Sentiment Analysis of Social Media for Airline Customer Service Improvement

Intermediate Report

CS-584 MACHINE LEARNING

Team Members:
Sujith Chandra Dara (A20528384)
Bhargava Ramani Gorle (A20528476)

(1) Introduction:

The proliferation of social media platforms has fundamentally reshaped the dynamics of customer engagement for businesses across industries. These platforms, particularly Twitter, have become indispensable channels for customers to voice their opinions, provide feedback, and share experiences in real-time. For service-oriented sectors like the airline industry, where customer satisfaction is intricately linked to brand reputation and operational success, harnessing the wealth of information embedded within social media conversations has become paramount.

In the contemporary digital landscape, sentiment analysis stands out as a pivotal tool for businesses to navigate the vast ocean of social media data effectively. By leveraging computational techniques to discern sentiments, opinions, and emotions expressed in textual data, sentiment analysis offers invaluable insights into the prevailing attitudes and perceptions of customers towards a brand, product, or service. In the context of airlines, sentiment analysis of tweets provides a nuanced understanding of customer sentiment, enabling airlines to gauge satisfaction levels, identify areas for improvement, and proactively address concerns or issues raised by passengers.

Against this backdrop, this project endeavors to explore the realm of sentiment analysis within the context of the airline industry, with a specific focus on Twitter data. By harnessing the power of machine learning algorithms and natural language processing techniques, the project seeks to develop robust models capable of accurately categorizing airline-related tweets into sentiment classes, including positive, neutral, and negative. Through this endeavor, the project aims to equip airlines with actionable insights derived from social media data, thereby empowering them to make informed decisions, enhance customer experiences, and fortify their brand reputation in an increasingly interconnected and digitally-driven world.

(2) Problem Description:

Despite the vast potential offered by sentiment analysis of social media data, analyzing tweets presents unique challenges. The brevity, informality, and linguistic nuances inherent in tweets pose obstacles to accurate sentiment classification. Factors such as sarcasm, slang, and the presence of mixed sentiments within individual messages complicate the task of automated sentiment analysis. Consequently, achieving high accuracy in sentiment classification requires navigating the intricate landscape of human language expressed through tweets.

This project seeks to address these challenges by leveraging advanced machine learning techniques to develop robust sentiment analysis models tailored specifically for airline-related tweets. By overcoming the complexities of natural language and effectively capturing the nuances of sentiment expressed in tweets, the project aims to provide airlines with actionable insights derived from social media data, facilitating informed decision-making and proactive customer engagement strategies.

(3) Description of the Data Used in the Project:

The dataset comprises tweets directed at various airlines, collected from Twitter. Each tweet is annotated with sentiments classified as positive, neutral, or negative. Key attributes include:

- tweet_id: Unique identifier for each tweet.
- airline_sentiment: Sentiment of the tweet (positive, neutral, negative).
- text: Text content of the tweet.
- airline: Airline the tweet is directed at.
- retweet_count: Number of retweets.

The dataset contains 14,640 entries, reflecting a diverse range of customer interactions and sentiments towards airlines.

(4) What Have We Done So Far:

Data Preprocessing:

Embarking on the journey towards effective sentiment analysis necessitates comprehensive data preprocessing - a vital step aimed at ensuring data quality and suitability for machine learning models. Our preprocessing pipeline comprises tailored stages designed to tackle the distinct challenges posed by text data, particularly from social media platforms like Twitter.

Cleaning and Normalization:

Initiating with data cleaning, we eliminate extraneous characters (e.g., symbols, numbers), standardize text cases, and rectify common typographical errors. This phase is pivotal for minimizing dataset noise and fostering consistency across entries.

Handling Missing Values:

Given the nature of the collected tweets, certain entries may lack information in specific fields, such as the 'negative reason' column. To mitigate this issue, we fill missing values with a designated placeholder (e.g., "None") or implement imputation techniques where applicable.

Text Vectorization:

Central to preprocessing text data for machine learning is the conversion of text into a numerical format comprehensible to models. We employ the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique, which elucidates the significance of words within a document in a collection or corpus. This method not only facilitates numerical conversion of text but also accentuates the relevance of words for sentiment analysis.

Feature Selection and Engineering:

Critical for model success is the judicious selection of features. While the primary focus remains on tweet text, we explore supplementary features like tweet length and the presence of hashtags or mentions as potential indicators of sentiment.

This bar chart illustrates the imbalance in sentiment distribution across the tweets in our dataset.

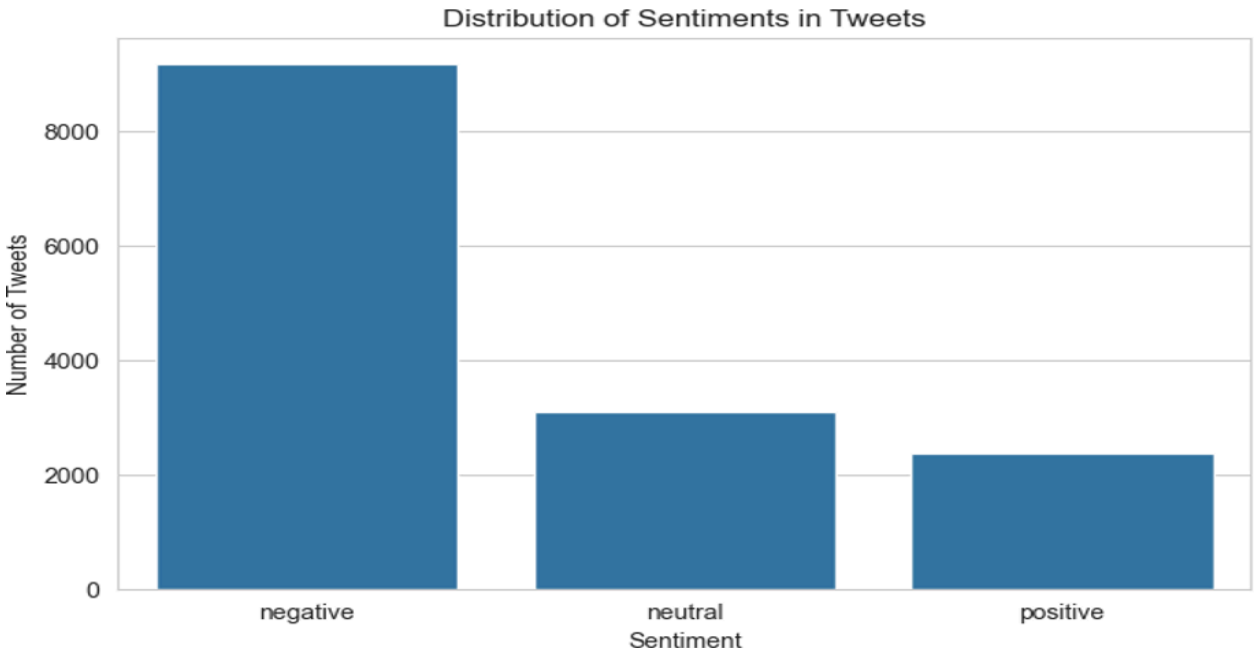

Figure 1: Distribution of Sentiments in Tweets

This boxplot compares the distribution of tweet lengths among different sentiment categories.

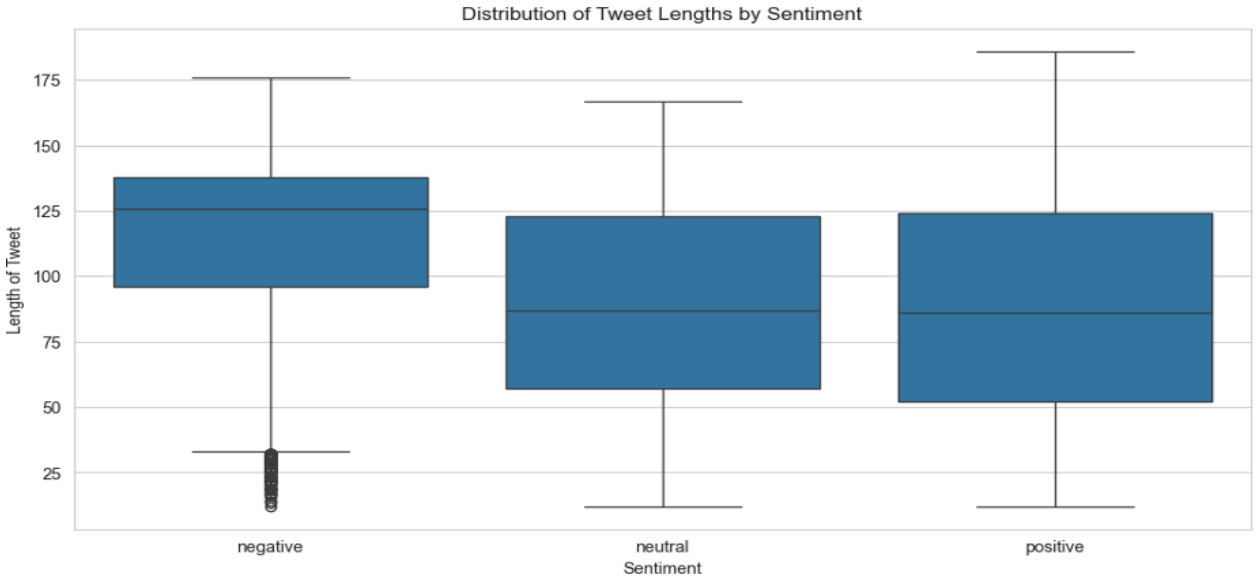

Figure 2: Distribution of Tweet Lengths by Sentiment

(5) Model Development and Training:

Finding the model that best reflects the complex nature of the data requires repeatedly developing and training machine learning models for sentiment analysis. To be able to advance to more complex algorithms, we began with a basic, easily understood model.

Logistic Regression Model: For sentiment analysis tasks, the logistic regression model is a good starting point because of its ease of use, effectiveness, and interpretability. With TF-IDF vectorized data, its linear structure enables fast training with the intent of reducing the difference between the anticipated and real sentiment labels. Through the utilization of logistic regression coefficients, interpretability is enhanced, providing valuable understanding of the impact of distinct features on sentiment classification. This method yields accessible results into the underlying mechanics guiding sentiment predictions, in addition to providing a trustworthy reference for assessing more complex models.

Model Evaluation: We looked at other performance indicators including confusion matrices and ROC curves in addition to accuracy and the metrics from the classification report. These extra insights offer a better comprehension of the model's performance in various sentiment categories and its capacity to manage unbalanced datasets. Through the integration of many assessment metrics, we guarantee a comprehensive evaluation of the model's overall efficacy and its feasibility for practical implementation.

Hyperparameter Optimization: Additionally, GridSearchCV made it possible for us to methodically investigate several hyperparameter setups, optimizing for precision, recall, and F1 scores in addition to accuracy. The model's parameters are carefully adjusted during this painstaking tuning process to achieve the ideal ratio of variance to bias, which eventually improves performance on unobserved data. With the help of GridSearchCV, we can effectively use a logistic regression model that has been precisely tuned to the specifics of our sentiment analysis task, maximizing its predictive ability and resilience in practical settings.

(6) Comparison with Other Models:

The comparison with alternative models, such as Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) networks, is aimed at exploring more sophisticated algorithms in order to potentially improve the performance of sentiment analysis. By evaluating these alternative models alongside the existing logistic regression model, we seek to identify which algorithm best captures the nuances of the data and provides the most accurate sentiment analysis results. This comparative analysis allows us to assess

the strengths and weaknesses of each model and determine which approach is most effective for the specific task of analyzing sentiments in airline-related tweets. Through this process, we aim to enhance the overall accuracy and reliability and reliability of our sentiment analysis methodology.

(7) Ongoing Tasks:

Further Model Training and Evaluation:

Our next steps involve the implementation and evaluation of Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) models. This phase aims to compare their performance against the baseline Logistic Regression model, providing insights into the efficacy of different algorithms for sentiment analysis.

Feature Engineering:

We plan to conduct experiments with various text preprocessing and feature extraction techniques to enhance model accuracy. By exploring different approaches, we aim to identify features that better capture the nuances of sentiment expressed in airline-related tweets.

Model Optimization:

Continued efforts in hyperparameter tuning and exploration of advanced techniques, such as ensemble methods and deep learning optimizations, are essential for improving model performance. This iterative process aims to refine model parameters and enhance its generalization ability to unseen data.

Analysis and Reporting:

A comprehensive analysis of model performances is imperative, extending beyond accuracy metrics to include precision, recall, and F1 scores. The culmination of this analysis will be a final report detailing our findings, model comparisons, and recommendations for practical applications in the airline industry.

(8) Conclusion:

Our intermediate project report marks significant progress in the realm of sentiment analysis for airline-related tweets. By leveraging machine learning techniques and rigorous data preprocessing, we've established a solid foundation for understanding public sentiment towards airline services, a critical factor in the service-oriented airline industry. Preliminary results indicate promising avenues for further exploration, including comparisons with more complex algorithms like Support Vector Machines (SVM) and

Long Short-Term Memory (LSTM) networks. Moving forward, our focus remains on refining models, experimenting with feature engineering, and deriving actionable insights to elevate customer satisfaction within the airline industry. Our ongoing efforts aim to deliver practical recommendations that enhance the overall customer experience and inform strategic decision-making for airlines.

(9) References:

- Twitter Inc. (Year). *Title of the dataset*. Retrieved from Twitter US Airline Sentiment (kaggle.com).
- Bird, S., Klein, E., C Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc. Retrieved from O'Reilly Media.
- Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python, Journal of Machine
- Learning Research, 12, 2825-2830. Retrieved from Journal of Machine Learning Research.
- Mikolov, T., Chen, K., Corrado, G., C Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Retrieved from arXiv.