**https://github.com/yasmeenha/NLP-project**

**CSCE 5290 Section 002 - Natural Language Processing**

**Project Proposal**

**Project Title:** Text Classification and Sentiment Analysis of Amazon Reviews

**GitHub Link:** https://github.com/yasmeenha/NLP-project

**Team members:**

**Sravani Katlaganti(11560617)**

**Panduga Raja Tejasvi Prasad (11414926)**

**Yasmeen Haleem (11462753)**

**Motivation:**

Online markets provide a huge variety of goods and services and have become an essential part of our lives in the current digital era. And amazon is a global leader among these platforms, bringing together millions of sellers and buyers from all over the world. To make wise purchasing decisions, shoppers frequently rely on the experiences and viewpoints of others, and Amazon reviews are crucial in influencing these decisions [1].

The inspiration for our project came from realizing the enormous value that can be found in the abundance of Amazon reviews. Through these evaluations, millions of customers each day share their opinions, suggestions, and sentiments, providing a wealth of data just waiting to be tapped into. However, manual analysis is impractical due to the overwhelming number of reviews. Hence, our project uses NLP, text classification and sentiment analysis to mine insightful information from Amazon reviews to empower both buyers and sellers.

Consumers are looking for effective solutions to traverse the enormous product environment in an age of information overload. Our initiative seeks to give customers a quick and thorough understanding of product experiences, assisting them in making informed decisions by automatically categorizing reviews and identifying feelings [1].

Amazon sellers are always looking for ways to enhance their goods and services. Sellers can improve their offers, maximize consumer satisfaction, and customize their marketing efforts by analyzing the feelings and categorizations of evaluations. Comprehensive datasets of classified and sentiment-tagged Amazon reviews are useful for researchers and analysts. Such information is useful for researching customer preferences, market trends, and the effects of new releases.

**Significance:**

Natural Language Processing (NLP) and e-commerce analysis have high importance in terms of text classification and sentiment analysis of Amazon reviews. Our project simplifies the categorization of huge volumes of textual data, such customer reviews on Amazon with respect to various products, into categories that reflect sentiments such as positive, negative, or neutral by using advanced NLP algorithms. The algorithmic classification speeds up the process of collecting insights from content generated by users which allows businesses and consumers to quickly assess the overall perception of goods and services and improve their quality.

It enables the businesses to manage their reputations proactively, track customer happiness, and pinpoint problem areas based on feedback from their customers. Essentially, it acts as a powerful tool for extracting insights from the large sea of unstructured textual data, advancing data-driven decision-making, and improving user experience in the e-commerce space. [2]

In conclusion, Text Classification and Sentiment Analysis of Amazon Reviews highlights the potential of advanced data analytics in the digital age and represents a critical part of business intelligence using NLP.

**Objectives:**

The objective of this project is to create a dependable and expandable system for classifying language and analyzing sentiment in Amazon reviews. While preserving the integrity and authenticity of the Amazon review platform, we seek to offer beneficial insights to buyers, sellers, and researchers by utilizing natural language processing (NLP) methods and machine learning models.

Create a text classification system that classifies Amazon reviews automatically into pertinent, predetermined categories, such as product features, customer service and delivery experience. The objective of this project is to create a dependable and expandable system for classifying language and analyzing sentiment in Amazon reviews.

Use sentiment analysis to identify the tone (positive, negative, or neutral) used in each review, giving users a better knowledge of the experiences of other people.

Create meaningful reports and visualizations that offer useful information to multiple stakeholders on the performance of the product, market trends, and customer sentiment. By meeting these goals, we hope to develop a useful tool that improves the Amazon shopping experience, gives sellers the power to upgrade their products, gives researchers the ability to analyze market trends, and ensures the openness and reliability of the Amazon review ecosystem.

**Features:**

The main features of this project are collection and cleaning the Amazon reviews dataset that contains the review text and their corresponding ratings. And then using this cleaned data performing exploratory data analysis for gaining insights into the trends of the ratings and text. Then using the feature engineering techniques for transforming any string data into numerical data so that it is suitable for classification and analysis. These features are then utilized as inputs in the modeling of the appropriate machine learning algorithms like transformer models of BERT. Then train these models using the train dataset and hyper tuning the model parameters to get optimum performance using cross-validation data. Then test the model using the test dataset and evaluate the model using various metrics of accuracy and precision. Thus, obtaining an optimal model for the purpose of text classification and sentiment analysis.

***Uniqueness:***

In this project we will try to provide customized product recommendations based on their sentiment preferences by combining collaborative filtering methods with sentiment analysis.

***Deliverables and Milestones:***

In the future we will be setting milestones below to complete the tasks.

* We will be drafting the project proposal document covering the goals, needs, and scope.
* Datasets that have been cleaned, preprocessed, and annotated with sentiment labels and specified categories are those that have been obtained from Amazon reviews.
* Trained NLP Models: Trained and improved NLP models for sentiment analysis, aspect-based sentiment analysis, and text classification.
* In text classification we will be classifying the Amazon reviews into the appropriate subject areas, such as product features, customer support, and delivery experience.
* Then will be performing sentimental analysis on the reviews to categorize it into positive, negative and neutral.
* A final presentation summarizing the project's goals, achievements, and future recommendations.

***Technologies***:

Advanced NLP tools and frameworks like NLTK and spaCy libraries are being used. These technologies make it possible to do crucial preprocessing tasks like text segmentation and part-of-speech labeling. Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, or DistilBERT, are state-of-the-art for NLP tasks like text classification and sentiment analysis.[2]

Text normalization, stop-word removal, stemming, and feature engineering are all part of a thorough data preprocessing workflow. This procedure boosts the precision of sentiment analysis while optimizing the textual material for model input.

***Data***:

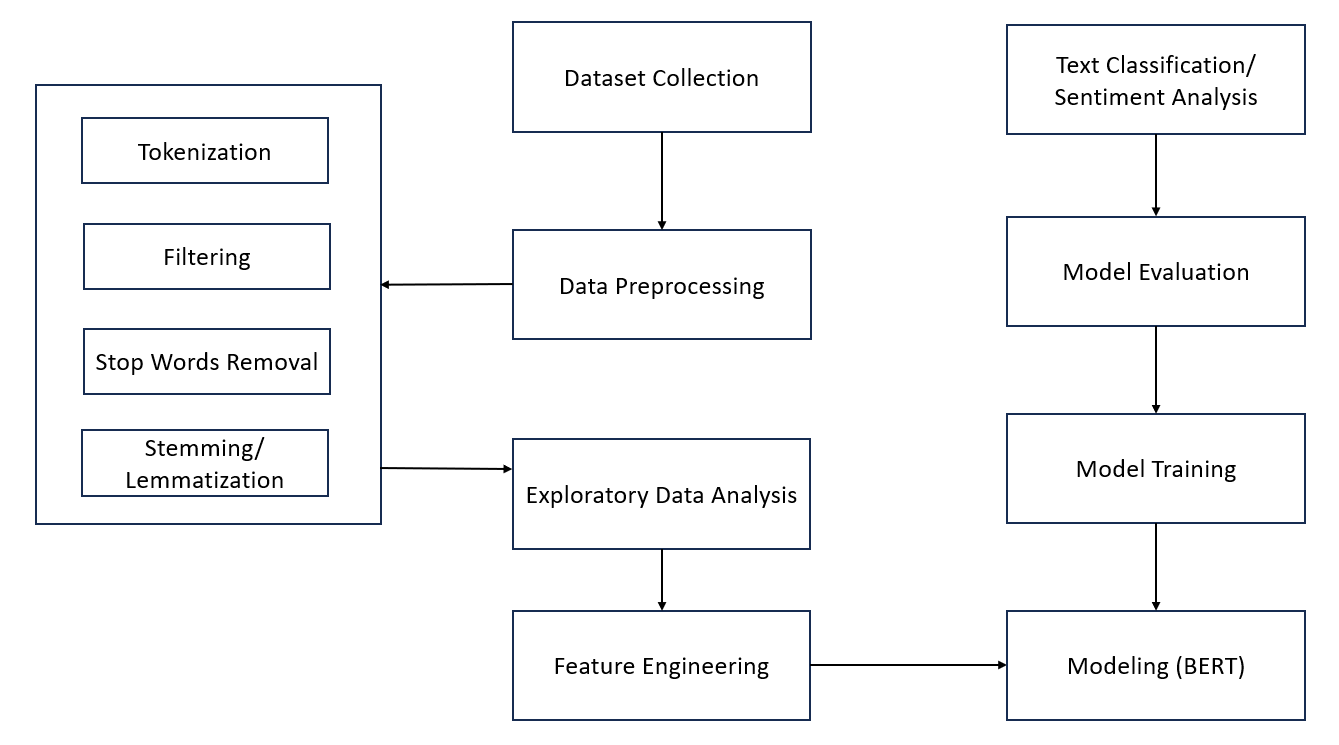
The dataset we are using is from the Kaggle website. It contains 34L Amazon reviews from amazon customers on 2 lakhs products of Amazon. This dataset contains around 2,500,000 training datasets and 300,000 testing datasets in each polarity sentiment. **Review bosy:** This column contains the actual review of the Amazon reviews.[3]

**Sentiment Label (Target):** This column contains the sentiment labels for each review (e.g., "positive," "negative," "neutral").

**Product Details:** This column may contain information about the products being reviewed, such as product IDs or names.

**Rating:** This column may contain the star rating given by the reviewer (1 to 5).

**Workflow:**



**Workflow of the Text Classification and Sentiment Analysis of Amazon Reviews**

The first step in the process of modeling any machine learning or natural language processing task is the collection of data. Using the unique and informative dataset is a very crucial step because data is very powerful and gathering informative data decides the overall performance and accuracy of the model. We have collected Amazon reviews dataset from Kaggle website which contains product details, text of the review and corresponding ratings. After the collection of data, the next step is to preprocess and clean the data. This step involves removal of any special characters, meaningless words, tags etc. [4] Then the process involves the text tokenization where all the individual words are extracted from each of the review texts. Next, from the words list removal of the stop words takes place. Then on this cleaned data lemmatization or stemming process occurs in order to transform the words into their root form.

The third step is the exploratory data analysis where we get some insights into the patterns of the data and draw meaningful conclusions on the trends of the reviews. This includes visualizing the commonly used words and the distribution of ratings corresponding to each of the products. This part of analysis is useful in modeling appropriate text classification model. And checking for any imbalances in the classes helps in the modeling of sentiment analysis. The next step of feature engineering involves transformations like changing some string data into digital values making them suitable for use in the NLP models. Some of these famous techniques are BERT, TF-IDF and some other word embeddings.[4]

The next step involves selection of suitable machine learning/ deep learning models for the modeling of sentiment analysis and text classification. Most popular models may include Support Vector Machines, Recurrent Neural Networks, or Transformed-based models like BERT which we are planning to use. Proceeding further, the next step is to train the selected and built model using the training dataset. It is also very important to maintain completely separate datasets for training, evaluating and cross validating the model to acquire better performance and robust models. After that hyperparameter tuning has to be applied to get optimal performance of the model. We can use cross-validation data at this stage for selecting optimal parameters. Next step is to evaluate the model using test dataset and some of the popular metrics used for evaluating the model are accuracy, F1-Score, Precision and Recall. Then analyze the model using these results and draw conclusions.

Finally, the best model obtained using hyperparameter tuning of optimal parameters is used for the classification of the review’s text and analysis of the sentiment present in the reviews. Thus, with the help of the accuracy and other metrics results we can show that the proposed and built models in this project perform well than the existing methods of text classification and sentiment analysis on Amazon reviews dataset.

**References:**

1.https://thesai.org/Paper\_3Sentiment\_Analysis\_on\_Amazon\_Product\_Reviews.pdf

2. <https://towardsdatascience.com/sentiment-analysis-on-amazon-reviews-45cd169447ac>

3. <https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews>

4. <https://ieeexplore.ieee.org/document/9402414>

5. [https://github.com/joshivaibhav/AmazonCustomerReview/blob/master/amazondata.csv](https://github.com/dataset)