Course Info: **CSCE 5290 – Natural Language Processing (Section 003 – Spring'23)**

Group: *Srikanth Reddy Dwarakapalli (SrikanthReddyDwarakapalli@my.unt.edu) (11650259)*

Sreevani Danthojue (sreevanidanthojue@my.unt.edu) (11643429)

Kishore Sai Lakshman Rangisetti (kishoresailakshmanrangisetti@my.unt.edu) (11666403)

Satvik Reddy Chittela (satvikreddychittela@my.unt.edu) (11657165)

GitHub Link: *https://github.com/Sreedanthojue2/NLP-project.git*

Instructor: **Dr. Zeenat Tariq (*Zeenat.Tariq@unt.edu*)**

Dataset: **Amazon and Walmart Common Products, Reviews, and Ratings Dataset**

Proposal: **Sentiment Analysis and Recommendation system on Amazon and Walmart Product Reviews (GROUP 14)**

INCREMENT -1

Contents

Motivation 3

Project Introduction 3

Objective 3

Features of Proposed System: 4

Dataset Description 4

Data Pre-processing 4

Exploratory Data Analysis 5

Sentiment Analysis 5

Profanity Checker 5

Future Work 6

Total Votes and Star Rating Analysis 6

Recommendation Generation using ML Models 6

Flow Chart 6

Conclusion 7

Related Work 7

Project Management 9

References 10

Motivation:

With the popularity of online marketplaces over the last few decades, online sellers and merchants now regularly seek customer feedback. As a result, millions of reviews are created every day, making it difficult for potential customers in making a purchasing decision. It takes time. This project considers the problem of classifying reviews according to their overall semantics (positive or negative). Two separate supervised machine-learning techniques will be used to conduct the study. Different category products from Amazon and Walmart will experiment with SVM, logistic regression, multinomial naive Bayes, decision trees, and ensemble classifiers.

Project Introduction:

People are purchasing goods through multiple e-commerce websites because the world commercial landscape almost entirely transitions to the online platform. And as a result, it's also a usual practice to read product reviews before making a purchase. So, to make the data more dynamic, it is now crucial to analyze the data from those customer evaluations. It takes a lot of time in this day and age of increasingly sophisticated machine learning-based algorithms to comprehend a product by reading thousands of reviews, though we can focus a review on a specific category to determine how well-liked it is among consumers everywhere.

This study classifies customer reviews of various goods into positive and negative feedback. It also develops a polarize a significant number of reviews using supervised learning. According to Amazon research last year, Online shoppers with 88% believe that online reviews are as reliable as personal recommendations. A convincing argument for the legitimacy of any internet product may be made for it by the quantity of positive reviews it has received. In contrast, a lack of feedback for any other online or books transaction makes prospective customers wary. It's just more credible to evaluate things with more numbers. It's important to respect other people's opinions and experiences, and the only way to learn what other people think of a product is to read reviews on it. User experiences with goods or topics, as well as opinions gathered from those experiences, directly affect future consumer purchasing decisions. Similarly, poor evaluations frequently result in decreased sales. Our goal is to get input from customer and use it to polarize a lot of data. There have been some comparable studies conducted using Walmart and Amazon datasets. To comprehend the polarized attitudes toward the products, we will conduct sentiment analysis on a small collection of datasets of Amazon and Walmart product reviews. We will contrast the performance of the two groups in several areas, including product rating and price.

Objective:

- The project uses sentiment analysis to analyze and compare customer reviews from Amazon and Walmart and prepare a word cloud to identify crucial keywords from the data in different categories.
- The objective is to develop a prediction system or a recommendation system that can be used to identify which category of products has better rating and pricing among Walmart and Amazon and it recommends the best possible outcome from the user request by suggesting the best product.

Features of Proposed System:

In our proposed approach we will use several Natural Language Processing Techniques like Pre-Processing and Sentiment Analysis and then we will use Machine learning algorithms like Logistic Regression (LR), Random Forest (RF), Support vector machine (SVM), TF - IDF matrix.

- First, we will take the Amazon product review dataset and other different e-commerce websites dataset containing different classes or products.
- We will filter the dataset according to requirements and create a new dataset that has attributes according to analysis to be done by combining all the different datasets alongside ratings and reviews.
- We will perform Pre-Processing on the dataset.
- We will Split the data into training and test it.
- We will train the model with training data and then analyze the testing dataset over the classification algorithm.
- Then, we will create a recommendation system from the positive and negative outcomes of the sentiment analysis.
- It will recommend products of a different kind from the user input by using considering different factors like rating and prices and review by the customers simultaneously as a priority.
- Also, it will classify the products based on the ratings of a product on different websites in different categories and recommend which website would be preferred.

Dataset Description:

The dataset for the Amazon and Walmart reviews for different category products are obtained from the Kaggle Website and a few other sources. We have processed the data so that we could get a merged dataset that includes data from Walmart and Amazon in a single dataset in the same rows.

The dataset used for this project consists of product reviews from two popular online shopping websites, Amazon, and Walmart. The dataset contains information about the product, customer reviews, and ratings. The reviews are written in the English language and cover a wide range of product categories. The link for the final dataset is given below:

- https://www.kaggle.com/datasets/sreevanidanthojue/combined-dataset-for-amazon-and-walmart-reviews

Data Pre-processing:

Data preprocessing is an essential step in natural language processing. The following preprocessing steps were performed on the dataset to make it suitable for further analysis:

- Removal of irrelevant information such as product details, customer information, etc.
- Removal of special characters, numbers, and punctuation marks.
- Conversion of text to lowercase.
- Tokenization of text.
- Removal of stop words such as 'a', 'an', 'the', etc.

Exploratory Data Analysis:

An essential phase is the analysis of exploratory data as it helps in understanding the data and identifying patterns and trends. The following analyses were performed on the preprocessed data:

- Distribution of star ratings across both Amazon and Walmart datasets.
- Analysis of the most common words used in the reviews.
- Visualization of the most common words used in the reviews using word clouds.

Sentiment Analysis:

Finding and extracting subjective information from text data is the process of sentiment analysis. Here in this project, this was performed using a Naive Bayes Classifier machine learning model. The following steps were performed for sentiment analysis:

- Cleaning the text data using the preprocessing techniques mentioned above.
- Generating an average star rating based on both Amazon and Walmart ratings.
- Defining good/bad products based on the star rating.
- Vectorizing the cleaned_review_body using CountVectorizer.
- Splitting the data into train and test sets.
- Training the Naive Bayes Classifier model.
- Predicting on the test dataset.
- Calculating the classification report for sentiment analysis.
- Visualizing the percentage of good vs bad product reviews.

Profanity Checker:

A profanity checker is used to identify the use of abusive or inappropriate language in text data. In this project, a profanity checker was implemented to check for the presence of profanity in product reviews. The following steps were performed for the profanity checker:

- Retrieving profanity wordlists from online sources.
- Removing duplicates and merging the profanity word data.
- Implementing a function to check for profanity in the text data.
- Setting a threshold for the profanity score.
- Checking for profanity in each review.
- Adding profanity scores as a new column in the data frame.
- Visualizing the percentage of profane vs non-profane reviews.

Future Work:

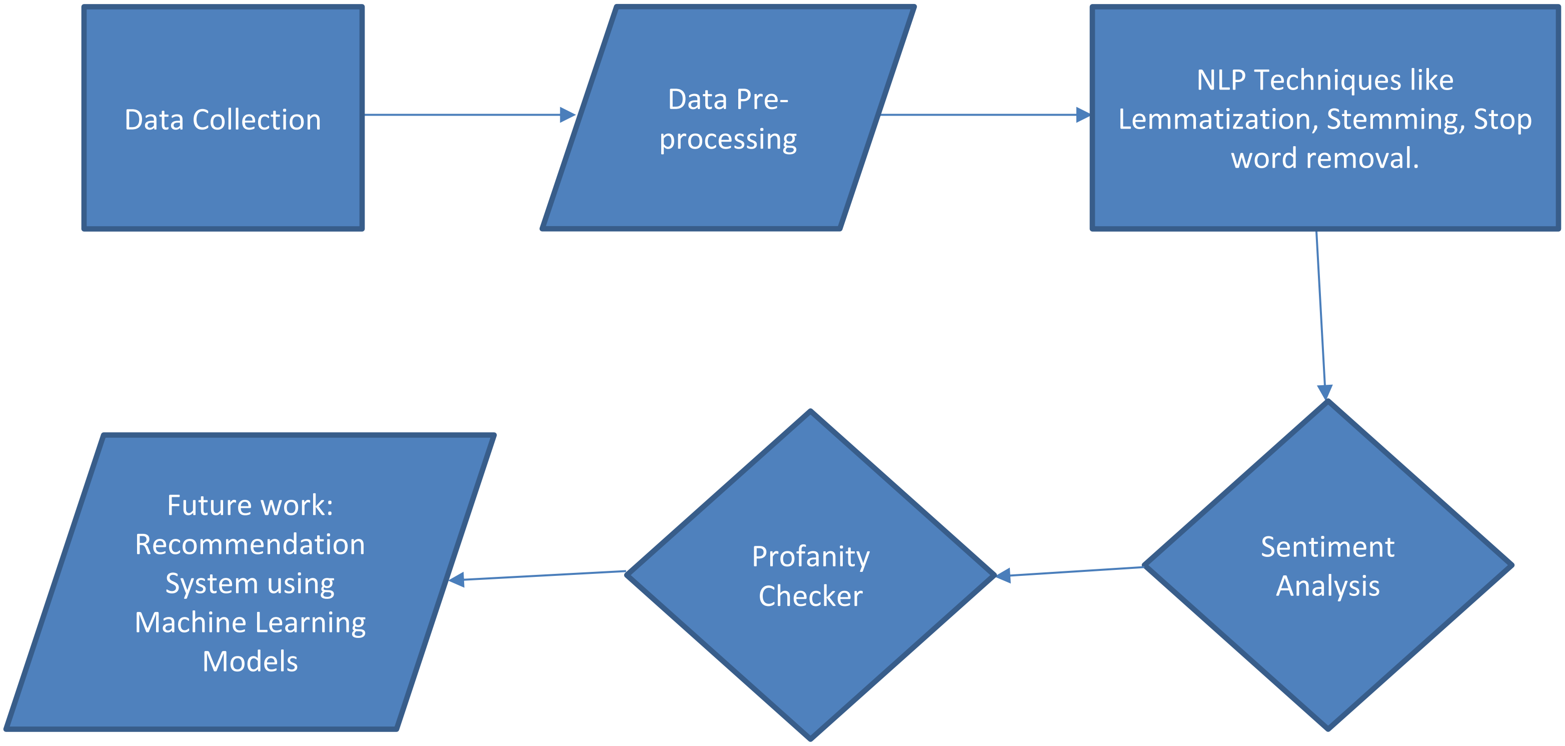
Total Votes and Star Rating Analysis:

We will also analyze the total votes and star ratings to determine the popularity and quality of a product on each platform. We will use the pandas library to compute various statistics such as the mode, mean, median, skewness, standard deviation, correlation and kurtosis.

Recommendation Generation using ML Models:

Based on the results of the sentiment analysis and the analysis of total votes and star ratings, we will generate a recommendation for the user on whether to buy the product from Amazon or Walmart. We can use a simple rule-based approach or a machine learning algorithm to generate the recommendation. But we plan to implement Machine Learning Models like Support Vector Machine (SVM)/Random Forest/Logistic Regression etc.

Flow Chart:


Data Collection
Data Pre-processing
NLP Techniques like Lemmatization, Stemming, Stop word removal.
Sentiment Analysis
Profanity Checker
Future work: Recommendation System using Machine Learning Models

Conclusion:

The project's use of natural language processing techniques to analyze the sentiments contained in product reviews was successful. The Naive Bayes Classifier model performed well in predicting the sentiment of the reviews. The profanity checker helped in identifying the presence of profanity in the reviews, which can be used to filter out inappropriate content. The results of the project can be used by businesses to improve their products and services and provide a better customer experience. Future work on Total Votes, Star Rating Analysis, and Recommendation generation would add more value to the project and could provide insights derived from Customer Feedback on E-commerce giant sites like Amazon and Walmart.

Related Work:

1. **Sentiment Analysis on Amazon Products - Kaggle:**

   https://www.kaggle.com/code/aherparesh/sentiment-analysis-on-amazon-product-rnn-97-acc

   Summary:

   Input: The input data consists of textual reviews that need to be preprocessed and cleaned to remove unwanted characters, links, and stopwords. The cleaned reviews are then tokenized and converted into integer sequences using the Keras Tokenizer class. These integer sequences are then padded to a fixed length using the Keras pad_sequences function. The input features are the padded integer sequences, and the maximum length of each sequence is set to a value of 8. The features seem appropriate for the sentiment analysis task as they capture the essence of the reviews in a numerical format.

   Output: The model output is a binary classification of sentiment, either positive or negative. It is desirable to have a balanced dataset with same number of positive and negative reviews, which is not explicitly stated in the code. If the dataset is imbalanced, the model's accuracy metric could be misleading as it would predict the majority class more frequently.

   Explanation: The chosen model architecture is a recurrent neural network (RNN) consisting of two LSTM layers and a dense layer with a ReLU activation function. The LSTM layers are used because they are good at handling sequential data and can learn long-term dependencies. The ReLU activation function is used in the dense layer because it has been shown to work well in deep neural networks. Using a sigmoid activation function, an output layer is constructed with a dense layer for binary classification. Overall, the chosen models seem appropriate for the sentiment analysis task, and the architecture can be further fine-tuned by adjusting hyperparameters such as the number of LSTM units and the learning rate of the optimizer. Alternatively, other models such as Convolutional Neural Networks (CNNs) or Transformers could also be used for this task.

(Continued on next page)

2. **Sentiment Analysis with Amazon Reviews:**

   https://www.kaggle.com/code/aaroha33/sentiment-analysis-with-amazon-reviews/

   Summary:

   Input:

   The input data consists of textual data represented as sequences of integers after tokenization and padding. The features used for training the models are the sequence of tokens that represent each input text. The tokenizer is used to convert the text into numerical sequences, and the pad_sequences method is used to make sure all the sequences have the same length.

   Output:

   The output of the models is a binary classification indicating whether the input text is positive or negative sentiment. If the output is not as desired, it could mean that the models are not learning the underlying patterns in the data or that the data is not properly represented. It could also be that the models are not complex enough to capture the nuances of the data.

   Explanation:

   Two models were used in the code blocks: a recurrent neural network (RNN) and convolutional neural network (CNN). For NLP tasks, the CNN model is a better pick because it can learn hierarchical representations of the input data by applying convolutional filters over the text. The RNN model, on the other hand, is a good pick because it can learn sequences of patterns in the input data. The models seem to be appropriate for the given task, but other models such as transformer-based models like BERT could be used to improve the performance. However, the selection of models also depends on the size of the dataset and the hardware available for training the models.

Presentation Video Link:

https://www.youtube.com/watch?v=RrMA4PHG_r8

Project Management:

We, as a group of 4 joined in a google meet and worked on the report individually and shared the report work. All of us participated equally in both the project code work and report work.

Name	Participation
Sreevani Danthojue	Collected the required dataset for the project from the Kaggle and merged them into a single dataset
Kishore Sai Lakshman Rangisetti	Did the pre-processing for the dataset such as stop word removal and generated the word cloud and bar chart
Srikanth Reddy Dwarakapalli	Generated the word cloud for positive and negative reviews and splitted the data into test and training set
Satvik Reddy Chittela	Trained the model and Generated the classification report, did the profanity check and generated a pie chart.

References:

1. Nina Isabel Holleschovsky, “The social influence factor: Impact of online product review characteristics on consumer purchasing decisions”, 5th IBA Bachelor Thesis Conference, Enschede, The Netherlands 2015
   https://essay.utwente.nl/67351/1/Holleschovsky_BA_MB.pdf
2. Elli, Maria Soledad, and Yi-Fan Wang. "Amazon Reviews, business analytics with sentiment analysis." 2016
   https://docplayer.net/151407565-Amazon-reviews-business-analytics-with-sentiment-analysis.html
3. Xu, Yun, Xinhui Wu, and Qinxia Wang. "Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews." (2015).
   https://cs229.stanford.edu/proj2014/Yun Xu, Xinhui Wu, Qinxia Wang, Sentiment Analysis of Yelp%27s Ratings Based on Text Reviews.pdf
4. Rain, Callen. "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning."Swarthmore College (2013).
   https://www.sccs.swarthmore.edu/users/15/crain1/files/NLP_Final_Project.pdf
5. D.M. Hussein, A Survey on Sentiment Analysis Challenges, Journal of King Saud University-Engineering Science, April 2016
   https://www.sciencedirect.com/science/article/pii/S1018363916300071
6. Singh and R. Sathyaraj, A Comparison between Classification Algorithms on Different Datasets Methodologies Using Rapidminer, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, May 2016
   https://www.ijarcce.com/upload/2016/may-16/IJARCCE 140.pdf