Tweet Sense: Analyzing Public Sentiment on X (Twitter)

Term Project3 – Final Submission

Shanthibooshan Subramanian

Bellevue University

DSC680

Amirfarrokh Iranitablob

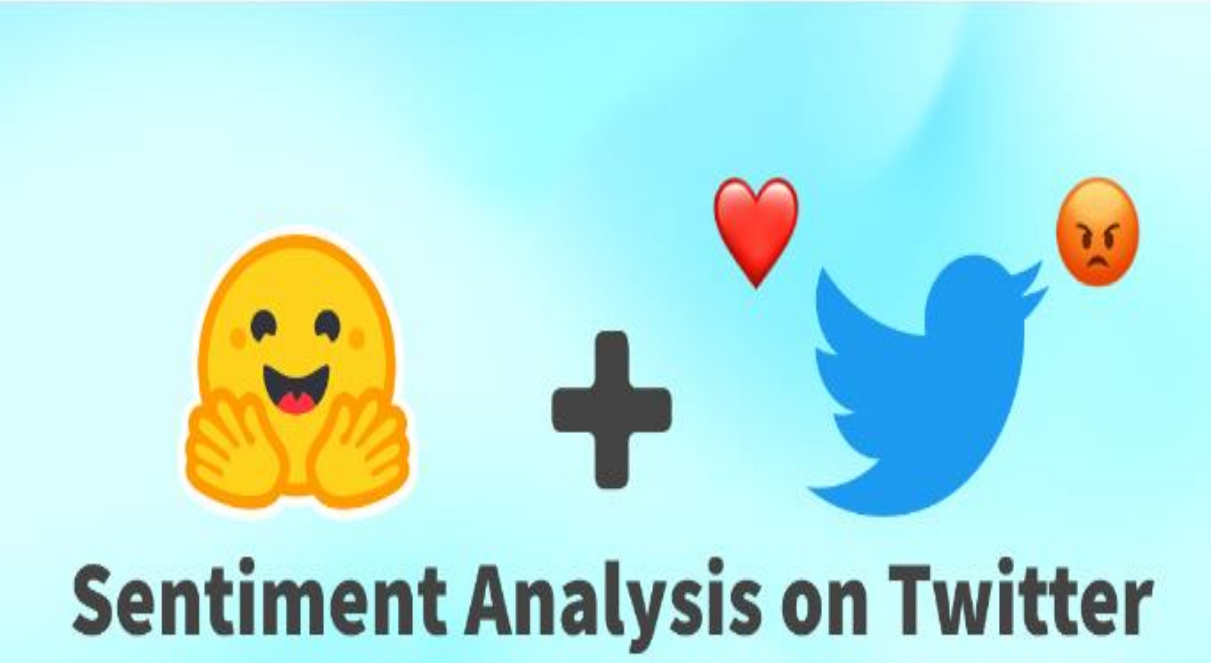
Table of Contents

1.0 Introduction 2
1.1 Problem Statement. 3
1.2 Importance/usefulness of solving the problem. 4
2.0 Dataset Overview 4
2.1 Dataset Details 4
2.2 Dataset Dictionary 4
2.3 Data Preprocessing for Analysis and Modeling 5
3.0 Comprehensive Analysis Summary 5
3.1 Data Exploration and Initial Insights 5
3.2 Data Preparation 11
3.3 Model Building and Evaluation. 13
4.0 Conclusion 14
4.1 Outcome of analysis and model building. 14
4.2 Model Deployment and Implementation Decision 15
4.3 Potential challenges and additional opportunities. 16
4.4 Conclusion. 16
4.5 Assumptions and Limitations 16
4.6 Recommendations 17
4.7 Ethical Assessment 17
5.0 References 18

1.0 Introduction

In today's digital age, social media platforms like X(Twitter) serve as valuable real-time public opinion and sentiment sources. Understanding how people feel about various topics, events, and entities can provide crucial insights for businesses, policymakers, and individuals. With millions of tweets generated daily, capturing and analyzing this vast amount of data presents both opportunities and challenges.

Sentiment analysis, or opinion mining, is a natural language processing (NLP) technique used to determine the sentiment expressed in text. It involves classifying text as positive, negative, or neutral and can be applied to various domains, including social media, customer reviews, and market research. By analyzing sentiments expressed in tweets, organizations can gain a deeper understanding of public opinion, monitor brand reputation, and make data-driven decisions.



The dataset for this project consists of labeled tweets with sentiments categorized into positive, negative, neutral, and irrelevant. Various sentiment analysis models will be evaluated for their effectiveness in classifying tweet sentiments. The results will provide actionable insights into public sentiment and guide strategic decision-making.

By exploring these aspects, the project aims to offer a comprehensive understanding of public sentiment on X(Twitter), empowering organizations to make informed decisions and enhance their engagement strategies

1.1 Problem Statement.

Understanding public sentiment is crucial for businesses, policymakers, and individuals. This project aims to provide insights into how the public perceives certain topics or events on X(Twitter), facilitating data-driven decisions and strategies. Accurate sentiment prediction will help organizations better understand customer feedback, monitor brand reputation, and gauge public reaction to events.

To achieve these goals, the project will address the following questions:

- Can we detect sentiment shifts in tweets before, during, and after significant events? This will help understand how public opinion evolves, offering valuable insights for event planning and crisis management.

- Which words, phrases, or hashtags are most associated with positive or negative sentiments?
  Identifying key sentiment drivers will enable organizations to tailor their messaging and campaigns more effectively.

- What metrics are used to evaluate the performance of our sentiment prediction model?
  Establishing robust evaluation metrics such as accuracy, precision, recall, and F1-score ensures the reliability and validity of the sentiment analysis.

- How can organizations use sentiment analysis to respond to public relations crises? Timely sentiment analysis can guide organizations in crafting appropriate responses to mitigate negative impacts and enhance their public image during crises.

- How can companies use sentiment data to improve products, services, and customer relations?
  Leveraging sentiment data to identify areas for improvement can enhance customer satisfaction and strengthen customer relationships.

- Can sentiment analysis predict future market trends or consumer behavior? Using sentiment data to anticipate future trends and consumer behavior can provide a competitive edge, allowing organizations to proactively adapt their strategies.

By exploring these questions, the project aims to deliver comprehensive insights into public sentiment on X(Twitter), enabling informed decision-making and effective strategy implementation.

1.2 Importance/usefulness of solving the problem.

Sentiment analysis of tweets is crucial for understanding public opinion and enhancing customer engagement. It helps businesses gauge market trends, manage reputations, and provide better customer service. By tracking and classifying sentiments, organizations can make informed decisions, address negative feedback, and stay responsive to emerging trends. Overall, solving this problem supports more effective communication and strategic planning in a digital landscape.

2.0 Dataset Overview

The dataset used for this term project can be accessed on Kaggle and includes tweets with 74,682 records for training and 1,000 records for validation. Each record represents a tweet categorized into different sentiment classes, with parameters used to assess the sentiment expressed in the tweet.

2.1 Dataset Details

The dataset used in this project comprises tweets collected from X(Twitter), labeled with sentiment categories such as Positive, Negative, Neutral, and Irrelevant. Each tweet includes metadata like the tweet id text, entity, and possibly user information.

2.2 Dataset Dictionary

The dataset dictionary provides a detailed description of each variable included in the dataset:

- **Tweet_ID**: A unique identifier for each tweet. It is used to differentiate between individual tweets within the dataset.
- **Entity**: The source or context of the tweet. This field typically indicates the subject or topic associated with the tweet, such as a game or brand name.
- **Sentiment**: The sentiment classification of the tweet. This field categorizes the sentiment expressed in the tweet into one of several predefined categories, such as Positive, Negative, Neutral, or Irrelevant.
- **Tweet_content**: The actual text of the tweet. This field contains the content of the tweet as posted by the user, which is the primary input for sentiment analysis.

This dataset consists of labeled tweets from both training and validation sets, which will be used for training and validating sentiment analysis models. Each tweet is tagged with

its sentiment and associated with an entity, providing both the text and context necessary for classification tasks.

2.3 Data Preprocessing for Analysis and Modeling

The data preprocessing steps involved several key actions to prepare the dataset for analysis and modeling. Data cleaning included removing URLs, mentions, hashtags, and special characters from the tweets. Tokenization was used to split tweets into individual words, followed by normalization, which converted text to lowercase and removed stop words. Vectorization transformed the cleaned text into numerical features using TF-IDF. Missing values were addressed to ensure completeness. The dataset was then divided into training and validation sets. Feature extraction utilized the TF-IDF vectorizer to convert tweet text into numerical features.

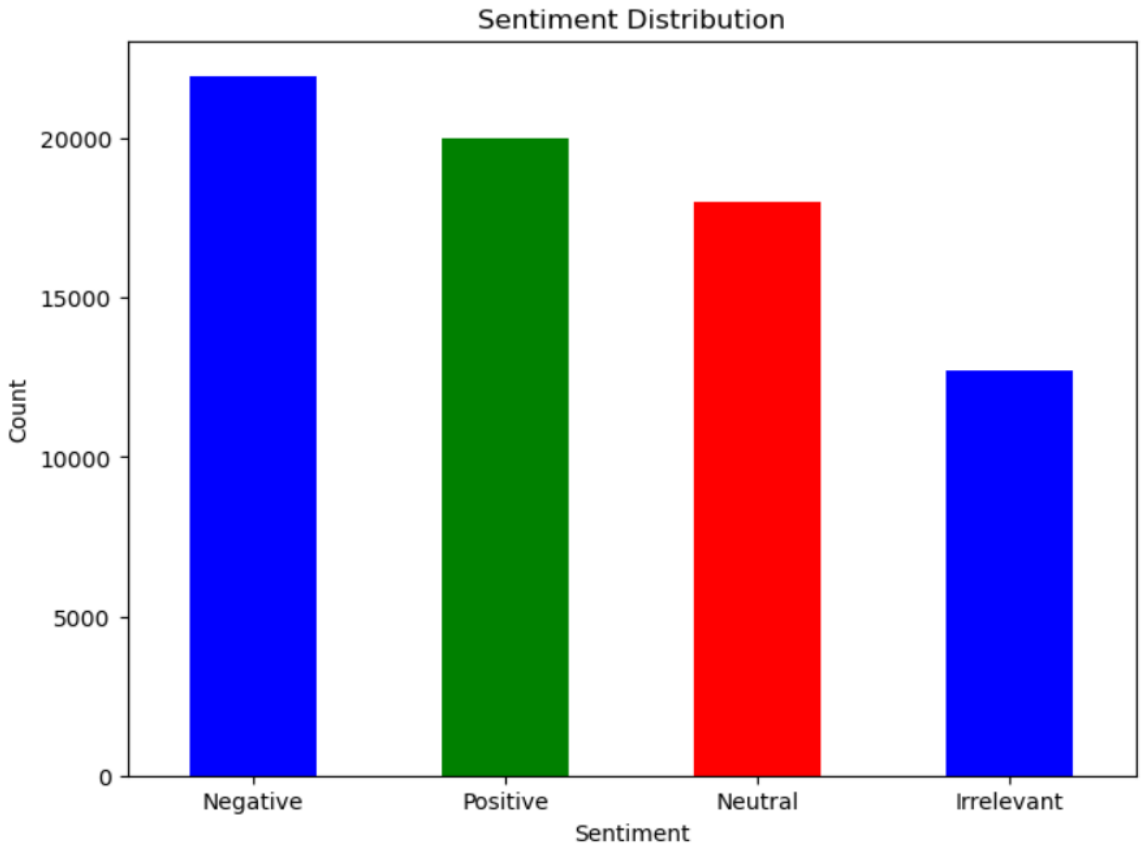
3.0 Comprehensive Analysis Summary

The dataset comprises 74,682 records in the training set and 1,000 records in the validation set, each with 4 columns. Within this dataset, the column labeled "Sentiment" serves as the target variable for this project, categorizing tweets into various sentiment classes: "Irrelevant," "Negative," "Neutral," and "Positive."

3.1 Data Exploration and Initial Insights

Bar chart – Sentiment Distribution Analysis

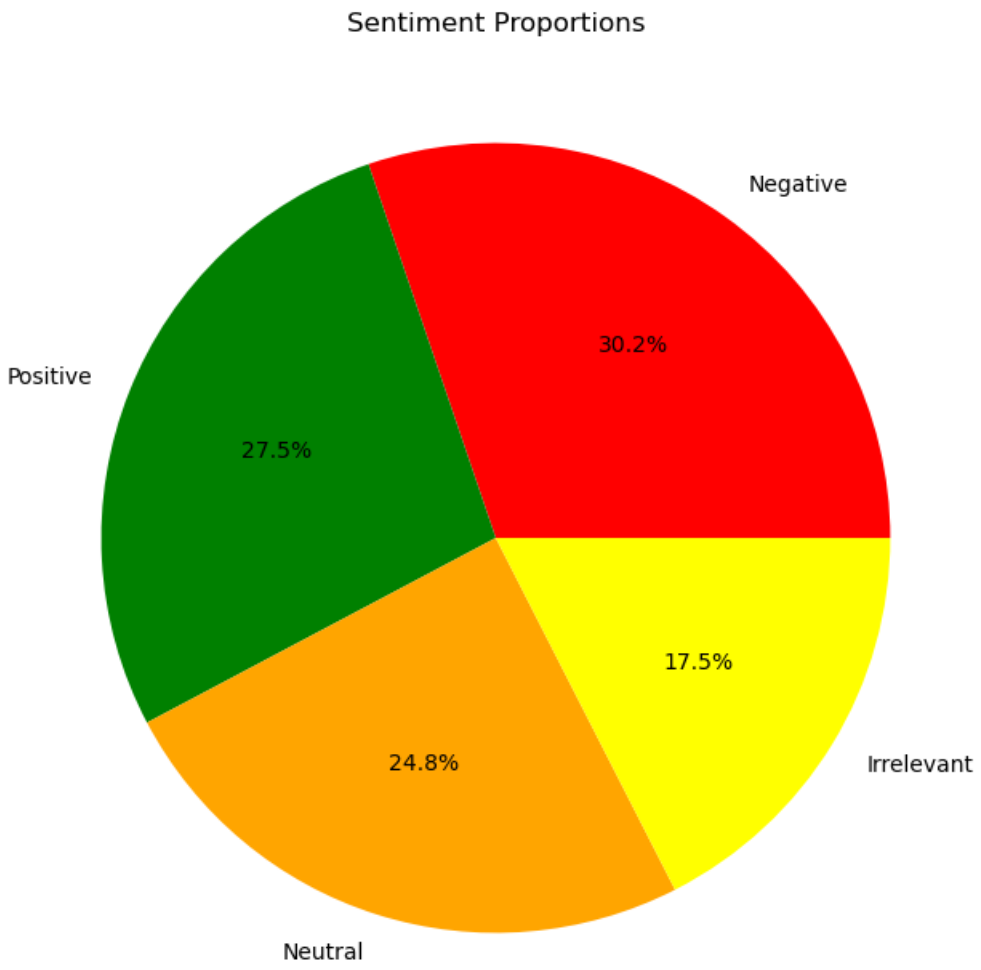
The bar chart reveals the distribution of sentiments within the dataset, illustrating that negative sentiment is the most prevalent among the tweets, followed by positive sentiment. Neutral and irrelevant sentiments are observed less frequently. This distribution highlights that the dataset predominantly features tweets with a negative emotional tone, which is crucial for understanding the overall sentiment landscape.



The predominance of negative sentiment suggests a significant presence of critical or dissatisfied content. This insight can guide further analysis, such as investigating the underlying causes of negative sentiment, monitoring sentiment changes over time, or comparing sentiment across different topics or user demographics. Additionally, the lower frequency of neutral and irrelevant sentiments indicates that most tweets convey clear emotional stances, providing a more focused basis for sentiment analysis. Overall, this exploration into sentiment distribution is instrumental in framing subsequent data analysis and modeling efforts, offering a clear view of the emotional dynamics present in the dataset.

Pie Chart Analysis of Sentiment Distribution

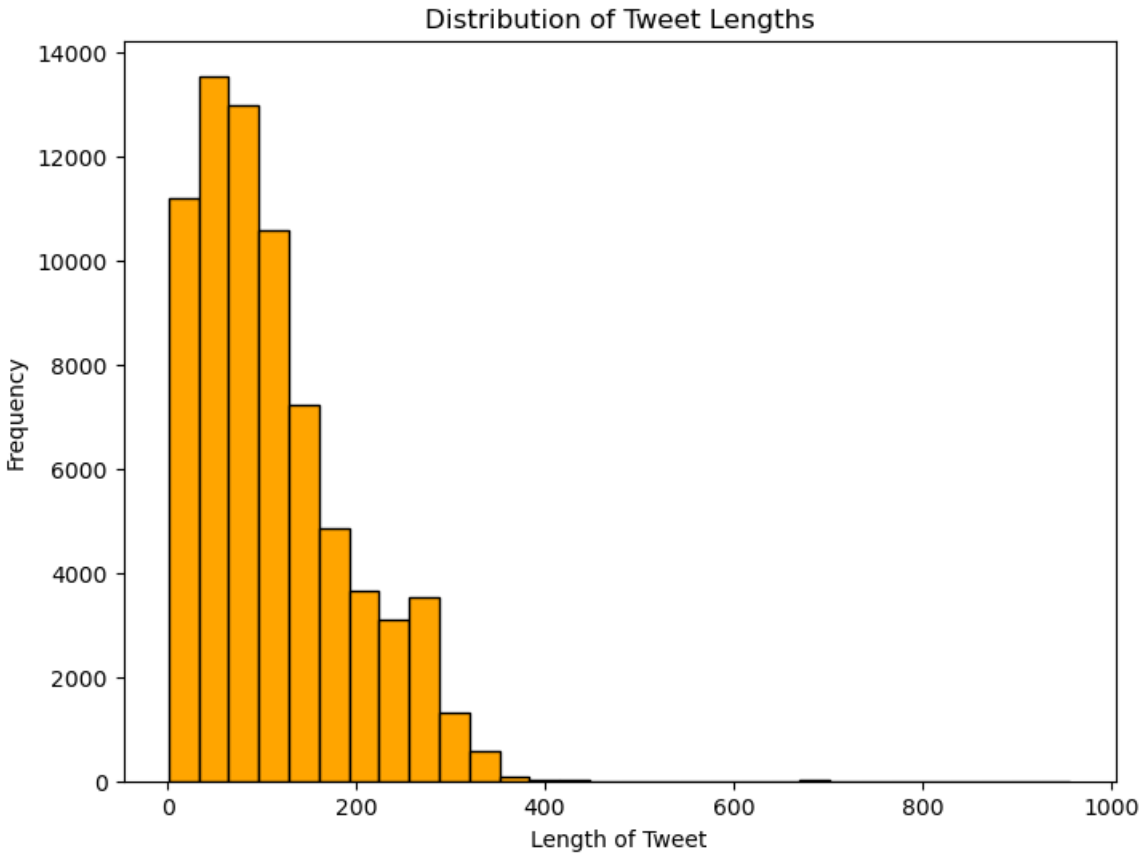
The pie chart reveals that negative sentiment is the most common, occupying 30.2% of the dataset. Positive sentiment follows, representing 27.5%, while neutral sentiment comprises 24.8% and irrelevant sentiment makes up 17.5%. This visualization is crucial for quickly grasping the overall sentiment distribution, which directly informs model training and data analysis.



Recognizing that negative sentiment is predominant can guide the focus of model development, emphasizing the need for robust detection and classification of negative sentiments. Understanding these proportions helps tailor the model to handle the most frequent sentiment types effectively, thereby improving overall predictive performance.

Distribution of Tweet Length Analysis

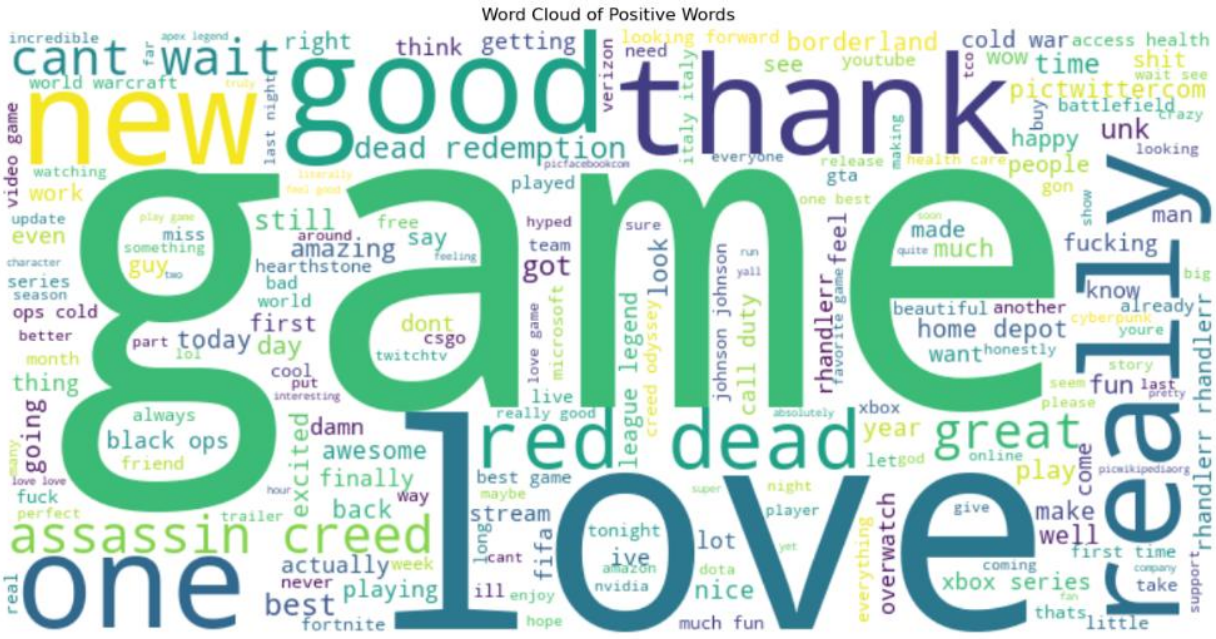
The histogram reveals a right-skewed distribution of tweet lengths, with most tweets being relatively short (under 100 characters). The distribution peaks in the shorter length range, while a long tail extends up to approximately 800 characters, reflecting a smaller number of significantly longer tweets. The tweet lengths span from 0 to 1000 characters.



This visualization is instrumental in understanding tweet length patterns, which is crucial for text preprocessing and model training. Highlighting that most tweets are short, it informs decisions on feature engineering, such as adjusting n-gram ranges or managing text truncation. This insight helps tailor preprocessing steps and feature extraction techniques to better accommodate the dataset's characteristics, ultimately improving the performance of sentiment analysis models.

Word Cloud Analysis of most Frequent Positive Terms

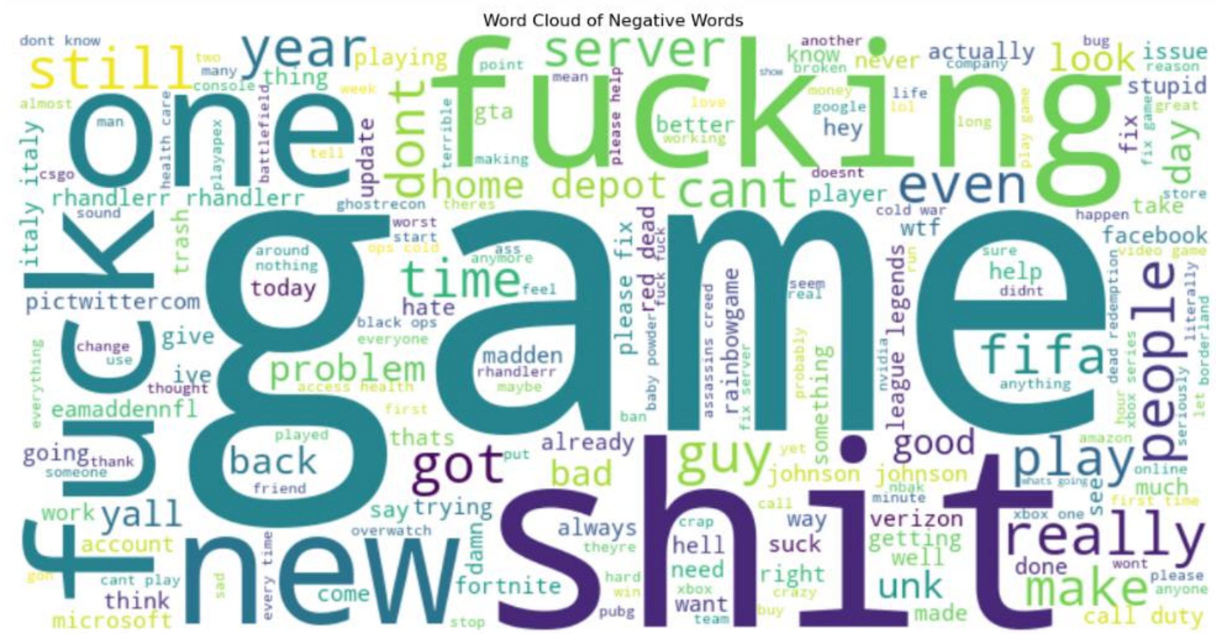
The word cloud visually represents the most frequently occurring positive terms in the dataset. Prominent themes include gaming-related terms (e.g., "game," "PlayStation," "Call of Duty"), expressions of positivity (e.g., "love," "great," "happy"), and technology-related words (e.g., "update," "technology"). Terms related to community and social interaction also feature prominently.



This visualization is valuable for quickly identifying key topics and sentiments within the dataset. For data analysis and model building, the word cloud aids in feature selection by highlighting prominent keywords and themes, ensuring that the most relevant terms are included in the analysis. It also provides context for understanding the dataset's focus, which can inform the design of sentiment analysis models and enhance the interpretation of their results. By emphasizing frequent positive terms, this visualization helps tailor the model to better capture and analyze the sentiment conveyed in the tweets.

Word Cloud Analysis of Negative Terms and User Frustrations

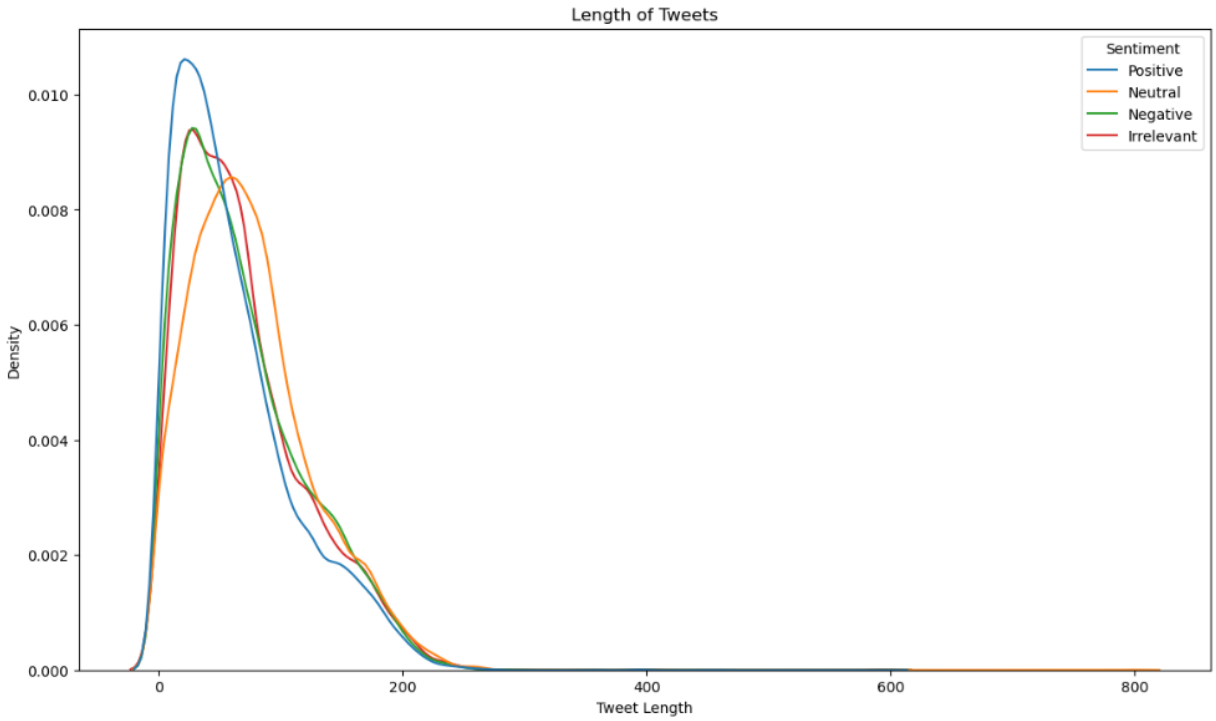
The word cloud highlights a strong presence of negative sentiments and user frustrations, reflecting frequent criticisms related to gaming issues, customer service, and general dissatisfaction. Prominent terms include expletives, game-related problems, and customer service complaints, underscoring the overall negative tone of the dataset.



This visualization is useful for understanding prevalent negative themes within the dataset. It can guide model development by identifying key negative terms and issues to focus on. For sentiment analysis, incorporating these terms can enhance model accuracy in detecting dissatisfaction and negative feedback, ensuring that the model is better equipped to handle and classify negative sentiments effectively.

Density Plot Analysis of Tweet Length Distribution by Sentiment

The density plot reveals that most tweets are relatively short, with positive tweets being slightly longer on average. Negative and neutral tweets exhibit similar length distributions, while irrelevant tweets are predominantly shorter.



This visualization is useful for understanding how tweet lengths vary by sentiment. By analyzing these patterns, data scientists can optimize sentiment analysis models to handle different tweet lengths more effectively. It provides insights into how the length of tweets might influence sentiment expression, which can guide feature engineering and model training to improve overall performance.

3.2 Data Preparation

In preparation for modeling and analysis, the original dataset underwent the following steps:

1. **Handling Missing Values**: The training dataset initially had 686 NaN values in both the Tweet_content and Tweet_length columns. These NaN values were addressed by replacing them with empty strings in the Tweet_content column. The validation dataset had no NaN values and was unaffected.

2. **Removing Duplicates**: The training dataset had 2,700 duplicated rows, which were removed to ensure the dataset's integrity. The validation dataset had no duplicated rows and remained unchanged.

3. **Data Cleaning**:
    - **Conversion to Strings**: The Tweet_content column in both the training and validation datasets was converted to strings, and NaN values were replaced with empty strings.
    - **URL Removal**: URLs were removed from the Tweet_content column, resulting in a Cleaned_Tweet column with tweets free from URLs.
    - **Removal of Mentions and Hashtags**: Mentions and hashtags were removed from the Cleaned_Tweet column, leaving only tweet text.
    - **Non-Alphabetic Characters**: Non-alphabetic characters were removed, ensuring that the Cleaned_Tweet column contains only alphabetic characters and spaces.

4. **Tokenization and Text Processing**:
    - **Tokenization**: The Cleaned_Tweet column was tokenized, creating a Tokens column with lists of individual words.
    - **Lowercasing**: Tokens were converted to lowercase to maintain uniformity.
    - **Stopword Removal**: Common stopwords were removed from the Tokens column, leaving only meaningful words.
    - **Lemmatization**: Tokens were lemmatized to their base or root form.
    - **Short Words Removal**: Words with fewer than three characters were removed, ensuring the Tokens column contains only longer, more meaningful words.

5. **Reassembly of Processed Text**: Tokens were joined back into strings, updating the Cleaned_Tweet column with fully processed tweets containing meaningful text.

6. **Combining Datasets**: The training and validation datasets were combined into a single dataset with labels indicating the source (Train and Validation). The combined dataset was reset to include these labels, and a sample of the data was made available for verification.

These data preparation steps ensure that the dataset is thoroughly cleaned, transformed, and optimized for analysis and modeling. By handling missing values, removing duplicates, and preprocessing text data including tokenization, stop word removal, and lemmatization. The dataset is now well-suited for building robust and accurate predictive models. These efforts enhance the dataset's integrity and reliability, laying a strong foundation for effective sentiment analysis and ensuring that the resulting models perform at their best.

3.3 Model Building and Evaluation.

A comprehensive approach was taken to model building and evaluation. Initially, several classification models were considered, including Logistic Regression, Bernoulli Naive Bayes, Multinomial Naive Bayes, and Linear SVC. Each model was assessed for its ability to classify tweets into sentiment categories effectively. Hyperparameter tuning was then performed to optimize the parameters of each model, aiming to enhance their performance by finding the most suitable configurations.

Model performance was rigorously evaluated using metrics such as precision, recall, F1-score, and accuracy, which provided a holistic view of how well each model performed across different sentiment categories. Additionally, confusion matrices were used to visualize the performance of each model, offering insights into the correct and incorrect predictions for each sentiment class. This visualization helped in understanding how well the models differentiated between various sentiment labels and highlighted areas for potential improvement.

Model Results:

Model	Accuracy	F1-score (Irrelevant)	F1-score (Negative)	F1-Score (Neutral)	F1-score (Positive)	Macro Avg F1 Score	Weighted Avg F1 Score
Logistic Regression	83%	0.81	0.86	0.81	0.84	0.83	0.83
Bernoulli Naive Bayes	80%	0.75	0.85	0.79	0.77	0.79	0.80
Multinomial Naive Bayes	86%	0.85	0.87	0.87	0.86	0.86	0.86
LinearSVC	89%	0.89	0.91	0.89	0.88	0.89	0.89

4.0 Conclusion

4.1 Outcome of analysis and model building.

The analysis revealed that the LinearSVC model emerged as the most effective, achieving an overall accuracy of 89%. It demonstrated high precision and recall across all sentiment categories, particularly excelling in identifying negative and irrelevant sentiments. The robustness of the LinearSVC model in distinguishing between different sentiments made it the standout choice among the evaluated models.

- **Logistic Regression**: The model achieved an overall accuracy of 83%. It performed best in classifying negative sentiments with an F1-score of 0.86 and had the lowest performance with irrelevant tweets, showing an F1-score of 0.81. The model demonstrated balanced performance across all sentiment categories, with macro and weighted average F1-scores of 0.83, indicating robust and consistent results.

- **Multinomial Naive Bayes**: This model achieved an accuracy of 80%, demonstrating strong precision for identifying "Irrelevant" tweets (0.98) but with lower recall (0.61), indicating it accurately classified irrelevant tweets but missed some instances. It performed well with "Negative" tweets, showing balanced precision (0.84) and recall (0.86). For "Neutral" tweets, the model achieved high precision (0.91) but lower recall (0.70). The model excelled in recalling "Positive" tweets with a high recall of 0.94 but had lower precision (0.66), suggesting it identified many positive tweets but with some misclassification.

- **Bernoulli Naive Bayes**: The model achieved an accuracy of 86%, demonstrating strong overall performance. It excelled in precision for "Irrelevant" tweets (0.93) and showed good recall (0.78). The model also performed well with "Negative" tweets, with high precision (0.83) and excellent recall (0.92). For "Neutral" tweets, it maintained high precision (0.89) and strong recall (0.84). "Positive" tweets were classified with balanced performance, achieving a precision of 0.85 and recall of 0.88.

- **LinearSVC**: Demonstrated excellent performance with an overall accuracy of 89%, achieving high precision and recall across all sentiment categories. It showed particularly strong results for negative tweets (precision: 0.92, recall: 0.91) and irrelevant tweets (precision: 0.93, recall: 0.85). For neutral tweets, it maintained high precision (0.91) and solid recall (0.88), while for positive tweets, it exhibited a precision of 0.84 and a strong recall of 0.93. The model's macro and weighted averages for precision, recall, and F1-score were consistently high, reflecting balanced performance.

In summary, the analysis highlights the superior performance of the LinearSVC model for sentiment classification tasks, with Multinomial Naive Bayes and Logistic Regression also demonstrating strong results. Bernoulli Naive Bayes, while effective in specific areas, showed room for improvement. These insights are valuable for refining sentiment analysis approaches and deploying the most effective models for accurate and reliable sentiment classification.

Model Results After Hyperparameter Tuning

We selected the LinearSVC and Multinomial Naive Bayes models for further tuning due to their strong pre-tuning performance. The tuning process aimed to refine these models, enhancing their ability to classify sentiments accurately.

- **LinearSVC**: After tuning, the LinearSVC model achieved an accuracy of 89.43%. It demonstrated balanced precision (89.67%), recall (89.43%), and F1 score (89.45%), confirming its effectiveness as the top-performing model.

- **Multinomial Naive Bayes**: The tuned Multinomial Naive Bayes model reached an accuracy of 86.37%, with precision, recall, and F1 score all around 86%, showing improved performance compared to its pre-tuning results.

These models' robustness and reliability post-tuning make them strong candidates for real-world sentiment analysis applications.

4.2 Model Deployment and Implementation Decision

The Linear SVC model is recommended for deployment due to its outstanding performance with an accuracy of 89%, demonstrating robust precision and recall across all sentiment categories. This model's balanced performance and high accuracy make it well-suited for real-time sentiment analysis applications, providing reliable and actionable insights from social media data. Its effectiveness in identifying positive and neutral sentiments positions it as a strong candidate for integration into social media monitoring tools and customer feedback systems.

The Linear SVC model's high accuracy and well-rounded performance across various metrics make it a reliable choice for practical implementation. Its suitability for real-time analysis ensures that it can deliver timely and valuable insights, enhancing the ability to respond promptly to public sentiment. Integrating this model into monitoring tools will optimize sentiment analysis capabilities, supporting better decision-making and engagement strategies based on real-time data.

4.3 Potential challenges and additional opportunities.

One of the primary challenges is handling the complexity of sarcasm and the use of emojis, which can obscure the true sentiment of tweets and lead to inaccuracies in analysis. The evolving nature of language on social media also complicates sentiment detection, as new slang and expressions constantly emerge. Additionally, real-time sentiment analysis requires substantial computational resources to process and analyze data efficiently.

On the other hand, there are significant opportunities to enhance the project. By integrating advanced deep learning models such as LSTM and BERT, the analysis can achieve a deeper understanding of context and meaning. Expanding the dataset to include a broader range of tweets from diverse sources will improve model performance and accuracy. Furthermore, developing scalable real-time analysis infrastructure will enable timely insights and more effective responses to emerging trends and issues.

4.4 Conclusion.

The sentiment analysis project successfully evaluated multiple models, with the **LinearSVC** model emerging as the most effective in classifying tweets into sentiment categories. Hyperparameter tuning was performed to optimize the model's performance, enhancing its accuracy and overall effectiveness. The project highlighted the importance of balancing performance metrics across different classes and addressed challenges such as class imbalance and the differentiation of similar sentiments. By deploying the LinearSVC model, the project is positioned to deliver high-quality, real-time sentiment analysis. Future work should focus on refining model performance further, exploring advanced techniques, and addressing challenges related to language diversity and class imbalance to enhance the accuracy and applicability of sentiment analysis tools.

4.5 Assumptions and Limitations

Assumptions: In developing the Tweet Sense project, I assumed that the tweets collected are broadly representative of public sentiment. This assumption underpins the validity of the sentiment analysis conducted, aiming to capture a snapshot of prevailing opinions and emotions within the X(Twitter) community.

Limitations: However, it's important to acknowledge several limitations. Firstly, the model's generalizability may be constrained by the dataset's primary focus on English-language tweets. Sentiment expression can vary significantly across languages and

regional dialects, potentially limiting the model's applicability in diverse linguistic contexts. Secondly, the size and diversity of the dataset utilized can impact the model's performance and robustness. A larger and more diverse data set would likely enhance the model's ability to capture broader sentiment trends more accurately.

4.6 Recommendations

To boost Tweet Sense's performance, we recommend using advanced deep learning techniques like Long Short-Term Memory (LSTM) networks and BERT (Bidirectional Encoder Representations from Transformers). These models excel in natural language processing tasks, including sentiment analysis, by effectively capturing contextual nuances and semantic relationships. Expanding the dataset to include more diverse and recent tweets across various languages and regions would improve the model's training data. This expansion can enhance the model's adaptability and accuracy in capturing global sentiment trends, making it more useful across different demographics and locations.

Developing a scalable infrastructure for real-time sentiment analysis is essential for timely insights and decision-making. Implementing technologies that support continuous data ingestion, processing, and analysis will enable Tweet Sense to provide up-to-date sentiment assessments efficiently. This capability is crucial for applications like market research, crisis monitoring, and brand reputation management.

4.7 Ethical Assessment

I prioritized the protection of user privacy by adhering to stringent data privacy regulations. All personal identifiers were removed from the dataset to ensure anonymization and data access was restricted to authorized personnel only. These measures help safeguard user information and comply with legal standards. Addressing potential biases in the dataset and model was a key focus. I employed techniques to detect and mitigate biases, ensuring that the sentiment analysis provided fair and accurate results. This included using a diverse dataset to avoid skewed representations and applying fairness metrics to evaluate the model's performance across different demographics. Additionally, continuous monitoring and adjustments were made to the algorithm to correct any identified biases, promoting an ethical and responsible approach to sentiment analysis.

5.0 References

- https://www.kaggle.com
- https://365datascience.com/
- https://towardsdatascience.com/
- https://www.analyticsvidhya.com/blog/2021/06/X(Twitter)-sentiment-analysis-a-nlp-use-case-for-beginners/
- https://www.techsalerator.com/post/top-X(Twitter)-sentiment-data-providers
- https://medium.com/@ubaidhaina/X(Twitter)-sentiment-analysis-05decd00a29f
- https://www.researchgate.net/publication/358439871_X(Twitter)_Sentiment_Analysis_using_Natural_Language_Processing
- https://twitter.com