GROUP 7

JANUARY 2024

TWITTER (X) SENTIMENT ANALYSIS

Introduction

In the ever-evolving digital landscape, our team embarked on a strategic journey to harness the power of Natural Language Processing (NLP). Our mission? To conduct meticulous sentiment analysis on over 9,000 manually annotated tweets, unveiling public sentiments towards tech giants, Apple and Google. This ambitious venture promises actionable insights for businesses, marketers, and researchers navigating the dynamic currents of the consumer electronics domain.

Business Understanding

In the realm of social media, how people perceive tech companies like Apple and Google, as reflected in user-generated content on platforms like Twitter, holds significant sway. For businesses, this presents a unique opportunity to gain actionable insights, refine marketing strategies, and stay attuned to evolving consumer sentiments. By tapping into online chatter, companies can effectively gauge public perception, identify trends, and make informed decisions to navigate the dynamic landscape of the tech industry.

Objectives

1. To identify unique words associated with positive, negative, and neutral sentiments in the dataset.
2. Initiate the project with a focused binary classifier to distinguish positive and negative sentiments. Gradually extend the model's capabilities to handle neutral sentiments, ensuring a robust multiclass classification system(Proof of Concept).
3. Develop a sophisticated NLP model capable of accurately classifying tweet sentiments as positive, negative, or neutral, providing nuanced insights into the public's opinions(Sentiment Analysis).

Data Cleaning

The data could not be used as it and some cleaning had to be done we gave the 'tweet_text' column a good scrub. The result? A neat dataset of 9,065 tweets was ready for some analysis.

This process aimed to strip the tweet down to its most important words that relay the meaning and most importantly the sentiment of the text

Removing Usernames

We started by removing usernames from the text. We were then left with mostly the main tweet text. Eg @johndoe. This ensured we kept the identify of the user private as well as removing junk text

Removing Links

The dataset still had links embedded in the text. We used regex expressions to remove them. As an example “http://j.mp/grN7pK)”

Recording Hashtags

We recorded hashtags as we would like to see the most used hashtags in the tweets.

Removing Stopwords

Stopwords increase the size of the dataset but do not represent the overall context of the sentence.

Removing Non English words

We then removed non-English words and with that, we had a clean tweet text column to work with

Data Analysis

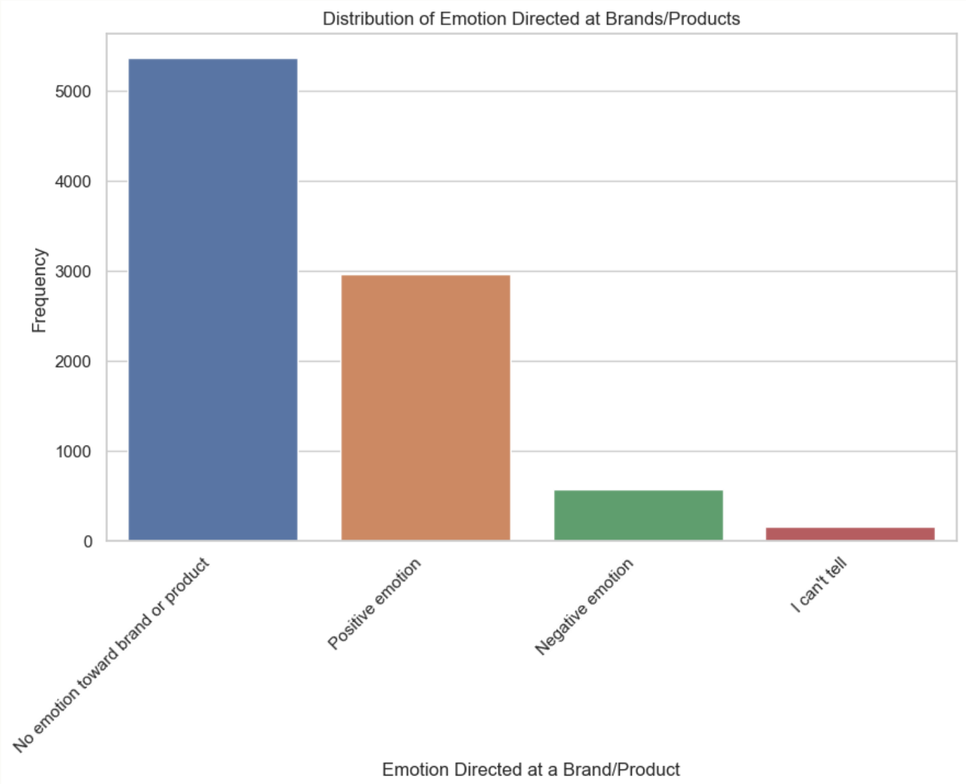
The data could not be used as it and some cleaning had to be done we gave the 'tweet_text' column a good scrub. The result? A neat dataset of 9,065 tweets was ready for some analysis.

This process aimed to strip the tweet down to its most important words that relay the meaning and most importantly the sentiment of the text

The chart depicts sentiment distribution in the dataset. "No emotion" dominates, followed by "Positive," while "Negative" is minimal.

The category "I can't tell" signifies indeterminate sentiments. This overview provides insights into the dataset's emotional composition.

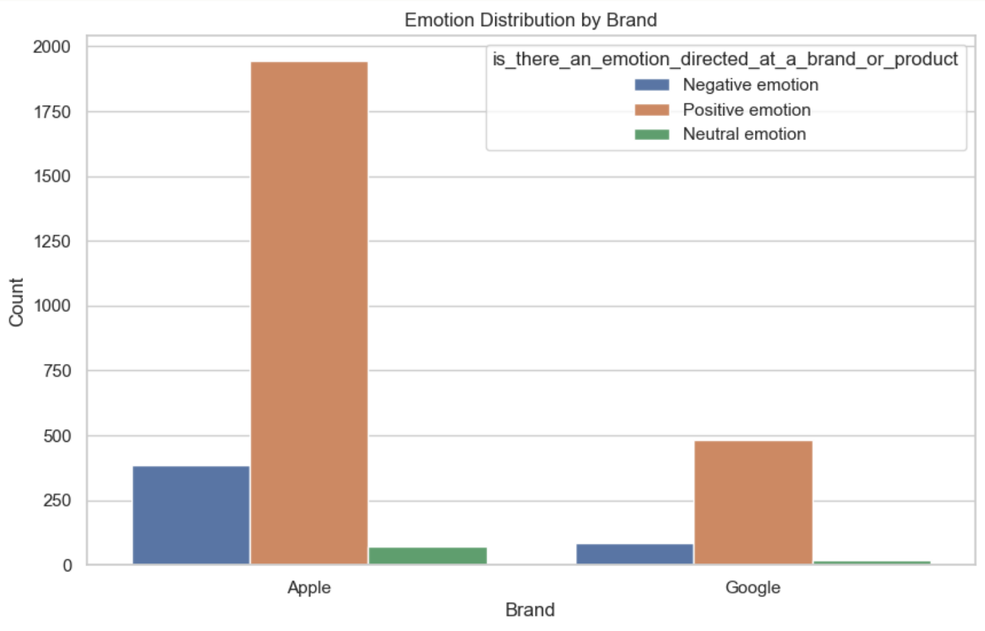
As you can see there is little representation in the Negative emotion target class



This chart presents the emotional distribution for two brands Apple and Google, based on collected tweets.

Positive emotions take the lead for both brands, followed by negative sentiments, and finally, neutral expressions.

It offers insights into the prevailing sentiments surrounding these brands on social media.



When analyzing the words found in each target class we found out that the words found in each class were as follows

Positive

- understanding
- proven,
- thought
- powerful,
- liking,
- optimum.

Negative

- disappointed,
- disliking,
- curse, sin,
- disgraceful

Methodology

1

- ***Binary Model***

We built a binary classifier to predict if a tweet is positive or negative. We used Naive Bayes

2

- ***Logistic Model***

This model could predict if a tweet was positive, negative, or neutral.

3

- ***Random Forest***

Building on from the second model. The Random Forest model was able to achieve higher accuracy.

Classification report

The overall accuracy score of 86.2% implies that the model correctly predicted the class labels for approximately 86.2% of the instances in the dataset.

Overall, the model exhibits good accuracy, particularly for classes 1 and 2, but may benefit from improvement in class 0 predictions.

```
Random Forest - Classification Report:
              precision    recall  f1-score   support

           0       0.26      0.27      0.26       119
           1       0.95      0.98      0.96      1112
           2       0.82      0.75      0.79       582

    accuracy                           0.86      1813
   macro avg       0.67      0.67      0.67      1813
weighted avg       0.86      0.86      0.86      1813

Random Forest - Accuracy Score: 0.8621070049641478
```

Assumptions

- Non-English words do not carry much weight as the dataset was primarily English-based.

- We merged the rows with the value “No emotion towards brand or product” with “I can’t tell into the Neutral class”

Conclusions

The Random Forest classification model impressively achieves an 86.2% accuracy score, showcasing strong overall performance. Classes 1 and 2 stand out with high precision, recall, and F1-score values, but struggle a bit short in determining Class 0 (Negative emotion).

1. **Additional Data Collection:** Gathering more data for the 'Negative' class can enhance the model's ability to generalize for this category.

2. **Model Optimization:** Exploring hyperparameter tuning and feature engineering can potentially elevate the model's discriminatory power.

Continuous monitoring and validation of new data are crucial. The model should undergo iterative refinement to ensure its effectiveness in real-world applications.
To access new tweet data, a request to X (formerly known as Twitter) is necessary, but the process might take up to 5 business days or longer based on their communication volume. Once approved, the API can be integrated with our model for fetching and analyzing new tweet data.

Next steps.

1. Consider additional data for the 'Negative' class to improve generalization.

2. Explore hyperparameter tuning and feature engineering for enhanced discrimination.

3. Address imbalanced classes through techniques like resampling or adjusting weights.

4. Ensure continuous monitoring for ongoing validation and refinement.

5. To access new tweet data:
   - Request API access from X (formerly Twitter).
   - Processing time is up to 5 business days or longer based on application volume.
   - Integrate the approved API for real-time data fetching and model execution.

GROUP 7

THANK YOU

Applause is not necessary but highly appreciated.