**ASPECT-BASED SENTIMENT ANALYSIS OF TWEETS USING NLP DATA REPORT**

**Title:** ASPECT-BASED SENTIMENT ANALYSIS OF TWEETS USING NLP DATA REPORT  
**Authors:** *Gabriel Tenesi, Sharon Wathiri, Celine Sitina, Wesley Kipsang.*  
**Date:** October 18, 2025.

# 1.Introduction

This report investigates why customers are leaving Syriatel and what can be done to keep them. Churn isn’t random. It’s linked to key features such as customer experience and usage habits. Identifying these features is a big step towards improving customer retention and predicting our company’s performance over time. In this analysis, we apply model evaluation to identify the most effective model for understanding churn and guiding future business decisions.

## 1.1 Problem Statement

# Apple and Google continuously monitor customer satisfaction to stay ahead in the technology market. However, given the massive volume and speed of data generated on Twitter, manual tracking of sentiment is impractical. Without automated systems, valuable insights into customer satisfaction, emerging issues, and product perception may be overlooked.

# This project aims to address this challenge by developing a machine learning model capable of classifying tweets related to Apple and Google as positive, negative, or neutral. The outcome will support organizations in understanding real-time consumer opinions, measuring brand perception, and identifying areas for improvement based on public feedback

## 1.2 Business objectives

### 1.2.1 Main objective:

To develop an NLP-based sentiment analysis model that automatically classifies tweets about Apple and Google into positive, negative, or neutral categories.

### 1.2.2 Specific objectives:

1. To explore and clean the tweet dataset, handling missing values, duplicates, and irrelevant characters.

2. To preprocess textual data through tokenization, stopword removal, and lemmatization.

3. To convert cleaned text into numerical features using appropriate vectorization techniques such as TF-IDF or Word2Vec.

4. To train and evaluate multiple classification algorithms (e.g., Logistic Regression, Naive Bayes, SVM) to identify the best-performing model.

5. To interpret and visualize model predictions, identifying which features most influence positive and negative sentiment.

6. To provide actionable insights that can guide Apple and Google in improving customer experience and brand perception.

### 1.2.3 Research question

1. How can the dataset be explored and cleaned to ensure data quality and reliability for sentiment analysis?

2. What preprocessing techniques are most effective for preparing Twitter text data for modeling?

3. Which text vectorization method (e.g., TF-IDF) produces better numerical representations for tweet classification?

4. Which classification algorithms yield the highest accuracy and robustness in predicting tweet sentiment?

5. Which textual features (words, phrases, or hashtags) most strongly influence model predictions of sentiment?

6. How can the resulting sentiment insights be applied by Apple and Google to improve customer satisfaction and brand reputation?

## 1.3 Success Criteria

1. Model Performance: Achieve at least 85% accuracy and a high AUC score (>0.85) in predicting churn.

2. Business Impact: Provide insights that reduce churn rates by enabling proactive retention strategies, targeting high-risk customers before they leave.

# 2.Findings & Analysis

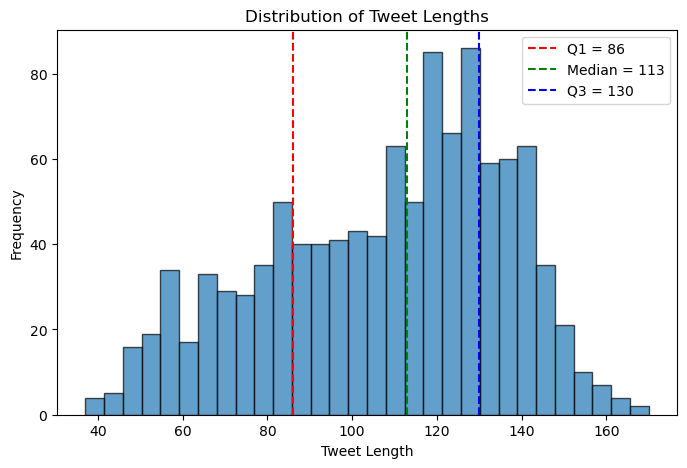
## 2.1 Sentiment Distribution by Brand

From the distribution of sentiments both brands get a mix of sentiments with the I can’t tell being the least, the positive sentiments dominate in both brands, but the Apple brand receives more positive sentiments compared to the Google brand.

## 

## 2.2 Tweet length Distribution

The distribution of tweet lengths is concentrated, with most tweets ranging between approximately 80 and 130 characters. This range corresponds to the interquartile range (Q1–Q3), indicating that most tweets are of moderate length.



## 2.3 World Cloud of Tweets



## 2.4 Proportion of sentiments by Vader.

Distribution of sentiments on a pie chart showing the positive class is dominating with 64.3% ,followef by the Neutral with 24.7% and the Negative being the least with 10.9%.

A pie chart with text on it

AI-generated content may be incorrect.

# 3.Model Selection

From our comparison of all models tested, SVM achieved the highest accuracy of 98.12%, slightly outperforming Logistic Regression of 97.55% and Naive Bayes of 93.97%.

This shows that SVM handled the high-dimensional TF-IDF features most effectively, making it the best overall model for this multiclass sentiment classification task.

A graph of a graph with blue and orange bars

AI-generated content may be incorrect.

# 4. Modelling & Evaluation

Creating the baseline model with the selected features as international plan, customer service calls, total day charge, total eve charge and the target variable is churn

The following comparison models were created:

1. Logistic regression
2. Naïve Bayes
3. Support Vector Machine

the following was how the models performed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | VADER | Logistic regression | NAÏVE-BAYES | SVM |
| Accuracy | 74.9% | 76% | 90% | 93% |
| Precision | 94% | 95% | 94% | 85% |
| Recall | 76% | 76% | 95% | 92% |
| F1 Score | 84% | 84% | 94% | 76% |

# 4.Recommendations

## 4.1 Conclusion

- This project explored different models for sentiment analysis, including VADER, Naive Bayes, Logistic Regression, and SVM. The SVM model performed best, achieving an accuracy of 0.9812, showing strong ability to understand patterns in the text. Overall, machine learning models, especially those using TF-IDF features, performed much better than rule-based approaches like VADER.

## 4.2 Recommendations

- Use SVM as the main model for sentiment classification as it perfomed better than other models.

- Add tools like LIME to explain model predictions.

- Keep improving the model with new data to maintain accuracy.

1. Adopt Random Forest as the Primary Model

* Random Forest achieved the highest ROC\_AUC (0.921) and Average Precision (0.851), making it the most reliable choice for predicting churn.
* **Business impact**: It can accurately flag customers most at risk, by an accuracy of 93%

2. Use Logistic Regression When Interpretability Matters

* Logistic Regression performed weaker with an ROC\_AUC of 0.815 and an Average Precision of 0.453, but it’s easy to interpret
* **Recommendation**: Use it alongside Random Forest in scenarios where transparency and stakeholder trust are more important than raw performance.

3. Treat Decision Trees as a Supporting Model

* Decision Tree had and ROC\_AUC of 0.889 and Average Precision of 0.776. These were solid, but still below Random Forest.
* **Recommendation**: They can be useful as a simple, explainable baseline, but not as the main production model. Monitor and retrain regularly using updated datasets to ensure the model adapts to evolving customer behavior.