Sentiment Analysis Project Report

1. Introduction

a. Project Overview:
This project focused on sentiment analysis of X (formerly Twitter) customer support interactions. The goal was to classify tweets into sentiment categories (positive, negative, neutral), aiding businesses in understanding customer sentiments. This analysis is pivotal for enhancing customer support strategies and improving overall customer experience.

b. Business Context:
In the digital era, customer sentiments, expressed through social media, significantly impact business strategies. Analyzing sentiments from customer interactions on platforms like X (formerly Twitter) provides valuable insights into customer satisfaction and brand perception.

c. Problem Statement:
The primary business challenge is managing customer satisfaction efficiently at scale for popular brands. By applying sentiment analysis to X data, we can predict the sentiment of customer interactions and identify areas for improvement in customer support which would impact directly their customer retention metrics.

2. Data Acquisition and Wrangling

Our initial challenge was the dataset's volume, spanning over half a gigabyte of customer interaction records. To manage this, we utilized PostgreSQL, a robust relational database system that facilitated advanced data manipulation and storage capabilities. This strategic decision enabled us to handle the data efficiently and perform initial sampling and load reduction for model development.

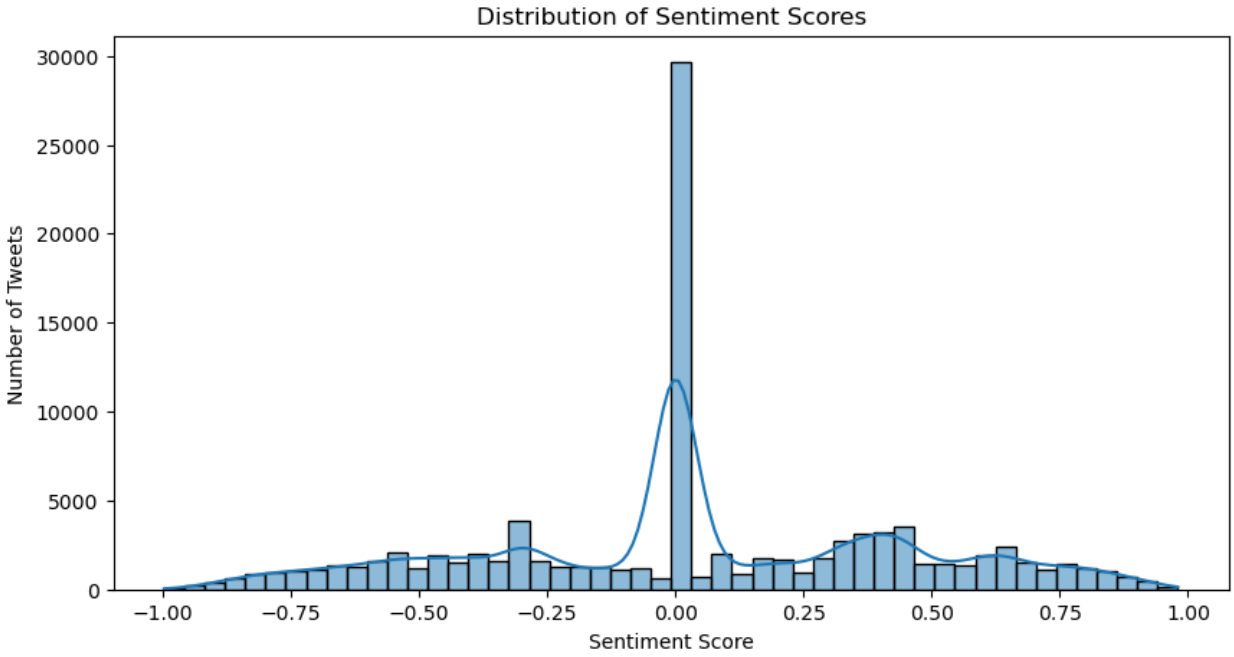
ETL Process:
To handle the large volume of data, an ETL process was designed to manage resources effectively. The process involved:

- Sampling and loading a subset of the data to prevent computational bottlenecks.
- Utilizing SQL queries to efficiently filter and preprocess records before analysis.
- Implementing data cleaning and transformation techniques to ensure data integrity.

Subsequent data wrangling efforts involved filtering to isolate initial customer queries and standardizing text data to ensure consistency across the dataset. Our cleaning processes included removing duplications, standardizing date formats, and employing regular expressions to cleanse the text. These steps were critical in improving data quality and preparing it for more nuanced analysis.

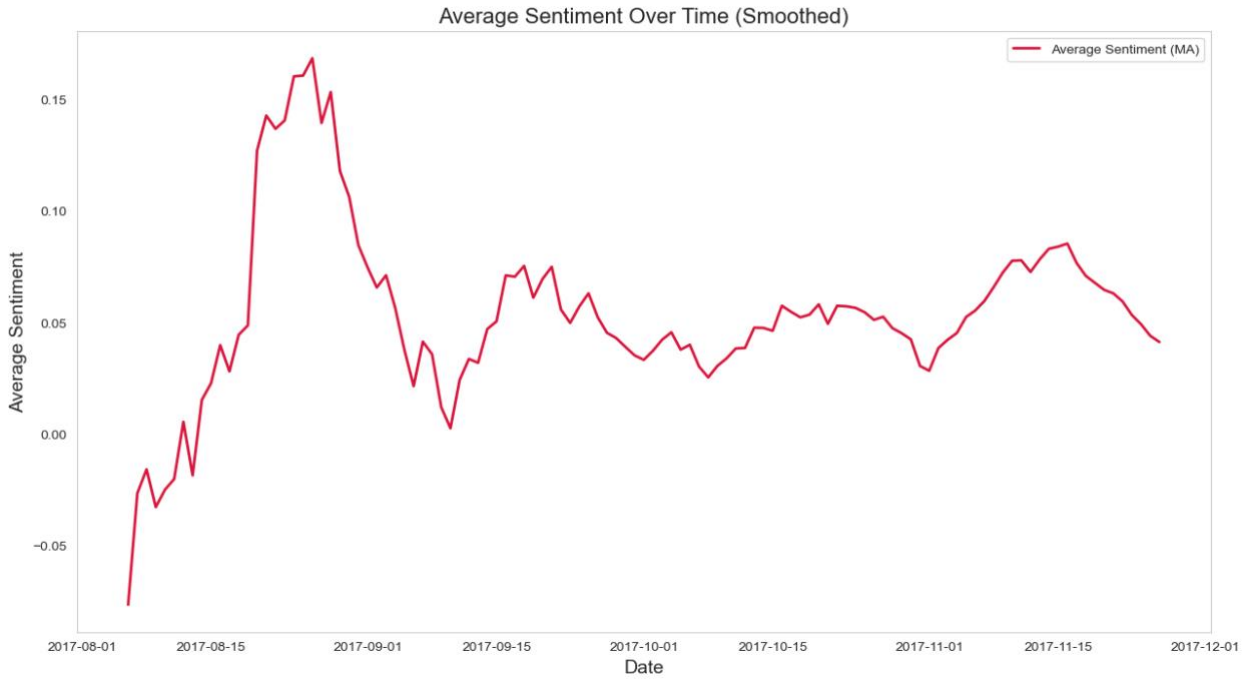
3. Exploratory Data Analysis (EDA)

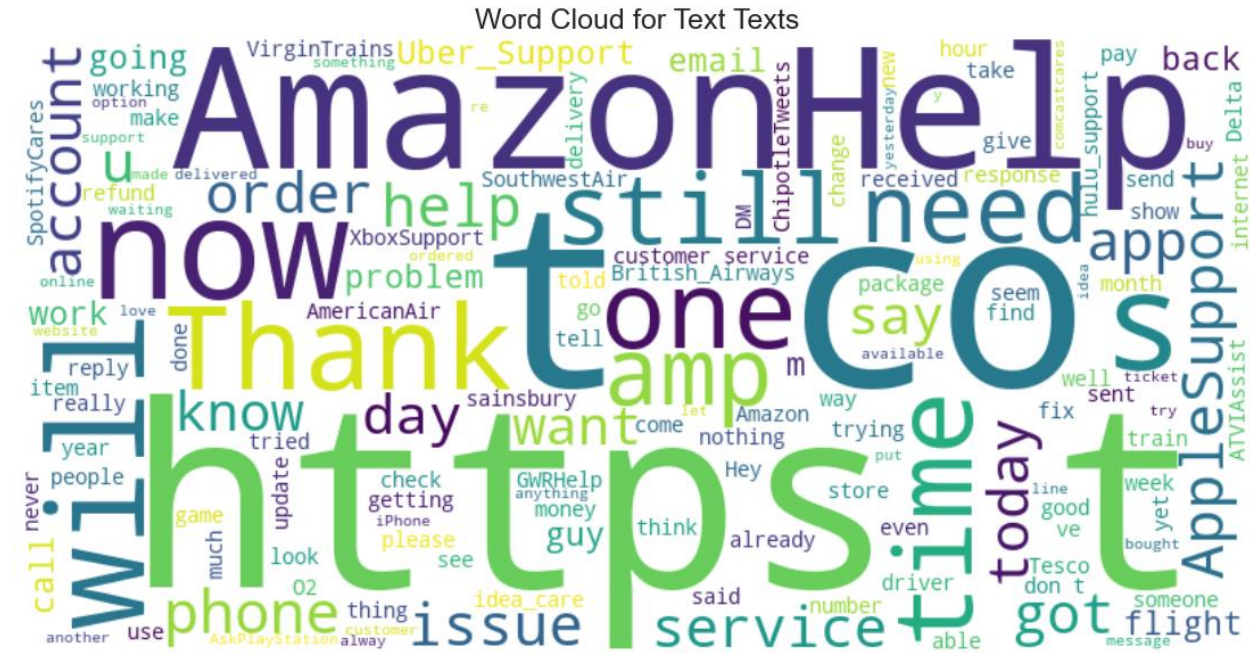
The EDA phase utilized visualizations to dissect sentiment distributions and message length, providing a lens into the customer service interaction landscape. We leveraged the Sentiment Intensity Analyzer from NLTK to assign sentiment scores, which became the basis for categorizing each interaction as positive, negative, or neutral. The visualization of sentiment scores over time revealed trends and patterns in customer satisfaction, while the common word analysis via word clouds and frequency distributions offered insights into prevalent customer concerns and experiences.



Key findings from EDA include:
- Common themes in customer queries were related to service issues and product inquiries.
- Sentiment trends over time, providing a macro view of customer satisfaction.
- Neutral sentiment score is dominant in the dataset. Sentiment distribution and common themes
- Significant peaks in tweet volumes correlated with specific events or promotional activities.
- Customer inquiry complexity through message length analysis.

Visualizations such as word clouds and sentiment distribution graphs were instrumental in revealing these insights.





4. Data Preprocessing and Feature Engineering

We engineered features to prepare the dataset, implementing TF-IDF vectorization to numerically represent text data and capture the relative importance of words within the corpus. The sentiment classification threshold was fine-tuned to distinguish between sentiments effectively, enhancing the classifier's interpretability.

Advanced NLP techniques like lemmatization and stopword removal were applied to refine the data. We performed feature engineering to extract meaningful attributes, including:

- Text length and complexity.
- Product mentions identified through NER leveraged by SpaCy module.
- Brand mentions identified through NER leveraged by SpaCy module.
- Part-of-speech tagging to understand the grammatical structure of sentences.
- Adjective counts derived from POS tagging.
- Noun counts derived from POS tagging.

Finally, we combined the TF-IDF features along with the engineered features in a sparse matrix and split the data to prevent data leakage, making it ready for the next stage.

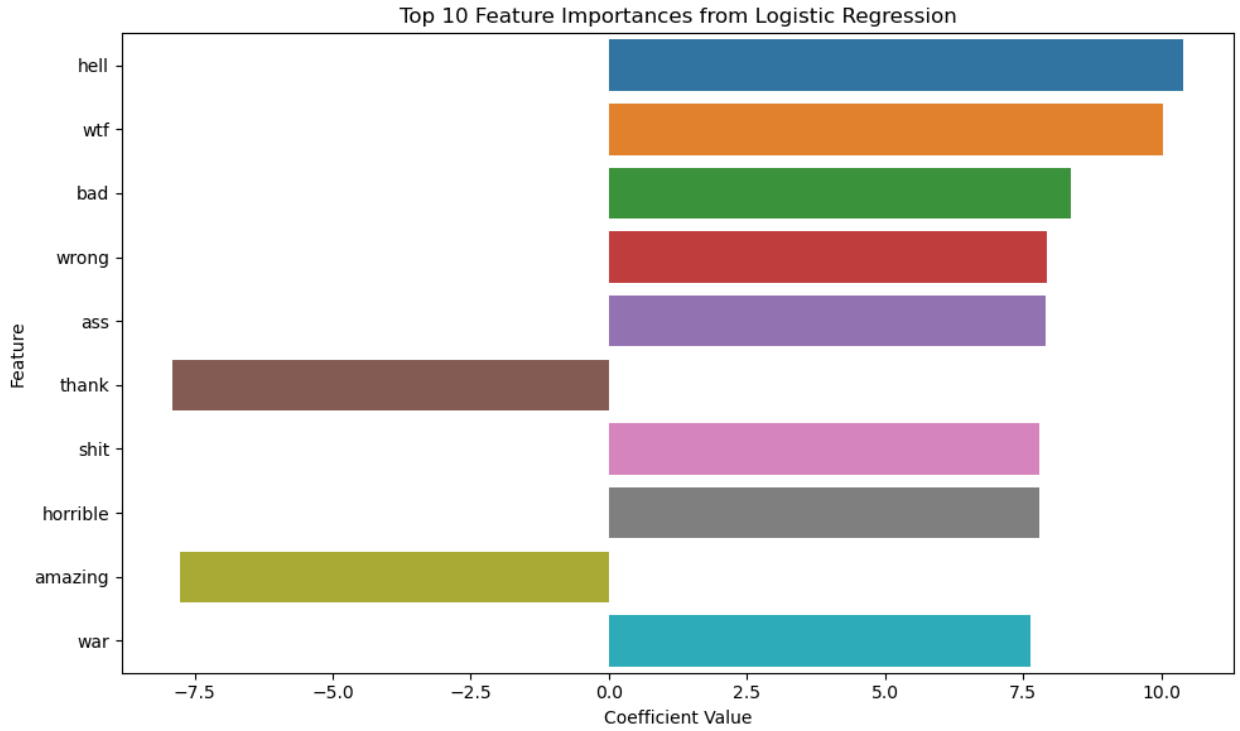
5. Model Building and Evaluation

We employed a range of metrics to evaluate the models, including accuracy, precision, recall, and the F1 score. These metrics provided a comprehensive view of each model's performance. The Logistic Regression model emerged as the most effective, offering a balance between predictive power and speed, essential for timely sentiment analysis in a customer support context.

a. Model Selection and Rationale:
We explored Logistic Regression, Random Forest, and XGBoost models. These were chosen for their diverse approaches to classification, from simple linear models to complex ensemble methods, providing a comprehensive understanding of the dataset.

b. Hyperparameter Tuning and Optimization:
Each model underwent hyperparameter tuning to optimize performance. The models were evaluated using accuracy, precision, recall, and F1-score. Logistic Regression, with its interpretability and efficiency, emerged as an ideal baseline model. XGBoost, known for its high performance in structured datasets, also showed promising results.

	Model	F1 Score	Accuracy	Recall	Precision
6	Logistic Regression (Tuned)	0.764294	0.768771	0.764212	0.771032
7	Random Forest Classifier (Tuned)	0.676217	0.680571	0.676561	0.680433
8	XGBoost Classifier (Tuned)	0.749329	0.754357	0.749883	0.758123



6. Conclusions and Recommendations

The project successfully categorized customer sentiments, revealing key factors influencing customer satisfaction. Recommendations include:
- Leveraging positive sentiments in marketing strategies.
- Addressing common negative sentiments to improve service quality.
- Continuously monitoring social media channels for real-time sentiment analysis.
- Further research of automation in customer support agents with chatbots.

7. Next Steps – Deployment

Given the time constraints to build the project, we did not follow any pattern. If given more time, further research into chatbot automation for customer support based on sentiment analysis could add significant value to reduce churn rate for companies which is one of the key KPIs in technology and information organizations.

8. References

Thoughtvector. (n.d.). *Customer Support on Twitter*. Kaggle. Retrieved from https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter