Prepared by group 3

Group Project

Sentiment Analysis on Apple-Twitter

31 March, 2025

Team Members

1. Felista Kiptoo
2. Habshi pedro
3. Amos Kipkoech
4. Gloryann Otieno
5. Stanley Macharia

Introduction

Social media web sites like **Twitter** are essential stores for public perspectives and remarks withinside the virtual age. Improving product development, advertising and marketing plans, and client relationships for companies like Apple relies upon on their understanding of the way clients view their items on Twitter.

Understanding sentiment with the aid of using manually analyzing heaps of tweets, though, is a Herculean effort. **Natural language processing (NLP)** can automate this, letting businesses unexpectedly compare public opinion on a huge scale.

This mission objective is to create a NLP version that could robotically categorize a tweet`s sentiment as **positive, negative**, or **neutral**. This will allow Apple and different fascinated events to base their selections on public opinion, consequently guiding them.

Business Understanding

Stakeholders

The following stakeholders can benefit from the sentiment analysis model:

Main Stakeholders:

- Apple (or comparable tech firms): - Marketing Teams: The model lets one monitor consumer responses to brand activities, product launches, and marketing initiatives.

- Real-time public sentiment allows product development teams to determine which features are well-received and which require improvement.

Secondary Stakeholders:

- Customers: - Customers can observe how their views are reflected in public sentiment trends, and how other users feel about Apple products.

Examining sentiment around Apple products helps rivals like Google to understand Apple's strengths and shortcomings, therefore informing their own product plans.

Our Solution

Automating sentiment analysis on Twitter helps Apple and other interested parties to acquire insightful information without requiring great manual effort.

Real-time tracking of public sentiment enables Apple to be more reactive to consumer input, maximize marketing strategies, and enhance product offers. In the end, this might result in more successful product launches and better consumer happiness.

Methodology

Data Understanding:

- The dataset used for this project is CrowdFlower, on data.world. The dataset has a lot of tweets that have been manually rated for sentiment by human annotators. Each tweet is labeled with one of three sentiment categories: positive, negative, or neutral.
- The dataset consists of over 5,000 tweets with the following key features:

1. **Tweet**: The textual content of the tweet (string).
2. **Sentiment**: The sentiment label assigned to each tweet (categorical: Positive, Negative, Neutral).

Data Preparation

- **Objective**: Remove irrelevant columns that won't be used for analysis, making the dataset cleaner and more focused.

1. Irrelevant columns: These columns provide metadata or details that do not contribute to the sentiment analysis model. Removing them streamlines the dataset and improves processing speed.
2. By dropping columns such as id, query, and _unit_id, we focus only on the features that matter: the tweet text and sentiment.

Modeling

- We will build multiple models iteratively, starting from a simple baseline model and progressively adding more complexity. Then compare the performance of each model and justify the improvements over the previous ones based on the results.

Methodology

Evaluation

- By comparing the results of each model; Logistic Regression, Random Forest, SVM, and XGBoost, we can view the progression in terms of accuracy and F1-score. Using these comparisons, the model with the best performance is selected for final evaluation and interpretation.
- After comparing model performance, model chosen is **Logistic Regression** as the final model. This decision is based on:

1. **Accuracy**: Logistic Regression achieved an accuracy of **0.7391**, the highest between all tested.
2. **Class Imbalance Handling**: Logistic Regression struggles with class 3 but performed well on the majority of classes, especially on class 1 (high recall of 94%).
3. **Model Simplicity**: Logistic Regression is a relatively simple model, making it easier to interpret and deploy in real-world applications.

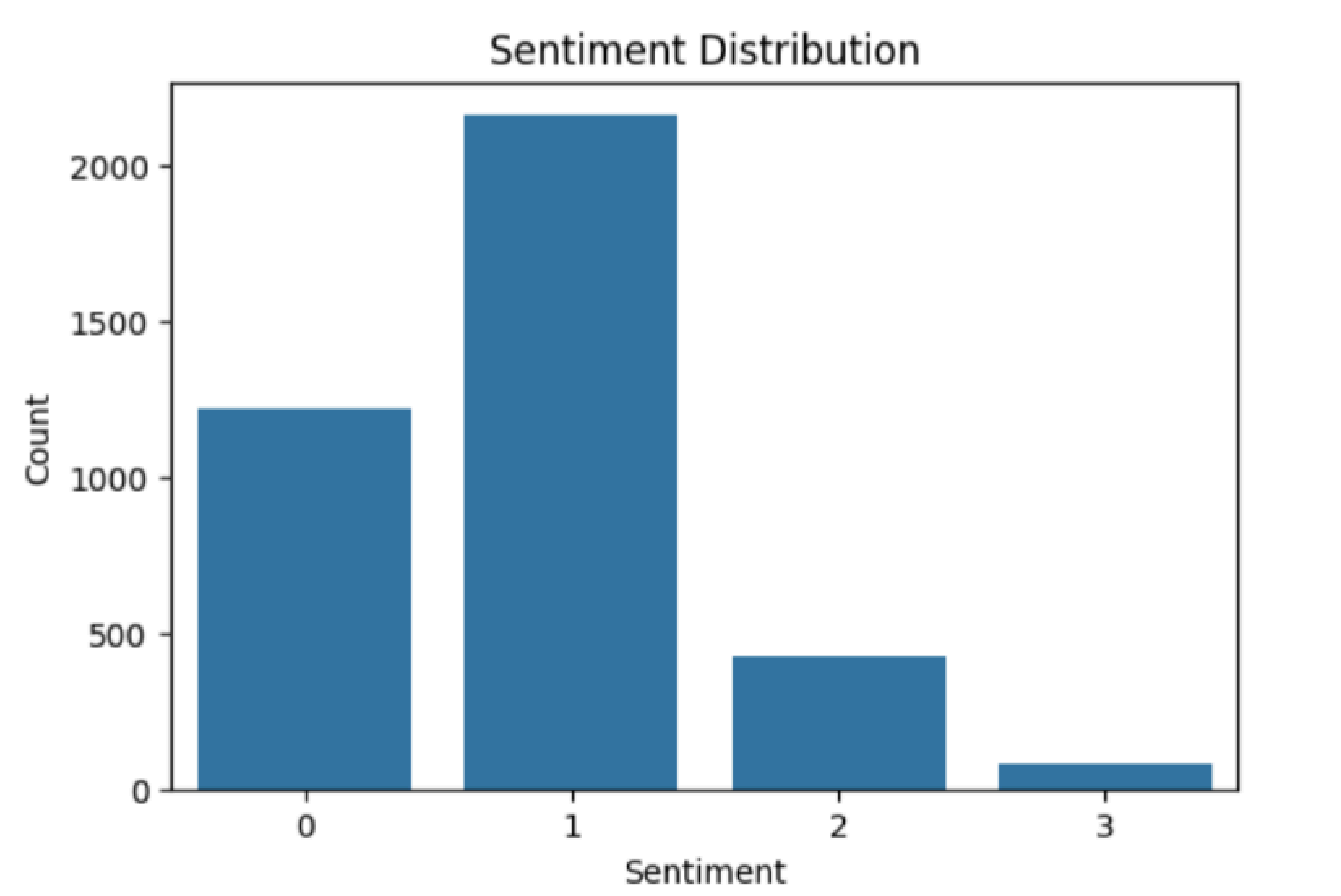
Looking at performance across all classes, it is justified that **Logistic regression** was the best.

Evaluation Metrics

We used these metrics for Logistic Regression model:

- **Accuracy**: Provides a general overview of model performance but may not be sufficient due to class imbalance.
- **Precision and Recall**: Helps evaluate how model distinguishes between classes.
- **F1-Score**: Balances precision and recall to give a better measure of the model effective ability.
- **Confusion Matrix**: Identifies misclassification patterns, which can be useful for refining the model.

Data Analysis



- This diagram shows that most tweets are **Neutral** on Apple. This indicates that the majority of users post factual or non-emotional content.
- **Negative tweets** are the second most common, highlighting a significant amount of criticism and complaints.
- **Positive tweets** are less common. This means that fewer users will express strong Apple approval.
- **Very positive tweets** are the least common and show that enthusiastic praise is relatively rare.
- Overall, the distribution of mood suggests that Apple receives a mix of opinions, but neutral arguments dominate more criticism than strong praise.

  - Sentiment (0): Negative sentiments
  - Sentiment (1): Neutral sentiments
  - Sentiment (2): Positive sentiments
  - Sentiment (3): Highly positive sentiments

Models Used

In this project we decided to use the following models:

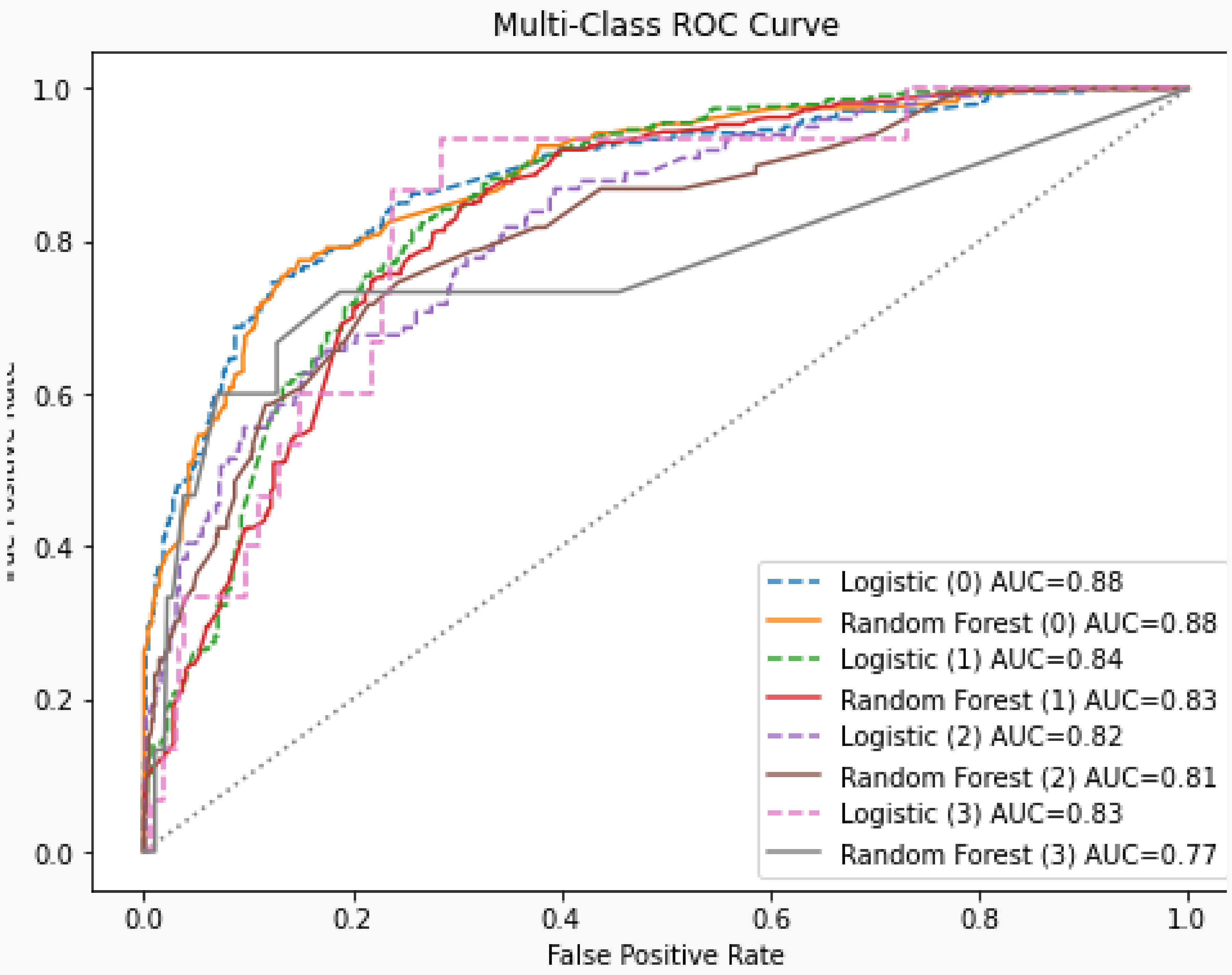
- **Logistic Regression model (Baseline Model).** Logistic regression is chosen for its simplicity and easy explanation. As the simplest model of the linear family classification, it acts as a good starting point.

- **Random Forest Classifier. Random forest** is known for its ability to manage more complex relationships and interactions between characteristics. It is also less excessive adjustment compared to personal decisions, making it a solid candidate to improve the performance of the model.

- **Support Vector Machine (SVM)**. SVM is known for its ability to find optimal decisions in large data. Using linear nucleus, we hope SVM will grasp complex models in data, capable of improving random forest model.

- **XGBoost**. XGBOOST uses enhanced use, combining weak learners (shallow trees) to create a strong learner. It is known for its high performance, especially on unbalanced data sets and will further improve the accuracy of the classification compared to previous models.

Model Selection



- The performance of the Random Forest and Logistic Regression classifiers across several categories is shown by the multi-class ROC curve. With AUC scores varying from 0.77 to 0.88 across various classes, both models demonstrate a high degree of predictive power.
- In most categories, Logistic Regression performs marginally better than Random Forest, but its AUC scores are higher in class (3). Both models are effective at differentiating between classes, as evidenced by their overall significantly higher performance compared to random guessing (AUC = 0.5).
- To enhance performance on the less ideal categories, future actions might include investigating ensemble techniques or fine-tuning hyperparameters.

- After comparing model performance, model chosen is Logistic Regression as the final model

Conclusion

This project demonstrates the performance of NLP in automating sentiment analysis, reducing the need for manual sentiment tagging every day and providing real-time insights into customer perceptions. By implementing a well-structured approach, this project is looking for a robust model that will help Apple and similar companies gain valuable business knowledge. Successful models like this lead to improved marketing strategies, improved customer service, and data-controlled product innovation, ensuring that companies stay in a competitive market.

Recommendations

1. **Improve Customer Support Response** by monitoring and addressing common complaints to reduce dissatisfaction and enhance Product Quality and Transparency.

2. Encouraging **user engagement** and promoting positive reviews by Featuring satisfied customer testimonials in marketing.

3. Improve data collection and sentiment monitoring by **collecting feedback from customers' reviews.**

4. Ensure **Real-Time sentiment analysis by Deploying sentiment tracking** to detect emerging issues early and respond quickly.

5. **Enhance AI-Driven Customer Insights** to detect sentiment in customer service chats and provide faster resolutions.

Thank you