Twitter Sentiment Analysis

Authored by:
1. Yusuf Ali
2. Gloria Ngure
3. Faith Watene
4. Abdirizark Jarso

Outline

1. Overview
2. Business Understanding
3. Data Understanding
4. Data Pre-processing
5. Exploratory Data Analysis
6. Modelling
7. Evaluation
8. Challenges Encountered
9. Conclusion
10. Recommendations

Overview

- ❖ Project Objective - Analyze customer feedback through Twitter to gauge sentiment toward products.

- ❖ Dataset - Over 9,000 tweets sourced from CrowdFlower.

- ❖ Sentiment Labels - Tweets are categorized as positive, negative, or neutral, facilitating targeted analysis.

Business Understanding

In today’s digital landscape, understanding customer feedback is a crucial step in maintaining brand reputation and enhancing customers reputation social media platforms like twitter provide an abundance of feedback where user express their opinions

Key Objectives:

- Determine sentiment around brands/products to enhance customer satisfaction.
- Address critical questions about negative themes and product perception.

Sub-objectives:

1. Gather and preprocess relevant data.
2. Conduct exploratory data analysis (EDA) to identify trends.
3. Develop and compare machine learning models for sentiment classification.
4. Evaluate model performance and provide actionable insights.

Data Understanding

The dataset consists of user tweets about specific products or brands, with labeled sentiments sourced from data world

- **tweet_text**: The raw text of the tweet, which serves as the primary source of information for understanding user sentiment.
- **motion_in_tweet_is_directed_at**: This column specifies which product or brand the tweet is referring to (e.g., iPhone, Google).
- emotion_in_tweet_is_directed_at: This column specifies which product or brand the tweet is referring to (e.g., iPhone, Google).

Data Pre-Processing

Text Cleaning Process:

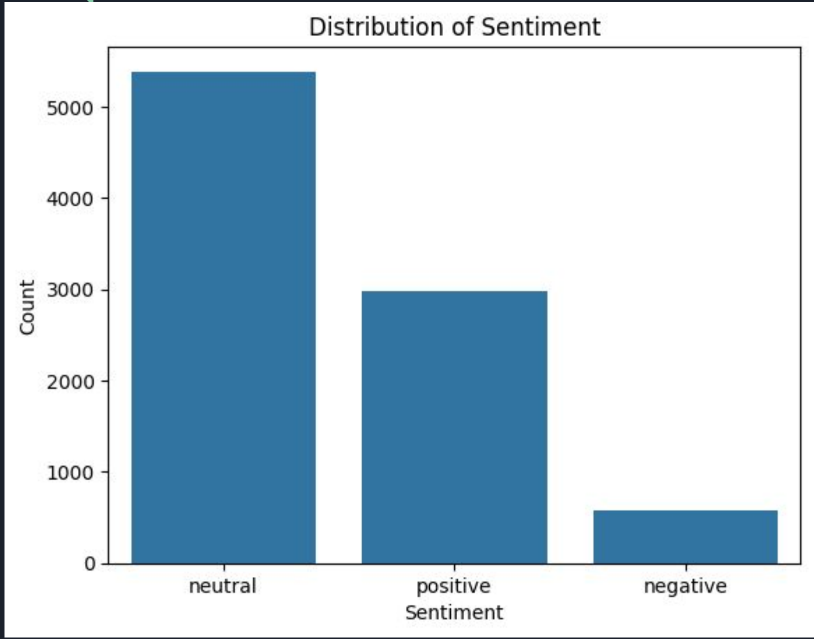
1. Removed irrelevant elements: mentions, URLs, hashtags, emojis, and special characters.
2. Tokenization: Split text into individual tokens for analysis.
3. Stopword Removal: Eliminated common words (e.g., "and," "the") using NLTK.
4. Lemmatization: Standardized word forms to reduce inflected words to their base form.
5. Result: Cleaned text stored in a new `cleaned_text` column for feature extraction.

Exploratory Data Analysis

Visualizations Conducted:

- Count of Sentiments: Distribution of sentiment types in the dataset to gauge overall public opinion.
- Sentiment Distribution by Product: Insights into how sentiments differ across products, revealing brand-specific perceptions.
- Average Tweet Length by Sentiment: Analyzed tweet lengths to identify potential patterns in sentiment expression.
- Most Common Tokens: Identified frequently used words to uncover prevalent themes in customer feedback.
- Word Cloud: Visual representation of common terms, emphasizing significant keywords in tweets.

Distribution of Sentiment



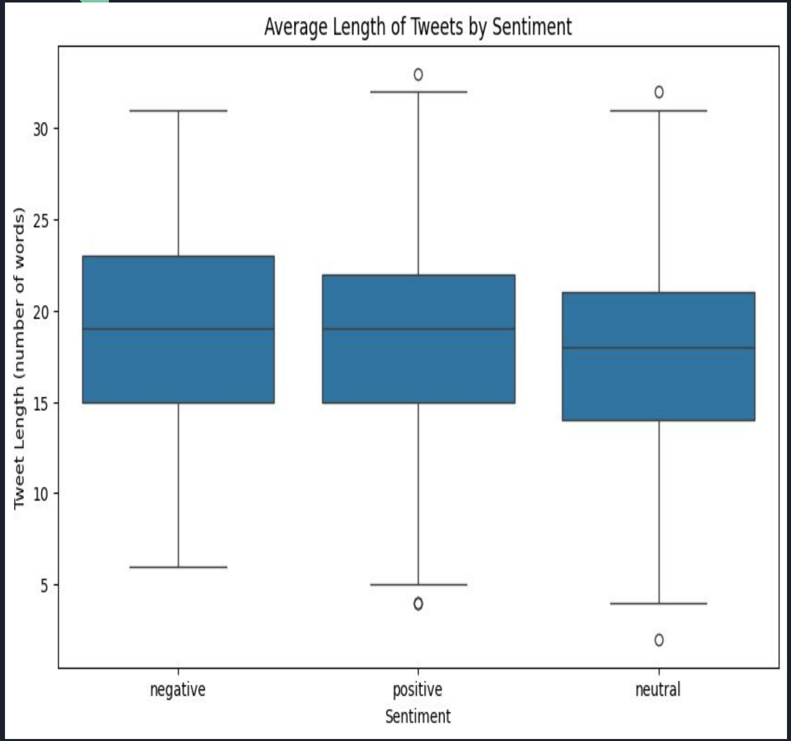
Neutral Sentiments

- ➢ The highest proportion of tweets is neutral, suggesting that many tweets provide informational content about the products.

Positive vs. Negative Sentiments

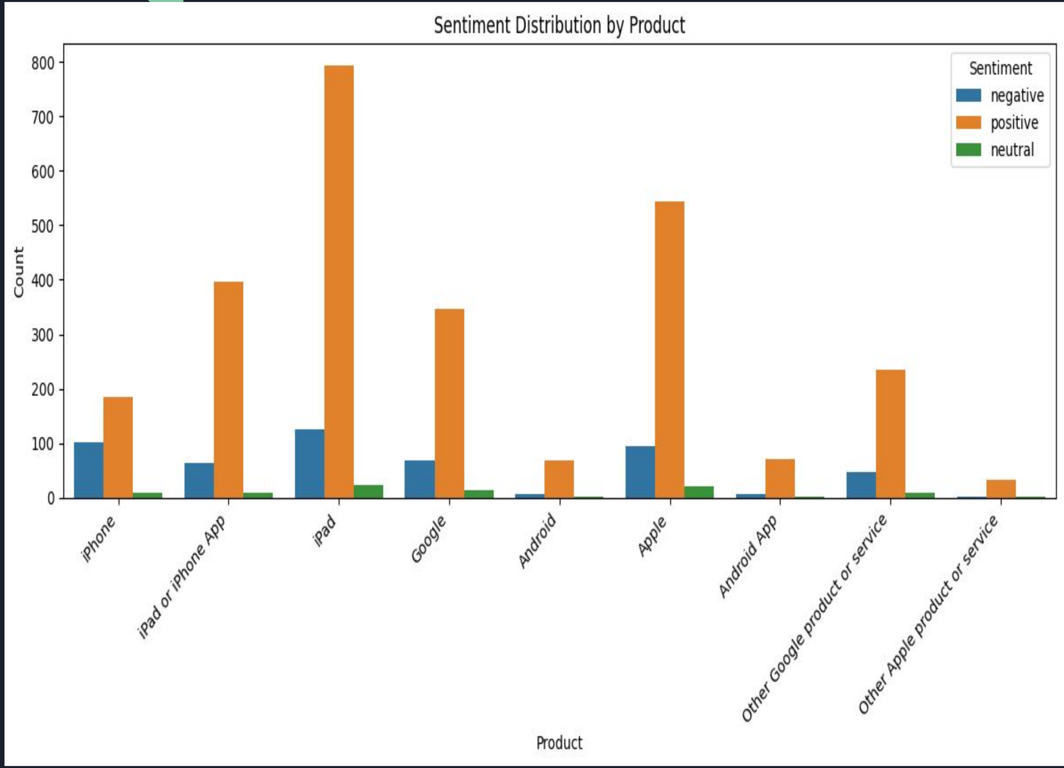
- ➢ Positive sentiments significantly **outweigh** negative sentiments.
- ➢ This indicates strong product performance and a high level of customer satisfaction.

Average Length of of tweets by Sentiment



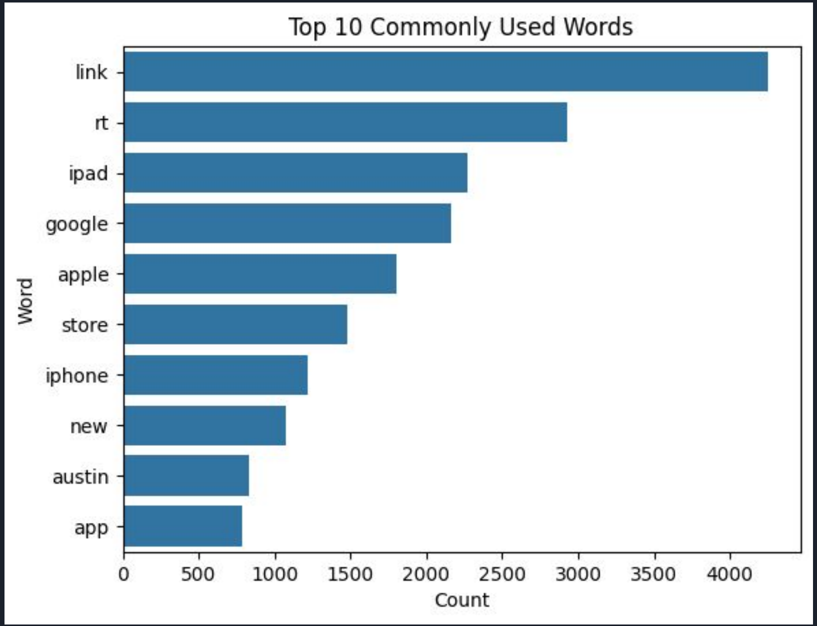
- ➢ Lengths of tweets are relatively similar across sentiments.
- ➢ Negative Tweets: Slightly longer on average, indicating users often provide more context or explanation when expressing criticism.

Sentiment Distribution by Product



- ➢ Apple Products - High positive sentiment indicates strong user satisfaction and excitement.
- ➢ iPad - Highest count of positive sentiment, suggesting it was a major topic of interest, possibly linked to a recent or anticipated release.
- ➢ Google Products - Positive sentiment is present but not as strong as Apple, hinting at less brand affinity among users.
- ➢ Android - Lower mention count compared to iOS devices but maintains a positive skew, indicating a generally favorable perception despite less buzz.
- ➢ App Ecosystem - Strong positive sentiment for "iPhone App" and "iPad or iPhone App," highlighting its importance in user satisfaction.

Top 10 Commonly Used Words



- ➢ link appears most frequently, suggesting sharing links was a common practice during discussions.
- ➢ High frequency of rt (retweet) indicates wide content sharing and amplification.
- ➢ iPad and iPhone are the most discussed Apple products.

Word Cloud

➢ Prominent Terms - The presence of Austin and SXSW indicates discussions were centered around the South by Southwest conference.

➢ Terms like app, Google, iPad, iPhone, launch, new, and party point to a focus on cutting-edge innovations during the event.

Modelling

Models Explored:

1. Baseline Model: Logistic Regression to establish a reference point for performance.
2. Tuned Model: Optimized Logistic Regression with hyperparameter tuning to improve accuracy.
3. Random Forest: Leveraged to handle complex relationships and interactions within the dataset.
4. Ensemble Model: Combined predictions from Logistic Regression and Random Forest for improved accuracy.
5. Neural Networks: Implemented deep learning techniques to capture intricate data patterns.

Evaluation

- Model Performance - Achieved an F1-score of approximately 0.6602, indicating solid predictive capability.

- Validation Approach - Utilized stratified train-test splits to maintain class distributions and ensure robust evaluation.

- Conclusion - The Logistic Regression model, after tuning, is superior in terms of compared to all other models explored. It offers a more reliable approach for accurately classifying the tweets.

Challenges Encountered

1. Data Imbalance: The dataset had imbalanced classes, which skewed the model’s performance.

2. Model Performance Variability: Different models performed variable across classes, making it a bit difficult to find one model that excelled in all aspects.

3. Resource Intensive: Training deep learning models was resource-intensive and time-consuming, requiring significant computational power.

Conclusion

- The project successfully analyzed Twitter sentiment toward products using various machine learning models. Logistic regression outperformed other models, including neural networks, in terms of accuracy and interpretability thus concluded as the best model for implementation.

- This suggests that for sentiment analysis tasks, simpler models can often provide robust results without the complexity of more advanced algorithms. However, challenges such as data imbalance and feature selection highlighted the need for careful consideration when building and evaluating models.

Recommendations

1. Feature Engineering - Explore additional feature engineering methods to enhance the dataset.

2. Ensemble Methods - Consider using other ensemble techniques, such as bagging or boosting, to improve model performance.

3. Hyperparameter Tuning - Perform more extensive hyperparameter tuning using techniques like Grid Search or Bayesian Optimization to optimize model performance further.

4. Model Interpretation - Invest time in other model interpretation tools like LIME to better understand how features influence predictions.