DATA REPORT

CRISP-DM METHODOLOGY

APPLE AND GOOGLE PRODUCTS TWITTER SENTIMENT ANALYSIS

GROUP 3 MEMBERS:

ELVIS WANJOHI

JESSICA GICHIMU

JESSE NGUGI

STEPHEN GACHINGU

LATIFA RIZIKI

OCTOBER 2025

TABLE OF CONTENTS

1. Business Understanding

1.1 Business Overview

Apple and Google are global technology companies whose growth depends on product adoption and user loyalty. In a competitive market, Twitter sentiment shapes brand reputation, purchase intent and revenue. This project classifies tweets about Apple and Google as positive, neutral or negative to identify issues and opportunities that guide product, marketing and support.

1.2 Problem Statement

Apple and Google receive a large volume of Twitter feedback that can affect brand reputation and product adoption. Without a consistent way to measure this sentiment, decisions are delayed. To address this, the project classifies tweets as positive, neutral or negative to provide insights that are clear and timely for action.

1.3 Business Objectives

1.3.1 Main Objective

The main objective of this project is to develop a sentiment classification model that analyzes tweets about Apple and Google products and classifies them as positive, negative or neutral.

1.3.2 Specific Objectives

To achieve the main objective, the project has the following specific objectives:

1. Determine the products and services from Apple or Google that have the largest negative, positive and neutral feedback.
2. Preprocess the data through processes such as vectorization and tokenization, handling missing values and creating new features with respect to user behavior.
3. Evaluate the model performance using Precision, Recall, F1score, Accuracy Score and ROC.
4. Compare different classification models to determine which performs best for this dataset.

These objectives guided the modeling and evaluation stages of the CRISP-DM process.

1.3.3 Research Questions

To ensure the analysis directly addresses the business problem, the following research questions were defined:

1. Which products and services from Apple or Google have the largest negative, positive and neutral feedback?
2. Which features influence user behavior?
3. Which classifier model had the best Precision, Recall, F1 score, Accuracy Score and ROC?
4. Which classification model performs best for this dataset?

The project answers these research questions through data exploration, Exploratory Data Analysis (EDA), data cleaning, preprocessing and vectorization, model training and evaluation, and model interpretation. This will provide the tech companies with data-driven insights which will in turn enhance long-term profitability.

1.4 Success Criteria

Project success will be measured by both model performance and business outcomes. The model should accurately classify tweets as positive, neutral or negative so actions can be taken in time. From a business perspective, success means clearer sentiment visibility, faster responses, improved messaging and higher user retention.

2. Data Understanding

2.1 Data Source

The Brands and Product Emotions dataset was collected from data.world. The dataset includes 9093 rows and 3 columns, which are all of object data type.

2.2 Data Description

The dataset contains two features and one target variable. They include:

- **tweet_text**: The text content of the tweet used for sentiment classification.
- **emotion_in_tweet_is_directed_at**: The target of the emotion expressed in the tweet (e.g., Apple or Google).
- **is_there_an_emotion_directed_at_a_brand_or_product**: Indicates whether the tweet expresses an emotion directed at a brand or product. This serves as the target label for classification.

2.3 Data Quality Checks

The dataset was assessed for quality before modeling.

- Checked for missing values in the dataset.
- Checked for duplicate rows and inconsistency in the dataset.
- Checked the data type of each column in the dataset.
- Checked for uniformity of data in the dataset.

3. Data Preparation

3.1 Normalizing and Data Cleaning

To prepare the data for modeling the following steps were taken:

- Removed URLs, user mentions and hashtag symbols.
- Dropped non-alphabetic characters and collapsed repeated punctuation.
- Lowercased text and trimmed extra whitespace.
- Created a clean_tweet column, dropped exact duplicates on this field and removed the original tweet column.

This resulted in 8,915 unique tweets prepared for processing.

3.2 Basic Linguistic Features

- This step computed per-tweet length metrics: Characters which were also stored as tweet_length, words and sentences.
- Across 8,915 tweets, the averages are approximately 105 characters, 24.4 words and 1.89 sentences per tweet.
- The interquartile range for characters is 86 to 126. This indicates that most tweets fall within a mid-length band.

3.3 Tokenization and Stopwords

- Removed common stopwords and very short tokens.
- Tokenized clean_tweet into tokenized_tweet for downstream steps.

3.4 POS-Aware Lemmatization

- Part-of-speech tags were assigned to the tokens. Each token was lemmatized to its base form. An example of this is tweeting to tweet.
- This helps keep the vocabulary compact and consistent without losing meaning.

3.5 Plotting the Distributions

Exploratory Data Analysis was performed by visualizing the relationships between the features and the target variable. Some of the key visualizations done include:

- **Sentiment Distribution Analysis:** This plot illustrates the distribution of the sentiments across the dataset. From this plot, the neutral emotion class had the highest number of tweets, with about 5531 tweets, while the positive emotion class had about 2970 tweets and the negative emotion class had about 569 tweets. This shows a huge imbalance in the classes.
- **Tweet Destination Distribution:** This plot illustrates the distribution of the entities to which the tweets were aimed at. From the visualization, it is noted that 5788 tweets were not directed towards a specific entity, while Apple and Google had about 659 and 428 tweets directed towards them.
- **Sentiment by Tweet Destination:** This plot illustrates the distribution of the sentiments for the top 5 tweet destinations: Apple, Google, 'Not Directed', iPad, and 'iPad or iPhone'. The tweets that were 'Not Directed' had the highest number of neutral tweets with about 5431 tweets. The iPad had the highest number of positive and negative emotion tweets with about 792 tweets for positive and about 125 tweets for negative.

4. Modeling

4.1 Overview of Models

Several text-classification models were evaluated for sentiment prediction:

- **Logistic Regression**: Baseline linear model that is interpretable and provided usable probability estimates for decision making.
- **Multinomial Naive Bayes**: Fast probabilistic baseline well suited to word-frequency features and efficient to train.
- **Linear Support Vector Classifier (LinearSVC)**: Large-margin linear classifier that performs strongly on high-dimensional TF-IDF text.
- **Random Forest**: Nonlinear tree ensemble included to test whether feature interactions add value in the multi-class setting.

These models offer a balanced comparison of interpretability, speed, and performance on sparse text features.

4.2 Model Training

4.2.1 Text Features

- Tweets were converted to numeric inputs using term frequency–inverse document frequency.
- Single words and two-word phrases were included.
- Vocabulary sizes of 3,000 and 5,000 terms were tested.

4.2.2 Data Splits

- **Binary Task**: Stratified split into 70% training and 30% test. Class imbalance was handled with SMOTE on the training split only. In addition, validation and test data were not resampled.
- **Multi-Class Task**: Stratified split into 70% training and 30% temporary. The temporary portion was divided evenly into validation and test.

4.2.3 Hyperparameter Tuning

- Three-fold cross-validation used validation accuracy to select settings.
- Logistic Regression and Linear Support Vector Classifier varied regularization strength.
- Multinomial Naive Bayes varied the smoothing parameter.
- Random Forest varied the number of trees and the maximum depth.
- All pipelines varied the TF-IDF vocabulary size.

5. Evaluation

5.1 Evaluate Results

- **Binary (Positive vs Negative)**: Linear Support Vector Classifier achieved test accuracy of about 87.85%, ahead of Logistic Regression at about 84.81% and Multinomial Naive Bayes at about 84.62%. Errors were balanced with low false positives and low false negatives.
- **Multi-class (Negative, Neutral and Positive)**: Linear Support Vector Classifier recorded the highest validation accuracy of about 67.2%. %. Logistic Regression and Naive bayes followed at about 67.2% and 64.9% validation accuracy respectively. Random Forest had high training accuracy of about 99.4% and a clear drop on validation, indicating overfitting. For probability-based evaluation, Logistic Regression produced the strongest ROC performance, with area under the curve around 0.772 for the Negative class and mid-0.7 values for Neutral and Positive. The most frequent confusion occurred between Neutral and Positive.

5.2 Review Process

The evaluation followed CRISP-DM standards. For both binary and multi-class tasks, class ratios were preserved using stratified splits. No resampling was applied. Text cleaning and TF-IDF vectorization, plus hyperparameter tuning, were fit on training folds only with 3-fold cross-validation to prevent leakage. Diagnostics from classification reports, confusion matrices and ROC analysis confirmed generalization and showed Neutral vs Positive as the main confusion. No methodological gaps were identified.

5.3 Determine Next Steps

- **Primary model**: Deploy LinearSVC for both binary and multi-class sentiment. It gives the best test accuracy with balanced errors.
- **Probability use cases**: Keep a Logistic Regression version for cases that need probability scores or a cutoff for actions.
- **Monitoring**: Track accuracy and macro-F1 per class, with a specific watch on Neutral vs. Positive confusions. In addition, trigger review and retraining if macro-F1 drops by three to five points or if class distributions shift.
- **Maintenance**: Refresh training data regularly with new tweets. For multi-class, use class weighting or targeted sampling if classes are uneven. Add a simple calibration step to LinearSVC only if probability scores are later required. In addition, consider lightweight transformer models only if higher accuracy is needed.

6. Deployment

6.1 Plan Deployment

The sentiment classifier was deployed as a Streamlit Cloud app so it can be used through a simple web page. The app loads the fitted TF-IDF vectorizer together with the trained models. LinearSVC as the primary option and Logistic Regression for probability-based use. The build is triggered from the GitHub repository and Streamlit installs the packages listed in requirements.txt, then starts the app and serves the prediction interface. During deployment, a build error was fixed caused by an invalid relative path to ../requirements.txt. Removing that line and redeploying resolved the issue.

The app is now live and accessible for testing and inference at: https://phase4projec-qkr7ewgse2npajgfsamzq3.streamlit.app/

6.2 Monitoring and Maintenance

After go-live, model quality will be tracked in production for dependable predictions. Monitoring will cover overall accuracy and macro-F1 by class, with specific attention on Neutral and Positive confusions. In addition, continuous checks for shifts in class mix or language that could indicate drift. If macro-F1 drops by about 3 to 5 points or drift is detected, the model will be retrained on recent tweets and redeployed with versioning and rollback controls. Routine maintenance will include pinned library versions and scheduled refreshes of the training data to reflect new topics and wording.

7. Conclusion and Recommendations

7.1 Conclusion

This project applied the CRISP-DM methodology to classify tweet sentiment for Apple and Google. Linear Support Vector Classifier emerged as the best performing model, delivering the highest accuracy on unseen data with balanced errors across classes. The project met its objectives by building a robust sentiment classifier, highlighting common confusion between Neutral and Positive, and providing insights that support brand monitoring, customer feedback analysis and product improvement.

7.2 Recommendations

1. Continuously monitor accuracy and macro-F1. If performance drops by 3 to 5 points or the class mix shifts, retrain with recent tweets and redeploy with versioning.
2. Deploy model for reliable positive and negative sensitive monitoring.

These actionable recommendations will help support informed, evidence-based decision-making as Google and Apple work to track customer sentiment in real time, prioritize product fixes and support issues, and plan marketing or feature updates with clearer signals from users.