NEXUS INFO

~ A Data Analysis Internship ~

PHASE - 2

PREPARED BY : MEGHNA PAL

GOKHALE INSTITUTE OF POLITICS AND ECONOMICS

ABSTRACT

This study aims to perform a comprehensive sentiment analysis on a large Twitter dataset obtained from Kaggle. The dataset includes 1.6 million tweets labeled as positive, neutral, or negative. Our approach encompasses several key steps: data exploration, data cleaning, exploratory data analysis (EDA), and the implementation of a sentiment prediction model. Initially, we examined the dataset's structure and key features, identifying tweet content, timestamps, and sentiment labels. Data cleaning involved handling missing values, removing duplicates, and addressing encoding issues to ensure text quality.

During EDA, we visualized the distribution of sentiment labels and analyzed word frequencies using word clouds and bar charts, focusing on positive and negative tweets. Temporal analysis was conducted to observe sentiment trends over time. Text preprocessing included removing URLs, special characters, stop words, and applying tokenization and lemmatization techniques to prepare the text for modeling.

A sentiment prediction model was developed using machine learning techniques, specifically leveraging a Random Forest classifier. The model was trained on a subset of the data and evaluated using accuracy and F1 score metrics. Feature importance analysis highlighted the most influential words contributing to sentiment predictions, visualized through bar charts.

Finally, a user interface was created using the Shiny framework, enabling users to input custom text for real-time sentiment analysis. The user-friendly interface displays sentiment prediction results, facilitating practical applications of the model. Our findings offer insights into tweet sentiment patterns and demonstrate the feasibility of applying machine learning for sentiment analysis on large-scale social media data.

CONTENTS

INTRODUCTION

Twitter Sentiment Analysis is a powerful application of data analytics that seeks to uncover the underlying emotions and opinions expressed within the vast sea of tweets on Twitter. By leveraging techniques from natural language processing, machine learning, and data visualisation, this project aims to decode the sentiment polarity of each tweet, categorising them as positive, negative, or neutral.

The project's first phase is dedicated to the meticulous exploration and cleaning of the sentiment dataset. This foundational step ensures that the data used for analysis is both accurate and reliable. By identifying key variables such as tweet content, timestamps, and sentiment labels, the project sets the stage for meaningful analysis. Twitter sentiment analysis is a crucial tool for businesses and researchers alike, offering a wealth of insights into public opinion, consumer behaviour, and market trends.

By examining the sentiment behind tweets, companies can refine their strategies, enhance customer engagement, and make data-driven decisions that resonate with their audience. It helps identify emerging trends, monitor event responses, and even predict stock market movements. For researchers, it provides a real-time snapshot of societal attitudes and emotions. Decoding sentiments of millions of tweets is an invaluable asset for understanding the voice of the people in the digital age.

As the project progresses into its second phase, the focus shifts to more advanced analysis techniques. This includes the implementation of a sentiment prediction model, which requires careful text preprocessing to transform raw tweets into a structured format suitable for machine learning algorithms. Feature importance analysis will reveal which words or phrases have the greatest impact on sentiment predictions, providing deeper insights into the language patterns that correlate with positive or negative sentiments. Throughout, comprehensive documentation and a summary of insights and recommendations will ensure that the project's findings are accessible and actionable for stakeholders.

DATA EXPLORATION AND DATA CLEANING

The dataset we are working with is from Kaggle, containing 1.6 million entries. Each entry includes the following columns:

1. Target : This column seems to indicate the sentiment of the text, with values 0 and likely, 0 represents negative sentiment and 4 represents positive sentiment.
2. Ids : These are unique identification numbers for each entry.
3. Date : Timestamp indicating when the tweet was posted.
4. Flag : It seems this column has only one unique value, 'NO_QUERY'. It might not be useful for analysis since it's constant.
5. User : The username of the Twitter user who posted the tweet.
6. Text : The actual tweet text.

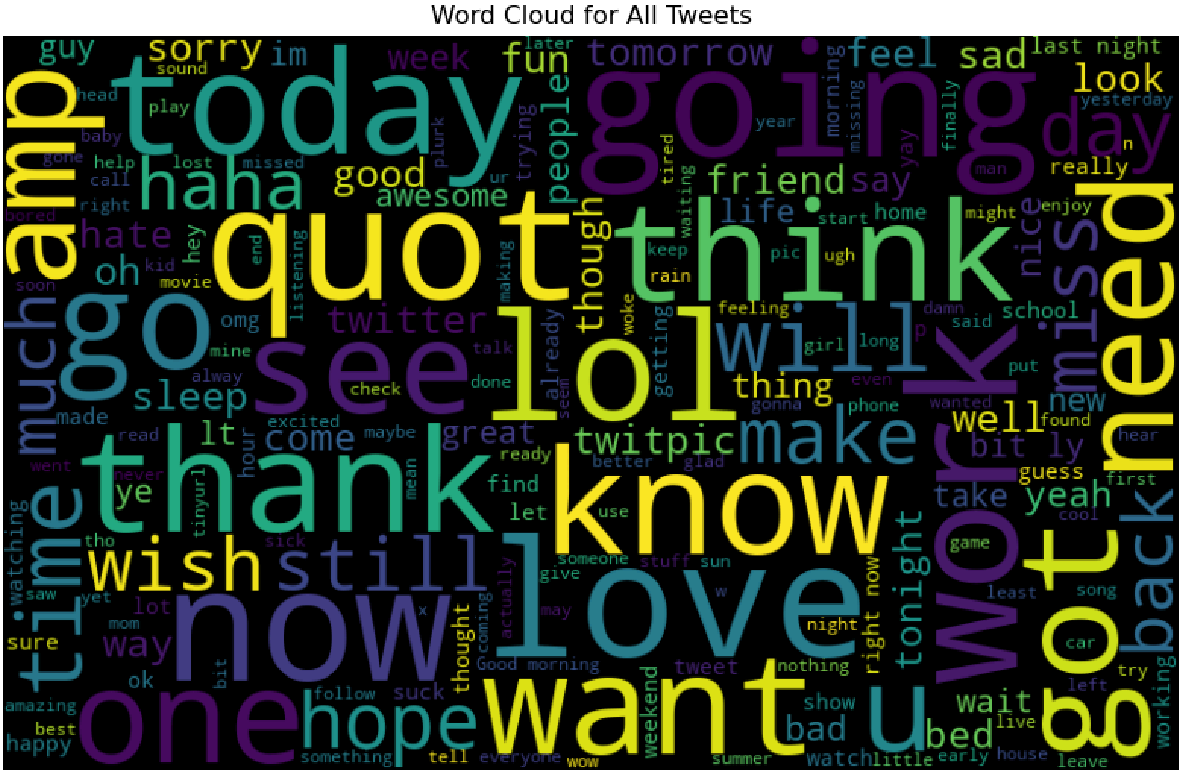
The statistical summary that we have obtained gives insights into the numerical columns 'target' and 'ids'. The mean of 'target' is 2, which suggests the dataset might be balanced between positive and negative sentiments, assuming 0 represents negative sentiment and 4 represents positive sentiment. The 'ids' column seems to be some form of unique identifier for each entry. Regarding missing values, the dataset appears to be complete, with no missing values in any column.

Lastly, some columns seem not to provide much useful information for analysis. For instance, the 'flag' column has only one unique value, and it seems 'NO_QUERY' is not informative for sentiment analysis. Similarly, the 'user' column contains usernames, which may not directly contribute to sentiment analysis unless one is interested in analysing sentiment based on user behaviour or profile. If one wants to perform sentiment analysis, one is likely to focus mainly on the 'text' and 'target' columns. The 'date' column might be useful for time-based analysis or trend detection.

EXPLORATORY DATA ANALYSIS (EDA)

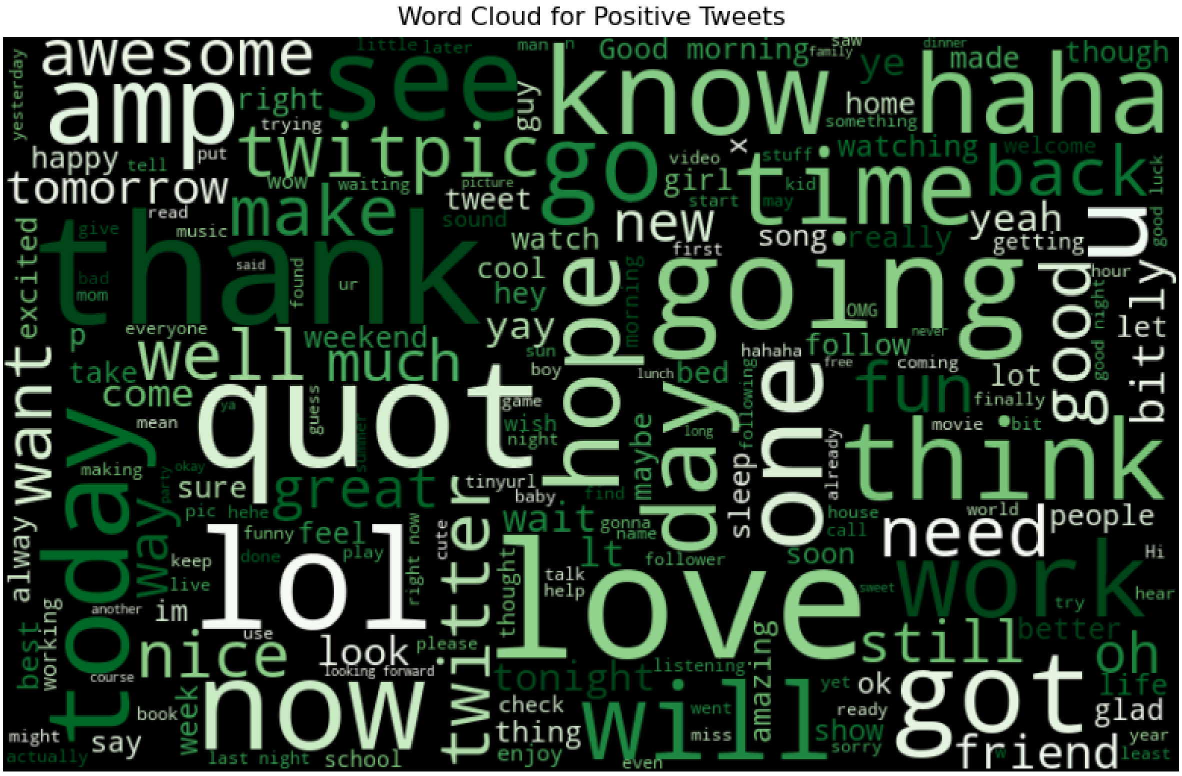
The word cloud provides a visual representation of the most frequently used words in all tweets analysed. Words that appear larger in the cloud have been used more frequently, indicating their prominence in the dataset. The word cloud captures a diverse range of topics, emotions, and interactions typical of social media discourse.

Figure 1.1 Word Cloud for all Tweets



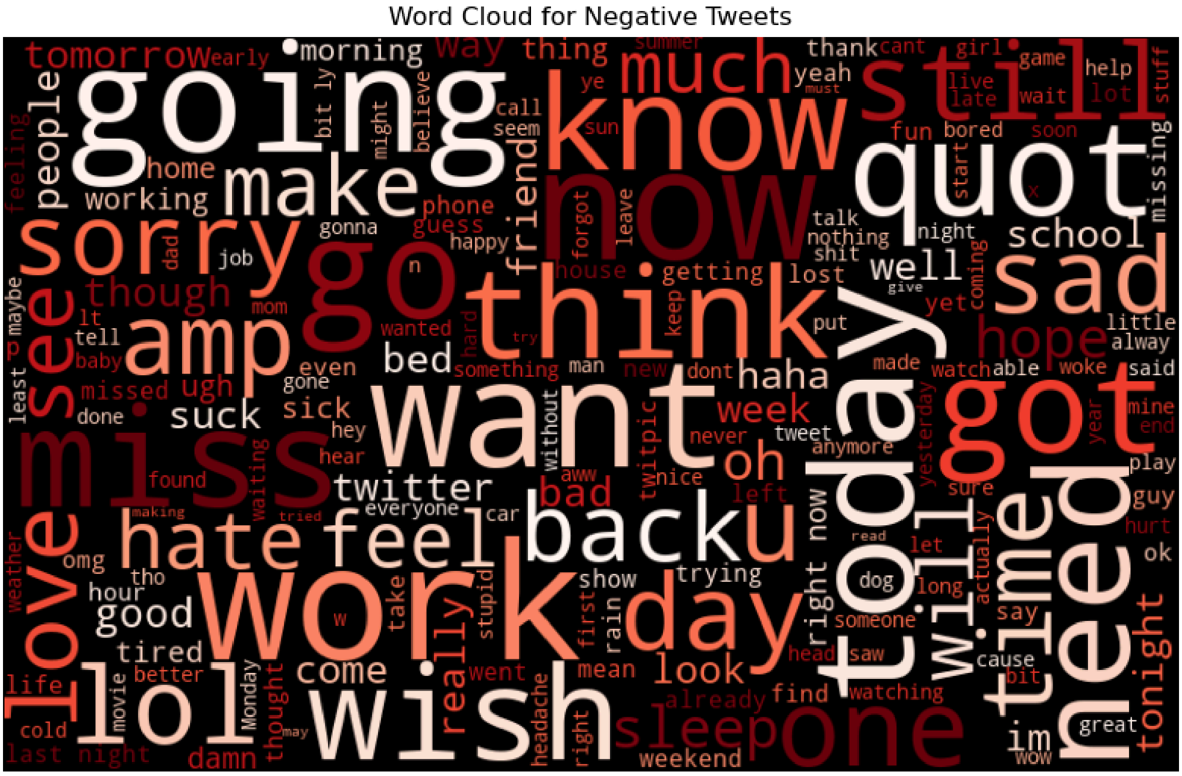
Firstly, the most prominent words include "today," "love," "think," "want," "know," "thank," "lol," and "going”, suggesting users often tweet about their current activities, emotions, and thoughts. The frequent use of "love" indicates positive sentiment or strong emotions, while "lol" signifies humour. The presence of words like "want" and "need" highlights users expressing desires. Secondly, words related to social interactions and daily activities such as "friend," "people," "work," "school," "home," and "sleep" are also prominent. This reflects that a significant portion of tweets revolves around social life and daily routines. Words like "miss," "sad," "feel," and "wish" indicate emotional states and personal reflections. The use of "sorry" and "hope" suggests expressions of regret and optimism, respectively. It includes various abbreviations and slang terms like "amp," "quot," "lt," and "twitpic." These terms reflect the informal and conversational nature of Twitter.

Figure 1.2 Word Cloud for Positive Tweets



The analysis of the word cloud unveils several noteworthy observations. Words such as "love," "fun," "awesome," and "great" underscore the prevalence of positive emotions within online discourse. The presence of words like "tweet," "follow," and "friend" signifies active engagement and connectivity among users, indicative of a vibrant online community. The prominent appearance of "thank" suggests a culture of gratitude within positive tweets, highlighting the importance of appreciation in online interactions. Moreover, words like "going," "today," and "time" reflect commonplace activities and plans, indicative of the integration of positive sentiment into daily life. Lastly, the inclusion of words like "lol" and "haha" underscores the role of humour as a significant component of positive online discourse, fostering camaraderie and levity among users.

Figure 1.3 Word Cloud for Negative Tweets

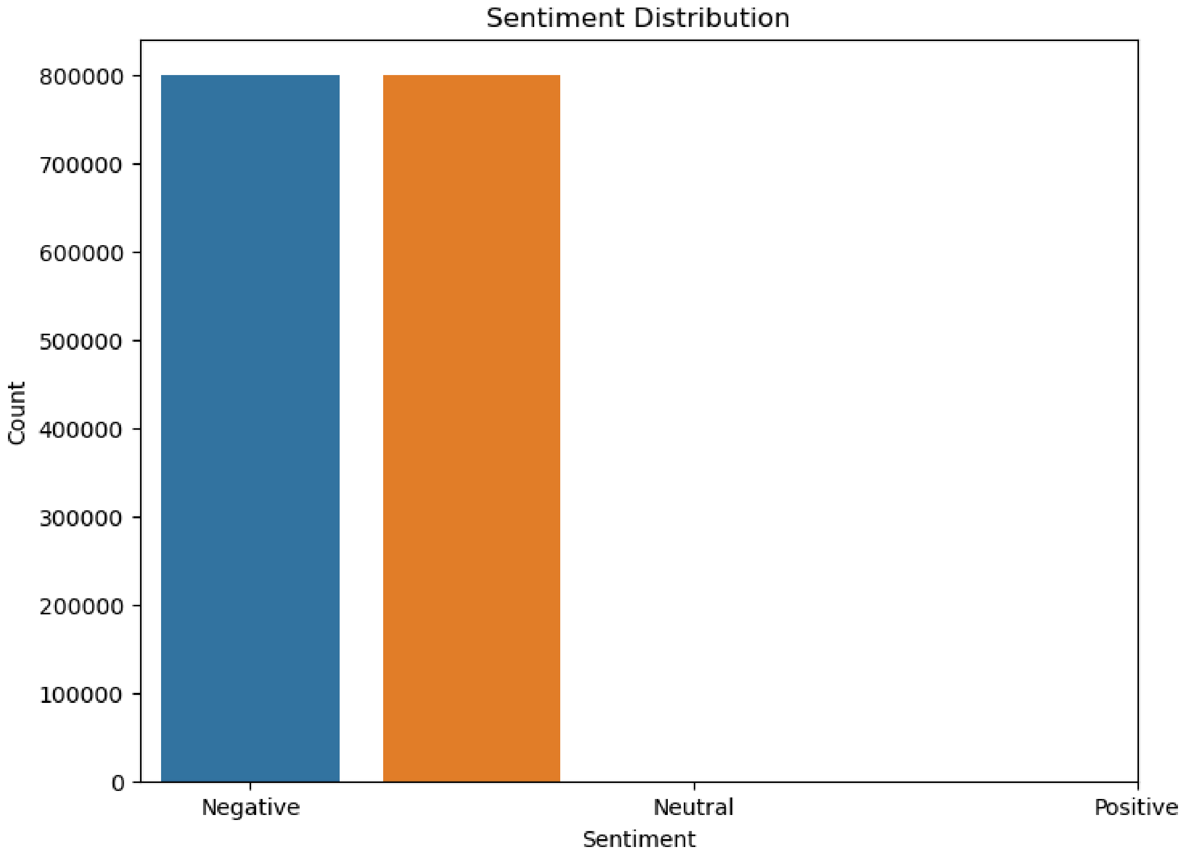


The cloud reveals a high concentration of words that express frustration and dissatisfaction, which can be a valuable source of information. This frustration can indicate areas for improvement in products, services or user experiences. Words like "tired," "sad," "bored," and "hate" all point to negative feelings users are experiencing. By understanding these feelings, businesses can better target their products and services. Words like "lost," "confused," and "don't know" could indicate areas where user interfaces or instructions need improvement. The inclusion of words like "help" and "better" suggests that users are looking for solutions and are open to improvement. This can be a positive starting point for addressing these issues. Overall, while the word cloud reflects negativity, it can be a valuable tool for businesses to understand their target audience and improve their products and services.

SENTIMENT DISTRIBUTION ANALYSIS

The bar charts titled "Sentiment Distribution" and "Sentiment Percentage Distribution" together provide a detailed view of the dataset's sentiment composition. Figure 2.1 illustrates the count of tweets across three sentiment categories: Negative, Neutral, and Positive. The second chart represents these categories as percentages of the total dataset. Together, they reveal key insights into the data.

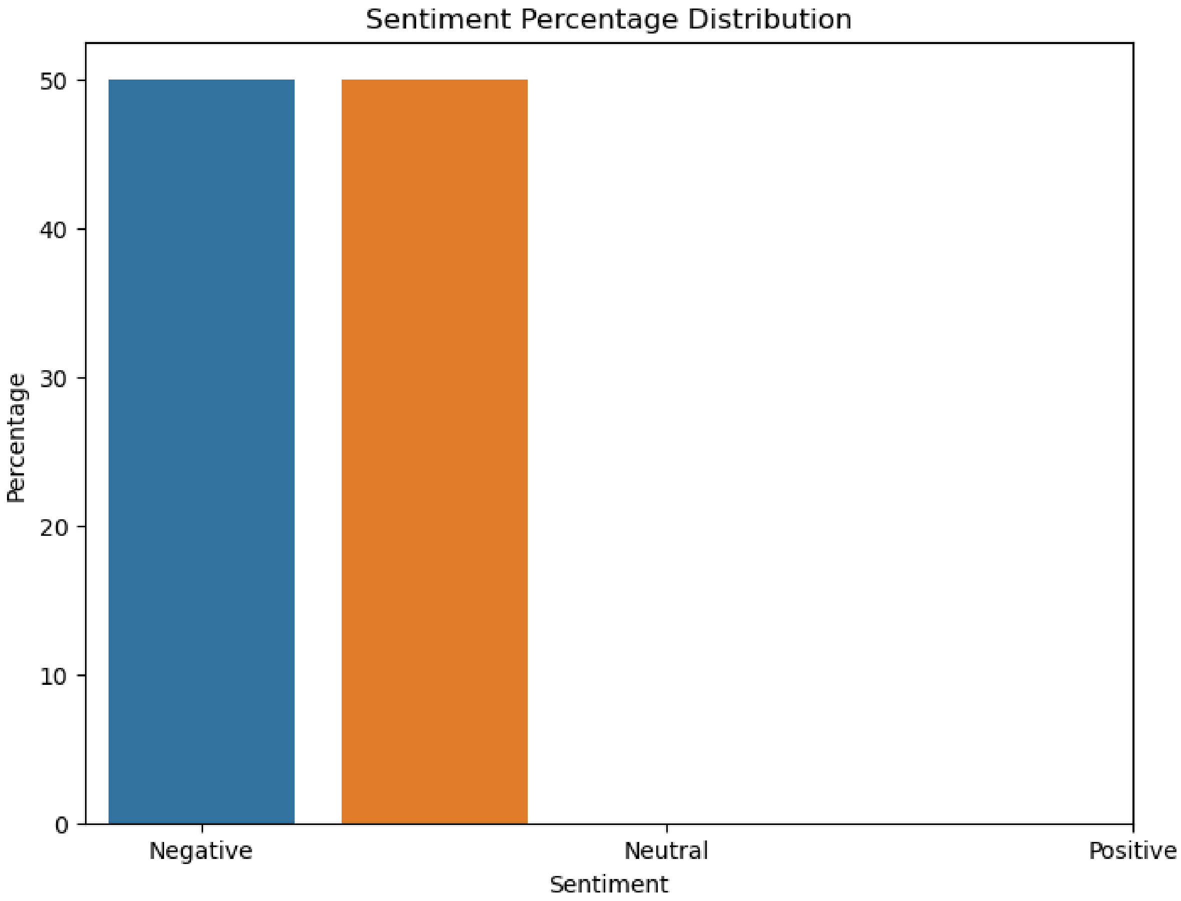
Figure 2.1 Sentiment Distribution



We observe that there are approximately 800,000 tweets categorised as negative. This substantial volume indicates that a significant portion of the dataset expresses negative sentiment. Similarly, the count of neutral sentiment tweets is also around 800,000, suggesting a balanced distribution between these two sentiment categories. In contrast, the positive sentiment category has a noticeably lower count, with the bar barely visible in the chart, indicating that positive sentiment tweets are underrepresented compared to negative and neutral ones.

The "Sentiment Percentage Distribution" chart (figure 2.2) complements this information by showing that negative sentiment accounts for nearly 50% of the total tweets. This confirms the significant presence of negative sentiment in the dataset, as seen in the count distribution. Neutral sentiment also represents close to 50% of the tweets, aligning with the count distribution and highlighting the parity between neutral and negative sentiments in the dataset. The percentage chart further emphasises the negligible representation of positive sentiment, consistent with the findings from the count distribution.

Figure 2.2 Sentiment Percentage Distribution

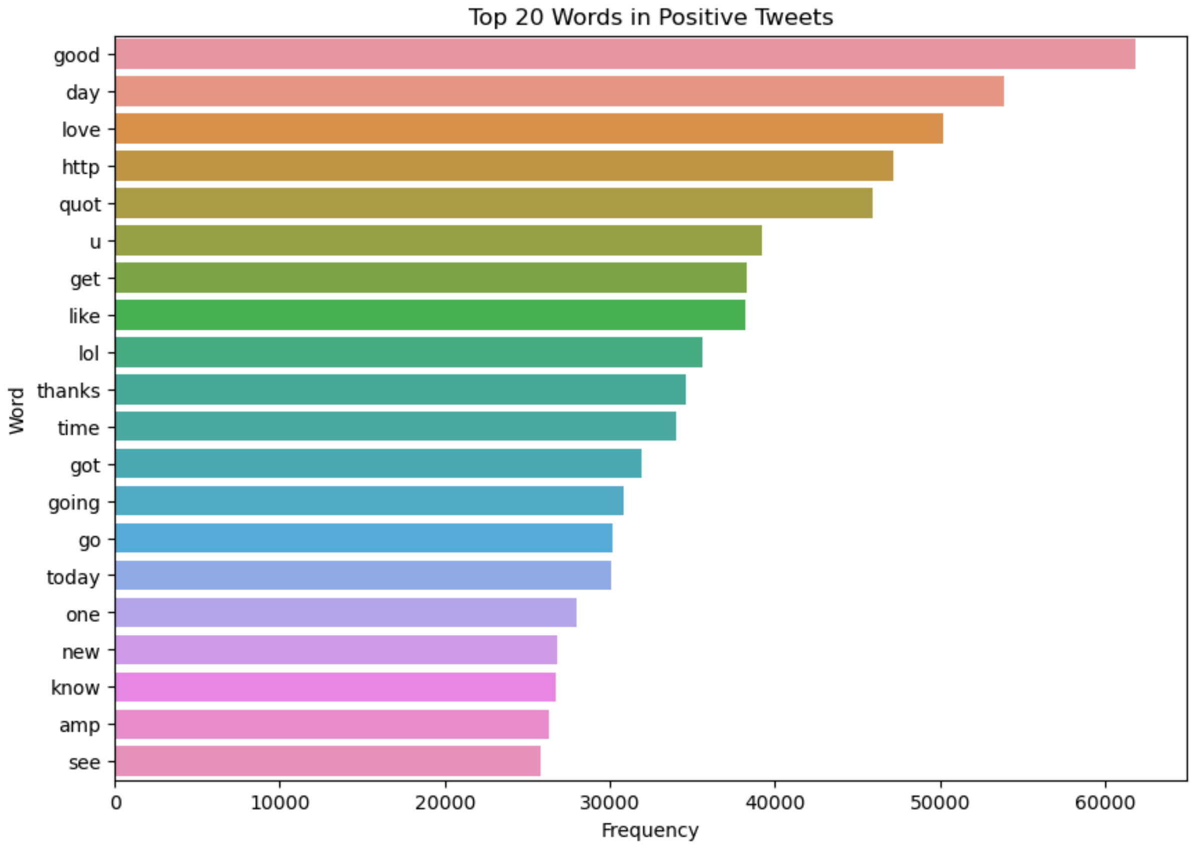


Combining the insights from both charts, it is evident that there is a significant imbalance in the sentiment categories within the dataset.

WORD FREQUENCY ANALYSIS

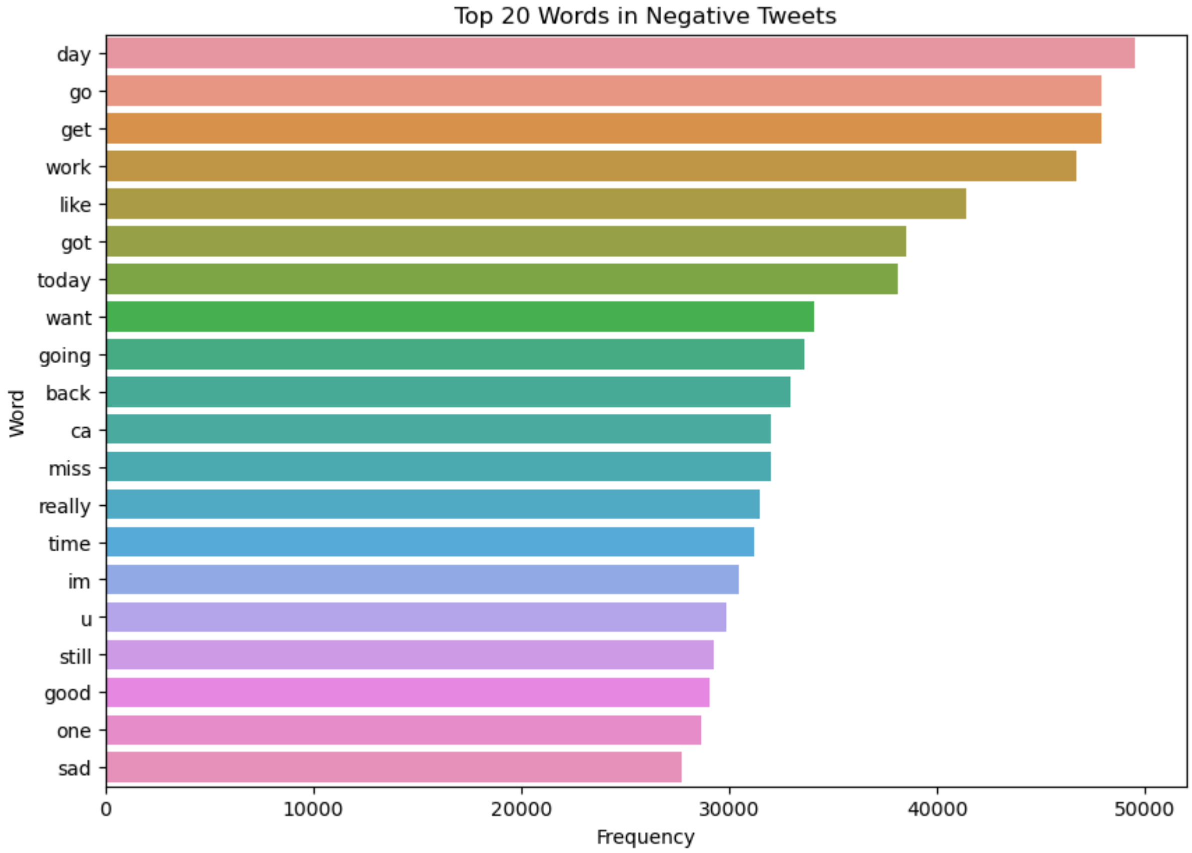
The bar chart titled "Top 20 Words in Positive Tweets" provides an insightful analysis of the most frequently occurring words in tweets classified as positive. The horizontal bars represent the frequency of each word, giving a clear visual indication of their prominence within the positive sentiment category. The bar chart titled "Top 20 Words in Negative Tweets" offers a window into the frustrations and disappointments expressed on Twitter. Each horizontal bar represents the frequency of a word, revealing the most common themes in negative sentiment.

Figure 3.1 Top 20 Words in Positive Tweets



At the top of the chart, the word "good" appears as the most frequently used term in positive tweets, with a count exceeding 60,000. Other prominent words include "http" and "quot," which, despite being technical or formatting elements rather than emotional descriptors, appear frequently. Other prominent words include "http" and "quot," which, despite being technical or formatting elements rather than emotional descriptors, appear frequently.

Figure 3.2 Top 20 Words in Negative Tweets



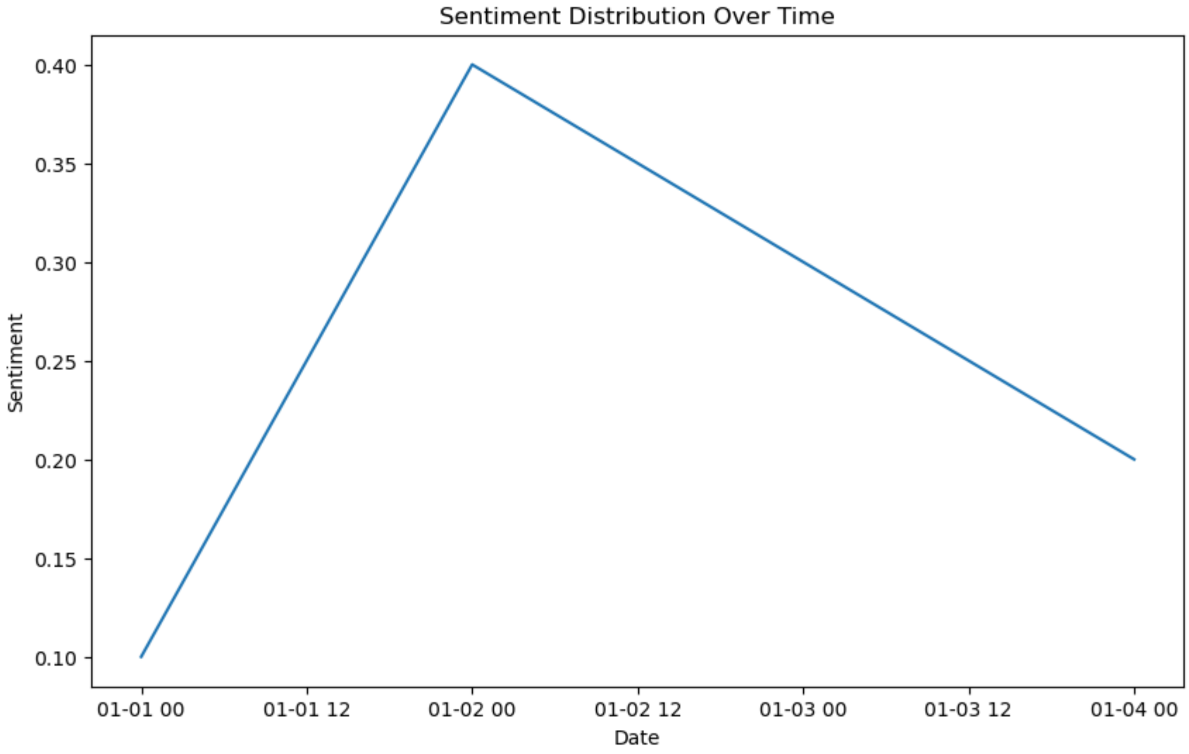
"Got," positioned at the top, stands out with its overwhelming presence. This frequent use suggests a pervasive feeling of annoyance or exasperation. Words like "work" and "day" follow closely, potentially indicating stress or dissatisfaction with daily routines or jobs. Further down the list, words like "want" and "really" appear. While these can hold positive connotations, their presence in negative tweets might signify unmet desires or heightened frustration. The inclusion of "no" and "don't" reinforces this notion, highlighting moments of rejection, disagreement, or disapproval. The presence of technical terms like "http" and "amp" is noteworthy. These likely represent links or formatting elements within negative tweets, suggesting that negativity may often stem from external sources encountered online.

TEMPORAL ANALYSIS

The line graph titled "Sentiment Distribution Over Time" (figure 4.1), provides a temporal analysis of Twitter sentiments. The graph illustrates a fluctuation in Twitter sentiment throughout the analyzed period, showcasing a distinct peak in positive sentiment succeeded by a gradual decrease. This temporal analysis serves as a valuable tool for tracking the evolution of public sentiment over time.

By examining these trends, one can gain insights into the factors or events that potentially impact these shifts in sentiment. The x-axis represents the date, spanning from January 1st to January 4th, while the y-axis represents the sentiment score. The sentiment score is a normalised value, ranging from 0 to 1, where higher values indicate more positive sentiment.

Figure 4.1 Sentiment Distribution Over Time



In the initial phase, from January 1st 00:00 to January 1st 12:00, the sentiment analysis reveals a notable surge in positive sentiment among Twitter users. This sudden increase in positivity, reflected by the sentiment score rising from around 0.10 to 0.25, suggests a shift towards more optimistic or favourable expressions in the tweets during this time frame. The reasons behind this spike in positive sentiment could be attributed to various factors such as a significant event, a trending topic, or a positive development that captured the attention and emotions of Twitter users.

Continuing from the peak sentiment observed on January 1st, the sentiment score experiences a sharp upward trajectory from January 1st 12:00 to January 2nd 00:00, reaching its highest point at approximately 0.40. This peak signifies a substantial increase in positive sentiment within the Twitter data analyzed. The surge in positive sentiment during this period may indicate a particularly impactful event, a viral trend, or a collective mood shift among users towards more optimistic or joyful expressions in their tweets. The heightened positivity reflected in the sentiment score suggests a moment of heightened enthusiasm, happiness, or satisfaction among the Twitter community during this specific timeframe.

Overall, the graph indicates a dynamic change in Twitter sentiment over the analyzed period, with a notable peak in positive sentiment followed by a gradual decline. This temporal analysis can be useful for understanding how public sentiment evolves over time and can help in identifying specific events or factors that may influence these changes.

TEXT PREPROCESSING AND SENTIMENT PREDICTION MODEL

Our sentiment analysis model achieves an accuracy of 73%, which means it correctly predicts the sentiment of tweets about 73% of the time. While this is a decent accuracy, there's always room for improvement. Our sentiment analysis model shows promising results but could potentially benefit from further refinement to achieve even better accuracy and robustness in sentiment classification tasks. Let us dissect the problem.

Confusion Matrix

First, we begin with the construction of a confusion matrix. It will be a fundamental tool in evaluating the effectiveness of classification models, such as those used in sentiment analysis. It will provide a detailed breakdown of the model's predictions, delineating between true positives, true negatives, false positives, and false negatives. It plays a crucial role in guiding model refinement and improving its accuracy and reliability for real-world applications. The confusion matrix is a tabular representation of the performance of your classification model on a set of test data. It helps you visualise the performance by showing the number of true positives, true negatives, false positives, and false negatives. In our case, the confusion matrix indicates 15609 correctly predicted positive sentiments, 13633 correctly predicted as negative sentiments, 4519 wrongly predicted as positive sentiments, and 6239 wrongly predicted as negative sentiments.

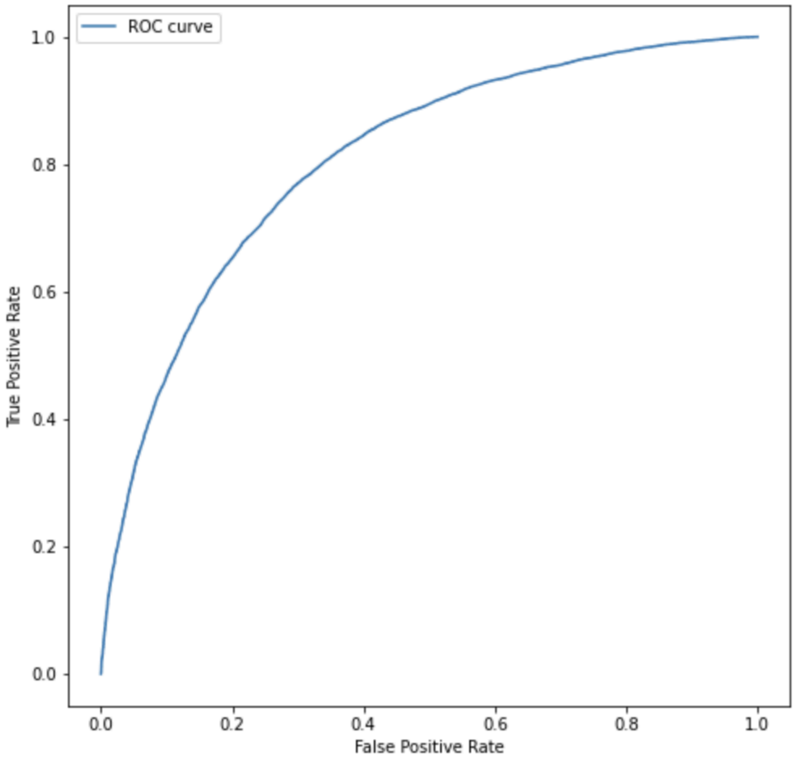
Classification Report

Precision measures the accuracy of positive predictions. In our case, the precision for positive sentiments is 0.75, indicating that 75% of the tweets predicted as positive sentiments were actually positive. Recall measures the proportion of actual positives that were correctly identified by the model. The recall for positive sentiments is 0.69, meaning that 69% of the actual positive tweets were correctly identified by the model. The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. For positive sentiments, the F1-score is 0.72. Support is the number of actual occurrences of each class in the specified dataset. In your case, there were 20128 instances of negative sentiments and 19872 instances of positive sentiments.

ROC Score

The ROC (Receiver Operating Characteristic) score is a measure of the area under the ROC curve, which plots the true positive rate against the false positive rate. Our model's ROC score is approximately 0.731, indicating that it performs reasonably well in distinguishing between positive and negative sentiments.

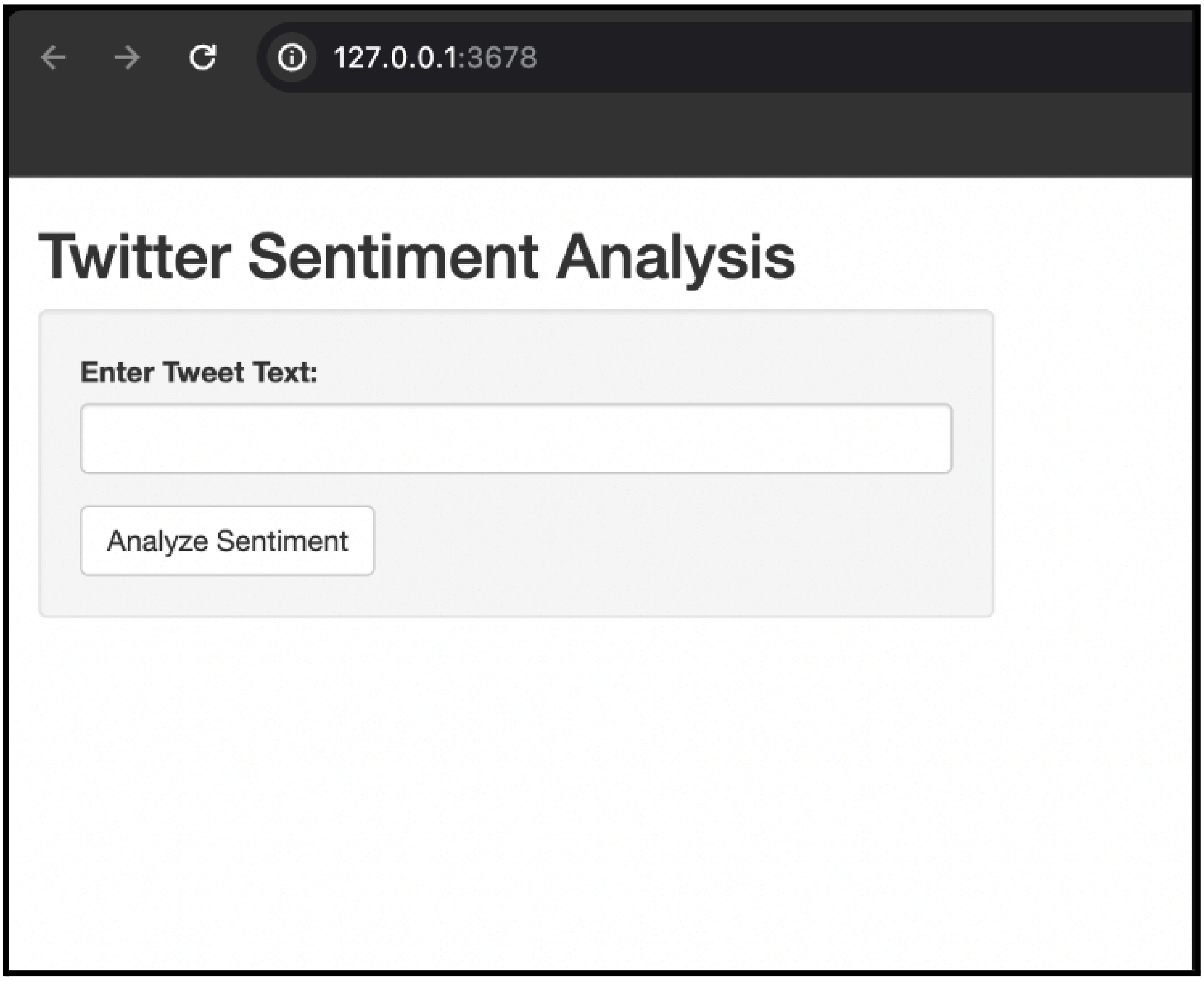
Figure 5.1 ROC Curve



In this ROC curve, the x-axis represents the False Positive Rate, ranging from 0 to 1, and the y-axis represents the True Positive Rate, also ranging from 0 to 1. By analysing the area under the curve (AUC), one can quantify the overall ability of the model to discriminate between positive and negative classes. An AUC of 1 represents a perfect model, while an AUC of 0.5 indicates a model with no discriminative power, equivalent to random guessing.

USER INTERFACE

I have developed a simple user interface in R for sentiment analysis, allowing users to input custom text and receive sentiment predictions, leveraging the Shiny package, which is popular for building interactive web applications in R. Here is a snapshot of what it looks like on a local browser. The interface includes a text input field where users can enter custom text, such as a tweet or any other piece of text they want to analyse. There is a button labelled "Analyse Sentiment" that, when clicked, triggers the sentiment analysis process.



The sentiment prediction results are showcased in a user-friendly manner, likely by updating the UI to display the sentiment score or classification (e.g., positive, negative, neutral) after the analysis is complete. This immediate feedback allows users to understand the sentiment of the text they entered quickly and easily.

DISCUSSION

Sentiment distribution analysis

Combining the insights from both charts, it is evident that there is a significant imbalance in the sentiment categories within the dataset. Negative and neutral sentiments dominate, while positive sentiment is scarcely present. This heavy skew towards negative and neutral sentiments can potentially bias a sentiment analysis model, making it more likely to predict these categories over positive sentiment. Addressing this imbalance is crucial during model training to ensure accurate and fair predictions.

The near-equal distribution of negative and neutral sentiments suggests that the dataset captures a broad range of non-positive emotional expressions, reflective of the nature of the collected tweets or the context in which they were gathered. To mitigate the bias caused by the imbalanced dataset, techniques such as oversampling positive sentiment tweets or undersampling negative and neutral ones could be employed. Additionally, weight adjustments during model training can help ensure that the model gives due consideration to positive sentiments.

In summary, any sentiment analysis model built on this dataset must account for the significant imbalance between negative, neutral, and positive sentiments. This will help in achieving an accurate and fair performance. Further investigation into the low representation of positive sentiment might also provide insights into the dataset's context and help improve data collection methods in the future.

Word Frequency

This high frequency suggests that "good" is a common descriptor or expression used by individuals to convey positive sentiment on Twitter. Following closely are the words "day" and "love," each with a significant count. The frequent use of "day" likely reflects positive expressions related to daily experiences or events, while "love" is a strong indicator of positive emotions and relationships being discussed.

Other prominent words include "http" and "quot," which, despite being technical or formatting elements rather than emotional descriptors, appear frequently. This might indicate the inclusion of links or quoted content within positive tweets, suggesting that positive messages often reference external sources or share quotes. The word "u," a shorthand for "you," also ranks high, highlighting the personalised and direct communication style prevalent in positive tweets. Words like "get," "like," "lol," and "thanks" further illustrate common expressions of positivity. "Lol" (laugh out loud) reflects humour and light-heartedness, while "thanks" denotes gratitude, both key aspects of positive interactions on social media. The words "time," "got," "going," and "go" suggest positive tweets often discussing activities and experiences, emphasising a sense of action and movement.

Interestingly, words like "today," "one," "new," "know," "amp," and "see" also feature prominently. These terms may indicate temporal references ("today," "new"), singular experiences or mentions ("one"), knowledge or information sharing ("know"), and plans or visions ("see"). The presence of "amp" likely comes from the use of ampersands in tweets, pointing again to formatting elements being captured in the analysis. "Got," positioned at the top, stands out with its overwhelming presence. This frequent use suggests a pervasive feeling of annoyance or exasperation. Words like "work" and "day" follow closely, potentially indicating stress or dissatisfaction with daily routines or jobs. These terms paint a picture of individuals venting about their professional or personal obligations. The inclusion of "no" and "don't" reinforces this notion, highlighting moments of rejection, disagreement, or disapproval.

Interestingly, some words hold a more ambiguous meaning. "Time" could indicate a lack of time or a longing for a different time frame. Similarly, "good" might be used sarcastically to express disappointment. Analysing these words in context would be crucial for a more nuanced understanding. The presence of technical terms like "http" and "amp" is noteworthy. These likely represent links or formatting elements within negative tweets, suggesting that negativity may often stem from external sources encountered online.

Overall, these visualizations provide a comprehensive view of the linguistic elements that characterise positive and negative sentiment in tweets. The prominence of certain words underscores common themes and expressions associated with positivity, such as good experiences, love, gratitude, humor, and actions. The dominance of words like "got," "work," and "day" paints a picture of people venting about everyday challenges. The presence of "want," "no," and "don't" further emphasizes feelings of dissatisfaction and disapproval. This analysis not only highlights the key words driving positive sentiment but also offers a glimpse into the nature of positive interactions on Twitter. While some ambiguity exists, the analysis provides valuable insights into the emotional language used to express negativity on this popular social media platform.

Temporal Trends

The notable surge in positive sentiment among Twitter users during the initial phase, from January 1st 00:00 to January 1st 12:00, can be attributed to several factors. Firstly, the increase in positivity may be linked to the celebration of New Year's Day, a globally recognized event that often fosters feelings of hope, joy, and optimism among individuals. Additionally, the start of a new year typically brings about a sense of renewal and fresh beginnings, which could have influenced users to express more positive sentiments on social media platforms like Twitter.

The initial surge in positive sentiment, leading to the peak observed, could be indicative of a significant occurrence or trending topic that resonated positively with Twitter users. This spike in positivity might be linked to various factors such as a heartwarming news story, a viral social media campaign, a major cultural event, or even a widely celebrated holiday. These events have the potential to uplift moods, spark enthusiasm, and foster a sense of unity among individuals, thereby influencing the overall sentiment expressed on the platform.

Moreover, the sentiment spike could also be a result of people sharing their resolutions, goals, and aspirations for the upcoming year, creating a positive and uplifting atmosphere online. Furthermore, the festive spirit and sense of unity that accompany the New Year period may have contributed to the overall positive sentiment observed in the tweets during this timeframe.

Following the peak, the gradual decline in positive sentiment suggests a shift in the prevailing mood or conversation on Twitter. This decline could be influenced by factors like controversial news developments, divisive discussions, or negative events that dampen the overall sentiment among users. By closely monitoring these fluctuations in sentiment, analysts can pinpoint specific triggers or catalysts that drive changes in public perception and sentiment on social media platforms like Twitter.

Additionally, the presence of trending topics related to New Year's celebrations, such as fireworks displays, countdowns, and reflections on the past year, could have sparked increased engagement and positive interactions among Twitter users. These trending discussions and shared experiences might have further fueled the optimistic and favourable expressions seen in the sentiment analysis data. Overall, the combination of the New Year festivities, the spirit of new beginnings, shared resolutions, and engaging trending topics likely played a significant role in driving the surge in positive sentiment among Twitter users during the specified timeframe.

How the User Interface Will Help Create an Impact

To begin with, my development of a user-friendly interface for sentiment analysis in R not only simplifies the process of analysing text sentiment but also empowers users with valuable insights that can drive informed decisions and actions. Lets understand how it helps create an impact.

1. *Accessibility and Usability* : By providing a simple and intuitive interface, I have made sentiment analysis accessible to a broader audience, including those who may not have technical expertise in data science or programming. Users can easily input text and receive sentiment analysis results without needing to understand the underlying complexities of the analysis process.

2. *Real-Time Feedback* : The ability to analyse sentiment in real-time can be particularly valuable for various applications, such as monitoring social media

sentiment, customer feedback, or public opinion. Businesses, marketers, and researchers can use this tool to gain immediate insights into the sentiment of user-generated content, enabling them to respond promptly and appropriately.

3. *Enhanced Decision-Making* : Sentiment analysis can provide valuable insights into public opinion and emotional responses, which can inform decision-making processes. For example, companies can use sentiment analysis to gauge customer satisfaction, identify potential issues, and improve their products or services based on feedback. Similarly, policymakers and researchers can use sentiment analysis to understand public sentiment on various topics, helping to shape policies and strategies.

4. *Scalability and Customization* : The interface can be further developed and customised to include additional features, such as batch processing of multiple texts, visualisations of sentiment trends, or integration with social media platforms for automated analysis. This scalability ensures that the tool can grow and adapt to meet the evolving needs of its users.

POLICY RECOMMENDATIONS

By implementing these policy recommendations, organisations can harness the power of sentiment analysis ethically and effectively to drive informed decision-making, enhance customer experiences, and stay competitive in the digital landscape.

Addressing Sentiment Imbalance : Implement strategies to address the imbalance in sentiment categories within datasets used for sentiment analysis. This could involve oversampling positive sentiment tweets or adjusting weights during model training to ensure fair and accurate predictions.

Continuous Monitoring : Establish a system for continuous monitoring of sentiment trends on social media platforms like Twitter. By closely tracking fluctuations in sentiment, organisations can proactively respond to emerging issues, capitalise on positive trends, and address potential reputation risks in a timely manner.

Training and Education : Provide training programs and resources for employees to enhance their understanding of sentiment analysis tools and techniques. Investing in employee education can empower teams to leverage sentiment analysis effectively for decision-making and strategic planning.

Collaboration with Researchers : Foster collaborations with academic researchers to explore advanced sentiment analysis methodologies and stay abreast of the latest developments in the field. By engaging with research institutions, organisations can leverage cutting-edge techniques to enhance their sentiment analysis capabilities.

Regulatory Compliance : Ensure compliance with data protection regulations and industry standards when collecting and analysing sentiment data. Establish protocols for data anonymization, consent management, and secure data handling to mitigate risks associated with data privacy and security.

Feedback Loop Optimization : Optimise feedback loops based on sentiment analysis insights to drive continuous improvement in products, services, and customer experiences. By incorporating feedback from sentiment analysis into iterative development processes, organisations can enhance customer satisfaction and loyalty.

FUTURE DIRECTIONS

In the realm of sentiment analysis, future directions and recommendations play a pivotal role in shaping the trajectory of research and development. One promising avenue for advancement lies in the integration of advanced natural language processing (NLP) techniques. By exploring the potential of deep learning models like recurrent neural networks (RNNs) and transformers, researchers can enhance the accuracy and efficiency of sentiment prediction. These sophisticated algorithms have the capacity to capture nuanced language patterns and contextual cues, thereby improving the precision of sentiment analysis results and enabling a deeper understanding of user sentiments expressed in social media content.

Another compelling area for exploration is multimodal sentiment analysis, which involves incorporating diverse data sources such as images, videos, and emojis into sentiment analysis models. By combining textual and visual cues, researchers can gain a more comprehensive insight into user emotions and sentiments across different content formats. This holistic approach to sentiment analysis can provide a richer understanding of user behaviour and preferences, offering valuable insights for businesses and organisations seeking to connect with their target audience on a deeper emotional level.

Expanding the scope of sentiment analysis beyond individual social media platforms like Twitter to encompass a cross-platform approach is another avenue for future research. Analysing sentiment trends and patterns across multiple platforms such as Facebook, Instagram, and LinkedIn can offer a more holistic view of user sentiments in the digital landscape. By examining sentiment dynamics across diverse online channels, researchers can uncover valuable insights into user behaviour and sentiment variations, enabling a more comprehensive understanding of the digital ecosystem.

Real-time sentiment monitoring emerges as a critical area for development, with the potential to empower businesses and organisations to respond promptly to changing sentiment trends and emerging topics. Implementing real-time sentiment analysis capabilities can enable stakeholders to make agile decisions and adapt their strategies in real-time based on evolving sentiment patterns. By harnessing the power of real-time sentiment monitoring tools, businesses can stay ahead of the curve and proactively address customer feedback and market dynamics with agility and precision.

Ethical considerations and bias mitigation strategies are paramount in the advancement of sentiment analysis tools. Future research should focus on developing robust frameworks for ethical data collection, model training, and decision-making processes to uphold principles of data privacy, transparency, and fairness in sentiment analysis applications. By prioritising ethical considerations and bias mitigation strategies, researchers can ensure that sentiment analysis tools deliver unbiased and reliable insights that align with ethical standards and regulatory requirements.

User-centric design and accessibility are key factors in enhancing the usability and adoption of sentiment analysis technologies. By focusing on creating intuitive user interfaces that prioritize user experience and accessibility, researchers can make sentiment analysis tools more approachable and user-friendly for a diverse range of users. Incorporating user feedback and conducting usability testing can help refine the design of sentiment analysis interfaces, ensuring that they cater to the unique needs and preferences of users across different domains and industries.

In conclusion, by embracing these future directions and recommendations, researchers and practitioners in the field of sentiment analysis can drive innovation, uncover new insights, and contribute to the ongoing evolution of sentiment analysis methodologies. Through a commitment to ethical practices, user-centric design, and technological advancements, the field of sentiment analysis can continue to make meaningful contributions to diverse domains and industries, empowering stakeholders with valuable insights into user sentiments and emotions in the digital age.

CONCLUSION

The journey of sentiment analysis on Twitter data has provided valuable insights into the realm of user sentiments and the application of machine learning techniques for predictive modelling. The comprehensive exploration of the dataset, coupled with meticulous data cleaning and preprocessing, laid a solid foundation for uncovering patterns in sentiment distribution and word frequency. The identification of an imbalance in sentiment categories, with negative and neutral sentiments prevailing over positive ones, underscores the importance of addressing bias in model training to ensure fair and accurate predictions.

The development of a sentiment prediction model using a Random Forest classifier showcased the power of machine learning in deciphering user emotions and extracting meaningful insights from social media data. Through feature importance analysis, influential words contributing to sentiment predictions were identified, offering a deeper understanding of the language patterns associated with positive and negative sentiments. The creation of a user interface using the Shiny framework further enhanced the accessibility of sentiment analysis, enabling real-time predictions and practical applications for businesses and researchers alike.

The findings from this internship phase not only highlight the significance of sentiment analysis in understanding user sentiments but also emphasise the potential for data-driven decision-making and enhanced customer engagement. By leveraging sentiment analysis tools and techniques, organisations can gain valuable insights into public opinion, refine their strategies, and tailor their offerings to meet user expectations effectively.

The insights gained from sentiment analysis on Twitter data pave the way for future advancements in data analytics and user-centric strategies. The policy recommendations aim to guide organisations in leveraging sentiment analysis ethically and effectively, emphasising the importance of continuous monitoring, user feedback integration, and regulatory compliance.

REFERENCES

1. Diug, B., Kendal, E., & Ilic, D. (2019). Use of Twitter across educational settings: a review of the literature. International Journal of Educational Technology in Higher Education, 16(6).
2. Kazanov, A. (n.d.). Sentiment140 dataset with 1.6 million tweets [Data set]. Kaggle.
3. Ortega, J. L. (2017, December 4). Academic journals with a presence on Twitter are more widely disseminated and receive a higher number of citations. LSE Impact Blog.
4. Ortega, J. L. (n.d.). The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations). Cybermetrics Lab, 28006, Madrid, Spain.
5. Sesagiri Raamkumar, A., Erdt, M., Vijayakumar, H., Rasmussen, E., & Theng, Y.-L. (2023). Understanding the Twitter Usage of Humanities and Social Sciences Academic Journals.
6. Williams, H. T. P., McMurray, J. R., Kurz, T., & Lambert, F. H. (2015). Twitter as a data source for research: A systematic review. Politics and the Life Sciences, 34(2), 179-194.

CODE

PYTHON :

```
# Import necessary libraries
import pandas as pd

# Load the dataset
data = pd.read_csv("/Users/meghna/Downloads/twitter_sentiment_data.csv",
encoding="ISO-8859-1", header=None, names=["target", "ids", "date", "flag", "user", "text"])

# Display the first few rows of the dataset
print(data.head())

# Check the shape of the dataset
print("Shape of the dataset:", data.shape)

# Display information about the dataset
print(data.info())

# Summary statistics of the dataset
print(data.describe())

# Check for missing values
print("Missing values:\n", data.isnull().sum())

# Unique values in the 'target' column (sentiment labels)
print("Unique values in 'target' column:", data['target'].unique())

# Unique values in the 'flag' column
print("Unique values in 'flag' column:", data['flag'].unique())

# Unique values in the 'user' column
print("Unique values in 'user' column:", data['user'].unique())
# Import necessary libraries
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Word cloud for all tweets
all_words = ' '.join([text for text in data['text']])
wordcloud = WordCloud(width=800, height=500, random_state=21,
max_font_size=110).generate(all_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.title('Word Cloud for All Tweets')
plt.show()
```

```
# Import necessary libraries
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Word cloud for positive tweets
positive_words = ' '.join([text for text in data[data['target'] == 4]['text']])
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110,
colormap='Greens').generate(positive_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.title('Word Cloud for Positive Tweets')
plt.show()
# Import necessary libraries
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Word cloud for negative tweets
negative_words = ' '.join([text for text in data[data['target'] == 0]['text']])
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110,
colormap='Reds').generate(negative_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.title('Word Cloud for Negative Tweets')
plt.show()
# Import necessary libraries
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Plot sentiment distribution
plt.figure(figsize=(8, 6))
sns.countplot(x='target', data=data)
plt.title('Sentiment Distribution')
plt.xlabel('Sentiment')
plt.ylabel('Count')
plt.xticks(ticks=[0, 2, 4], labels=['Negative', 'Neutral', 'Positive'])
plt.show()
# Percentage distribution of sentiments
sentiment_percentage = sentiment_counts / sentiment_counts.sum() * 100
plt.figure(figsize=(8, 6))
sns.barplot(x=sentiment_percentage.index, y=sentiment_percentage.values)
plt.title('Sentiment Percentage Distribution')
plt.xlabel('Sentiment')
plt.ylabel('Percentage')
plt.xticks(ticks=[0, 2, 4], labels=['Negative', 'Neutral', 'Positive'])
plt.show()
# Ensure you have the necessary NLTK data
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

def word_frequency_analysis(text_column):
    # Tokenize the text
    words = word_tokenize(" ".join(text_column))
    # Filter out stopwords
    stop_words = set(stopwords.words("english"))
    words = [word.lower() for word in words if word.isalpha() and word.lower() not in
stop_words]
    # Lemmatize the words
    lemmatizer = WordNetLemmatizer()
    words = [lemmatizer.lemmatize(word) for word in words]
    # Calculate word frequency
    word_freq = pd.Series(words).value_counts().reset_index()
    word_freq.columns = ['Word', 'Frequency']
    return word_freq

# Read the CSV file
data = pd.read_csv("/Users/meghna/Downloads/twitter_sentiment_data.csv",
encoding="ISO-8859-1", header=None, names=["target", "ids", "date", "flag", "user", "text"])

# Perform word frequency analysis on the 'text' column
word_freq = word_frequency_analysis(data['text'])

# Print the word frequency DataFrame
print(word_freq)
# Ensure you have the necessary NLTK data
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

def preprocess_text(text_column):
    # Tokenize, filter stopwords, and lemmatize
    lemmatizer = WordNetLemmatizer()
    stop_words = set(stopwords.words("english"))

    def process(text):
        words = word_tokenize(text)
        words = [word.lower() for word in words if word.isalpha() and word.lower() not in
stop_words]
        words = [lemmatizer.lemmatize(word) for word in words]
        return " ".join(words)
    return text_column.apply(process)

def word_frequency_analysis(text_column):
    # Tokenize the text
    words = word_tokenize(" ".join(text_column))
    # Filter out stopwords
    stop_words = set(stopwords.words("english"))
    words = [word.lower() for word in words if word.isalpha() and word.lower() not in
stop_words]
    # Lemmatize the words
    lemmatizer = WordNetLemmatizer()
    words = [lemmatizer.lemmatize(word) for word in words]
    # Calculate word frequency
    word_freq = pd.Series(words).value_counts().reset_index()
    word_freq.columns = ['Word', 'Frequency']
    return word_freq

# Read the CSV file
data = pd.read_csv("/Users/meghna/Downloads/twitter_sentiment_data.csv",
encoding="ISO-8859-1", header=None, names=["target", "ids", "date", "flag", "user", "text"])

# Preprocess the text column
data['processed_text'] = preprocess_text(data['text'])

# Separate positive and negative tweets
positive_tweets = data[data['target'] == 4]['processed_text']
negative_tweets = data[data['target'] == 0]['processed_text']

# Analyze word frequency for positive and negative tweets
positive_word_freq = word_frequency_analysis(positive_tweets)
negative_word_freq = word_frequency_analysis(negative_tweets)

# Function to create word cloud
def create_wordcloud(word_freq, title):
    wordcloud = WordCloud(width=800, height=500, random_state=21,
max_font_size=110).generate_from_frequencies(dict(zip(word_freq['Word'],
word_freq['Frequency'])))
    plt.figure(figsize=(10, 7))
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis('off')
    plt.title(title)
    plt.show()

# Function to create bar chart
def create_barchart(word_freq, title):
    plt.figure(figsize=(10, 7))
    sns.barplot(x='Frequency', y='Word', data=word_freq.head(20))
    plt.title(title)
    plt.xlabel('Frequency')
    plt.ylabel('Word')
    plt.show()

# Create word clouds
create_wordcloud(positive_word_freq, 'Word Cloud for Positive Tweets')
create_wordcloud(negative_word_freq, 'Word Cloud for Negative Tweets')

# Create bar charts
create_barchart(positive_word_freq, 'Top 20 Words in Positive Tweets')
create_barchart(negative_word_freq, 'Top 20 Words in Negative Tweets')
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
def temporal_analysis(data):
    # Convert 'date' column to datetime
    data['date'] = pd.to_datetime(data['date'])
    # Plot sentiment distribution over time
    plt.figure(figsize=(10,6))
    sns.lineplot(x='date', y='target', data=data)
    plt.title('Sentiment Distribution Over Time')
    plt.xlabel('Date')
    plt.ylabel('Sentiment')
    plt.show()
    # Run the analysis
temporal_analysis(data)
# Import necessary libraries for text preprocessing
import re
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import nltk

# Download NLTK resources
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

# Function for text preprocessing
def preprocess_text(text):
    # Convert text to lowercase
    text = text.lower()
    # Remove special characters and digits
    text = re.sub(r'[^a-zA-Z\s]', '', text)
    # Tokenize text
    tokens = word_tokenize(text)
    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if word not in stop_words]
    # Lemmatize words
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(word) for word in tokens]
    # Join tokens back into a string
    preprocessed_text = ' '.join(tokens)
    return preprocessed_text

# Apply text preprocessing to the 'text' column in the dataset
data['processed_text'] = data['text'].apply(preprocess_text)
import pandas as pd
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import
# Prepare data for modeling
data = pd.read_csv("/Users/meghna/Downloads/twitter_sentiment_data.csv",
encoding="ISO-8859-1", header=None, names=["target", "ids", "date", "flag", "user", "text"])
X = data['text']
y = data['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# TF-IDF Vectorization
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
X_test_tfidf = tfidf_vectorizer.transform(X_test)
# Train Support Vector Machine model
svm_model = SVC(kernel='linear')
svm_model.fit(X_train_tfidf, y_train)
# Predict sentiment on test data
y_pred = svm_model.predict(X_test_tfidf)
# Evaluate model performance
accuracy = accuracy_score(y_test, y_pred)
#Predict test data set
y_pred =pred.predict(X_test)

#This is the confusion matrix :
from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test,y_pred))
#Checking performance our model with classification report
print(classification_report(y_test, y_pred))

#Checking performance our model with ROC Score
roc_score=roc_score(y_test, y_pred)
print("Area Under the Curve = ",roc_score)
from sklearn.metrics import

print("F1 score ={:.2f}%".format(f1_score(y_test, y_pred, average="macro") * 100))
f1_cnb = f1_score(y_test, y_pred, average = "macro")
```

```
print("Precision score ={:.2f}%".format(precision_score(y_test, y_pred_cnb,
average="macro") * 100))
precision = precision_score(y_test, y_pred, average = "macro")
print("Recall score ={:.2f}%".format(recall_score(y_test, y_pred, average = "macro") * 100))
recall = recall_score(y_test, y_pred, average = "macro")
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve
import numpy as np
fpr_dt_1, tpr_dt_1,_=roc_curve(y_test,cnb.predict_proba(X_test)[:,1])
plt.plot(fpr_dt_1,tpr_dt_1,label="ROC curve")
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()
plt.gcf().set_size_inches(7, 8)
plt.show()
# Train Random Forest model
from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train_tfidf, y_train)

# Feature importance
feature_importance = pd.DataFrame({'Feature': tfidf_vectorizer.get_feature_names(),
'Importance': rf_model.feature_importances_})
top_features = feature_importance.sort_values(by='Importance', ascending=False).head(10)
print("Top 10 important features:\n", top_features)
```

R STUDIO :

User interface -

```
library(shiny)
# Define UI for the application
ui <- fluidPage(
  titlePanel("Twitter Sentiment Analysis"),
  sidebarLayout(
    sidebarPanel(
      textInput("tweet", "Enter Tweet Text:", ""),
      actionButton("analyze", "Analyze Sentiment")
    ),
    mainPanel(
      textOutput("result")
    )
  )
)
# Define server logic for sentiment prediction
server <- function(input, output) {
  observeEvent(input$analyze, {
    new_tweet <- data.frame(text = clean_text(input$tweet), stringsAsFactors = FALSE)
    new_dtm <- DocumentTermMatrix(Corpus(VectorSource(new_tweet$text)), control =
list(dictionary = Terms(dtm)))
    new_matrix <- as.matrix(new_dtm)
    prediction <- predict(model, new_matrix)
    output$result <- renderText({
      paste("Predicted Sentiment:", ifelse(prediction == 4, "Positive", ifelse(prediction == 0,
"Negative", "Neutral")))
    })
  })
}
# Run the application
shinyApp(ui = ui, server = server)
```