PHASE 4 PROJECT

Sentiment Analysis Report

BUSINESS UNDERSTANDING

In this section, we are discussing the project overview, problem statement, objectives of the project, the methodology used, and the success metrics.

PROJECT OVERVIEW

In today's digital age, social media platforms such as Twitter have become powerful sources of real-time customer feedback and opinions. Understanding the sentiment expressed by customers toward specific brands and products is essential for businesses to make informed decisions, enhance customer satisfaction, and maintain a positive brand reputation.

The goal of this project is to develop a sentimental analysis model specifically tailored to analyze Twitter data related to Google, Apple, and other products.

By building this sentimental analysis model, the business aims to achieve several objectives:

1. Customer sentiment Analysis
2. Brand MonitoringProduct Feedback and Enhancement
3. Competitive Analysis
4. Marketing Campaign Evaluation

Problem Statement

As a consulting firm, Twitter has assigned us the task of building a model that can rate the sentiment of a Tweet based on its content that can correctly categorize Twitter sentiment about Apple and Google products into positive, negative, or neutral categories and gain valuable insights into public perception, that will be used for informed decision-making in business strategies and customer satisfaction enterprise

Objectives for the project:

1. To develop a binary classifier that can classify tweets into positive or negative sentiment categories. This will serve as a proof of concept and provide a foundation for further analysis. The classifier will be a Logistic regression model and the benchmark accuracy will be 85%.
2. To expand to a multiclass classifier, thereby incorporating the neutral tweets to create a multiclass classifier that can accurately classify tweets as positive, negative, or neutral. This will provide a more comprehensive sentiment analysis of the tweets.
3. To compare sentiment between Apple and Google products by analyzing the sentiment distribution of tweets mentioning Apple, Google, and other products.

Methodology

The tweet text will be our primary feature for modeling. The text data will be subject to the usual preprocessing steps for NLP operations. For this purpose, we will use custom Python functions that:

Searches a document for regular expressions and creates tokens, using NLTK's RegexpTokenizer. In this case, combinations of alphanumeric characters that immediately follow instances of an @ symbol were replaced with 'user'. Punctuation and other stray characters are removed during tokenization.

Standardizes the tokens by converting all (alphabetical) characters to lowercase.

Discards stopwords, which have negligible semantic value, from the list of tokens, using NLTK's corpus. stopwords module.
Note: the stopwords list was also programmatically updated to exclude location- and platform-specific strings like "SXSW" or "RT".

Reduces instances of related words/tokens to common roots, using NLTK's lemmatizer.

To prepare the data (i.e. strings of text) for modeling, we used two vectorization techniques - a simple bag-of-words approach with sklearn's CountVectorizer approach using the same library. With this vectorization method, we tested for accuracy using four classification metrics:

MultinomialNB (Multinomial Naive Bayes)

Success Metrics
Since the goal is to achieve high precision (minimize false positives), and high recall (minimize false negatives):

1. **Accuracy** test to measure the proportion of correctly classified instances. A benchmark accuracy of 70% will be the aim.

Data Understanding

The dataset is from CrowdFlower via Data.world. The dataset is about 9000 tweet sentiments about Apple and Google that were classified as either 'positive', 'negative', 'I don't know', or 'no emotion'. The dataset has 3 columns namely: **tweet_text, emotion_in_tweet_is_directed_at, and is_there_an_emotion_directed_at_a_brand_or_product**. The dataset has 9093 rows and 3 columns

Missing Values: The dataset has 5802 rows (63.81%) with missing values in the emotion_in_tweet_is_directed_at column and 1 row (0.01%) missing value in the tweet_text column.
Duplicates: The dataset has 22 duplicated rows that constitute 0.242% of the data set.
Dealing with Missing values: The 1 row in the tweet_text column was dropped since we could not pervade the missing tweet. The 63.81% of missing value in the emotion_in_tweet_is_directed_at column was filled with "none" to avoid dropping the whole column which had valuable data.
Dealing with Duplicates: The duplicated rows were all dropped.

Data Preparation

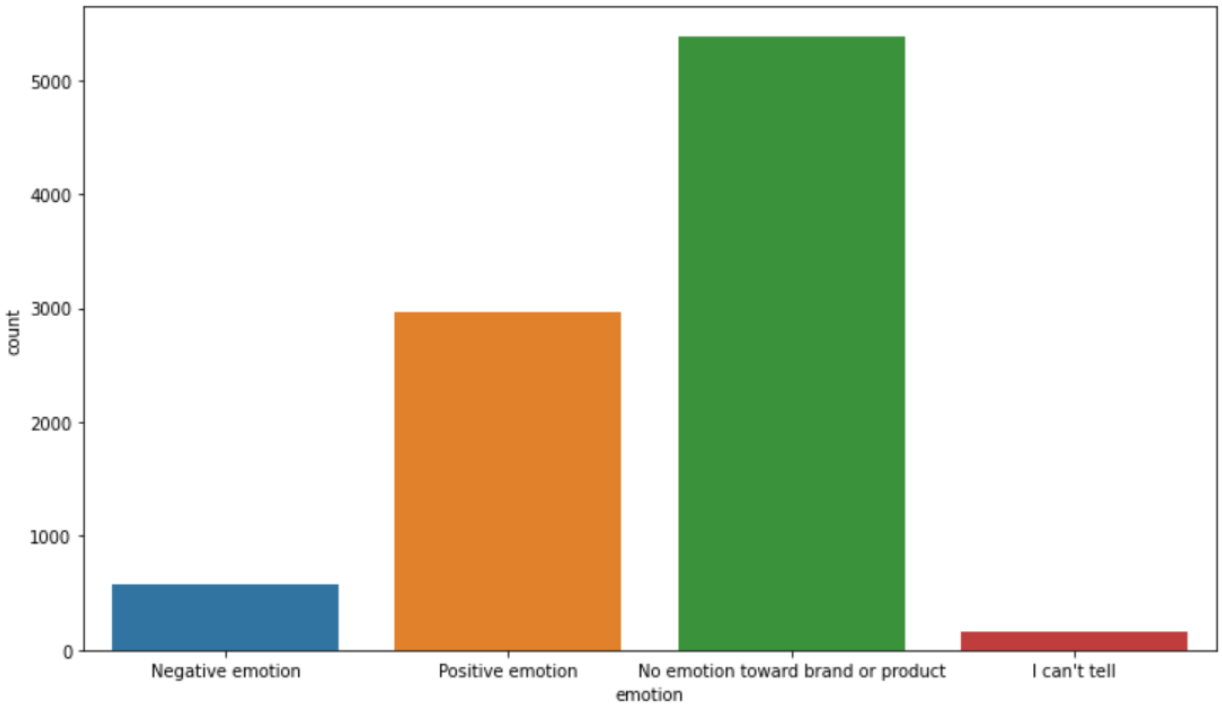
To prepare the data, we performed various preprocessing steps.

These included; removing duplicates, handling missing values by filling them with "none" for the emotion_in_tweet_is_directed_at column, and dropping the row with a missing tweet_text. We also employed tokenization, lowercase conversion, stopword removal, and lemmatization to refine the text data.

4. Exploratory Data Analysis (EDA)¶

We conducted Univariate and Bivariate analysis of the sentiments and created visualizations to see how they relate with each other and individually.

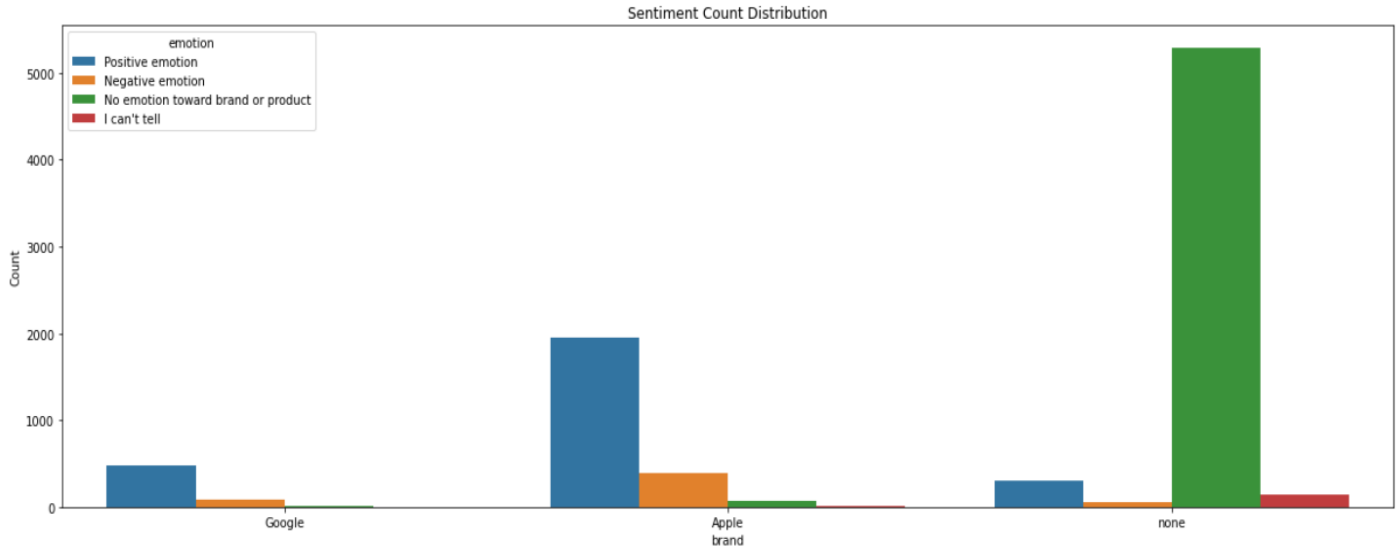
Univariate Analysis



Observation¶

We see that majority of people had No emotion toward brand or product at 59.26% followed by people with Positive emotions at 32.75%, I can't tell makes up less than 2% of our dataset, and doesn't offer much more information in the way of word significance than the tweets labeled No emotion toward brand or product.

Bivariate analysis



Interpretation

From the graph above Apple had the highest positive emotions compared to Google

5. Preprocessing

In this section we;

- Converted the tweet text to lowercase
- Removed html tags
- Removed the Url
- Expanded the contractions
- Removed the punctuations
- Tokenized
- Removed stopwords
- Lemmatized the tweet

So as to prepare the data for modeling

Modeling

- The problem at hand was a classification problem.
- We explored 2 models: a binary logistic regression model, a multi-class XGBoost model and MultinomialNB.
- Model accuracy was the metric for evaluation.
- Justification: Accuracy to get a verdict if a tweet is positive or negative.
- Accuracy of 70% was the threshold to deem the model as successful.

Binary classification

- In this section we created a base model to identify if a tweet was 'Positive' or 'Negative'.
- LogisticRegression was used for the classification.
- The normal preprocessing of vectorization and train test split was implemented.

Machine Learning Communication

Rationale why modeling was implemented.¶

- While simpler forms of data analysis, such as descriptive statistics or basic data visualization, can provide initial insights, they may not be sufficient for complex problems or large datasets. Machine learning leverages advanced algorithms to uncover hidden patterns.

Results.

  - Training accuracy: 96%
  - Testing accuracy: 90%
- The accuracy means that the model can predict with an accuracy of 90% whether a tweet is positive or negative.
- The current model is fit for prediction since it is generalizing well to new data even with high accuracy.

Limitations of binary model.

- Not fit for multiclass datasets.

Multiclass Classifier

- Here we work with the original dataset.
- We build a multi-class classifier.
- MultinomialNB and XGBoost model will be tested.

Machine Learning Communication XGBoost.¶

Rationale why ensemble modeling was implemented.

- While simpler forms of data analysis, such as descriptive statistics or basic data visualization, can provide initial insights, they are not sufficient for complex problems or large datasets such as this one. Ensemble models leverage advanced algorithms to uncover hidden patterns, make accurate predictions.

Results.

- Accuracy on the training set: 78%
- Accuracy on the testing set: 68%
- The accuracy means that the model can predict with an accuracy of 78% whether a tweet falls within the specified labels.

Limitations.

- The current model is not fit for prediction since it is overfitting.
- This we see from the difference between train and test accuracy.

Final note on modeling

- Despite implementing resampling techniques to address class imbalance, the model's accuracy has not improved significantly. This suggests that other factors within the dataset may be limiting predictive performance.
- The binary logistic regression model performs best.
- For Multiclass XGBoost was better.

Limitation and Challenges

- Class Imbalance Issue: The dataset suffers from class imbalance, where one sentiment class is dominant while others are underrepresented. This can result in biased models that are more accurate for the majority class but perform poorly on the minority classes. Addressing this issue is important to ensure fair and balanced sentiment analysis.
- Limited Dataset Size: The dataset used for sentiment analysis is relatively small, which can limit the model's ability to capture the full complexity of sentiments expressed in text. A larger and more diverse dataset would provide a broader representation of sentiments and improve the model's performance and generalization.
- Language Ambiguity and Sarcasm Detection: Language can be inherently ambiguous, and detecting sarcasm in text adds an extra layer of complexity. Sarcasm detection is challenging due to the subtleties and nuances involved. Developing robust strategies to handle language ambiguity and detect sarcasm is crucial for accurate sentiment analysis

7.Findings

1. Most tweets were directed to 'None' brand category. This may indicate that customers were not engaging with the brand.
2. Positive sentiments had the highest count compared to Negative sentiments, indicating that most people in general liked respective brands(Google and Apple)
3. Most of the positive tweets were directed to Apple brands
4. In the field of sentiment analysis, one of the significant challenges is dealing with language ambiguity and sarcasm detection. Natural language is complex and often subjective, making it difficult to accurately interpret sentiments from text.
5. On average most of the tweets were 10-15 words long - more words increase ambiguity.
6. NLP is a difficult task to model accurately.

8.Recommendations

We recommend that there be more customer engagement.

Probably check on this areas;

- Churn ratio - rate at which customers discontinue their relationship with a product company within a given time period
- Social media influencers through brand or product endorsement
- Customer feedback - The brands can introduce a rating system to accurately capture the sentiments of their customers

9. Next steps

- In our future work, we plan to explore advanced techniques such as incorporating attention mechanisms, using ensemble methods to further enhance the model's performance by incorporating domain-specific and fine-tuning the model on industry-specific datasets could improve its accuracy and adaptability.
- By considering these evaluation metrics, addressing limitations, and planning for future improvements, we aim to develop a robust NLP sentiment analysis solution that effectively captures sentiment