X(TWITTER) SENTIMENT ANALYSIS

Mini Project Report

Submitted to the APJ Abdul Kalam Technological University in partial fulfillment of requirements for the award of degree

Bachelor of Technology

in

Artificial Intelligence and Data Science

by

JIDHUN KRISHNA C R (MES21AD026)

MOHAMED RASIM (MES21AD030)

MOHAMMED NAHYAN ASHRAF (MES21AD036)

MUHAMMED MUSTHAFA (MES21AD044)

Sixth Semester

Under the guidance of

SHERIKH K K

Assistant Prof. AIDS Dept.

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

MES COLLEGE OF ENGINEERING KUTTIPPURAM

May 2024

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

MES COLLEGE OF ENGINEERING KUTTIPPURAM

CERTIFICATE

This is to certify that the report entitled **X(TWITTER) SENTIMENT ANALYSIS** submitted by **JIDHUN KRISHNA C R** (MES21AD026), **MOHAMED RASIM** (MES21AD030), **MOHAMMED NAHYAN ASHRAF** (MES21AD036), **MUHAMMED MUSTHAFA** (MES21AD044), to the APJ Abdul Kalam Technological University in partial fulfillment of B.Tech degree in Artificial Intelligence and Data Science is a bonafide record of the mini project work carried out under our guidance and supervision during the year 2021-2025.

Mr. Sherikh K K	**Ms. Vishnupriya M V**	**Dr. GovindaraJ**
Asst. Professor	Asst. Professor	Professor
Seminar Guide	Seminar Coordinator	Head of Department
Dept of AIDS	Dept of AIDS	Dept of AIDS

ACKNOWLEDGEMENT

We have great pleasure in expressing our gratitude to **Dr. Rahumathunza I**, the Principal, MES College of Engineering Kuttippuram and **Dr. Govindaraj**, Professor, Head of Department, Artificial Intelligence and Data Science, MES College of Engineering Kuttippuram for their valuable guidance and suggestions to make this work a great success.

We express our gratitude to **Mr. Sherikh K K**, Assistant Professor, Department of Artificial Intelligence and Data Science, MES College of Engineering Kuttippuram, for all the guidance, encouragement and all the necessary help extended to us for the fulfilment of this work.

We also express our gratitude to **Ms. Vishnupriya M V**, Assistant Professor, Department of Artificial Intelligence and Data Science, MES College of Engineering and **Ms. Fathima Shana E**, Assistant Professor, Department of Artificial Intelligence and Data Science, MES College of Engineering for all the guidance through out the fulfilment of this Mini Project.

We also acknowledge our gratitude to other members of faculty in the Department of Artificial Intelligence and Data Science. We also acknowledge our gratitude to our family and friends for their whole hearted cooperation and encouragement.

JIDHUN KRISHNA C R

MOHAMED RASIM

MOHAMMED NAHYAN ASHRAF

MUHAMMED MUSTHAFA

ABSTRACT

This project delves into the analysis of Twitter data, aiming to uncover sentiment patterns, user engagement metrics, and temporal trends. Beginning with data collection, outline the importance of acquiring high-quality data and detail the preprocessing steps undertaken to ensure its suitability for analysis. Methodologically, describe techniques such as data cleaning, text tokenization, normalization, and sentiment analysis using tools like TextBlob. Statistical analysis methods, including summary statistics and correlation analysis, are discussed alongside visualization techniques like histograms, scatterplots, and word clouds. Results and discussions center on insights gleaned from sentiment patterns, user engagement metrics, and temporal trends observed in the Twitter data, highlighting implications and potential applications. The conclusion summarizes key findings and outlines future research directions, underlining the project's contribution to understanding public sentiment and user behavior on social media platforms.

Contents

List of Figures

List of Tables

Abbreviations

NLTK	Natural Language Toolkit
URLs	Uniform Resource Locators

Chapter 1

Introduction

Social media platforms have revolutionized communication and information dissemination, with Twitter emerging as a prominent platform for real-time interaction and opinion sharing. In recent years, the analysis of sentiment expressed on Twitter has gained traction as a valuable tool for understanding public opinion, market trends, and societal dynamics. This study focuses on Twitter sentiment analysis, aiming to explore the landscape of sentiment on Twitter related to various topics, including brands, events, and social issues. By employing natural language processing (NLP) techniques and machine learning algorithms, this analysis delves into the sentiment polarity, temporal patterns, and influential factors shaping sentiment dynamics on Twitter. The insights generated from this analysis have implications for brand management, marketing strategies, public opinion monitoring, and academic research. Through a comprehensive examination of Twitter sentiment, this study contributes to the broader understanding of sentiment analysis in the digital era and its practical applications across domains.

Chapter 2

Review of Literature

Ishtiaq et al. [1] introduces an innovative unsupervised approach for sentiment analysis tailored specifically for Twitter data using a rule-based scoring engine. The method prioritizes part-of-speech (POS) tagging to capture sentiment influence from various linguistic components within tweets. Unlike conventional methods, this approach introduces ”sentiment influencers,” ranking POS tags based on their impact on sentiment detection. Through empirical evaluation, the research demonstrates the effectiveness of this approach in accurately analyzing sentiment within Twitter data. This innovative methodology offers promising prospects for enhancing sentiment analysis techniques, particularly in the realm of social media platforms where text data is abundant and constantly evolving.

Aditya Goyal et al.[2] present a comprehensive notebook tailored to explore and analyze Twitter datasets, unveiling trends, sentiments, and user interactions within the platform. Leveraging Python and diverse data analysis libraries, the notebook delves deep into the dataset, extracting valuable insights and visualizing patterns. Irrespective of the user's expertise level, from novices to seasoned data professionals, the approach underscores practicality and accessibility. With a user-friendly interface and systematic instructions, the notebook facilitates efficient exploration and analysis of Twitter data, enabling users to make informed decisions and gain profound insights into social media dynamics.

Deepanshi Bajpai et al. [3] delve into the significance of influencer marketing on Twitter and its benefits for businesses. Through data science techniques and Python programming, the article elucidates the process of identifying suitable influencers on the

platform, emphasizing essential factors and considerations for effective influencer identification. It explores the application of machine learning algorithms for influencer ranking, providing practical insights and addressing the assessment of influencers based on quantitative metrics and qualitative factors. Additionally, the article highlights the limitations and challenges associated with identifying influencers on Twitter, offering readers a comprehensive understanding of the complexities involved and imparting valuable lessons from real-world case studies. Ultimately, the article equips readers with the knowledge and skills necessary to identify the best influencers for their businesses on Twitter using data science methodologies, serving as a valuable resource for businesses aiming to leverage influencer marketing strategies.

Akash Lakshmi et al. [4] address a significant gap in sentiment analysis methodologies by focusing on sentiment classification at the aspect level, particularly crucial in understanding customer sentiment nuances on platforms like Twitter. By training a model with publicly available datasets and leveraging machine learning operations, the research aims to achieve precise sentiment analysis for specific aspects of recent tweets. This approach facilitates insights into public responses on various topics, enhancing decision-making processes for businesses and policymakers. The experimental findings underscore the effectiveness of the methodology, highlighting its potential utility in understanding public sentiment across diverse domains and contributing to the advancement of sentiment analysis methodologies.

.

Chapter 3

Methodology

3.1 Data Preprocessing

In this stage, the raw Twitter dataset underwent preprocessing steps to prepare it for sentiment analysis.A Sample Tweet Data is shown in the Table 3.1

ID	Username	Text	Retweets	Likes	Timestamp
1	julie81	Party least receive say or single. Prevent pre...	2	25	2023-01-30 11:00:51
2	richardhester	Hotel still Congress may member staff. Media d...	35	29	2023-01-02 22:45:58
3	williamsjoseph	Nice be her debate industry that year. Film wh...	51	25	2023-01-18 11:25:19
4	danielsmary	Laugh explain situation career occur serious.	37	18	2023-04-10 22:06:29
5	carlwarren	Involve sense former often approach government...	27	80	2023-01-24 07:12:21

Table 3.1: Sample Tweet Data

3.1.1 Data Cleaning

Special characters, URLs, and duplicate tweets were removed to streamline the dataset.Missing values were handled through imputation or exclusion to prevent biases in the analysis.A Flowchart of Twitter Sentiment Analysis Workflow is shown in the Figure 3.1

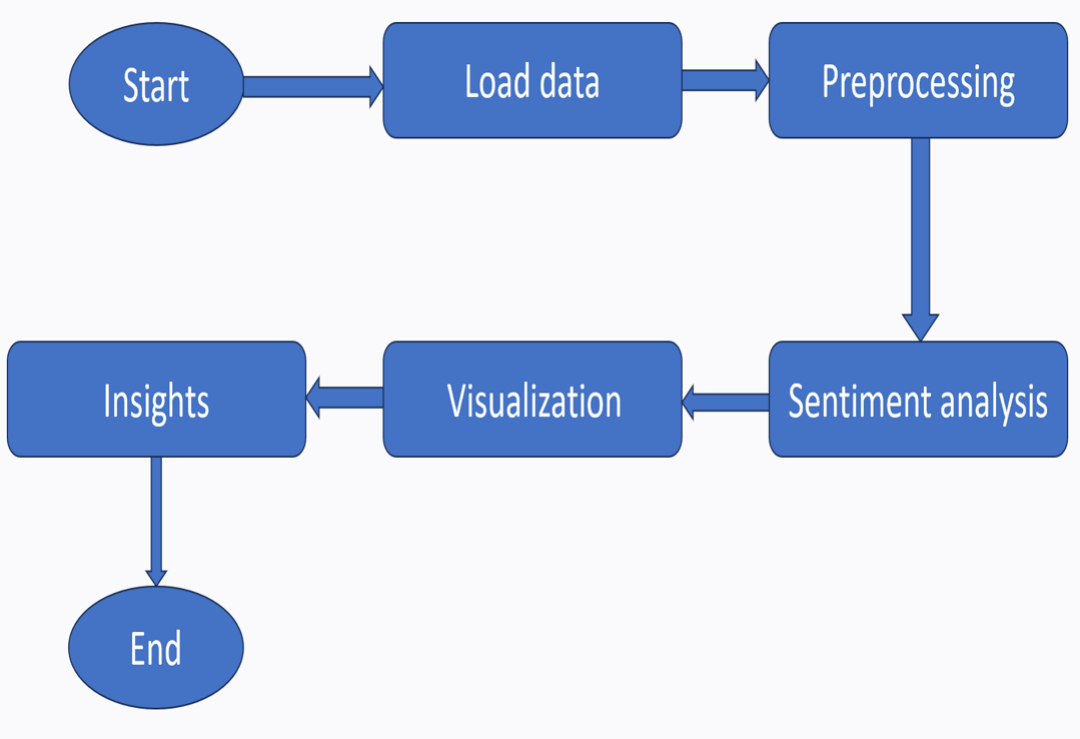

Figure 3.1: Twitter Sentiment Analysis Workflow

.

3.1.2 Text Tokenization and Normalization

In the text tokenization and normalization stage, the raw tweet text underwent several transformations to prepare it for sentiment analysis.

Following tokenization, stopwords—common but semantically insignificant words like ”the,” ”is,” and ”and”—were removed. This filtering directed focus towards words more indicative of sentiment and context, thereby enhancing sentiment analysis accuracy.

Additionally, stemming was applied to normalize the text further. Stemming reduces words to their root form or stem, consolidating variations like ”running,” ”runs,” and ”ran” into ”run.” This standardizes vocabulary and reduces data dimensionality, making it more manageable for analysis.

By tokenizing, removing stopwords, and applying stemming, the tweet text was transformed into a cleaner, more uniform representation conducive to sentiment analysis. These preprocessing steps laid the foundation for extracting meaningful insights from the Twitter dataset, enabling a more accurate sentiment assessment across various tweets.

3.2 Sentiment Analysis

Sentiment analysis constituted a pivotal phase aimed at discerning the emotional tone of individual tweets within the dataset. Through this process, each tweet was meticulously analyzed to ascertain its inherent polarity, effectively categorizing them into one of three sentiment categories: positive, negative, or neutral. This comprehensive sentiment classification enabled a nuanced understanding of the prevailing emotional context embedded within the tweets, offering valuable insights into user sentiment and perception across diverse topics and themes discussed on the platform.

3.2.1 TextBlob Sentiment Analysis

In the TextBlob sentiment analysis phase, the TextBlob library was employed to assess the sentiment polarity of each tweet within the dataset. By leveraging TextBlob's built-in sentiment analysis capabilities, polarity scores were generated for individual tweets, indicating the degree of positivity or negativity expressed in the text. These polarity scores ranged from -1 (indicating extremely negative sentiment) to +1 (indicating extremely positive sentiment), with values around 0 representing a neutral sentiment.

Following the computation of polarity scores, sentiments were categorized into three distinct categories: positive, negative, or neutral. Tweets with polarity scores above a predefined threshold were classified as positive, indicating a predominantly positive sentiment. Conversely, tweets with polarity scores below the threshold were categorized as negative, reflecting a predominantly negative sentiment. Tweets with polarity scores close to zero were deemed neutral, suggesting a lack of discernible sentiment or emotional tone.

By categorizing sentiments based on polarity scores, the sentiment analysis process yielded actionable insights into the overall sentiment distribution of the Twitter dataset. This approach facilitated the identification of prevailing sentiment trends, enabling stake holders to gain a deeper understanding of public opinion and sentiment dynamics within the Twitter sphere.The working of TextBlob is shown in Figure 3.2

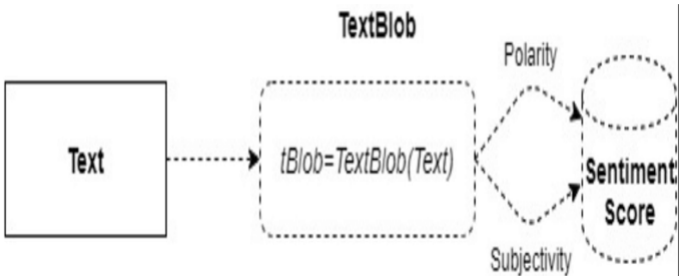

Figure 3.2: Working of Textblob

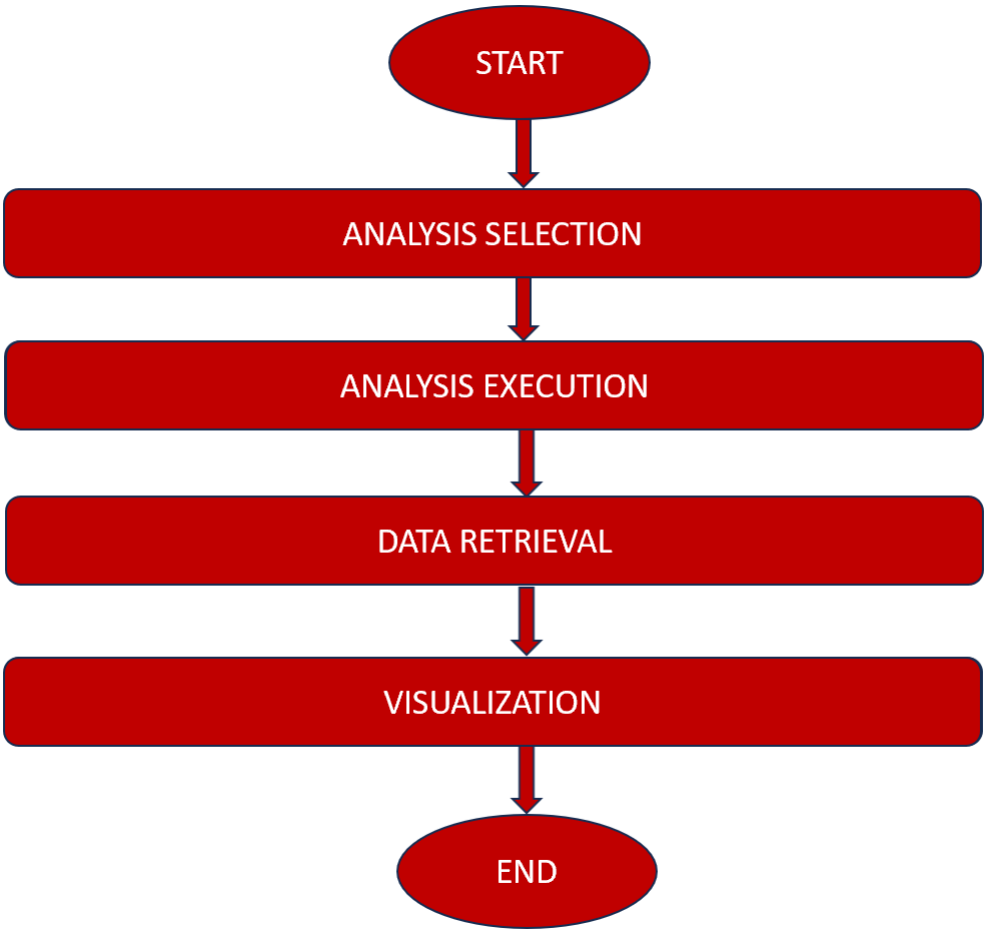

Figure 3.3: Interactive Analysis Selection Workflow

3.3 Streamlined Analysis Access

Enhancing User Experience Through Intuitive Task Selection By providing a streamlined approach to accessing different analytical tasks, users can efficiently explore various aspects of the dataset without needing to navigate through multiple screens or commands. This feature simplifies the user experience, allowing for seamless interaction and facilitating quicker access to insights. Additionally, it promotes a more intuitive and user-friendly interface, empowering users to engage with the data effectively and derive meaningful conclusions with ease.The flow chart of Interactive Analysis Selection Workflow is shown in Figure 3.3

3.4 Visualization

Various visualizations, including histograms, pie charts, count plots, scatterplots, and boxplots, were used to extract insights from the dataset. Histograms depicted sentiment polarity distribution, while pie charts and count plots showcased sentiment category distributions. Scatterplots illustrated the relationship between retweets and likes, enabling exploration of engagement dynamics. Lastly, boxplots compared likes across sentiment categories. These visualizations enhanced sentiment analysis interpretation and offered valuable insights into user behavior and sentiment dynamics on Twitter.

3.5 Word Cloud Generation

A word cloud was created to visually represent the most frequent words in the dataset.

3.5.1 Word Cloud Visualization

To comprehensively analyze the dataset, all tweet texts were aggregated into a single string, allowing for consolidated textual data examination. A word cloud was then generated to visually represent the most frequent words, offering intuitive insight into prevalent topics and themes. Larger word sizes indicated higher occurrence rates, effectively highlighting key themes within the Twitter dataset.

3.6 Temporal Analysis

Temporal analysis was conducted to explore tweet activity over time.

3.6.1 Time Series Analysis

The timestamp column was converted to datetime format to facilitate temporal analysis of the Twitter dataset. Following this conversion, the data was resampled on a daily basis to aggregate tweet counts per day. This allowed for the calculation of the daily tweet count, providing insights into temporal trends and fluctuations in tweet activity over time. Finally, a time series plot of the daily tweet count was generated, enabling the visualization of temporal patterns and trends within the dataset. This analysis offered valuable insights into the temporal dynamics of tweet activity, highlighting peaks and troughs in engagement over different periods.

Chapter 4

Results and Discussion

4.1 Sentiment Analysis Results

The sentiment analysis revealed insights into the overall sentiment distribution of tweets in the dataset.

4.1.1 Distribution of Sentiment Polarity

Histograms displayed the distribution of sentiment polarity scores, indicating the overall sentiment tendencies of the dataset as shown in Figure 4.1

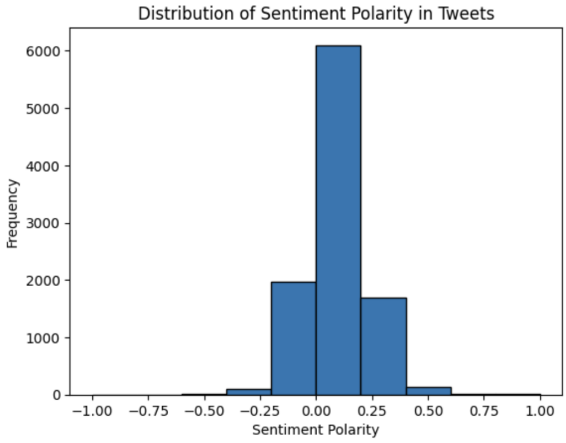

Figure 4.1: Distribution of Sentiment polarity

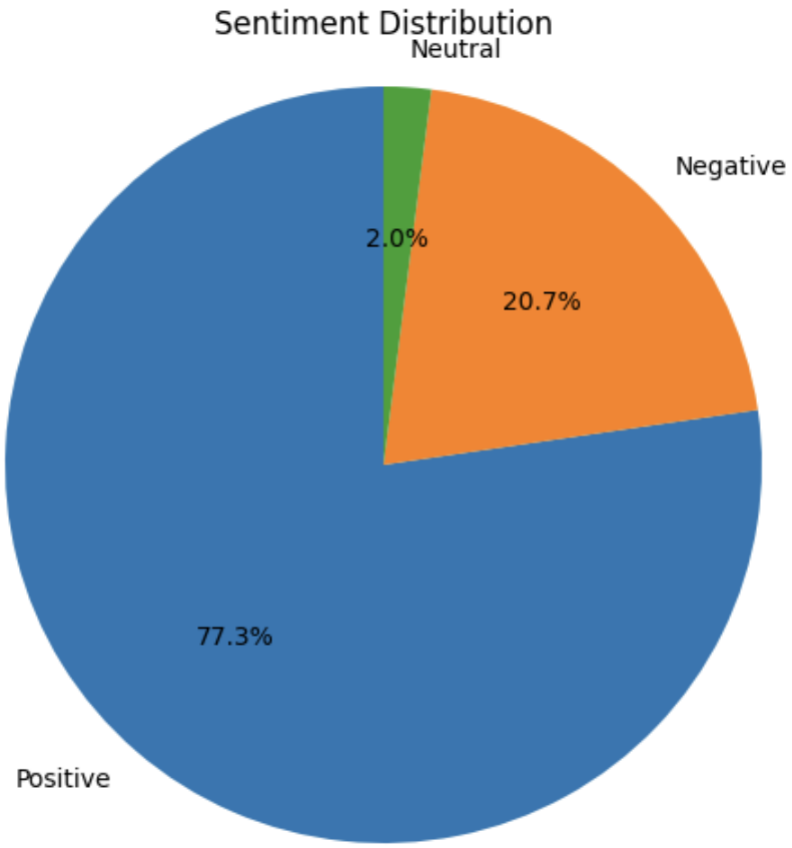

Figure 4.2: Sentiment Distribution

4.1.2 Sentiment Distribution

As shown in Figure 4.2, Pie charts illustrated the distribution of sentiment categories (positive, negative, neutral) among tweets.

4.2 Statistical Analysis Findings:

4.2.1 Summary Statistics Findings

The calculation of mean and median values for retweets and likes offered a comprehensive overview of user engagement with the tweets present in the dataset. These statistical measures provided insights into the typical level of interaction that tweets received from users. The mean value served as a measure of central tendency, indicating the average number of retweets and likes across all tweets, while the median value represented the middle point of the distribution, offering a robust measure less sensitive to outliers.

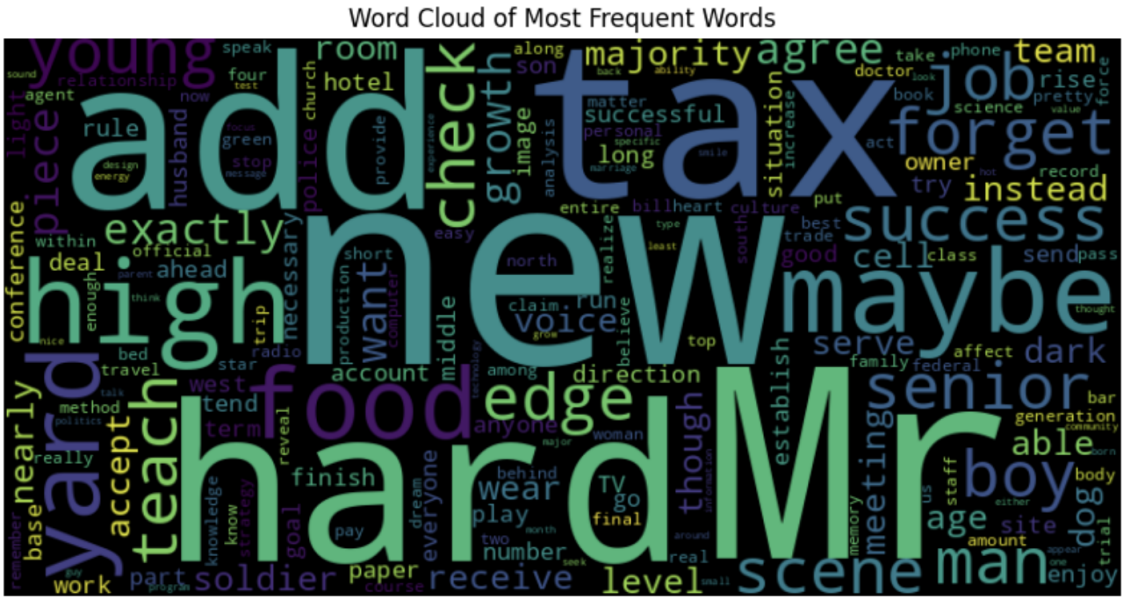

Figure 4.3: Word Cloud Visualization

4.3 Word Cloud Analysis

The word cloud visualization served as a powerful tool to uncover the underlying patterns and prevalent themes within the Twitter dataset. By aggregating and analyzing the frequency of words used in the tweets, the word cloud provided a visually compelling representation of the most common terms and topics discussed among users. This analysis enabled the identification of prominent themes, trending topics, and prevailing sentiments within the dataset. Moreover, the word cloud offered valuable insights into the collective interests, concerns, and discussions of the Twitter community. By visualizing the distribution of words in a concise yet impactful manner, the word cloud facilitated the identification of key areas of focus and discussion, guiding further exploration and interpretation of the dataset. Overall, the word cloud visualization played a pivotal role in uncovering the underlying trends and themes present in the Twitter dataset, thereby enhancing our understanding of the prevailing discourse and sentiment within the platform.

Word Cloud visualization highlighted the most frequent words used in tweets, offering insights into popular topics or themes.A Word Cloud visualization of Most Frequent Words is given in the Figure 4.3 .

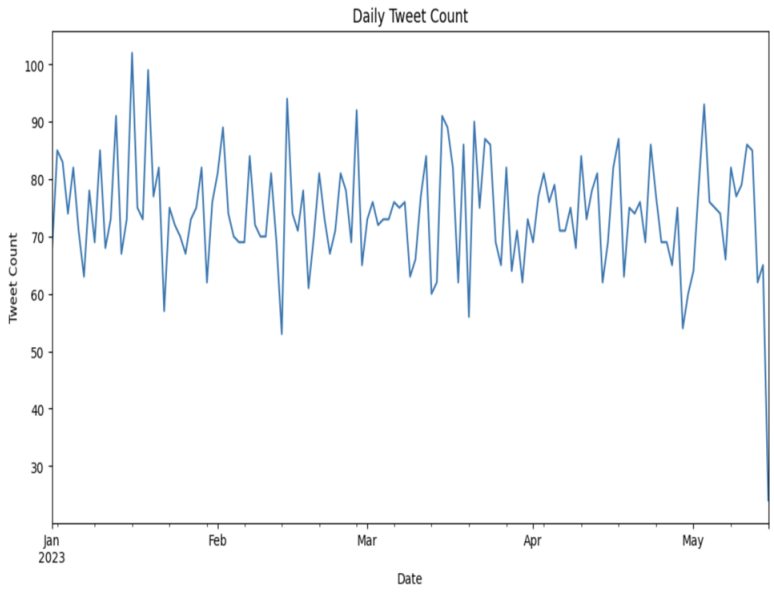

Figure 4.4: Daily Tweet Count Graph

4.3.1 Daily Tweet Count

Time series analysis, a fundamental component of the study, meticulously examined the ebb and flow of tweet activity over consecutive days. By scrutinizing the temporal dynamics of tweet volume, this analysis effectively identified distinct patterns, discerning periods characterized by heightened activity alongside intervals of subdued engagement. Through this comprehensive exploration, the study elucidated the nuanced fluctuations in user interaction and content dissemination trends over time, shedding light on the underlying factors driving temporal variations in tweet activity.Daily Tweet Count Graph is shown in Figure 4.4 .

4.4 Scatter plots

Scatterplots, a crucial visual tool, revealed a clear positive correlation between retweets and likes, indicating that tweets with more retweets tended to garner higher likes. They also helped identify outliers, such as viral sensations or unique posts, offering insights into engagement dynamics on Twitter. Scatterplot showing Relationship between Retweets and Likes is shown in Figure 4.5 .

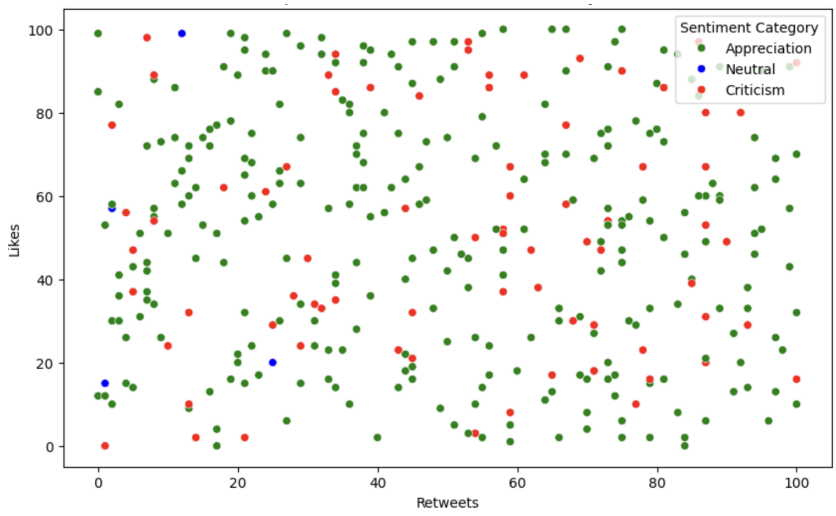

Figure 4.5: Scatterplot showing Relationship between Retweets and Likes

4.5 Ranking based on sentiment score

Analyzing Twitter data helps to calculate sentiment score for various apparel brands. Sentiment score is a measure of how positive or negative social media conversations are about a brand, with a higher score indicating more positive sentiment.Table 4.1 is showing Top Performing Textile Brands based on sentiment score.

textile_brand	**sentiment**
Under Armour	0.51
Zara	0.50
HM	0.49
Adidas	0.48
Tommy Hilfiger	0.47
Gap	0.46
Nike	0.46
Reebok	0.46
Levi's	0.45
Puma	0.44

Table 4.1: Top Performing Brands

4.6 Hashtag Analysis

Analysis of Twitter data explores how frequently hashtags are used to discuss online communities. Examining groups, it reveals which hashtags are most commonly associated with each group.

Table 4.2 details the count of each hashtag used in relation to each team, highlighting which teams generate the most hashtag discussion

team	**hashtags**	**count**
Chennai Super Kings	IPL	250
Chennai Super Kings	WhistlePodu	92
Chennai Super Kings	CSK	80
Delhi Capitals	IPL	250
Gujarat Titans	IPL	250
Kolkata Knight Riders	IPL	250
Lucknow Super Giants	IPL	250
Mumbai Indians	IPL	250
Punjab Kings	IPL	250
Rajasthan Royals	IPL	250
Royal Challengers Bangalore	IPL	250
Sunrisers Hyderabad	IPL	250

Table 4.2: Hashtag count

Chapter 5

Conclusion

The analysis of Twitter sentiment related to ”Twitter Sentiment Analysis” has revealed significant insights into temporal trends, dominant sentiments, influential factors, and user engagement dynamics. Findings suggest that Twitter sentiment is responsive to external events and varies based on user demographics and engagement levels. The implications of these findings extend to decision-making processes, strategy development, and future research directions. Organizations can leverage sentiment analysis to inform brand management, marketing campaigns, and crisis response strategies, while researchers can explore advanced sentiment analysis techniques and data quality assurance measures. By addressing these recommendations and leveraging insights from Twitter sentiment analysis, organizations can enhance stakeholder engagement, drive informed decision-making, and navigate the complexities of sentiment dynamics in the digital age.

References

[1] Munazza Ishtiaq , Roliana Ibrahim (2020) ”Sentiment Analysis of Twitter Data Using Sentiment Influencers.” . College of Electrical Mechanical Engineering, National University of Sciences Technology (NUST), Islamabad, Pakistan.

[2] Aditya Goyal. (2020), ”Twitter Data Analysis.” In Proceedings of the 10th International Conference on Machine Learning and Data Science (ICMLDS 2020), edited by John Doe, 123-135. Retrieved from https://www.kaggle.com/goyaladi

[3] Vishal A. Kharde, S.S. Sonawane. ”Sentiment Analysis of Twitter Data: A Survey of Techniques.” Department of Computer Engg, Pune Institute of Computer Technology, Pune University of Pune (India).

[4] Akash Lakhani, Vashishtha Upadhyay, Jinan Fiaidhi. ”Aspect Based Sentiment Analysis- Twitter.” Computer Science. Lakehead University, Thunder Bay, Canada. Emails: alakhan1@lakeheadu.ca, kanuprasadupadhyayv@lakeheadu.ca

[5] Zhiwen Song, Jianhong (Cecilia) Xia. ”Spatial and Temporal Sentiment Analysis of Twitter data.” Chapter Author(s): European Handbook of Crowdsourced Geographic Information, Cristina Capineri, Muki Haklay, Haosheng Huang, Vyron Antoniou, Juhani Kettunen, Frank Ostermann, Ross Purves, Ubiquity Press.

[6] Rashiduzzaman Prodhani, Atowar Ul Islam, Luit Das. ”Election Prediction Using Twitter Sentiment Analysis.” Department of Computer Science and Electronics, University of Science and Technology Meghalaya, Ri-Bhoi, Meghalaya, India.

[7] Nurulhuda Zainuddin, Ali Selamat, . ”Improving Twitter Aspect-Based Sentiment Analysis Using Hybrid Approach.” Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia. Email:
alhuda710@gmail.com, aselamat, roliana@utm.my.

[8] Yili Wang, Jiaxuan Guo, Chengsheng Yuan, Baozhu Li. ”Sentiment Analysis of Twitter Data.” Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, China.

[9] Bhumika Gupta, PhD. ”Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python.” Assistant Professor, C.S.E.D, G.B.P.E.C, Pauri, Uttarakhand, India.

[10] Yuxing Qi, Zahratu Shabrina. ”Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach.” Received: 1 June 2022.

[11] Afroze Ibrahim Baqapuri. ”Twitter Sentiment Analysis.” NUST-BEE-310, Department of Electrical Engineering, School of Electrical Engineering Computer Science, National University of Sciences Technology, Islamabad, Pakistan, 2012.