SENTIMENTAL ANALYSIS ON CUSTOMER REVIEWS

AIM

The aim of this project is to utilize Selenium for web scraping to extract product reviews from various e-commerce platforms, focusing on Flipkart. The collected reviews will undergo meticulous preprocessing to refine raw data for sentiment analysis. Advanced deep learning techniques, particularly BERT, will be employed to conduct nuanced sentiment analysis on the reviews. The ultimate goal is to provide actionable insights through visual representations of sentiment trends, aiding businesses in strategic decision-making to enhance customer satisfaction and competitiveness in the market.

INTRODUCTION

In the digital age, understanding customer sentiment is vital for businesses. E-commerce platforms provide a plethora of reviews, crucial for this understanding. Leveraging Selenium, this project scrapes reviews from diverse platforms for a targeted product. Through meticulous preprocessing, raw data is refined for sentiment analysis. Advanced deep learning, particularly BERT, is utilized for nuanced analysis. Visual representations of sentiment trends offer actionable insights, aiding businesses in strategic decision-making to enhance customer satisfaction and competitiveness in the market.

DATASET DESCRIPTION

The dataset comprises approximately 8000 product reviews sourced exclusively from the e-commerce platform Flipkart using web scraping techniques. Each review entry primarily includes the following attributes: **Review Text**: The main body of the review, containing the customer's sentiments, opinions, and feedback regarding the product. **Rating**: The numerical rating assigned by the reviewer to indicate their satisfaction level with the product, typically on a scale of 1 to 5. **Date**: The date when the review was posted or submitted by the reviewer.

OBJECTIVE

DATA COLLECTION AND STORAGE PROCESS:

The objective of this phase is to collect product reviews from various e-commerce websites using web scraping techniques and store them efficiently for further analysis. This involves:

1. **Utilizing Selenium for Web Scraping**: Employ the Selenium framework to automate web browsers, enabling the extraction of reviews from e-commerce websites. Utilize WebDriverWait to handle dynamic elements that load asynchronously on web pages, ensuring comprehensive data collection.

2. **Scraping and Storing Reviews**: Scrape the reviews from the website and store them in a list format. Utilize the pandas library convert the list into a DataFrame. This DataFrame serves as a database for organizing the collected reviews.
3. **Converting to CSV Format**: Convert the DataFrame containing the reviews into a CSV (Comma Separated Values) file format. This format facilitates easy access and manipulation of the data for subsequent analysis and visualization tasks.

DATA PREPROCESSING:

1. **Removing duplicates**: Check for and remove any duplicate reviews to ensure data integrity and avoid bias in analysis
2. **Handling missing values**: Inspect the dataset for any missing values in review text, rating, or date fields. Handle these missing values appropriately, either by imputation or removal.

3. **Text cleaning:**
   1. Removal of special characters: Remove special characters, punctuation, and non‑alphanumeric characters from the review text.
   2. Removal of emojis: Remove emojis which is unnecessary noise in the dataset to predict accurate results.
   3. Removal of stop words: Remove stop words (commonly occurring words like 'and', 'the', 'is') that do not contribute much to the sentiment analysis.
   4. Tokenisation: NLTK provides a word tokenizer that splits text into words based on whitespace and punctuation(if needed).
   5. Converting to CSV Format: Convert the DataFrame containing the reviews into a CSV (Comma Separated Values) file format. This format facilitates easy access and manipulation of the data for subsequent analysis and visualization tasks.

SENTIMENT ANALYSIS PROCESS:

The primary objective is to conduct a comprehensive analysis of sentiment using pre-trained models. The focus will be on exploring the capabilities of these models in understanding and classifying sentiment in textual data. The specific goals include:

1. **Model Implementation**: Implement sentiment analysis using both BERT and DistilBERT models on a relevant dataset. Preprocess the data and fine-tune the models for the sentiment classification task.
2. **Performance Evaluation**: Evaluate the performance of the BERT and DistilBERT models in terms of accuracy, precision, recall, and F1 score. Compare their efficiency and computational requirements.
3. **Feature Comparison:** Investigate the key features and aspects of BERT and DistilBERT that contribute to their performance in sentiment analysis. Explore how the models handle nuances and context in sentiment-laden text.
4. **Computational Efficiency**: Assess the computational efficiency of DistilBERT compared to the original BERT model. Analyze the trade-offs between model complexity and performance.

SENTIMENTAL ANALYSIS ON CUSTOMER REVIEWS

MODEL ARCHITECTURE AND EVALUATION

MODEL ARCHITECTURE:

1. The sentiment analysis model architecture employed in this project leverages the BERT (Bidirectional Encoder Representations from Transformers) model, a state-of-the-art deep learning architecture for natural language processing tasks. BERT is a transformer-based model that utilizes a bidirectional approach to capture contextual information from both left and right context in a sentence. This allows the model to understand the nuances and semantics of language more effectively
2. The BERT model is fine-tuned on the collected dataset of product reviews from Flipkart. Fine-tuning involves adjusting the pre-trained BERT model parameters to adapt it to the specific task of sentiment analysis on the review data. This process typically involves adding an additional classification layer on top of the BERT model, which is trained to predict the sentiment polarity (positive, negative, or neutral) of each review.
3. During training, the model learns to map input review text sequences to corresponding sentiment labels by minimizing a predefined loss function, such as cross-entropy loss. The training process involves optimizing the model's parameters using gradient descent-based optimization algorithms, such as Adam.

EVALUATION:

The sentiment analysis model is evaluated using standard evaluation metrics to assess its performance and generalization capability. Common evaluation metrics for sentiment analysis tasks include accuracy, precision, recall, F1-score, and confusion matrix analysis.

1. Accuracy: The percentage of correctly predicted sentiment labels over the total number of reviews.
2. Precision: The proportion of true positive predictions (correctly predicted positive sentiments) out of all positive predictions.
3. Recall: The proportion of true positive predictions out of all actual positive sentiments in the dataset.
4. F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
5. Confusion Matrix: A table summarizing the model's predictions versus the actual sentiment labels, facilitating a deeper analysis of classification errors.

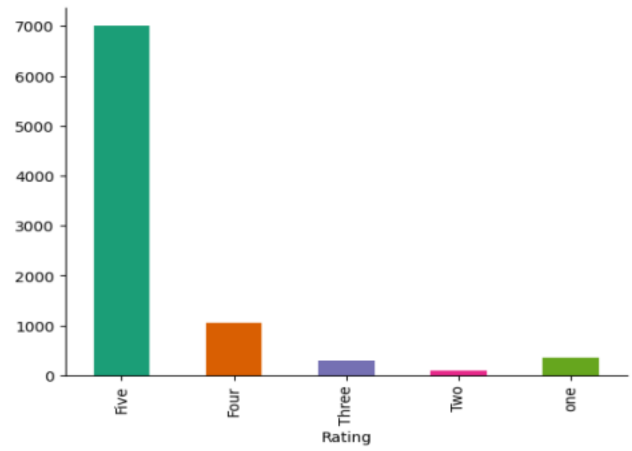
Additionally, the model's performance may be visualized using precision-recall curves, ROC curves, and confusion matrix heatmaps to provide insights into its strengths and weaknesses across different sentiment classes. By evaluating the model's performance on a separate test dataset, the project assesses its ability to generalize to unseen data and provides insights into its effectiveness in accurately predicting sentiment polarity in product reviews from Flipkart.

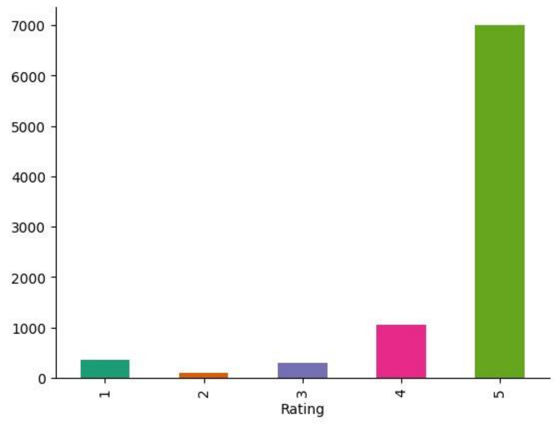
CONCLUSION

In conclusion, our project successfully utilized Selenium for web scraping to extract iPhone 13 reviews from Flipkart. Meticulous data preprocessing prepared the information for analysis, and the implementation of the BERT model facilitated insightful sentiment analysis. This synergy of web scraping, data preprocessing, and advanced sentiment analysis with BERT not only demonstrates technical proficiency but also underscores the effectiveness of cutting-edge natural language processing. The project provides valuable insights into customer sentiments about the iPhone 13 on Flipkart, benefiting businesses, consumers, and researchers

RESULT

In summary, our analysis using the BERT model, coupled with evaluation methods like bar graphs, revealed a notable correlation between 5-star ratings and highly positive sentiments in iPhone 13 reviews on Flipkart. The concentration of positive opinions was distinctly higher among users who gave a 5-star rating.





This finding underscores the reliability of the BERT model and provides valuable insights into the alignment between numerical ratings and the expressed sentiments in textual reviews. This correlation highlights a practical understanding of how users convey their satisfaction through rating systems, offering meaningful implications for businesses and consumers alike.

REFERENCE

BERT MODEL:

https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

https://metatext.io/models/nlptown-bert-base-multilingual-uncased-sentiment

https://medium.com/@sreevatsavthamminana/sentiment-analysis-with-hugging-face-transformers-unleashing-the-power-of-nlp-2749fba08e02

SELENIUM:

https://www.selenium.dev/documentation/