Contents

Project Title: Sentiment Analysis of Tweets on Tech Brands: Apple And Google

Project Overview:

The project aimed to perform sentiment analysis on tweets related to tech brands such as Apple and Google. The goal was to develop a model capable of accurately classifying the sentiment expressed in tweets as positive, negative, or neutral.

Business Understanding:

Implementing sentiment analysis for tech brands offers valuable insights into customer perceptions, preferences, and sentiments expressed on social media platforms. By leveraging these insights, businesses can make data-driven decisions to enhance brand reputation, improve customer satisfaction, and gain a competitive edge in the market. The primary objective of implementing sentiment analysis for tech brands is to gain actionable insights into customer perceptions, opinions, and sentiments expressed on social media platforms. By analyzing tweets related to tech brands such as Apple and Google, businesses can understand how their products are being perceived by the public, identify areas for improvement, and make informed decisions to enhance brand reputation and customer satisfaction.

Research Question:

- **What is the overall sentiment towards our brand/products?**

Understanding the general sentiment can help assess brand perception and identify potential areas for improvement or areas of strength.

- **Which specific products or features are receiving positive/negative feedback?**

Identifying sentiments towards specific products or features allows businesses to prioritize areas for enhancement or capitalize on strengths.

- **How does sentiment vary across different customer segments or demographics?**

Analyzing sentiment variations across different customer segments provides insights into audience preferences and helps tailor marketing strategies accordingly.

- **Are there any emerging trends or topics driving sentiment?**

Monitoring emerging trends and topics influencing sentiment enables businesses to stay ahead of market shifts and adapt their strategies accordingly.

- **How does our brand sentiment compare to competitors?**

Benchmarking brand sentiment against competitors helps identify competitive advantages and areas where improvements are needed.

Problem Statement:

We aim to address the need for actionable insights into customer sentiments towards tech brands, particularly Apple and Google, as expressed on social media platforms. Despite the abundance of user-generated content on platforms like Twitter, businesses often struggle to extract meaningful insights from this data due to its unstructured nature and sheer volume. By addressing the challenges associated with extracting insights from social media data, our proposed solution aims to empower businesses in the tech industry to gain a deeper understanding of customer sentiments, identify areas for improvement, and make data-driven decisions to drive growth and maintain competitive advantage.

Objectives:

Main Objective:

The main objective of the project is to analyze the sentiment of tweets towards various brands and products. By categorizing the sentiment as positive, negative, or neutral and identifying the specific targets of emotions, the aim is to gain insights into consumer perceptions and attitudes.

Specific Objectives:

1. Classify each tweet as expressing positive, negative, or neutral sentiments towards brands and products.
2. Identify the specific brands or products targeted by the emotional content in each tweet.
3. Analyze the overall sentiment distribution across brands and products.
4. Explore patterns and trends in consumer sentiment over time or in response to specific events or marketing campaigns.

Data Understanding:

The dataset contains 9093 rows of tweets related to multiple brands and products. Each tweet is annotated by contributors to indicate the sentiment expressed (positive, negative, or neutral) and specify the brand or product targeted by the emotion. The dataset used for this project was sourced from CrowdFlower via data.world. It contained 9000 tweets mentioning various tech brands and their associated sentiments.

Data Description:

The dataset included features such as tweet text, brand/product mentioned, and sentiment expressed. We had a column with about 5500 missing values which we reduced to about 700 then dropped

Data Preparation:

Data preparation is a crucial step in any data analysis or machine learning project, ensuring that the data is in a suitable format for analysis and modeling. In this project, the data preparation involved several steps, including loading the data, handling missing values, cleaning text data, and renaming columns to improve readability and consistency.

1. Loading the Data:

- The project started with loading the dataset containing tweets mentioning tech brands like Apple and Google from a CSV file into a pandas DataFrame. The pd.read_csv() function was used for this purpose.

2. Handling Missing Values:

- After loading the data, the presence of missing values was checked using the .isnull() method. For columns with missing values, appropriate strategies were applied, such as imputation or removal, depending on the context and impact on the analysis.

3. Renaming Columns:

- Some columns in the dataset had long or ambiguous names, making them less intuitive to work with. As part of data preparation, column names were renamed to improve clarity and consistency.
- For example, columns like 'emotion_in_tweet_is_directed_at' and 'is_there_an_emotion_directed_at_a_brand_or_product' were renamed to 'emotion_towards' and 'emotion_reaction', respectively, for better readability and understanding.

4. Text Data Cleaning:

- Text data in the tweets needed cleaning to remove noise, punctuation, special characters, and irrelevant information that could affect analysis and modeling.

5. Feature Engineering:

- In some cases, additional features were engineered from the existing data to enhance the analysis or modeling process. For example, sentiment scores could be calculated based on the cleaned text data using sentiment analysis techniques.
- Sentiments like “I can’t tell” were dropped and our sentiments to either positive, negative, or neutral

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is a critical phase in any data analysis project, aimed at understanding the structure, patterns, and relationships within the data. In the context of this project, which involves analyzing sentiment trends towards tech brands on social media, EDA plays a crucial role in uncovering insights that can inform business decisions and model development.

1. Distribution of Sentiment:

- Visualizing the distribution of sentiment labels (positive, negative, neutral) across the dataset provided an initial understanding of the overall sentiment trends towards tech brands.
- Bar charts were used to display the distribution of sentiment labels, to allow grasping of the proportion of positive, negative, and neutral sentiments.

2. Most Mentioned Brands or Products:

Most mentioned brands or products in the dataset were identified helping prioritize analysis efforts and understand which brands attract the most attention or discussion on social media.

3. Sentiment Distribution for Each Brand/Product:

Analyzing the sentiment distribution for each brand or product provided insights into how sentiments vary across different brands and products.

4. Emotion Reaction Distribution:

Explored the distribution of emotion reactions (e.g., positive emotion, negative emotion, neutral) to help understand how users respond to brands or products on social media.

Visualizations Used in EDA:

- Bar Charts: Used to visualize the distribution of sentiment labels, most mentioned brands/products, sentiment distribution for each brand/product, and emotion reaction distribution.
- Pie Charts: Used to visualize the proportion of sentiment labels or emotion reactions.
- Word Clouds: Used to visually represent the frequency of brand mentions, with larger words indicating higher frequencies.
- Line Charts/Time Series Plots: Used to visualize sentiment trends over time.
- Heat maps/Correlation Matrices: Used to visualize correlations between variables and identify patterns or relationships in the data.

Data Preprocessing:

Text data preprocessing techniques were applied, including:

- Lowercasing text.
- Removing punctuation, numbers, and special characters.
- Tokenization and removal of stop words.
- Stemming or lemmatization to normalize text.

Converted text data into numerical representations using techniques like TF-IDF Vectorization. When tokenizing we realized that we had a high count of some words that were not useful for our analysis e.g.......SXSW

Model Building:

- Trained multiple machine learning models, including Logistic Regression, Support Vector Machines (SVM), Naive Bayes, Random Forest, and Gradient Boosting.
- Applied hyper parameter tuning using techniques like Grid Search to optimize model performance.
- Utilized ensemble methods like Random Forest to improve classification accuracy.
- Experimented with multi-class classification using Logistic Regression and evaluated performance metrics.

Model Evaluation:

- Evaluated models using standard classification metrics such as accuracy, precision, recall, and F1-score.
- Conducted cross-validation to assess model generalization performance.
- Plotted ROC curves and calculated AUC scores to evaluate binary classification models.
- Compared the performance of different models to select the best-performing one.

Results and Insights:

1. Naive Bayes Model:

- Accuracy: Achieved an accuracy of approximately 84.5% on the testing set, indicating that 84.5% of the predictions made by the model were correct.
- Precision: Precision measures the proportion of true positive predictions out of all positive predictions made by the model. The Naive Bayes model achieved a precision score of around 85.2% on the testing set.
- Recall: Recall, also known as sensitivity, measures the proportion of true positive predictions out of all actual positive instances in the dataset. The Naive Bayes model achieved a recall score of approximately 98.3% on the testing set.
- F1 Score: The F1 score is the harmonic mean of precision and recall and provides a balanced measure of the model's performance. The Naive Bayes model achieved an F1 score of about 91.3% on the testing set.
- ROC AUC Score: The ROC AUC score measures the area under the Receiver Operating Characteristic (ROC) curve and indicates the model's ability to distinguish between positive and negative classes. The Naive Bayes model achieved an ROC AUC score of around 57.8% on the testing set.

2. Random Forest Model:

- Accuracy: Achieved an accuracy of approximately 86.3% on the testing set, indicating slightly better performance compared to the Naive Bayes model.
- Precision: The Random Forest model achieved a precision score of around 86.4% on the testing set, slightly higher than that of the Naive Bayes model.
- Recall: Achieved a recall score of approximately 99.1% on the testing set, indicating that the model effectively captured the majority of positive instances in the dataset.
- F1 Score: The Random Forest model achieved an F1 score of about 92.3% on the testing set, indicating good overall performance.
- ROC AUC Score: Achieved an ROC AUC score of approximately 61.6% on the testing set, indicating moderate discriminative ability

3. SVM Model:

- Accuracy: Achieved an accuracy of approximately 85.9% on the testing set, similar to that of the Naive Bayes model but slightly lower than the Random Forest model.
- Precision: The SVM model achieved a precision score of around 85.8% on the testing set, consistent with the performance of the Naive Bayes model.
- Recall: Achieved a recall score of approximately 99.5% on the testing set, indicating high sensitivity to positive instances.
- F1 Score: The SVM model achieved an F1 score of about 92.1% on the testing set, similar to that of the Naive Bayes model.
- ROC AUC Score: Achieved an ROC AUC score of approximately 59.7% on the testing set, indicating fair discriminative ability.

Overall Insights:

- The Random Forest model achieved the highest accuracy and F1 score among the three models, indicating better overall performance.
- All models demonstrated high recall scores, indicating their ability to capture positive instances effectively.
- However, the ROC AUC scores suggest that the models' ability to distinguish between positive and negative classes may be limited, especially for the Naive Bayes and SVM models.
- Further optimization and fine-tuning of the models may be necessary to improve their performance and discriminative ability.

In conclusion,

While the sentiment analysis models achieved relatively high accuracy and recall scores, there is still room for improvement in terms of discriminative ability and overall performance. Further analysis and experimentation may help uncover insights to enhance the models' effectiveness in sentiment classification tasks.

Challenges Faced:

- Limited dataset size for certain brands/products.
- Imbalanced class distribution in the sentiment labels.
- Difficulty in accurately determining sentiment from short and informal text data.
- Difficulties in distinguishing advertising tweets, mentions, and retweets

RECOMMENDATION REGARDING BUSINESS:

1. Customer Feedback Analysis: By analyzing sentiments expressed in customer feedback, businesses can gain deep insights into customer satisfaction levels, product/service performance, and areas for improvement**.**
2. Brand Reputation Management: Monitoring sentiment towards their brand on social media platforms helps businesses understand public perception. Positive sentiment indicates brand loyalty and satisfaction, while negative sentiment may highlight issues requiring attention.
3. Product Development: Identifying patterns in sentiment towards specific products or features enables businesses to tailor their product development strategies to meet customer needs and preferences more effectively.

Recommendations for Model:

- Collect more data to improve model performance, especially for less-represented brands/products.
- Experiment with advanced NLP techniques like word embedding’s or pre-trained language models.

- Explore techniques for handling imbalanced datasets, such as oversampling or using weighted loss functions.
- Implement sentiment analysis in real-time to monitor brand sentiment on social media platforms.

Conclusion:

The project successfully developed a sentiment analysis model capable of classifying tweets related to tech brands into positive, negative, or neutral sentiments. Through thorough data preprocessing, model building, and evaluation, valuable insights were gained into the sentiment trends associated with various tech brands. Moving forward, further enhancements and refinements to the model could lead to more accurate sentiment analysis and valuable insights for brand monitoring and decision-making purposes.

Model Deployment:

- We successfully deployed the sentiment analysis model, developed using a Random Forest classifier, for real-world use.
- Using the Joblib library, we saved the trained model (rf_sentiment_model.pkl) and the TF-IDF vectorizer (tfidf_vectorizer.pkl).
- The saved model and vectorizer were loaded into memory to make predictions on new data.

Prediction Function:

- We created a function, predict sentiment (tweet), to pre-process input tweets and predict their sentiment using the deployed model.
- The function preprocesses the tweet by stemming words and removing stop words, then victories it using the loaded TF-IDF vectorizer.
- Finally, it predicts the sentiment using the loaded Random Forest model.

Demonstration:

- We validated the deployment by inputting an example tweet ("I hate Google") into the predict sentiment () function.
- The function accurately predicted the sentiment of the tweet as "negative," demonstrating the model's real-time effectiveness.

Validation and Performance:

- The successful deployment and accurate predictions validate the effectiveness of our sentiment analysis model.

- Performance metrics obtained during model development further reinforce the model's quality and reliability.