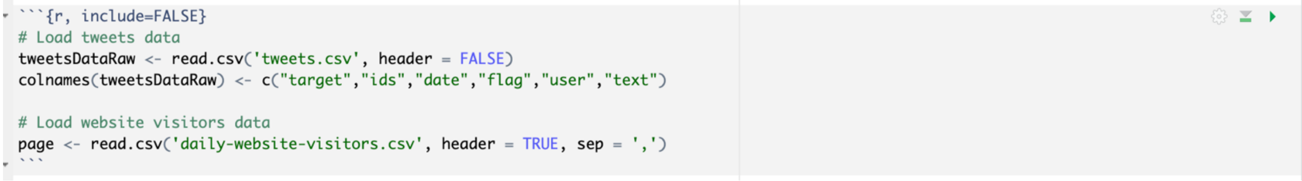
  We examined the number of tweets per day over a three-month period to understand the temporal dynamics of tweet activity.

Chapter 4

Implementation

4.1 Data Preparation

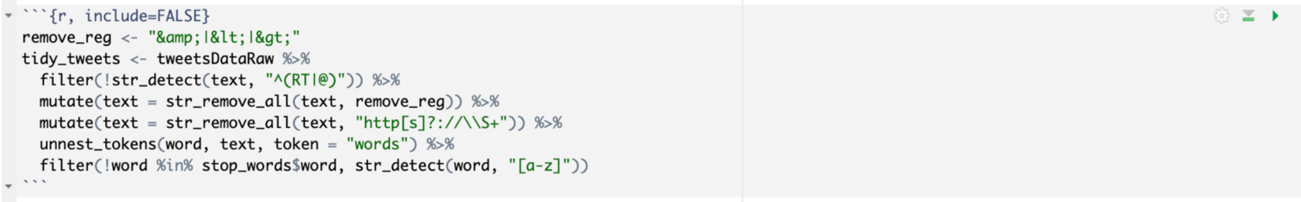
Data Acquisition: Data is acquired from two CSV files: tweets.csv for social media data and daily-website-visitors.csv for website traffic. The data is loaded and columns are renamed for better readability.



~~~
```{r, include=FALSE}
# Load tweets data
tweetsDataRaw <- read.csv('tweets.csv', header = FALSE)
colnames(tweetsDataRaw) <- c("target","ids","date","flag","user","text")

# Load website visitors data
page <- read.csv('daily-website-visitors.csv', header = TRUE, sep = ',')
```
~~~

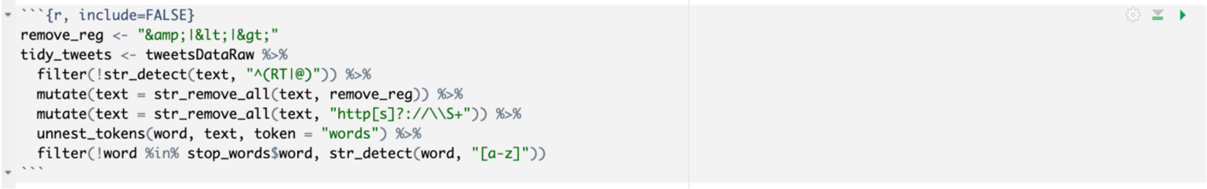
Data Preview: Previewing a subset of the data from both sources helps to verify the structure and content.



~~~
```{r, include=FALSE}
remove_reg <- "&amp;|&lt;|&gt;"
tidy_tweets <- tweetsDataRaw %>%
  filter(!str_detect(text, "^(RT|@)")) %>%
  mutate(text = str_remove_all(text, remove_reg)) %>%
  mutate(text = str_remove_all(text, "http[s]?://\\S+")) %>%
  unnest_tokens(word, text, token = "words") %>%
  filter(!word %in% stop_words$word, str_detect(word, "[a-z]"))
```
~~~

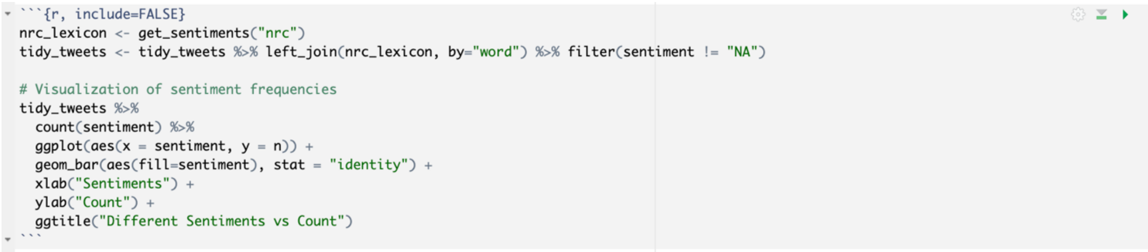
4.2 Text Mining

Data Cleaning and Tokenization: The tweet text is cleaned by removing URLs, special characters, and stop words. The text is then tokenized into individual words for analysis.



~~~
```{r, include=FALSE}
remove_reg <- "&amp;|&lt;|&gt;"
tidy_tweets <- tweetsDataRaw %>%
  filter(!str_detect(text, "^(RT|@)")) %>%
  mutate(text = str_remove_all(text, remove_reg)) %>%
  mutate(text = str_remove_all(text, "http[s]?://\\S+")) %>%
  unnest_tokens(word, text, token = "words") %>%
  filter(!word %in% stop_words$word, str_detect(word, "[a-z]"))
```
~~~

Sentiment Analysis: Sentiment analysis is performed using the NRC lexicon to categorize the words into different sentiments. The frequency of each sentiment is then visualized. [3]

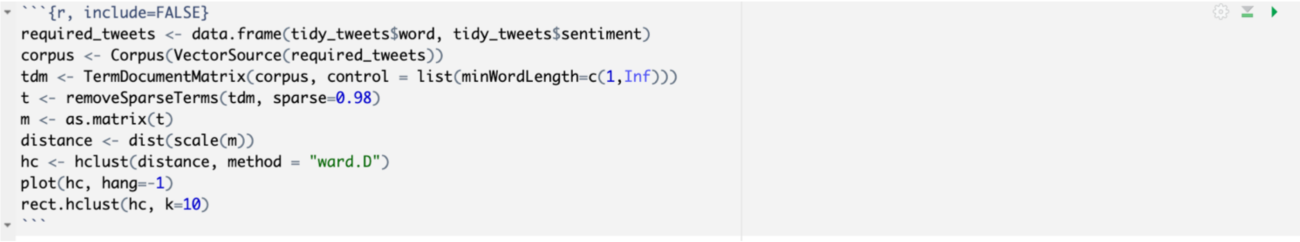


~~~
```{r, include=FALSE}
nrc_lexicon <- get_sentiments("nrc")
tidy_tweets <- tidy_tweets %>% left_join(nrc_lexicon, by="word") %>% filter(sentiment != "NA")

# Visualization of sentiment frequencies
tidy_tweets %>%
  count(sentiment) %>%
  ggplot(aes(x = sentiment, y = n)) +
  geom_bar(aes(fill=sentiment), stat = "identity") +
  xlab("Sentiments") +
  ylab("Count") +
  ggtitle("Different Sentiments vs Count")
```
~~~

4.3 Clustering Analysis

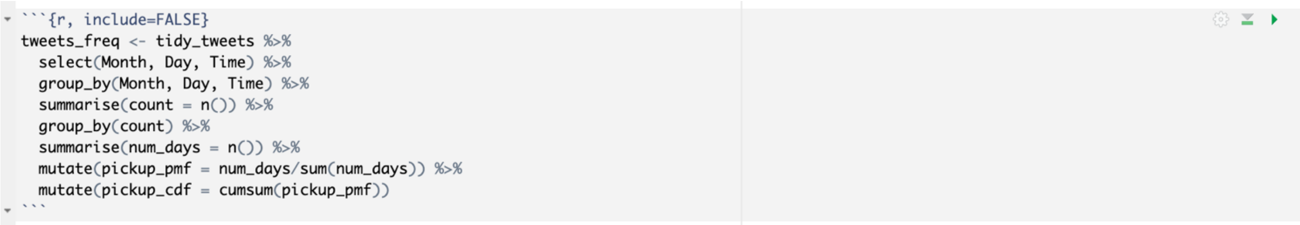
Hierarchical Clustering: Hierarchical clustering is applied to group words based on their sentiments. A term-document matrix is created, and clustering is visualized with dendrograms. [4]



~~~
```{r, include=FALSE}
required_tweets <- data.frame(tidy_tweets$word, tidy_tweets$sentiment)
corpus <- Corpus(VectorSource(required_tweets))
tdm <- TermDocumentMatrix(corpus, control = list(minWordLength=c(1,Inf)))
t <- removeSparseTerms(tdm, sparse=0.98)
m <- as.matrix(t)
distance <- dist(scale(m))
hc <- hclust(distance, method = "ward.D")
plot(hc, hang=-1)
rect.hclust(hc, k=10)
```
~~~

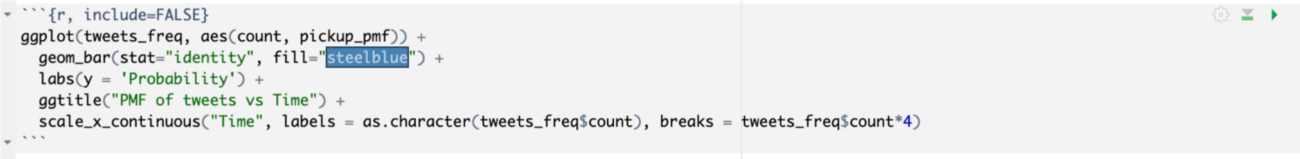
4.4 Probability and Statistical Analysis

PMF and CDF Calculation: The Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) of tweet frequencies are calculated to analyze their distribution.



~~~
```{r, include=FALSE}
tweets_freq <- tidy_tweets %>%
  select(Month, Day, Time) %>%
  group_by(Month, Day, Time) %>%
  summarise(count = n()) %>%
  group_by(count) %>%
  summarise(num_days = n()) %>%
  mutate(pickup_pmf = num_days/sum(num_days)) %>%
  mutate(pickup_cdf = cumsum(pickup_pmf))
```
~~~

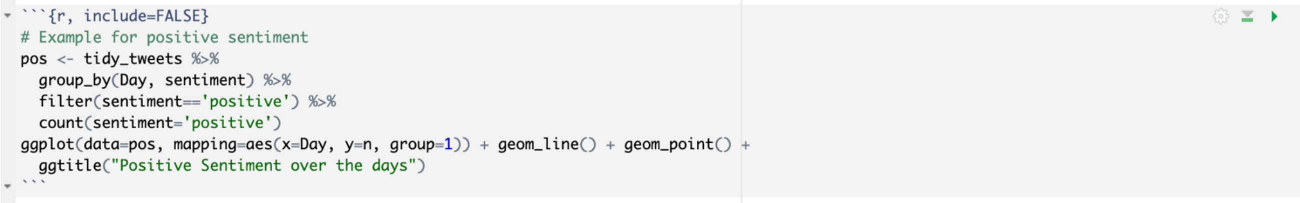
PMF Visualization: The PMF is visualized over time to understand the distribution of tweet counts.



~~~
```{r, include=FALSE}
ggplot(tweets_freq, aes(count, pickup_pmf)) +
  geom_bar(stat="identity", fill="steelblue") +
  labs(y = 'Probability') +
  ggtitle("PMF of tweets vs Time") +
  scale_x_continuous("Time", labels = as.character(tweets_freq$count), breaks = tweets_freq$count*4)
```
~~~

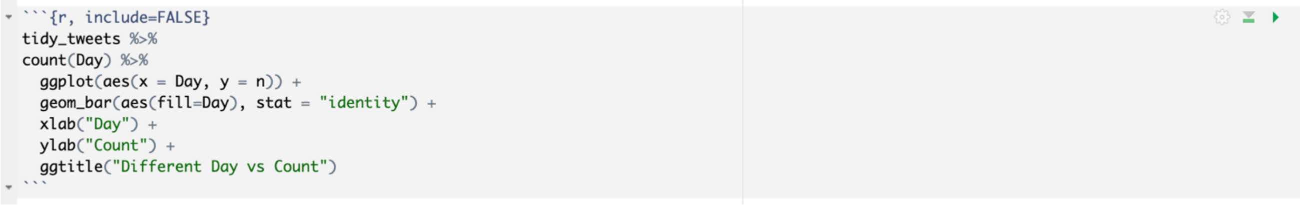
4.5 Time Series Analysis

Sentiment Trends: Trend analysis is conducted to observe sentiment variations across different days of the week. Various sentiments such as positive, negative, and joy are analyzed and visualized.



~~~
```{r, include=FALSE}
# Example for positive sentiment
pos <- tidy_tweets %>%
  group_by(Day, sentiment) %>%
  filter(sentiment=='positive') %>%
  count(sentiment='positive')
ggplot(data=pos, mapping=aes(x=Day, y=n, group=1)) + geom_line() + geom_point() +
  ggtitle("Positive Sentiment over the days")
```
~~~

Tweet Count Analysis: The number of tweets per day is analyzed to observe daily variations in tweet activity.



~~~
```{r, include=FALSE}
tidy_tweets %>%
count(Day) %>%
  ggplot(aes(x = Day, y = n)) +
  geom_bar(aes(fill=Day), stat = "identity") +
  xlab("Day") +
  ylab("Count") +
  ggtitle("Different Day vs Count")
```
~~~

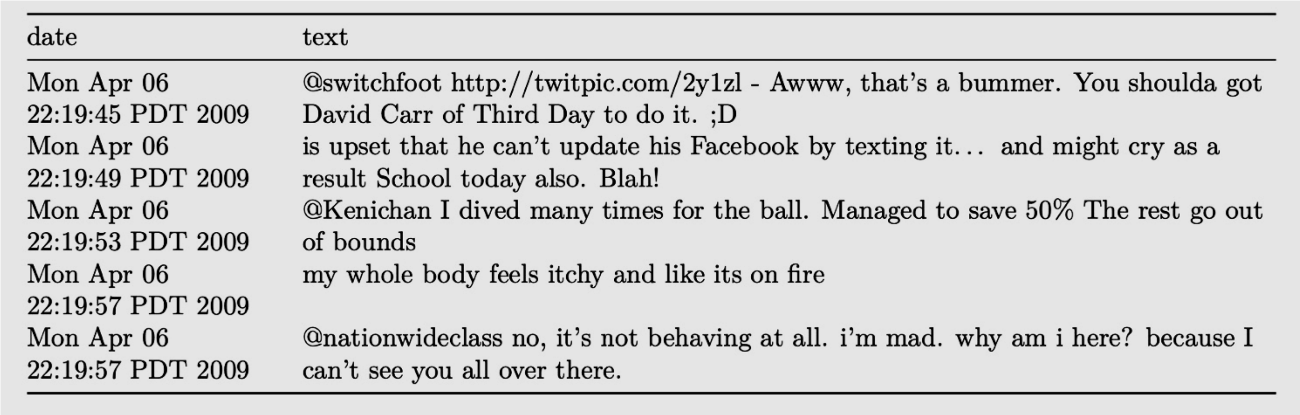
Chapter 5

Results

In this chapter, we present the findings derived from the analysis of social media and website visitor data. The results encompass a range of analytical techniques including text mining, sentiment analysis, clustering, probability and statistical analysis and time series analysis. Each section provides a detailed account of the methods applied, key observations, and visualizations that highlight significant patterns and insights from the data. This comprehensive analysis aims to uncover trends, behaviours, and underlying structures within the datasets, offering valuable insights into user interactions and sentiment dynamics over time.

5.1 Data Acquisition Results

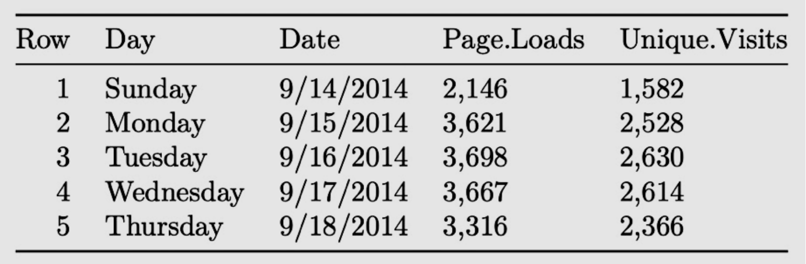
Tweets Data Preview: The tweets.csv file containing social media data was successfully loaded and previewed. The dataset includes columns such as date, text, and user information.



date	text
Mon Apr 06 22:19:45 PDT 2009	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
Mon Apr 06 22:19:49 PDT 2009	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
Mon Apr 06 22:19:53 PDT 2009	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
Mon Apr 06 22:19:57 PDT 2009	my whole body feels itchy and like its on fire
Mon Apr 06 22:19:57 PDT 2009	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.

Figure 5.1 Previewing few columns of Twitter user data set

Website Visitors Data Preview: The daily-website-visitors.csv file containing website traffic data was successfully loaded and previewed. This dataset includes columns such as date, page loads, and unique visits.



Row	Day	Date	Page.Loads	Unique.Visits
1	Sunday	9/14/2014	2,146	1,582
2	Monday	9/15/2014	3,621	2,528
3	Tuesday	9/16/2014	3,698	2,630
4	Wednesday	9/17/2014	3,667	2,614
5	Thursday	9/18/2014	3,316	2,366

Figure 5.2 Previewing few columns of Daily time series data set.

5.2 Text Mining Results

Frequently Used Unique Words

After cleaning and tokenizing the tweet text, the most frequently used unique words were identified and visualized.

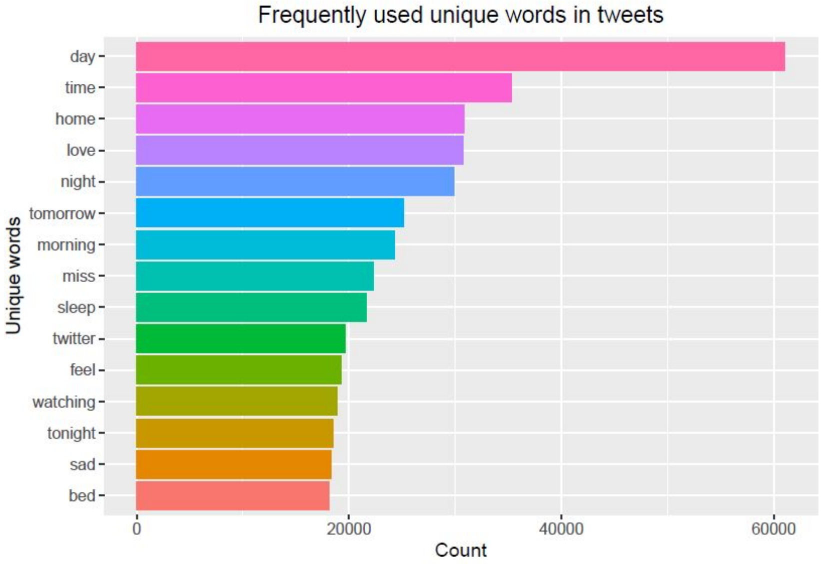


Figure 5.3 Finding the frequently used unique words

Sentiment Analysis

Sentiment analysis was performed to categorize words into different sentiments. The frequency of each sentiment was visualized.

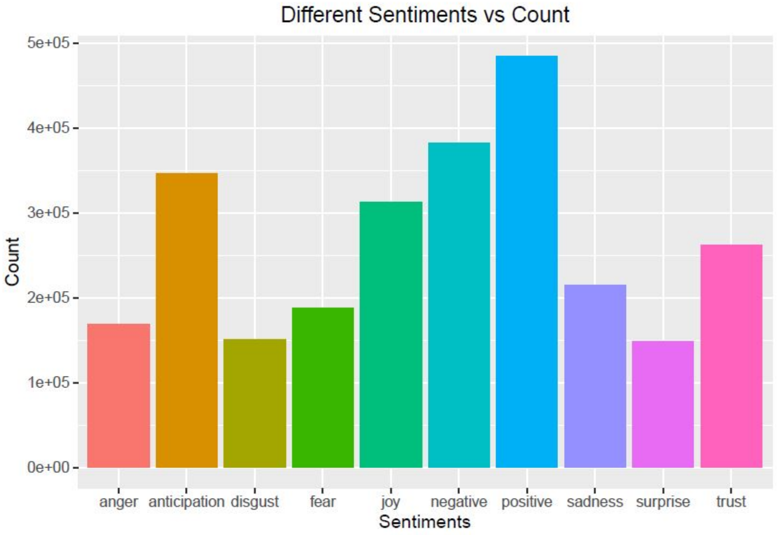


Figure 5.4 Sentimental Trends of Tweets

5.3 Clustering Analysis Results

Hierarchical Clustering

Hierarchical clustering was applied to group words based on their sentiments. The clustering results were visualized using a dendrogram

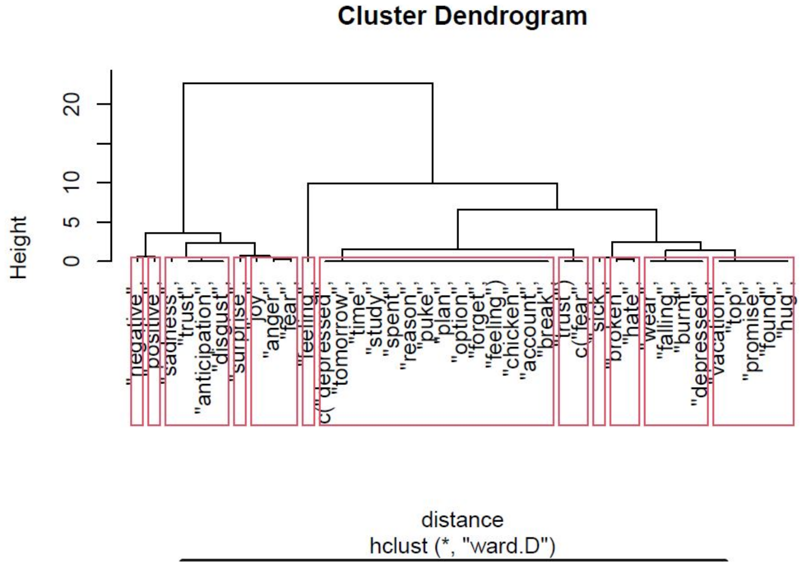
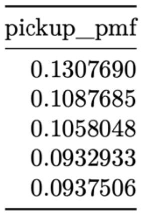


Figure 5.5 Hierarchical clustering words by sentiments

5.4 Probability and Statistical Analysis Results

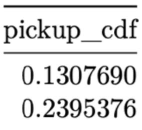
PMF and CDF Calculation

The Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) of tweet frequencies were calculated and the first few records were presented.

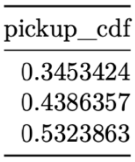


pickup_pmf
0.1307690
0.1087685
0.1058048
0.0932933
0.0937506

Figure 5.6 First 5 records of PMF of the tweet frequency



pickup_cdf
0.1307690
0.2395376



pickup_cdf
0.3453424
0.4386357
0.5323863

Figure 5.7 First 5 records of CDF of the tweet frequency

PMF Visualization

The PMF of tweet frequencies over time was visualized to understand the distribution of tweet counts.

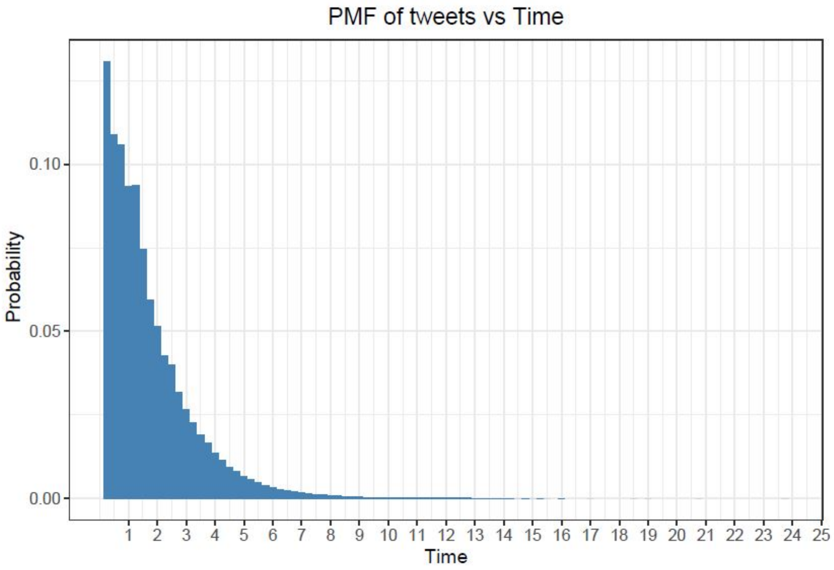


Figure 5.8 Probability Mass Function over Time

5.5 Time Series Analysis Results

Sentiment Trends

Sentiment trends were analyzed for different days of the week, and the results were visualized for sentiments such as positive, negative, anticipation, joy, and trust.

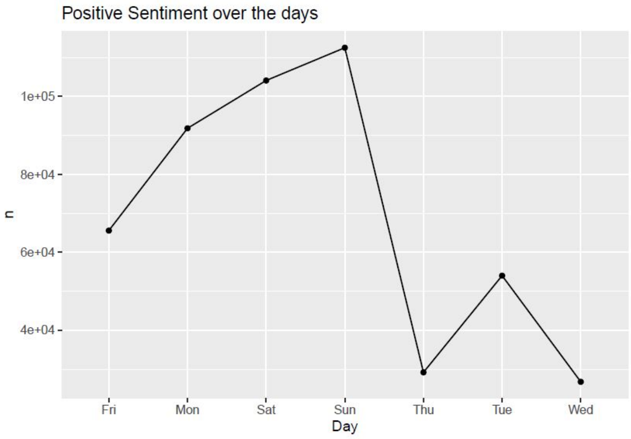


Figure 5.9 Positive Sentiment over the days

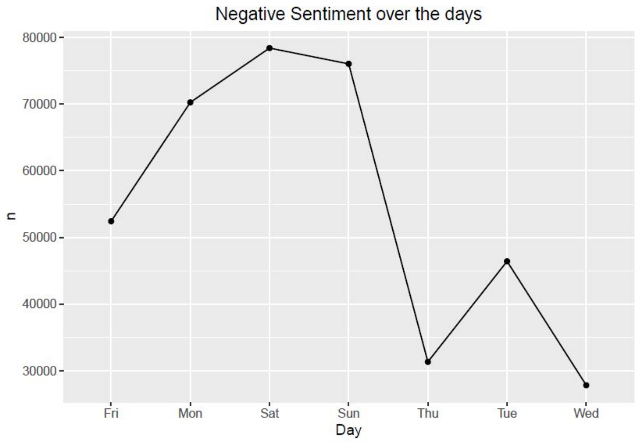


Figure 5.10 Negative Sentiment over the days

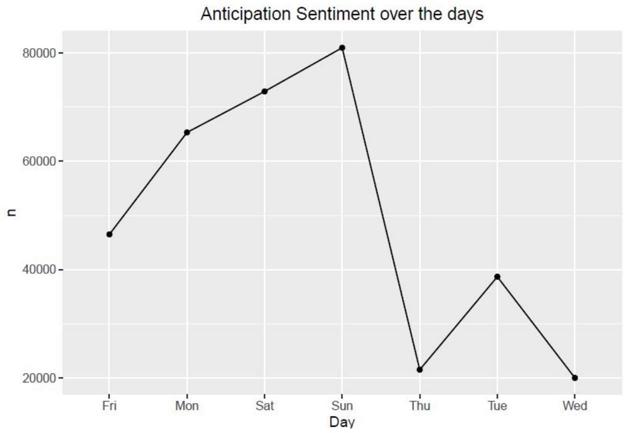


Figure 5.11 Anticipation Sentiment over the days

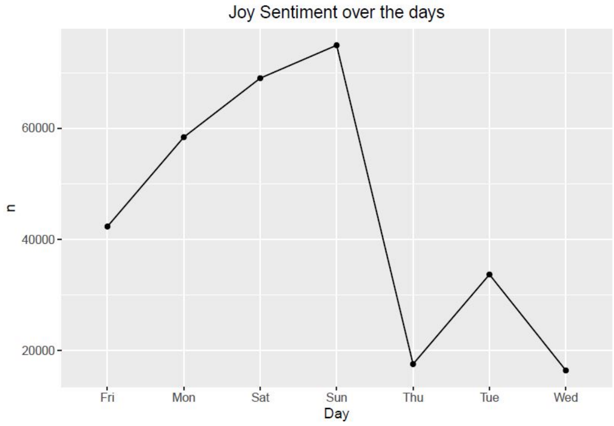


Figure 5.12 Joy Sentiment over the days

Tweet Count Analysis

The number of tweets per day was analyzed to observe daily variations in tweet activity.

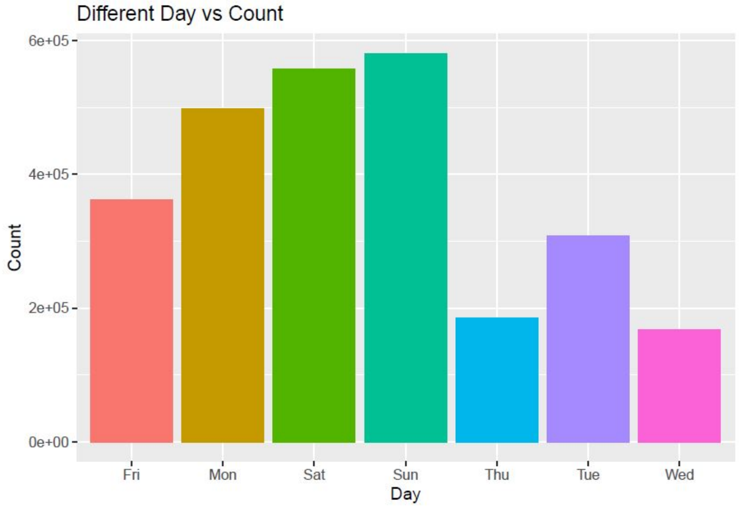


Figure 5.13 Tweet Count Analysis

Chapter 6

Challenges

6.1 Data Collection Issues

One of the primary challenges faced during this project was the acquisition of high-quality and relevant data. Social media platforms often restrict access to their APIs, limiting the amount and type of data that can be retrieved. Additionally, the data collected sometimes contained missing values, inconsistencies, or irrelevant information that needed extensive preprocessing to ensure accuracy and reliability in the analysis.

6.2 Data Preprocessing and Cleaning

The raw data from social media and website visits required significant preprocessing. This included handling missing values, removing duplicates, normalizing text data, and converting data into suitable formats for analysis. The preprocessing step was particularly challenging due to the unstructured nature of social media text, which included slang, emojis, and various shorthand notations that complicated the tokenization and sentiment analysis processes.

6.3 Sentiment Analysis Limitations

Sentiment analysis using predefined lexicons such as NRC has inherent limitations. These lexicons might not fully capture the context-specific nuances of the language used in social media posts. Sarcasm, irony, and slang can lead to misclassification of sentiments, thereby affecting the accuracy of the results. Additionally, multilingual data posed another layer of complexity as sentiment analysis tools are often optimized for English text.

6.4 Clustering Complexity

Performing clustering analysis, particularly hierarchical clustering, presented challenges in terms of determining the optimal number of clusters and interpreting the clusters meaningfully. The high dimensionality of text data and the need for dimensionality reduction techniques added to the complexity. Ensuring that the clusters were both statistically valid and contextually relevant required careful selection of clustering parameters and validation techniques.

6.5 Time Series Analysis Difficulties

Time series analysis of social media and website visitor data involved dealing with irregular time intervals and missing timestamps. Aggregating data to meaningful time units while preserving the integrity of trends and patterns was challenging. Additionally, differentiating between seasonal effects and actual trends required sophisticated time series decomposition techniques.

6.6 Computational Constraints

The analysis involved processing large volumes of data, which demanded substantial computational resources. Limited computational power sometimes led to longer processing times and memory constraints, particularly when applying complex algorithms such as running extensive clustering operations. Optimizing code and leveraging efficient computational techniques was essential to overcome these constraints.

6.7 Interpretability of Results

Ensuring the interpretability of complex analytical results for stakeholders was another significant challenge. Visualizing high-dimensional data, making clustering results understandable, and presenting statistical findings in an accessible manner required careful consideration of the audience's background and the use of intuitive visualizations and clear explanations. Balancing technical accuracy with simplicity was crucial to effectively communicate insights.

Chapter 7

Conclusion & Future Enhancement

7.1 Conclusion

In conclusion, this project has demonstrated the profound potential of leveraging data analytics and machine learning techniques to extract valuable insights from social media and web traffic data. By employing a comprehensive approach encompassing data acquisition, preprocessing, sentiment analysis, clustering, probability analysis, and time series analysis, we successfully unveiled patterns and trends that offer significant implications for businesses and researchers alike. The challenges encountered, such as data collection constraints, preprocessing complexities, and computational limitations, highlighted the necessity for meticulous planning and robust methodologies to ensure data quality and analytical accuracy. Despite these challenges, the integration of advanced text mining techniques, including the identification of frequently used words and sentiment trends, provided a nuanced understanding of user interactions and sentiments. The hierarchical clustering of words by sentiment, along with probability mass function analysis, further enriched our comprehension of user behavior over time. Additionally, time series analysis revealed critical insights into daily sentiment variations, contributing to a holistic understanding of user engagement patterns. Throughout the project, the importance of interpretability and effective communication of results was emphasized, ensuring that complex findings were presented in an accessible manner to stakeholders. This project not only underscores the value of data-driven decision-making but also paves the way for future research and applications in the domain of social media and web analytics. By continually refining analytical techniques and addressing the evolving challenges of data analysis, we can unlock deeper insights and foster more informed decision-making processes in various fields.

7.2 Future Enhancement

The project has laid a strong foundation for analyzing social media data and website visitor trends, but there are several avenues for future enhancements. One significant improvement could be the integration of more advanced machine learning models, such as transformer-based models like BERT or GPT-3, for more nuanced sentiment analysis and text classification. These models can better understand context, sarcasm, and complex language patterns, providing more accurate insights. Additionally, expanding the data sources to include other social media platforms and integrating multimedia content analysis (such as images and videos) could offer a more comprehensive view of user engagement and sentiment. Implementing real-time data processing pipelines would enable continuous monitoring and quicker response to emerging trends, enhancing the project's applicability in dynamic environments. Furthermore, incorporating more sophisticated time series analysis methods, such as ARIMA or LSTM models, could improve the accuracy of trend predictions and anomaly detection. Enhancing the clustering techniques with more robust validation metrics and incorporating user feedback loops could refine the clustering results, making them more actionable. To address computational constraints, leveraging cloud computing resources and distributed processing frameworks can scale the analysis to handle larger datasets more efficiently. Finally, improving the visualization and reporting aspects by using interactive dashboards and more intuitive visual representations will make the insights more accessible to a broader audience, including non-technical stakeholders. By addressing these enhancements, the project can evolve into a more powerful tool for social media analytics and web traffic analysis, providing deeper insights and more actionable recommendations.

References

[1] Tweet.csv: https://www.kaggle.com/kazanova/sentiment140

[2] Daily-website-visitors.csv: https://www.kaggle.com/bobnau/daily-website-visitors

[3] "The NRC Emotion Lexicon: A Semantic Resource for Affect Analysis in Text" by Saif M. Mohammad and Peter D. Turney. Introduces the NRC Emotion Lexicon and its application in emotion and sentiment analysis.

[4] “Data Mining - Concepts & Techniques” by Jiawei Han, Micheline Kamber and Jian Pei - Hierarchical clustering groups data points into nested clusters arranged in a tree-like structure, represented as a dendrogram.