PRODUCT REVIEW SENTIMENT ANALYSIS

Abstract:

Over the last three decades, technological advancements have increased the importance of data of all sorts and manners. It has become one of the essential raw materials that lay the foundation of any strong business. Data, in various forms, is now utilized daily to accomplish tasks, ranging from improving customer experience to predicting whether it will rain ten days later.

However, this is easier said than done. The collection of data is not taxing. Processing, analyzing, and understanding the meaning of what the data represents and deriving helpful information and knowledge from it is challenging and the part which matters. This process is most commonly known as Data Mining.

In this project, we have used a set of Data Mining techniques and their derivatives to develop a sentiment analysis project with numerous real-life applications.

Problem definition:

In today's age, where e-commerce websites are on the rise, and there is an abundance of data, customer reviews take the limelight. There is a plethora of the same, and every day, companies like Amazon, Snapdeal, and Flipkart, to name a few, are overwhelmed with the number of reviews they receive. An important aspect of processing and usefully utilizing this data is sorting these reviews into positive and negative reviews.

Since many customer reviews are being registered daily, it is paramount to automate classifying them into the said categories. It can significantly reduce the amount of time and effort that goes into analyzing these reviews to critique the quality of products being supplied. It would enable the product suppliers to take adequate steps and implement the changes that would mold the product as per the customer's requirements.

The solution and the dataset:

We have curated an Amazon product review dataset containing about 3.6 million extensive product reviews, each associated with a positive or negative sentiment. All the reviews are in the English language.

We used this dataset to train a model which uses Natural Language Processing (NLP) techniques, count vectorization, tokenization, and a Naïve Bayes classifier to classify any given product review as positive or negative.

Following is the link to the Kaggle dataset:

https://www.kaggle.com/datasets/nabamitachakraborty/amazon-reviews

The outcome of the project:

The classifier that we built from scratch works as expected and consistently classified testing product reviews under their correct sentiment with around 85% accuracy over several tests.

We verified the correctness of our NLP that we built from scratch, against that of the inbuilt NLP library. Both were found to give the same result, hence establishing the truth value of our NLP.

Novelty in our project:

The unique factor surrounding our project is its scalability. By substituting the Naïve Bayes classifier with Neural networks and improving upon the currently implemented NLP, we can significantly improve the classification accuracy with significant ease. However, these methods were beyond the scope of this project and course.

Team members and Contributions:

1. Harshit Verma – 2020A7PS0041H; Implemented NLP
2. Rahil Sanghavi – 2020A7PS2052H; Implemented Naïve Bayes Classifier
3. Mohit Agarwal – 2020A7PS0189H; Explored and cleaned the dataset