**DISCUSSION ON JOB TRAINING – 1**

**AMAZON REVIEW CLASSIFICATION**

**Abstract:**

Sentiment analysis, a branch of natural language processing (NLP), plays a crucial role in understanding customer opinions and attitudes towards products or services. In this report, we present a sentiment analysis model developed specifically for Amazon reviews. The model utilizes logistic regression coupled with TF-IDF vectorization for feature extraction.

The dataset comprises Amazon reviews across various products, including their corresponding sentiment scores. Initially, data preprocessing techniques were applied, including handling missing values and removing stop words. Sentiment labels were then dichotomized into positive and negative sentiments for binary classification.

Exploratory data analysis (EDA) was conducted to visualize the distribution of sentiments within the dataset. Word clouds were generated to visually represent the most frequent words associated with positive and negative sentiments, offering intuitive insights into the underlying sentiment trends.

To address the issue of class imbalance inherent in sentiment analysis tasks, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to augment the minority class samples. This technique ensures a balanced representation of both positive and negative sentiments, enhancing the model's predictive performance.

The logistic regression model was trained on the preprocessed data, achieving competitive accuracy in sentiment classification. Furthermore, the model was serialized using pickle for future deployment and integration into applications.

In conclusion, this report demonstrates an effective approach to sentiment analysis of Amazon reviews, leveraging logistic regression and SMOTE to handle class imbalance. The developed model provides a valuable tool for businesses to gain insights into customer sentiments and make data-driven decisions to enhance customer satisfaction and product quality.

**Introduction:**

As the commercial site of the world is almost fully undergone in online platform people is trading products through different e-commerce website. And for that reason reviewing products before buying is also a common scenario. Also now a day, customers are more inclined towards the reviews to buy a product. So analyzing the data from those customer reviews to make the data more dynamic is an essential field nowadays. In this age of increasing machine learning based algorithms reading thousands of reviews to understand a product is rather time consuming where we can polarize a review on particular category to understand its popularity among the buyers all over the world.

The objective of this paper is to categorize the positive and negative feedbacks of the customers over different products and build a supervised learning model to polarize large amounts of reviews. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations. Any online item with large amounts of positive reviews provides a powerful comment of the legitimacy of the item. Conversely, books, or any other online item, without reviews puts potential prospects in a state of distrust. Quite simply, more reviews look more convincing. People value the consent and experience of others and the review on a material is the only way to understand others impression on the product. Opinions, collected from users’ experiences regarding specific products or topics, straightforwardly influence future customer purchase decisions. Similarly, negative reviews often cause sales loss. For those understanding the feedback of customers and polarizing accordingly over a large amount of data is the goal. There are some similar works done over amazon dataset. In did opinion mining over small set of datasets of Amazon product reviews to understand the polarized attitudes towards the products.

In our model, we used both manual and active learning approach to label our datasets. In the active learning process different classifiers are used to provide accuracy until reaching satisfactory level. After getting satisfactory result we took those labeled datasets and processed it. From the processed dataset we extracted features that are then classified by different classifiers. We used combination of two kinds of approaches to extract features: the bag of words approach and tf-idf & Chi square approach for getting higher accuracy.

**Aim:**

The world we see nowadays is becoming more digitalized. In this digitalized world e-commerce is taking the ascendancy by making products available within the reach of customers where the customer doesn’t have to go out of their house. As now a day’s people are relying on online products so the importance of a review is going higher. For selecting a product, a customer needs to go through thousands of reviews to understand a product. But in this prospering day of machine learning, going through thousands of reviews would be much easier if a model is used to polarize those reviews and learn from it. We used supervised learning method on a large scale amazon dataset to polarize it and get satisfactory accuracy.

**Objective:**

* The objective of this research is to develop a supervised learning model that efficiently categorizes customer feedback into positive and negative sentiments for various products. In an era dominated by online commerce, where consumer decisions are heavily influenced by reviews, the need to streamline the analysis of vast amounts of feedback is paramount. By automating the sentiment polarization process, this study aims to provide consumers with valuable insights for informed purchasing decisions and assist businesses in understanding the reception of their products in the market.
* To achieve this objective, a combination of manual and active learning approaches will be utilized for labeling datasets. Active learning involves iteratively employing different classifiers to improve accuracy until a satisfactory level is attained. The labeled datasets will then undergo processing to extract relevant features. Feature extraction will utilize two approaches: the bag of words approach and tf-idf & Chi-square approach. These features will serve as input for various classifiers, ensuring accurate sentiment classification.
* Ultimately, the goal is to develop a robust model capable of efficiently categorizing customer feedback, thereby empowering both consumers and businesses in their decision-making processes within the dynamic landscape of online commerce.

**Purpose:**

* The purpose of this research is to address the pivotal role of online product reviews in influencing consumer behavior and purchasing decisions in the digital marketplace. With the increasing reliance on e-commerce platforms for trading goods, the significance of customer feedback has surged, as shoppers often turn to reviews to guide their buying choices. However, the sheer volume of reviews poses a challenge for effective analysis and interpretation.
* This study seeks to develop a supervised learning model that can automatically categorize customer feedback into positive and negative sentiments for diverse products. By leveraging machine learning algorithms, the aim is to streamline the process of sentiment analysis, making it more efficient and scalable. The ultimate objective is to empower both consumers and businesses with actionable insights derived from large-scale review data.
* Through a combination of manual and active learning approaches, the research endeavors to label datasets accurately, ensuring the training of robust classifiers. Active learning methodologies will be employed iteratively to enhance classification accuracy, culminating in a model capable of effectively polarizing sentiments across a wide array of products. Feature extraction techniques, including the bag of words approach and tf-idf & Chi-square approach, will be utilized to capture the nuanced characteristics of customer feedback and improve classification performance.
* By understanding and categorizing the sentiments expressed in customer reviews, this study aims to provide valuable intelligence to consumers, enabling them to make informed purchasing decisions. Likewise, businesses stand to benefit from insights into the reception of their products in the market, facilitating strategic decision-making and product development efforts. Ultimately, the research endeavors to contribute to the optimization of online shopping experiences and the enhancement of consumer satisfaction in the digital age.

**Scope:**

The scope of this research encompasses the development and implementation of a supervised learning model aimed at categorizing customer feedback into positive and negative sentiments across various products within the realm of e-commerce. Given the widespread reliance on online platforms for trading goods and the growing importance of customer reviews in influencing purchasing decisions, the study focuses on analyzing large volumes of review data to derive actionable insights.

**Methodology:**

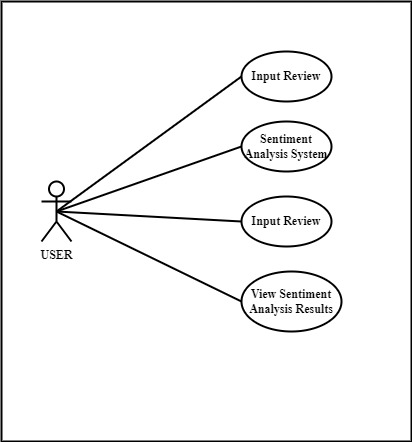
1. **Data Collection:**
   * Description of the dataset used (AmazonReview.csv).Preprocess the data to handle missing values, outliers, and inconsistencies.
   * Explanation of the features (columns) in the dataset.
   * Any preprocessing steps applied during data collection.
2. **Feature Selection and Engineering:**
   * Select relevant features that have a significant impact on rainfall predictions.
   * Engineer new features or transform existing ones to improve the predictive power of the models.
   * Use domain knowledge and statistical techniques to guide feature selection and engineering.
3. **Data Preprocessing:**
   * Handling missing values.
   * Removing stopwords.
   * Tokenization.
   * Lowercasing.
   * Other text normalization techniques applied.
4. **Exploratory Data Analysis (EDA):**
   * Basic statistics of the dataset (e.g., number of samples, distribution of sentiments).
   * Visualizations.
     + Distribution of sentiments (bar plot, pie chart).
     + Word clouds for positive and negative sentiments.
5. **Feature Engineering:**
   * TF-IDF Vectorization:
     + Explanation of TF-IDF.
     + Implementation using TfidfVectorizer.
     + Limiting features to top N terms.
6. **Model Building:**
   * Splitting the dataset into train and test sets.
   * Implementation of logistic regression model.
   * Model training and evaluation:
     + Training accuracy.
     + Test accuracy.
     + Confusion matrix (if applicable).
7. **Model Performance Improvement:**
   * Addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).
   * Explanation of SMOTE.
   * Implementation and impact on model performance.
8. **Model Deployment:**
   * Saving the trained model and vectorizer for future use.
   * Deployment options:
     + Local deployment for individual predictions.
     + Web service deployment for bulk predictions.

**Advantages:**

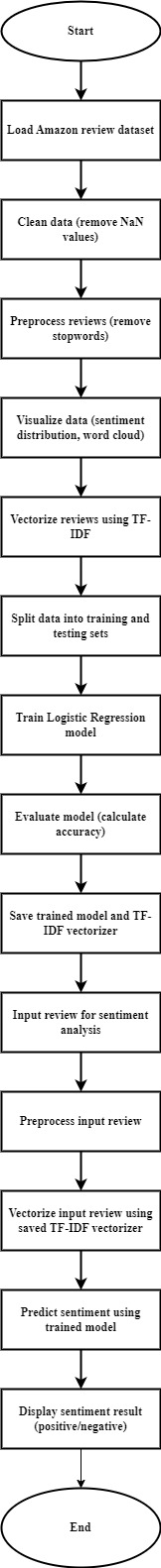
**ADVANTAGES:**

1. The script preprocesses the raw text data by removing stop words and cleaning the reviews, which helps in improving the quality of the input data for analysis.
2. Through bar charts and pie charts, the script provides a visual representation of the distribution of positive and negative sentiments in the dataset, allowing for quick insights into sentiment proportions.
3. The word cloud visualizations generated for both positive and negative sentiment categories offer a concise representation of the most frequent words used in reviews, aiding in understanding the key themes and sentiments expressed by customers.
4. By using TF-IDF vectorization, the script transforms text data into numerical vectors, capturing the importance of words in reviews relative to the entire corpus. This approach helps in representing text data effectively for machine learning algorithms.
5. The logistic regression model is a simple yet effective algorithm for binary classification tasks like sentiment analysis. It offers interpretability and can handle large feature spaces efficiently.
6. The script evaluates the performance of the sentiment analysis model using accuracy score, providing a quantitative measure of how well the model performs on unseen data.
7. The trained logistic regression model is saved using pickle, allowing for easy deployment and reuse without the need for retraining.
8. The script utilizes the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the dataset, ensuring better generalization of the model by generating synthetic samples for the minority class

**UseCase Diagram:**

****

**Activity** **Diagram:**

****

**Conclusion:**

In conclusion, as online commerce continues to dominate global trade, the significance of customer reviews in influencing purchasing decisions has never been greater. Analyzing vast amounts of feedback to categorize and polarize sentiments towards products is essential in this landscape. With over 88% of online shoppers trusting reviews as much as personal recommendations, the volume and sentiment of reviews directly impact consumer trust and purchasing behavior. This study aimed to develop a supervised learning model to categorize positive and negative feedback, leveraging both manual and active learning approaches to label datasets. By employing various classifiers and feature extraction techniques, such as the bag of words and tf-idf & Chi-square methods, the model achieved satisfactory accuracy levels. Understanding and effectively polarizing customer feedback is crucial for businesses in maintaining consumer trust and competitiveness in the online marketplace.