**Data Report Phase 4**

**Contributors**

Grace Mwega

Boniface Thuo

Doreen Chepkong’a

Linda Tomoet

**DECODING BRAND SENTIMENTS: A MACHINE LEARNING APPROACH**

Background Information

In today's digital era, Twitter and other social media platforms profoundly influence consumer opinions about brands and products. Tweets reflect public sentiment—positive, negative, or neutral—toward specific brands or products. These sentiments significantly shape brand reputation, consumer trust, and buying decisions.

**Challenges in Sentiment Analysis**

* The overwhelming volume of tweets makes manual analysis impractical and time-consuming.
* Tweets often contain slang, emojis, and abbreviations, complicating text analysis.
* Accurately distinguishing between similar sentiments (e.g., sarcasm vs. genuine praise) poses a challenge.

**Stakeholders**

* **Brands and Marketing Teams**: To understand public perception, adjust marketing strategies, and improve product offerings.
* **Product Managers**: To gather insights into consumer pain points and develop solutions.
* **Competitors**: To benchmark performance and identify competitive advantages.

**Proposed Solution (Analysis & Modeling)**The project proposes a machine learning-based sentiment analysis system using Natural Language Processing (NLP) techniques. The model will classify tweets about Apple and Google products into sentiment categories (positive, neutral, negative), providing actionable insights for marketing, product development, and competitive benchmarking.

**Objective**

The aim of this project is to design and implement a sentiment analysis system that accurately classifies tweets into various sentiment categories, providing actionable insights for brands and product stakeholders.

* Develop an NLP model to classify tweets about Apple and Google products as positive, negative, or neutral, providing actionable insights into customer sentiment.
* Optimize the model's performance using text preprocessing, feature engineering, and iterative evaluations to ensure high accuracy and reliability.
* Analyze sentiment trends to support strategic decision-making for marketing, product development, and competitive benchmarking.
* Compare the two brands' perceptions by people by analyzing which brand has more positive tweets and negative tweets.

**Metrics of Success**

**Accuracy**: Achieve ≥85% correct sentiment classifications.

**Precision**: Attain ≥80% accuracy in predicting positive/negative sentiments.

**Recall**: Ensure ≥80% identification of actual positive/negative sentiments.

**F1-Score**: Maintain ≥80% balance between precision and recall.

**DATA UNDERSTANDING**

The data for this analysis is extracted from CrowdFlower via data.world. The dataset contains over 9,000 tweets that reflect emotional reactions to brands or products. It includes columns for the text of the tweet, the targeted brand, product, or event, and the sentiment direction (positive or negative). Analyzing this data will help in developing models to predict sentiment and assess user engagement with specific products or events. The tweets primarily mention popular tech products like the iPhone, iPad, specific applications, and events like SXSW and RISE Austin. It contains three columns that hold the following meaning:

1. **Tweet\_text**: This column contains the text of the tweet itself. It includes the user's thoughts, comments, and opinions, often mentioning specific products, brands, or events.
2. **Emotion\_in\_tweet\_is\_directed\_at**: This column indicates which brand, product, or entity the emotional tone in the tweet is directed toward. It helps identify whether the tweet is talking about the iPhone, iPad, or a specific app or event.
3. **Is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product**: This column categorizes whether the sentiment in the tweet is positive or negative toward the mentioned product or brand. It helps classify the emotional tone of the tweet.

**DATA PREPARATION & ANALYSIS**

1. **Shape**: The dataset has 9993 rows (observations or data points) and 3 columns (features or variables).
2. **Data Types**: All columns are of type objects, which usually represent text data in pandas.
3. **Missing Values**: There are missing values in the tweet text and emotion\_in\_tweet\_is\_directed\_at columns. This will need to be addressed during data cleaning or preprocessing.

We have completed **data cleaning**, which involved removing irrelevant characters (e.g., URLs, emojis), handling missing values, and normalizing text by converting it to lowercase, removing stopwords, and applying lemmatization or stemming. The dataset was also filtered for relevant sentiment labels to ensure modeling consistency.

Key issues addressed include improving **accuracy** by removing noise, ensuring **validity** by maintaining relevant data, achieving **completeness** by handling missing values, and enforcing **uniformity** through standardization. This thorough cleaning process ensures the dataset is optimized for reliable and effective analysis and modeling.

**Data Analysis**

We applied preprocessing techniques such as tokenization, stop word removal, and lemmatization to clean and prepare the text data in the DataFrame. We focused on columns like *tweet\_text*, *emotion\_in\_tweet\_is\_directed\_at*, and *is\_there\_an\_emotion\_directed\_at\_a\_brand\_or\_product* to ensure consistency. For **Exploratory Data Analysis (EDA)**, we analyzed each brand's review sentiments (positive, negative, neutral), visualizing positive reviews with bar charts. Additionally, we identified the most frequent words per brand using frequency tables and word clouds. This analysis provided valuable insights into brand perceptions and customer sentiment trends, guiding future strategies.

**MODELING**

1. **Logistic Regression**

The Logistic Regression model exhibits strong performance in sentiment classification. It achieves an accuracy of 88.9%, reflecting a high percentage of correct predictions. Precision is 87.8%, indicating reliable identification of positive tweets, while recall is also 88.9%, showcasing the model's sensitivity. The F1 score of 87.4% balances precision and recall, confirming robust classification capabilities. With low rates of false positives and negatives, this model proves reliable in identifying sentiments accurately. Its consistent performance across metrics highlights its ability to effectively classify tweet sentiments, making it a strong choice for this task.

1. **Decision Tree Classifier**

The Decision Tree Classifier achieves a solid 86.1% accuracy after hyperparameter optimization (maximum depth: 30, minimum samples split: 5). The confusion matrix indicates 540 true positives and 46 true negatives, with 35 false negatives and 55 false positives. Precision is 84.4%, and recall is 86.1%, demonstrating a good balance between sensitivity and prediction reliability. The F1 score of 84.9% confirms a strong trade-off between precision and recall. Overall, the Decision Tree Classifier effectively identifies sentiments while maintaining good performance through optimized parameters for better generalization.

1. **Random Forest Classifier**

The Random Forest model delivers moderate performance, with 69.6% accuracy and a 67.4% F1 score. While the confusion matrix shows balanced precision (69.2%) and recall (69.6%), the model struggles with classifying "positive" sentiments, resulting in higher false negatives. It performs well on the "neutral" class but has difficulty with other labels. This indicates potential room for improvement. Strategies such as hyperparameter tuning, advanced feature engineering, or addressing class imbalance (e.g., oversampling or undersampling) could enhance model performance and improve the classification of challenging categories.

1. **XGBoost**

The XGBoost model achieves 67% accuracy and an F1 score of 63.28%, reflecting moderate performance. The confusion matrix reveals reliable predictions for the "neutral" class but difficulties with "positive" and "negative" classes, showing higher false positives and false negatives. To improve results, hyperparameter tuning (e.g., adjusting learning rate, number of estimators, or max depth), feature engineering, and addressing class imbalance (e.g., SMOTE or undersampling) could be explored. While the current performance is sufficient for basic classification, further optimization would be necessary to enhance its ability to classify sentiments effectively

**EVALUATION**

**Evaluation** varies based on the models used:

**Logistic Regression** demonstrated strong performance with **88.9% accuracy**, high precision (**87.8%**), and an **F1 score of 87.4%**, ensuring balanced classification. Its reliability in identifying positive tweets highlights its effectiveness for sentiment analysis.

**Decision Tree Classifier** achieved **86.1% accuracy**, with good balance across precision (**84.4%**) and recall, supported by hyperparameter tuning.

**Random Forest** showed moderate performance with **69.6% accuracy** and struggled with "positive" classifications, indicating a need for optimization.

**XGBoost** performed similarly with **67% accuracy**, requiring feature engineering and hyperparameter adjustments to enhance sensitivity and reduce misclassifications.

**CONCLUSIONS**

The models, including Logistic Regression, Decision Tree, Random Forest and XGBoost, showed moderate performance with accuracies ranging from 67% to 88%. All models exhibited a good balance between precision and recall, indicating their ability to identify positive cases while minimizing false positives. The confusion matrices revealed that all models struggled to a varying degree with the "positive" and "negative" classes, exhibiting higher rates of false positives and false negatives for these classes compared to the "neutral" class.

**RECOMMENDATIONS**

1. Use the Logistic Regression model to monitor and respond to sentiment trends for improved customer engagement.
2. Address negative feedback to resolve consumer pain points and enhance products or services.
3. Benchmark against competitors by analyzing comparative sentiment data for Apple and Google.
4. Refine the model's accuracy by improving feature engineering to handle nuanced tweets like sarcasm and abbreviations

NEXT STEPS

1. Hyperparameter tuning is performed on models like Random Forest and XGBoost to improve classification accuracy.
2. Expand the dataset by including recent tweets for a more comprehensive analysis of brand sentiments.
3. Implement the best-performing model (Logistic Regression) in a real-time sentiment monitoring system.