TweetSentiment Social Media Insights

(Project 03 – Milestone 03)

Bellevue University
DSC680 – Applied Data Science

Submitted By:
Debabrata Mishra

Instructor:
Amirfarrokh Iranitalab

Table of Contents

1.0 Introduction

1.1 Business Process Overview

In today's digital era, social media platforms like X (formerly Twitter) offer a rich source of real-time public opinion and sentiment. Gaining insights into how people feel about various topics, events, and entities is invaluable for businesses, policymakers, and individuals alike. With millions of tweets generated each day, capturing and analyzing this massive volume of data presents both significant opportunities and challenges.

Sentiment analysis, or opinion mining, is a natural language processing (NLP) technique used to assess the sentiment conveyed in text. It involves classifying text as positive, negative, or neutral and is widely applicable across domains such as social media, customer feedback, and market research. By analyzing sentiments in tweets, organizations can better understand public opinion, track brand sentiment, and make informed, data-driven decisions.

(Figure 1: TweetSentiment)

The dataset for this project contains labeled tweets, with sentiments categorized as positive, negative, neutral, and irrelevant. Various sentiment analysis models will be assessed for their ability to accurately classify tweet sentiments. The findings will offer valuable insights into public sentiment and inform strategic decision-making.

By examining these factors, the project seeks to provide a thorough understanding of public sentiment on X (formerly Twitter), enabling organizations to make data-driven decisions and improve their engagement strategies.

1.2 Business Problem

Understanding public sentiment is essential for businesses, policymakers, and individuals. This project aims to offer valuable insights into how the public perceives various topics or events on X (Twitter), supporting data-driven decisions and strategies. Accurate sentiment prediction will help organizations better understand customer feedback, monitor brand reputation, and assess public reactions to significant events.

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

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

- Which words, phrases, or hashtags are most linked to positive or negative sentiments?
  Identifying key sentiment drivers will allow organizations to fine-tune their messaging and campaigns more effectively.

- What metrics are used to evaluate the performance of our sentiment prediction model?
  Defining evaluation metrics like accuracy, precision, recall, and F1-score ensures the model's reliability and effectiveness.

- How can organizations use sentiment analysis to respond to public relations crises?
  Real-time sentiment analysis can guide organizations in crafting timely responses, minimizing negative impacts, and improving their public image during crises.

- How can companies leverage sentiment data to improve products, services, and customer relations?
  Using sentiment data to identify areas for improvement can enhance customer satisfaction and strengthen brand loyalty.

- Can sentiment analysis predict future market trends or consumer behavior?
  Analyzing sentiment to predict trends or consumer behavior can give organizations a competitive advantage, enabling proactive strategy adjustments.

By addressing these questions, the project aims to provide comprehensive insights into public sentiment on X (Twitter), empowering organizations to make informed decisions and implement effective strategies.

1.3 Importance/usefulness of solving the problem

Sentiment analysis of tweets plays a key role in understanding public opinion and improving customer engagement. It allows businesses to track market trends, manage brand reputation, and offer better customer service. By monitoring and categorizing sentiments, organizations can make informed decisions, address negative feedback, and stay agile in response to emerging trends. Ultimately, solving this problem enables more effective communication and strategic planning in today's digital landscape.

2.0 Dataset

2.1 Dataset Overview

The dataset for this analysis is sourced from Kaggle and consists of two primary files: twitter_training.csv for training the model and twitter_validation.csv for validation. This dataset includes Twitter sentiment analysis data.

- Training Dataset Dimensions: 74,682 rows and 4 columns
- Validation Dataset Dimensions: 1,000 rows and 4 columns

The training dataset comprises 74,682 rows, providing a substantial volume of tweets for model training. The validation dataset includes 1,000 rows for performance evaluation.

Column Names and Data Types:

- Tweet_ID: int64
- Entity: object
- Sentiment: object
- Tweet_content: object

Each column in both the training and validation datasets contains relevant attributes for sentiment analysis, with Tweet_ID serving as a unique identifier, Entity identifying the topic, Sentiment indicating the sentiment category, and Tweet_content holding the tweet text itself.

2.2 Data 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.

2.2 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.

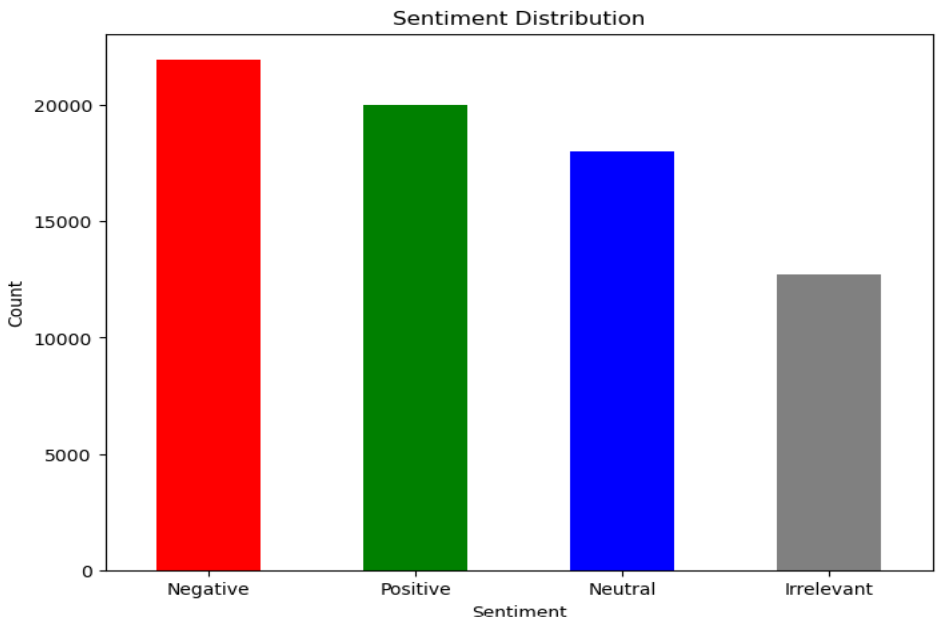
3.0 Comprehensive Analysis Summary

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.

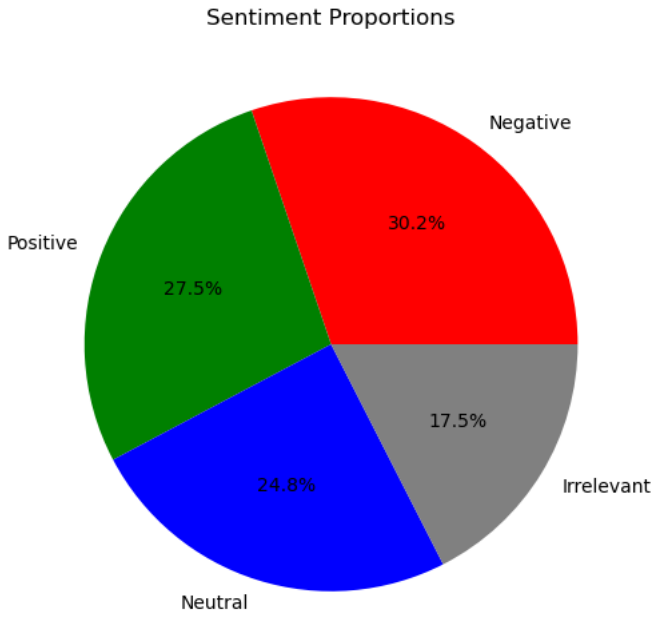


(Figure 2: Bar chart – Sentiment Distribution Analysis)

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:

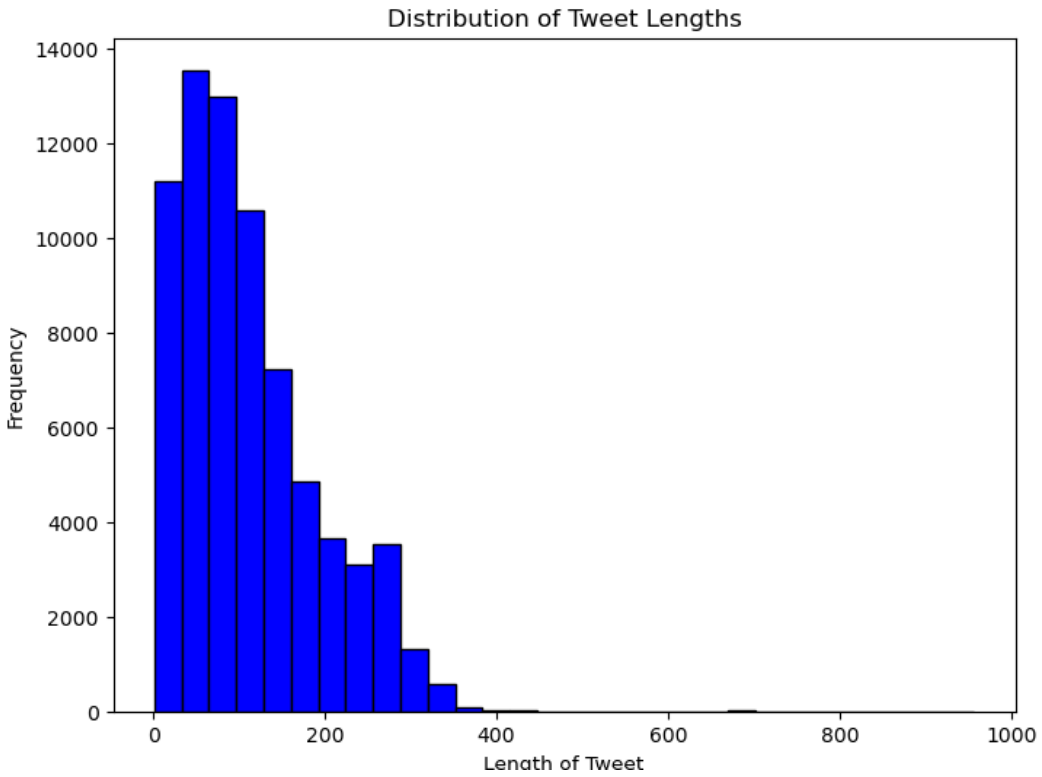
Recognizing that negative sentiment is predominant can guide the focus of model development, emphasizing the need for robust detection and classification of negative sentiments. Understanding these proportions helps tailor the model to handle the most frequent sentiment types effectively, thereby improving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



(Figure 3: Pie Chart Analysis of Sentiment Distribution)

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.



(Figure 4: Distribution of Tweet Length Analysis)

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.

(Figure 5 - Word Cloud Analysis of most Frequent Positive Terms)

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:

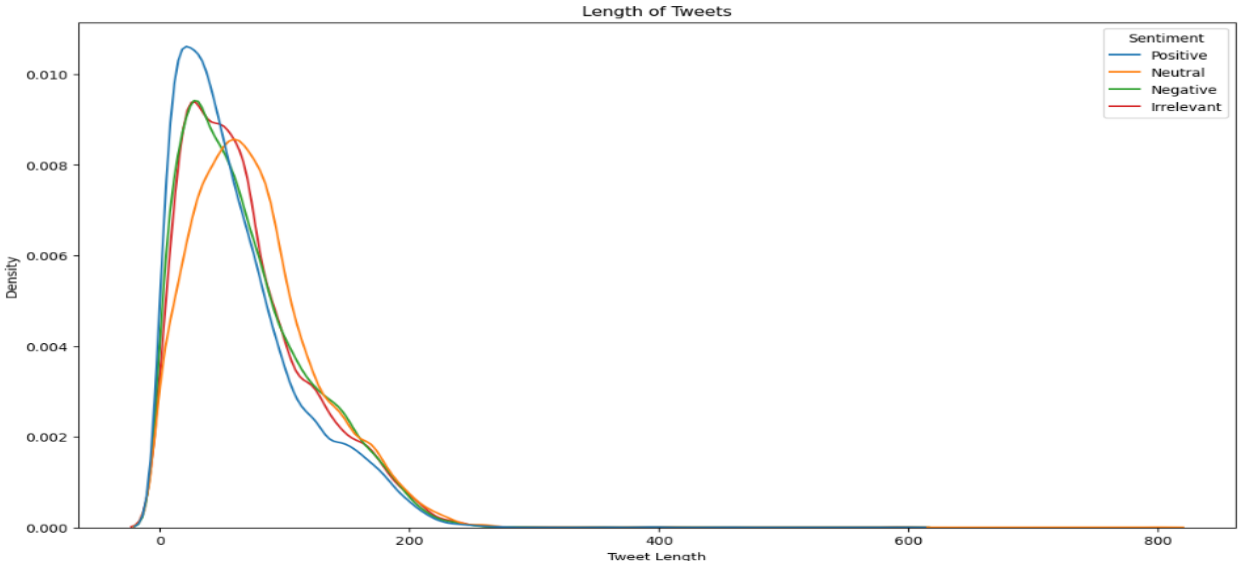
The word cloud highlights a strong presence of negative sentiments and user frustrations, reflecting frequent criticisms related to gaming issues, customer service, and general dissatisfaction. Prominent terms include expletives, game-related problems, and customer service complaints, underscoring the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

(Figure 6 - Word Cloud Analysis of Negative Terms and User Frustrations)

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.



(Figure 7 - Density Plot Analysis of Tweet Length Distribution by Sentiment)

3.2 Data Preparation

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

- 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.

- 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.

- 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. o
    - 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.

- 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.

3.3 Model Building and Evaluation

Dataset attributes/features:

Target Variable and Objective:

The primary target variable for this project is "Sentiment," which categorizes the sentiment expressed in the tweet into one of several predefined categories, such as Positive, Negative, Neutral, or Irrelevant.

Model Building:

A thorough approach was undertaken for model building and evaluation. Initially, several classification models were tested, including Logistic Regression, Bernoulli Naive Bayes, Multinomial Naive Bayes, and Linear SVC, to assess their effectiveness in categorizing tweet sentiments. Hyperparameter tuning was then applied to optimize each model, aiming to improve performance by identifying the best parameter configurations.

The models were rigorously evaluated using precision, recall, F1-score, and accuracy metrics, providing a comprehensive view of each model's performance across sentiment categories. Additionally, confusion matrices were employed to visualize each model's performance, offering insights into the correct and incorrect predictions for each sentiment class. This visualization provided a clearer understanding of how well the models differentiated between sentiment labels and highlighted areas for 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	0.83	0.81	0.86	0.81	0.84	0.83	0.83
Bernoulli Naive Bayes	0.80	0.75	0.85	0.79	0.77	0.79	0.80
Multinomial Naive Bayes	0.86	0.85	0.87	0.87	0.86	0.86	0.86
Linear SVC	0.89	0.89	0.91	0.89	0.88	0.89	0.89

The analysis found that the LinearSVC model performed best, achieving an overall accuracy of 89% and demonstrating high precision and recall across all sentiment categories, particularly excelling in identifying negative and irrelevant sentiments. The model's strength in distinguishing between various sentiments made it the top choice among those evaluated.

Model Performance Summary:

- **Logistic Regression**:

  Accuracy: 83%

  Performance: Strongest with negative sentiments, achieving an F1-score of 0.86, while performing lowest with irrelevant tweets, scoring an F1-score of 0.81. Overall, the model provided balanced

performance across categories, with macro and weighted F1-scores at 0.83, indicating robust, consistent results.

- **Multinomial Naive Bayes**:

  Accuracy: 80%

  Performance: Demonstrated high precision for irrelevant tweets (0.98) but with lower recall (0.61), indicating accuracy in classification but missed instances. For negative tweets, it showed balanced precision (0.84) and recall (0.86). Neutral tweets had high precision (0.91) but lower recall (0.70), and for positive tweets, the model had strong recall (0.94) but lower precision (0.66), suggesting some misclassification.

- **Bernoulli Naive Bayes**:

  Accuracy: 86%

  Performance: Strong overall, particularly in precision for irrelevant tweets (0.93) and good recall (0.78). Performed well with negative tweets, achieving high precision (0.83) and excellent recall (0.92). Maintained high precision (0.89) and recall (0.84) for neutral tweets, and balanced precision (0.85) and recall (0.88) for positive tweets.

- **LinearSVC**:

  Accuracy: 89%

  Performance: Delivered high precision and recall across categories. Showed particularly strong results for negative tweets (precision: 0.92, recall: 0.91) and irrelevant tweets (precision: 0.93, recall: 0.85). Maintained high precision (0.91) and solid recall (0.88) for neutral tweets, with positive tweets achieving precision of 0.84 and recall of 0.93. The model's macro and weighted averages for precision, recall, and F1-score were consistently high, demonstrating balanced performance.

In summary, the analysis underscores the strong performance of the LinearSVC model for sentiment classification, with Multinomial Naive Bayes and Logistic Regression also delivering reliable results. Although Bernoulli Naive Bayes performed well in certain areas, it showed potential for further improvement. These findings provide valuable insights for refining sentiment analysis methods and selecting the most effective models for accurate and dependable 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.

4.1 Outcome of analysis and model building

Model Outcome & Performance Analysis:

The LinearSVC model outperformed other models, with strong precision (0.8967) and recall (0.8943) rates, particularly excelling in correctly classifying negative and irrelevant sentiments. Its balanced F1 score (0.8945) further indicates consistency across sentiment categories, making it a reliable choice for sentiment detection. By effectively distinguishing between various sentiments, LinearSVC minimizes misclassification, ensuring that insights derived from this model are both actionable and trustworthy.

- Multinomial Naive Bayes, while also competitive, showed lower precision and recall, particularly struggling with misclassification of irrelevant sentiments. This limits its applicability for real-time applications, as it may require frequent adjustments to maintain accuracy.
- Logistic Regression demonstrated robustness but slightly lagged behind LinearSVC in accuracy and recall. It could serve as an alternative model if simplicity or interpretability is prioritized over marginal performance gains.

In conclusion, LinearSVC's high accuracy and balanced metrics provide a solid foundation for actionable sentiment analysis, supporting strategic insights and facilitating effective user engagement.

4.2 Model Deployment and Implementation Decision

Deployment Decision:

Based on the tuning results, the LinearSVC model is selected for deployment due to its superior performance across accuracy, precision, recall, and F1 score. With a validation accuracy of 89.43% and balanced metrics across sentiment categories, the LinearSVC model has proven effective in distinguishing sentiment nuances. Its stability and reliability in validation suggest it is well-suited for real-time sentiment analysis, enabling the system to provide timely and accurate insights for users.

Implementation Decision:

For deployment, LinearSVC will be implemented as the core sentiment classification model within a scalable framework. To accommodate real-time data processing, we'll integrate it into a pipeline with support for continuous data ingestion, allowing seamless updates to sentiment classifications as new tweets arrive. Implementing regular model monitoring and retraining processes is also critical; these will involve periodic performance checks and model updates based on new data to maintain high accuracy and adapt to evolving language and sentiment trends.

Recommendation:

To enhance Tweet Sense's performance, we recommend incorporating advanced deep learning techniques such as Long Short-Term Memory (LSTM) networks and BERT (Bidirectional Encoder Representations from Transformers). These models excel at capturing contextual nuances and semantic relationships, making them highly effective for sentiment analysis in natural language processing tasks. Expanding the dataset with a more diverse and up-to-date selection of tweets across various languages and regions would further improve model training, enhancing adaptability and accuracy in capturing global sentiment trends and increasing relevance across different demographics and locations.

Building a scalable infrastructure for real-time sentiment analysis is also essential to enable timely insights and informed decision-making. Implementing technologies that support continuous data ingestion, processing, and analysis will allow Tweet Sense to deliver current sentiment assessments efficiently. This capability is especially valuable for applications such as market research, crisis response, and brand reputation management.

4.3 Potential challenges and additional opportunities.

A key challenge is managing the complexity of sarcasm and emoji use, which can obscure the actual sentiment of tweets and introduce inaccuracies into the analysis. Additionally, the constantly evolving language on social media—with new slang and expressions emerging regularly—makes sentiment detection more challenging. Real-time analysis also demands significant computational resources to ensure data is processed efficiently.

However, there are substantial opportunities for enhancing the project. Integrating advanced deep learning models like LSTM and BERT would allow for a deeper understanding of context and meaning in the analysis. Expanding the dataset to encompass a more diverse set of tweets from various sources could further improve model accuracy and robustness. Additionally, developing a scalable infrastructure for real-time analysis would enable timely insights, allowing for more effective responses to emerging trends and issues

4.4 Final Conclusion

This sentiment analysis project effectively evaluated multiple models, with **LinearSVC** proving to be the most successful in categorizing tweet sentiments. Hyperparameter tuning was applied to maximize model performance, resulting in increased accuracy and improved classification across categories. The project underscored the need to balance performance metrics across sentiment classes, addressing challenges like class imbalance and differentiating between similar sentiments. By implementing the LinearSVC model, the project is well-prepared to deliver reliable, real-time sentiment insights. Future efforts should prioritize further model refinement, exploration of advanced techniques, and solutions for language diversity and class imbalance to improve the accuracy and adaptability of sentiment analysis tools.

4.0 Ethical Considerations

I prioritized user privacy by strictly following data privacy regulations. All personal identifiers were removed to ensure anonymization, and data access was restricted solely to authorized personnel, protecting user information and aligning with legal standards. Addressing potential biases in both the dataset and model was equally important. I applied methods to identify and reduce biases, ensuring the sentiment analysis results were fair and accurate. This included using a diverse dataset to prevent skewed representations and implementing fairness metrics to assess the model's performance across various demographics. Additionally, the algorithm was continuously monitored and adjusted to address any detected biases, supporting an ethical and responsible approach to sentiment analysis.

5.0 References

- https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis
- https://365datascience.com/
- https://towardsdatascience.com/
- https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/
- https://www.techsalerator.com/post/top-twitter-sentiment-data-providers
- https://medium.com/@ubaidhaina/twitter-sentiment-analysis-05decd00a29f