Sentiment Analysis on Spotify App Reviews: Implementation Report

Implementation Overview:

In this project, sentiment analysis was performed on Spotify app reviews using natural language processing (NLP) techniques and machine learning models. The dataset included user reviews, ratings, and engagement metrics. The analysis covered data cleaning, visualization, text preprocessing, feature engineering, and the implementation of classification models.

Challenges Faced:

Imbalanced Data: The dataset had a significant class imbalance, with a higher number of positive reviews. Addressing this imbalance was crucial to prevent biased model training.

Text Preprocessing: Cleaning and transforming textual data presented challenges, such as dealing with stopwords, punctuation, and ensuring uniformity in text formats.

Model Selection: Choosing between Random Forest and Multinomial Naive Bayes classifiers required careful consideration of their strengths and weaknesses.
Implementation Steps:

Data Cleaning: Irrelevant columns (Time_submitted, Reply) were dropped due to missing values and lack of impact. The 'Rating' column was transformed into three sentiment classes (Good, Neutral, Bad).

Exploratory Data Analysis (EDA): Visualization techniques, including pie charts, histograms, and scatter plots, were used to understand the distribution of ratings, review lengths, and their relationships.

Text Preprocessing: Lowercasing, punctuation removal, tokenization, stopword removal, and lemmatization were applied to clean and normalize the text data.

Feature Engineering: The dataset was split into training and testing sets, label-encoded, and transformed using Count Vectorization and TF-IDF Vectorization.

Classification Models: Random Forest and Multinomial Naive Bayes classifiers were trained on both Count Vectors and TF-IDF Vectors to predict sentiment labels.

Insights Gained:

The majority of reviews were positive (Good), indicating a high overall satisfaction with the Spotify app. Review length varied widely, with a mean length of 163 characters and some reviews exceeding 3753 characters.
Machine learning models achieved satisfactory accuracy, with Random Forest on Count Vectors performing the best (77.8% accuracy).

The choice of vectorization method impacted model performance, highlighting the importance of feature representation.

Conclusion:

The implementation provided valuable insights into user sentiments towards the Spotify app. Challenges in handling imbalanced data and preprocessing text were successfully addressed. The models demonstrated the ability to classify sentiments effectively. Further refinements and explorations could enhance the model's accuracy and uncover deeper patterns in user reviews.