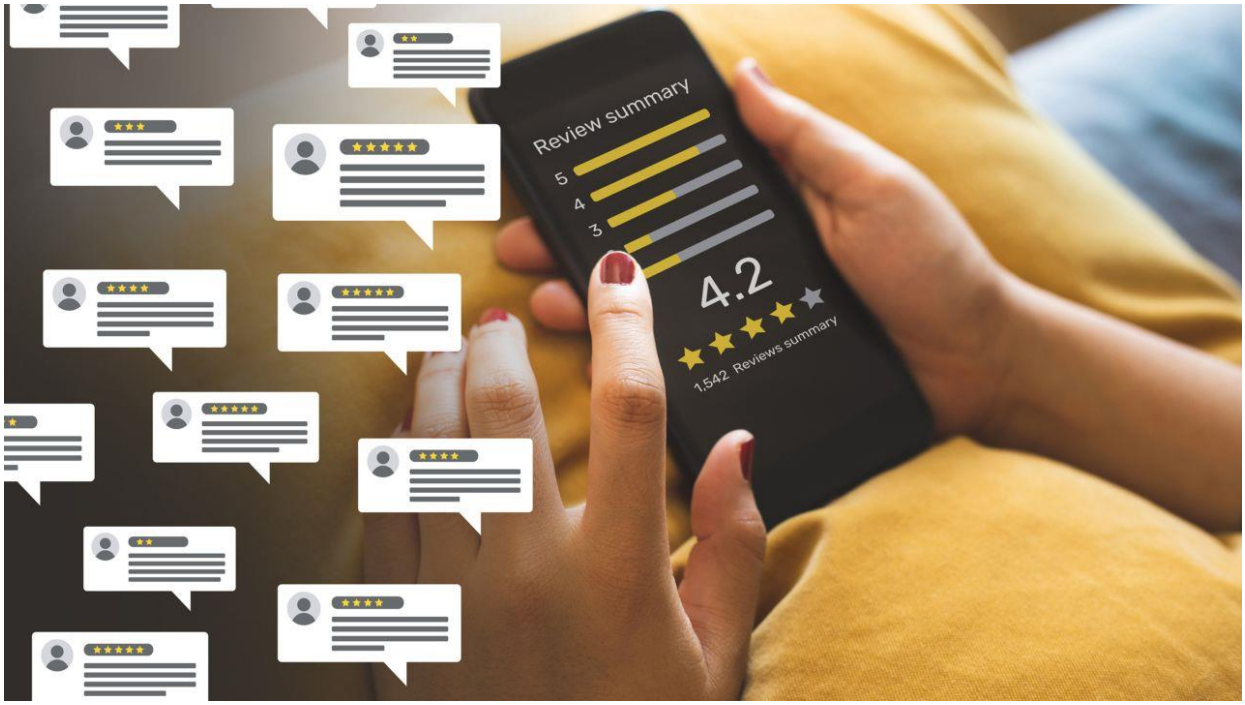

SENTIMENT ANALYSIS

OF

AMAZON PRODUCT REVIEWS

ABSTRACT

The project aims to analyze Amazon reviews' sentiment using machine learning techniques, classifying customer feedback as positive, negative, or neutral, to provide valuable insights into customer opinions and sentiments towards Amazon products.

PROJECT REPORT

KANCHAN CHOWDHURY

Introduction:

Sentiment analysis, also known as opinion mining, is a crucial tool in online shopping that allows businesses to understand customer feedback, identify trends, and predict satisfaction. By analyzing the sentiment of Amazon reviews, businesses can make data-driven decisions and enhance their products. This report provides a detailed description of the methodology used for sentiment analysis, including data cleaning, exploratory data analysis, N-gram analysis, stemming, TF-IDF method, and model-building. The report concludes with key findings and recommendations.

Methodology:

The dataset used for sentiment analysis was obtained from Kaggle in the form of a .csv file. The dataset contained ratings and sentence reviews. To analyze the sentiment of the sentence reviews, Natural Language Processing (NLP) techniques were employed. The following steps were taken in the methodology:

1.Data Cleaning:

- **Converting review texts to lowercase:** This step standardized the text across the dataset, eliminating variations in capitalization.
- **Punctuation Cleaning:** Thorough cleaning of punctuation was performed to ensure data accuracy and readability.
- **Removing Stopwords:** Common words that do not carry significant meaning in sentiment analysis, such as "the" or "is," were removed to improve accuracy.

2.Exploratory Data Analysis (EDA):

- EDA was conducted to identify patterns and trends in the data. Visualizations like plots, charts, and graphs were used to analyze the data.
- Key findings from EDA included the relationship between positive reviews and helpful rate, the rise and dip in positive reviews over time, and the distribution of review counts.
- N-Gram Analysis: N-gram analysis was performed to identify common words and phrases in positive, neutral, and negative reviews.
- **Wordclouds:** Wordclouds were created for negative, neutral, and positive reviews to visualize the most frequently occurring words in each sentiment category.

3.Stemming:

Stemming was applied to reduce words to their root form, allowing for the grouping of words with similar meanings. This technique helped capture the overall positive sentiment and improve the accuracy of sentiment analysis.

4.TF-IDF Method:

The TF-IDF (Term Frequency-Inverse Document Frequency) method was used to encode words and their sentiment.This method assigned weights to words based on their frequency and importance in the document.Bigrams (two words) were considered, and only the top 5000 words from the reviews were used.

5.Model-Building:

- The sentiment of the reviews was taken as the target feature, and the reviews themselves were used as the input feature.
- As there were more positive sentiments compared to neutral and negative sentiments, the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset.
- The dataset was divided into a 75:25 ratio for training and testing.
- Several classification models, including Logistic Regression, Decision Tree Classifier, KNN Classifier, Support Vector Machine Classifier, and Naïve Bayes, were used to classify sentiment.
- Cross-validation was employed to identify the best performing model, with Logistic Regression we achived 94% accuracy in the test set.

Conclusion:

This report provides the detailed methodology of performing sentiment analysis on Amazon reviews, including data cleaning, exploratory data analysis, N-gram analysis, stemming, TF-IDF method, and model-building. Key findings emphasize the importance of N-gram analysis, manual stop word checking, and neutral reviews for product improvement.

The report provides a detailed analysis of Amazon reviews, aiming to offer businesses valuable insights into customer satisfaction and areas for improvement, using various techniques for enhanced accuracy and effectiveness.