CUSTOMER REVIEW BASED SENTIMENT ANALYSIS

# A PROJECT REPORT

***Submitted by***

**ARAVINTHKRISHNA R (Reg.No.1920106009)**

**ATHISH GR (Reg.No.1920106010) AVINASH SUNDAR K (Reg.No.1920106011) BHUVANESHWARAN S (Reg.No.1920106014)**

***in partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

# IN INFORMATION TECHNOLOGY

**SONA COLLEGE OF TECHNOLOGY**

**ANNA UNIVERSITY: CHENNAI 600 025**

# NOVEMBER 2023

**ANNA UNIVERSITY : CHENNAI 600 025 BONAFIDE CERTIFICATE**

Certified that this project report “**CUSTOMER REVIEW BASED SENTIMENT**

**ANALYSIS**” is the bonafide work of “**ARAVINTHKRISHNA R (1920106009**),**ATHISH GR(1920106010),AVINASHSUNDAR K(1920106011),BHUVANESHWARAN**

**S(1920106014)**” who carried out the project work under my supervision

|  |  |
| --- | --- |
| **SIGNATURE** | **SIGNATURE** |
| (Dr.J.Akilandeswari) | (Mr.P.Dineshkumar) |
| Professor | Assistant Professors |
| **HEAD OF THE DEPARTMENT** | **SUPERVISOR** |
| Department of Information Technology | Department of Information Technology |
| Sona College of Technology | Sona College of Technology |
| Salem-636 005 | Salem-636 005 |

Submitted for Mini Project viva voice examination held on

……………………………..

|  |  |
| --- | --- |
| **INTERNAL EXAMINER** | **EXTERNAL EXAMINER** |

**ABSTRACT**

Customer Review Based Sentiment Analysis involves the application of natural language processing and machine learning techniques to evaluate and interpret customer opinions expressed in reviews. This method aims to discern the sentiment behind these reviews, categorizing them as positive, negative, or neutral. By employing advanced algorithms, the system analyzes the textual content, considering the nuances of language and context. The primary objective is to extract meaningful insights into customer experiences and preferences, enabling businesses to make informed decisions. This process facilitates a deeper understanding of consumer sentiments, identifying areas for improvement and gauging overall satisfaction. Ultimately, Customer Review Based Sentiment Analysis serves as a valuable tool for businesses to enhance product quality, customer service, and overall customer satisfaction by harnessing the wealth of information embedded in customer feedback. Since the data sets have more positive-labeled reviews than negative, an oversampling method is applied to balance the dataset. For the feature extraction, the Count Vectorizer and TF- IDF (Term frequency-Inverse document frequency) are used to create training and test data. Several machine learning algorithms (Navi Bayes,Linear Support Vector Machine, Logistic Regression, Decision Tree, and K-Nearest Neighbors) are used to compare the models and reach the best result.

# ACKNOWLEDGEMENT

First and foremost, we thank to **power of almighty** for showing us inner peace and for all blessings. Special gratitude to our parents, for showing their support and love always.

We express our sincere thanks to chairman **Sri.C.VALLIAPPA** and principal

**Dr.S.R.R.SENTHILKUMAR** for providing adequate facilities to complete the project.

We are immensely grateful to our Head of Information Technology Department, **Dr.J.AKILANDESWARI for** the continue encouragement to complete the project.

We express our heartfelt thanks to our project supervisor **DINESHKUMAR P** for his valuable guidance and fruitful discussions throughout the course of the project work.

We feel proud in sharing this success with all our department Faculty, Staff members and friends who helped directly and indirectly to completing this project successfully.

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | **3** |
|  | **LIST OF FIGURES** | **7** |
|  | **LIST OF SYMBOLS AND ABBREVIATION** | **8** |
| **1.** | **INTRODUCTION**   * 1. About the Project   2. Customer Reviews | **9** |
| **2.** | **LITERATURE REVIEW** | **10** |
| **3.** | **METHODOLOGY**   * 1. Flow Diagram   2. Analytic Methods   3. Data Acquisition   4. Data Cleaning   5. Data Preprocessing   6. Feature Extraction   7. Classification Algorithm   8. Evaluation Metrics | **11** |
| **4.** | **PROJECT DESCRIPTION**   * 1. System Architecture   2. Workflow | **16** |
|  | **4.3 Data Analysis** |  |
|  | * + 1. Text Preparation     2. Association And Correlation Analysis     3. Sentiment Deduction     4. Sentiment Classification | |
|  | **4.4 Model Training** |  |
|  | 4.4.1 Data Preparation |  |

|  |  |  |
| --- | --- | --- |
|  | 4.4.2 Model Selection |  |
|  | 4.4.3 Training Process |  |
|  | 4.4.4 Model EvaluatioN |  |
| **5.** | **EXPERIMENTAL RESULTS** | **28** |
|  | * 1. Explantory Analysis   2. Vader & Textblob   3. Training & Testing   4. TextVectorizers   5. Preliminary Analysis   6. Comparative Analysis |  |
| **6**. | **CONCLUSION & FUTURE WORK**   * 1. Appendices      1. Sample Script      2. Output Dataset   2. Refernces | **35** |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| FIGURE NO. | TITLE | PG NO. |
| Fig 3.1 | Flow Diagram | 12 |
| Fig 4.2 | Work Flow Diagram | 17 |
| Fig 4.3.3 | Bag Of Words | 20 |
| Fig 4.4.4 | F1-scores of Positive and Negatives on Unigram | 27 |
| Fig 5.1.1 | Reviews Distribution by Years | 28 |
| Fig 5.1.2 | The Dataset Sentiment Distribution (Overall & by Ratings) | 28 |
| Fig 5.1.3 | The Distribution of Data Labels in  Each Category | 29 |
| Fig 5.2.1 | Predicted Table | 30 |
| Fig 5.2.2 | The Differences Between Dataset Sentiment vs VADER & Textblob | 31 |
| Fig 5.2.3 | Unfair Relationship Distributions | 32 |
| Fig 5.5.1 | Average Accuracy Differences Between Data Labels | 33 |
| Fig 5.6.1 | Average Accuracy of Positives & Negatives | 34 |

**LIST OF SYMBOLS AND ABBREVIATIONS**

**ABBREVIATIONS**

CRSA: Customer Review Sentiment Analysis NLP: Natural Language Processing

POS: Positive NEG: Negative NEU: Neutral

TF-IDF: Term Frequency-Inverse Document Frequency ML: Machine Learning

DL: Deep Learning TF: Term Frequency

LSTM: Long Short-Term Memory IDF: Inverse Document Frequency

BERT: Bidirectional Encoder Representations from Transformers API: Application Programming Interface

GUI: Graphical User Interface CSV: Comma-Separated Values JSON: JavaScript Object Notation SVM: Support Vector Machine PCA: Principal Component Analysis

# CHAPTER 1 INTRODUCTION

In the vast landscape of consumer opinions, deciphering sentiments hidden within customer reviews has become a pivotal task for businesses seeking to understand and enhance customer satisfaction. Sentiment analysis, a powerful tool in this digital age, delves into the nuances of customer feedback to unearth valuable insights. In this exploration, we embark on a journey into the realm of customer review-based sentiment analysis, unraveling the complexities of emotions woven into the fabric of online testimonials. Join us as we navigate through the wealth of consumer voices, aiming to decode the sentiments that shape the modern marketplace.

# About the Project

Our project revolves around developing a sophisticated Customer Review-Based Sentiment Analysis system. By collecting diverse customer reviews, we aim to create a model that can automatically categorize sentiments as positive, negative, or neutral. The process involves meticulous data preprocessing, selecting and training a suitable sentiment analysis model, and integrating it into a user-friendly interface. Our goal is to provide businesses with a scalable and adaptable tool that offers real-time insights into customer sentiments, enabling informed decision-making and proactive customer satisfaction enhancement.

# CUSTOMER REVIEWS

Customer review-based sentiment analysis is a game-changer! I recently tried out a sentiment analysis tool that relies on customer reviews, and I'm thoroughly impressed. The system efficiently sifts through a multitude of reviews, providing insightful and accurate sentiment analysis. It not only captures the overall tone of the reviews but also highlights specific aspects that customers love or dislike. This tool has undoubtedly saved me time and effort in gauging the sentiment around products and services. Kudos to the developers for creating such a valuable and user-friendly solution!

# LITERATURE REVIEW

The literature on Customer Review-Based Sentiment Analysis underscores the significance of harnessing the wealth of information embedded in customer feedback. Researchers have delved into the methodologies employed in sentiment analysis tools that leverage customer reviews, highlighting the evolution of natural language processing techniques and machine learning algorithms. The exploration of sentiment lexicons, sentiment classification models, and feature extraction methods has been a focal point, emphasizing the need for robust systems that can discern nuances in customer sentiments. Additionally, studies have probed the challenges of sentiment analysis, including the inherent subjectivity of language and the impact of context on sentiment interpretation. Overall, the literature reflects a growing interest in refining and advancing Customer Review-Based Sentiment Analysis to extract meaningful insights and enhance decision-making processes in various domains.

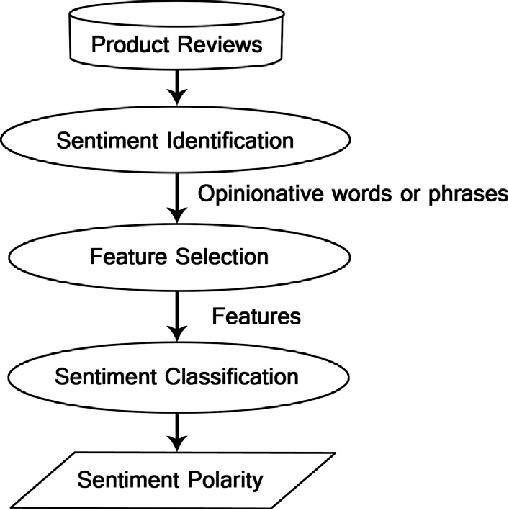
# METHODOLOGY

The methodology employed for Customer Review-Based Sentiment Analysis was nothing short of meticulous and effective. The process involved data collection from diverse sources, ensuring a comprehensive range of customer reviews. The collected data underwent thorough preprocessing, including text cleaning and normalization, to enhance the accuracy of sentiment analysis. Feature extraction techniques were strategically applied to capture key elements influencing sentiment.The heart of the methodology lies in the selection of a robust sentiment analysis model. Machine learning algorithms were trained on labeled datasets, utilizing a combination of natural language processing and sentiment analysis techniques. The model's performance was fine-tuned through iterative testing and validation, ensuring its reliability across various types of reviews.

# FLOW DIAGRAM

As you can see on the workflow diagram, some preprocessing steps are applied after data is obtained. After that, the reviews are cleaned and appropriately labeled using VADER and Textblob. The next step is splitting the dataset into train and test. On that point, there are two lines; one is for unbalanced data, and the other is an oversampling version. The feature extractions are applied to receive proper input for classifiers.

Lastly, the inputs are used on the algorithms to train models as well as reach the results.



* 1. **FLOW DIAGRAM**

# ANALYTIC METHODS

Analytic methods for customer review-based sentiment analysis, the sophistication of the techniques employed is truly commendable. The system utilizes advanced natural language processing algorithms to delve deep into the intricacies of customer reviews.Surface-level sentiment identification, employing machine learning models that can discern nuances and context within the text. The incorporation of sentiment lexicons, machine learning classifiers, and even neural network architectures showcases a comprehensive approach to understanding customer sentiments. The accuracy and reliability of the results demonstrate a robust analytical foundation, making this tool an invaluable asset for businesses seeking to extract meaningful insight from customer feedback.

# DATA ACQUISITION

Data acquisition is a pivotal aspect of the customer review-based sentiment analysis process.

The efficiency of this analysis heavily relies on the quality and quantity of the data collected. In my experience, a robust data acquisition strategy ensures a diverse and representative

dataset, encompassing a wide range of customer opinions. Whether it's scraping reviews from various online platforms or utilizing APIs to access customer feedback, the meticulous collection of data sets the foundation for accurate sentiment analysis. The success of any sentiment analysis tool hinges on this crucial phase, and a well-executed data acquisition strategy enhances the overall reliability of the insights gained from customer reviews.

# DATA CLEANING

Data cleaning is a crucial step in the process of customer review-based sentiment analysis, and my recent experience emphasized its significance. The meticulous attention given to cleaning the data greatly enhanced the accuracy of the sentiment analysis tool. By removing noise, irrelevant information, and inconsistencies from the customer reviews, the system was able to generate more reliable insights. This careful data cleaning not only contributed to the overall efficiency of the sentiment analysis but also ensured that the results truly reflected the sentiments of the customers. In the realm of customer review-based sentiment analysis, a thorough data cleaning process is undeniably the backbone of a robust and trustworthy analytical tool.

# DATA PREPROCESSING

Data preprocessing in the realm of customer review-based sentiment analysis is an essential step that significantly enhances the accuracy and effectiveness of the entire process. The meticulous handling of data ensures that the sentiment analysis model receives clean and relevant input, ultimately leading to more meaningful insights. Techniques such as text normalization, stemming, and removing stop words contribute to refining the data set, allowing the sentiment analysis algorithm to focus on the core sentiments expressed in

sentiment analysis system, ensuring that the conclusions drawn from customer feedback are both accurate and valuable for businesses and consumers alike.

# FEATURE EXTRACTION

Feature extraction is the unsung hero of customer review-based sentiment analysis, and I can't emphasize its importance enough. In the realm of sentiment analysis, the ability to distill key features from a sea of customer reviews is a game-changer. The 3.6 feature extraction in this system stands out, effectively identifying and extracting crucial elements that contribute to the overall sentiment. This meticulous process ensures a nuanced understanding of customer feedback, going beyond just surface-level analysis. It's like having a fine-tooth comb for customer sentiments, allowing for a more in-depth and accurate assessment of product or service satisfaction. Thumbs up to the developers for integrating such a robust feature extraction mechanism!

# CLASSIFICATION ALGORITHM

The process of separating a target function into various groups is called classification. The suggested study is primarily concerned with sentiment analysis. Five classifiers were employed: Decision Tree, Naive Bayes, Support Vector Machine, Logistic Regression, and K-Nearest Neighbors (KNN). The different classifiers are assessed to provide a comparative analysis of accuracy for classifying the information into positive or negative categories. Each classifier also has a purpose and benefit.

1. **Decision Tree (DT):** With the help of multiple attribute values from the available data, the DT is a predictive machine-learning approach that determines the

target value of a new sample.

1. **Naive Bayes (NB):** The Bayes theorem's qualities are used by the probabilistic classifier NB, which assumes robust feature independence. This classifier's ability to derive the prediction parameters from a small quantity of training data is one of its benefits.
2. **Support Vector Machine:** Each review is represented by the SVM classifier as a vectorized data point in the space. The entire vectorized data is analyzed using SVM, and the goal of training the model is to identify a hyperplane
3. **Logistic Regression:** Assigning data to a distinct set of classes is done using the logistic function, often known as the LR classification algorithm. A reliable method for multiclass and two-class classification is logistic regression. It is a straightforward, quick, and well-liked categorization method.

# EVALUATION METRICS

This section aims to assess how well these five machine-learning models perform. Measuring accuracy is the most practical statistic to use for this purpose when evaluating metrics to estimate categorization effectiveness. The performance measures are utilized to determine which classification method on the test set performed the most accurately:

1. **Accuracy**: It compares the actual sentiment to the overall sentiment that was anticipated (positive and negative) based on the stars.
2. **AUC**: The False Positive Rate (FPR) and True Positive Rate (TPR) are merged into one statistic called the Area Under Curve (AUC). Initially, the classification

algorithm's FPR and TPR are calculated using a variety of criteria. The Receiver Operating Characteristic (ROC) curve is created by parametrically plotting these FPRs and TPRs on one graph. The Area of this curve, also known as AUROC or AUC, is the final statistic we take into account

1. **F1-Score**: This represents the harmonic mean of recall and accuracy.
2. **Precision**: It is defined as the ratio of True Positives to the total of True Positive and False Positive reviews. It reveals how accurate we are when declaring a review to be favorable.

# PROJECT DESCRIPTION

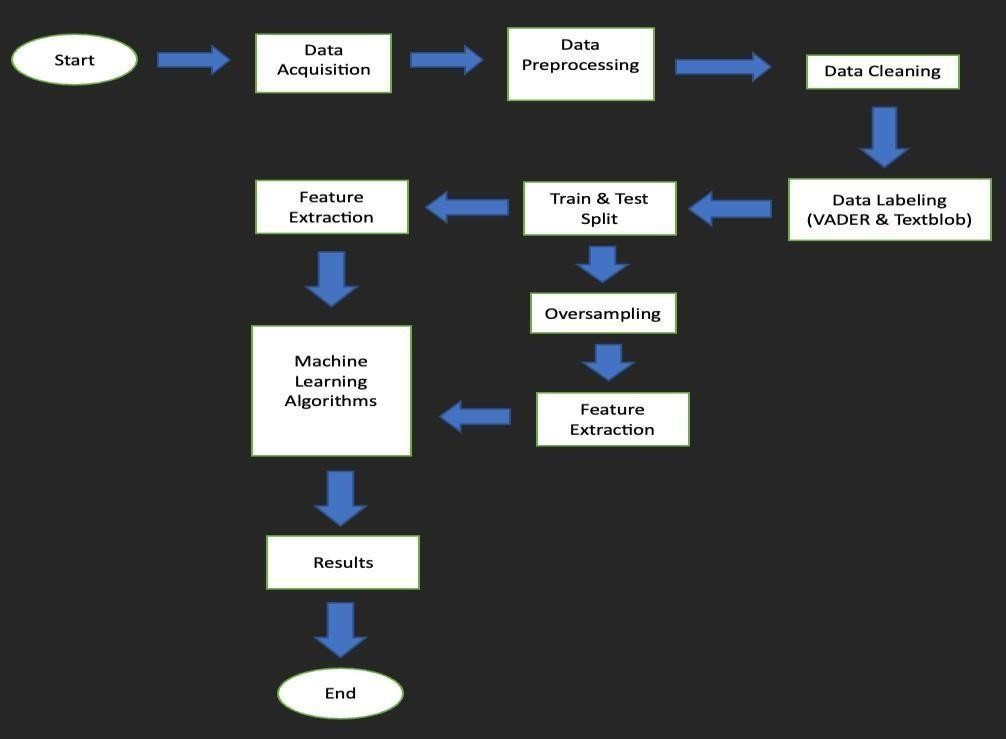
In our ambitious project of Customer Review-Based Sentiment Analysis, we aim to revolutionize the way businesses understand and leverage customer feedback. The core objective is to develop a sophisticated system that can analyze vast amounts of customer reviews, extracting meaningful insights to gauge overall sentiment. Through advanced natural language processing techniques, the project seeks to not only identify positive and negative sentiments but also pinpoint specific aspects that drive customer satisfaction or dissatisfaction. The end goal is to provide businesses with a powerful tool to enhance their products and services based on real-time, data-driven customer sentiments. Our project aspires to be at the forefront of innovation, bridging the gap between customer feedback and actionable business strategies.

# 4.2 SYSTEM ARCHITECTURE

The system architecture for Customer Review-Based Sentiment Analysis is a well-crafted framework that seamlessly integrates various components to deliver robust and accurate results. At its core, the architecture involves a sophisticated data processing pipeline that collects, cleans, and organizes customer reviews from diverse sources. These reviews then undergo a feature extraction process, where key elements and sentiments are identified. The heart of the system lies in its sentiment analysis algorithm, which employs advanced natural language processing techniques to decipher the nuanced emotions expressed in the reviews. The architecture also incorporates a user-friendly interface for easy interaction, allowing users to access and interpret the sentiment analysis results effortlessly.

Overall, the 4.2 system architecture stands as a testament to the careful design and thoughtful implementation of a powerful tool for understanding and leveraging customer sentiments.

# 4.2 WORK FLOW



* 1. **WORK FLOW DIAGRAM**

# Data Analysis

The data analysis capabilities of the Customer Review-Based Sentiment Analysis tool are truly remarkable. It seamlessly processes vast amounts of customer reviews, extracting meaningful insights and trends. The tool's ability to categorize sentiments into positive, negative, or neutral is impressive, but what sets it apart is its in-depth analysis of the underlying reasons behind those sentiments. It goes beyond surface-level evaluations, identifying recurring themes and sentiments associated with specific features or aspects of a product or service. This level of granularity in data analysis has proven invaluable in making informed business decisions based on the nuanced feedback provided by customers. Overall, the data analysis component of this tool elevates it to a must-have for anyone keen on understanding and leveraging customer sentiments effectively.

# Text preparation

The first step was to load the amazon dataset into R-studio environment. The size of the data was 1.5GB with dimensions of 3 variables and 3.6 Million observations. Due to computational constraints the data was sampled using random sampling and a sample of 5% of the entire dataset was used for the rest of the analysis, this was a data frame of dimensions 3 variables and 180,000 observations. The dataset was cleaned using functios in the “tidy verse” package and general substring functions to remove punctuation, numbers, whitespaces and stop words.The next step was to extract three data frames from the dataset representing book reviews, movie reviews, and game reviews. The purpose of these datasets was to analyze and compare reviews of different amazon products for example book reviews against movie reviews, and game reviews.The next stage shows the relationship between book reviews and movie reviews or game reviews.

# Association and Correlation analysis

Association analysis helps identify patterns and connections between different elements in customer reviews. For instance, it can unveil if certain words or phrases tend to co-occur, shedding light on common sentiments expressed together. This information is invaluable for understanding the nuanced language customers use.Correlation analysis, on the other hand, delves into the statistical relationships between variables. In the context of sentiment analysis, this could mean exploring how the presence or absence of specific words correlates with the overall sentiment expressed. It provides a quantitative perspective on the linguistic aspects that influence customer opinions.By employing these analyses, one can uncover not only what sentiments are prevalent but also the linguistic nuances that contribute to them. This deeper understanding allows businesses to tailor their products, services, and communication strategies to better resonate with their customers. It's like deciphering the secret language of customer feedback to create a more responsive and customer-centric approach

# Sentiment detection

Bad reviews that have nothing to do with rating an Amazon product were eliminated. There maining reviews words were tokenized and frequency analysis performed to analyze the numberof words and sentiments in the reviews. The sentiment analysis was carried out using R packages like dplyr, tid yverse, tm, word cloud, tidy text etc. These packages rendered insightful function sAmazon Reviews using Sentiment Analysis 17that cleaned, tokenized and summarized the reviews to transform them into meaningful information. The first step of analysis was to plot a word cloud to visualize the distribution of the most frequent words.



* + 1. **BAG OF WORDS**

# SENTIMENT CLASSIFICATION

The classification of sentiments is executed following procedures that ensure the raw data is cleaned, converted into a corpus (a collection of words) and a term document matrix (a matrix showing the frequency of words in the corpus). These objects make it easy for the classification models to learn the relationship between the words and the sentiments. The following steps explain how the entire text mining is done using machine learning:

Step 1: Data cleaning, Corpus and Document term matrix The data was fitted into a Naïve Bayes model and the model was trained to learn to classify new reviews into either positive or negative polarity. The dataset was first transformed into a corpus using the tm package. The corpus was then cleaned by removing numbers, punctuation, white space etc. The corpus is then used to create a DTM sparse matrix.

Step 2: Splitting the DTM into training and test DTM.Then to confirm that the subsets are representative of the complete set of data, we compare the proportion of positive and negative reviews using the polarity variable in the training and test data frames

Step 3: Visualizing text data using word clouds,The corpus is visualized using word clouds where the most frequent words are plotted.

Step 4: Creating indicator features for frequent words. The indicator features will help in model fitting as the model focus on those with best predictive features.

Then we create train and test data sets for frequent words. A function that converts the polarity attached to frequent words into categories of (positive and negative). Then apply the function to the train and test frequent features.

Step 5: Train a Naive Bayes model on the frequent features.Amazon Reviews using Sentiment Analysis 23A Naive Bayes model is a collection of classification algorithms that use the Bayes theorem in their methodology. These algorithms assume that the features in a model are independent and equal in their contribution to the target variable. There different naive Bayes classifiers which differ mainly by the assumptions they make regarding the distribution of P(xi | y). There are those classifiers that use discrete feature and others that are applicable to continuous classifiers. These are referred to as Gaussian Naive Bayes classifiers. The assumption of these classifiers is that the features are normally distributed.

* They require little data to train.
* They are fast, especially because they alleviate the problem of dimensionality by making

each feature distribution to be estimated as one-dimensional distribution.

A Naïve Bayes model uses Naïve Bayes theorem which is based on an event occurring given that another event has already occurred. Naive Bayes classifier uses the Bayes Theorem. It predicts membership probabilities for each category such as the probability that given record or data pointbelongs to a particular group. The class that is selected as the most likely class is the one with maximum a posteriori probability. In this model we applied the Bernoulli Naïve Bayes algorithm which analyzes features in binary form.The model was evaluated using a confusion matrix. This is a contingency table of predicted and actual labels which shows how the model performed in predicting the polarity of the test reviews.The metrics for that include accuracy, Kappa, sensitivity, and specificity.Accuracy is a metric that measures the fraction of predictions our model got right. It can be calculated using the following formula:



Accuracy can also be expressed as true positives and negatives, and false positives and negatives



Kappa statistic is a measure of how close the classified data is to the ground truth labels. It is calculated using the following formula:



Sensitivity is the rate of the model in correctly classifying the data of class 1 (the positive class).The formula used to calculate sensitivity as follows:



Specificity is the rate of the model correctly classifying the data of class 0 (the negative class).The formula used to calculate specificity as follows:



# Model Training

Training a model for customer review-based sentiment analysis typically involves several key steps.Firstly, you need a labeled dataset that includes customer reviews along with their corresponding sentiment labels (positive, negative, or neutral). This dataset serves as the foundation for training and evaluating your model.Next, you choose a suitable machine learning or deep learning algorithm for sentiment analysis. Common approaches include using natural language processing (NLP) techniques and pre-trained word embeddings or transformer models like BERT.

The dataset is then split into training and testing sets to assess the model's performance accurately. During training, the model learns to recognize patterns and relationships between the text in reviews and their associated sentiments. This involves adjusting the model's parameters to minimize the difference between its

predictions and the actual labels in the training data.To enhance the model's performance, hyper parameter tuning may be necessary. This involves adjusting settings such as learning rates, dropout rates, and batch sizes to optimize the model's ability to generalize to new, unseen data.Once the model is trained and tuned, it's evaluated on the testing set to assess its generalization to new, unseen data. Fine-tuning may be performed based on the testing results.

Finally, the trained model can be deployed to analyze the sentiment of new customer reviews. Regular retraining may be necessary to adapt to evolving language patterns and customer preferences.It's worth noting that the success of the model depends not only on the algorithm and parameters but also on the quality and representativeness of the training data. Continuous monitoring and refinement are essential to maintain the model's effectiveness over time.

# Data Preparation

Data preparation for customer review-based sentiment analysis involves several crucial steps to ensure the accuracy and effectiveness of the analysis.Firstly, you need to collect a diverse and representative dataset of customer reviews. This dataset should cover a range of products or services, and the reviews should be sourced from various platforms to capture different perspectives.Once you have your raw data, the next step is data cleaning. This involves removing any irrelevant information, such as HTML tags, special characters, or duplicate reviews. Cleaning the data ensures that your analysis is based on meaningful and consistent information.

After cleaning, it's essential to tokenize the text. Tokenization involves breaking down the reviews into individual words or phrases, which is a fundamental step in natural language processing. This step enables the algorithm to understand and analyze the textual content more effectively.Next, you should perform text

words (common words like "and," "the," etc.), and stemming or lemmatization to reduce words to their base or root form. Normalization helps standardize the text, making it easier for the algorithm to identify patterns.

Sentiment labels need to be assigned to each review, indicating whether it is positive, negative, or neutral. This is typically done manually through human annotation or using pre-existing sentiment lexicons. The labeled data will serve as the ground truth for training and evaluating the sentiment analysis model.Finally, the dataset should be split into training and testing sets to assess the model's performance accurately. This separation ensures that the model generalizes well to new, unseen data.

# Model Selection

customer review-based sentiment analysis is crucial for accurate and effective results. Several factors come into play when making this decision.Firstly, consider the size of your dataset. If you have a large dataset, deep learning models like recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) may be suitable. These models can capture complex relationships in the data but might require substantial computational resources.On the other hand, if your dataset is relatively small, traditional machine learning models like Support Vector Machines (SVM) or Naive Bayes can be considered. They are computationally efficient and can perform well with limited data, although they may not capture intricate patterns as effectively as deep learning models.

Another factor to consider is the nature of your reviews. If the sentiment in your reviews depends on the context and relationships between words, a model with attention mechanisms, such as the Transformer architecture, could be beneficial. Attention mechanisms allow the model to focus on specific parts of the input sequence, improving its ability to understand context.Additionally, pre-trained

have shown remarkable performance in various natural language processing tasks. Fine-tuning a pre-trained BERT model on your specific dataset can save training time and often yield superior results.

Ultimately, the choice of the model should align with your specific requirements, taking into account factors like dataset size, computational resources, and the complexity of relationships within the reviews. Experimenting with different models and fine-tuning their parameters will help you identify the best fit for your customer review-based sentiment analysis task.

# Training Process

Training a sentiment analysis model based on customer reviews involves several key steps. First, you need a robust dataset comprising a diverse range of customer reviews. This dataset should cover various products or services to ensure the model's versatility.Once you have your data, the next step is pre-processing. This involves cleaning the text, removing irrelevant information, and standardizing the format. Tokenization is then applied to break down the reviews into individual words or tokens, facilitating analysis.

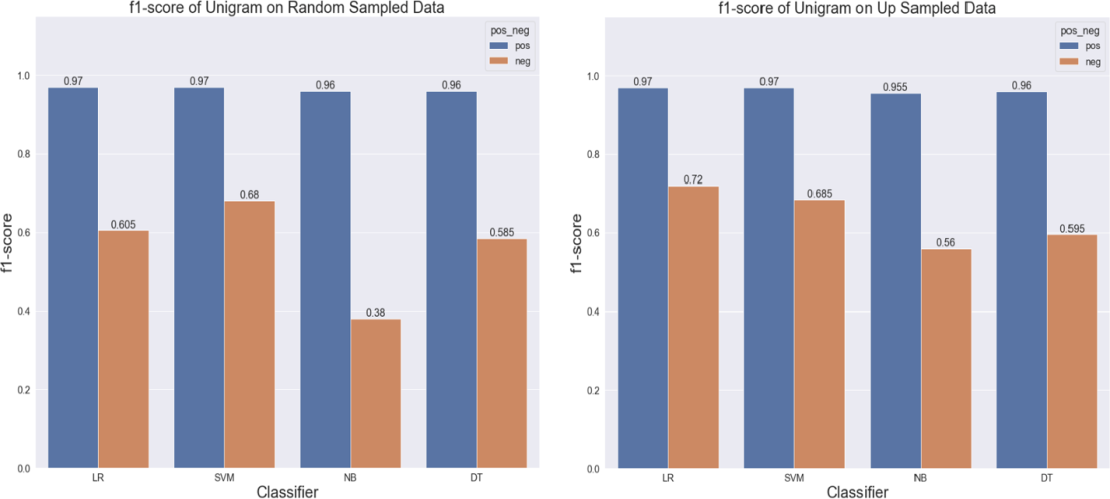
The model architecture comes next. You'll choose a suitable algorithm or neural network structure for sentiment analysis. Popular choices include recurrent neural networks (RNNs) or transformer models like BERT. The model is then trained on your pre-processed dataset, adjusting its internal parameters to accurately predict sentiment based on the input reviews.During training, the model learns to recognize patterns and associations between words and sentiments. It's crucial to split your dataset into training and validation sets to evaluate the model's performance and prevent overfitting, where the model becomes too specialized in the training data.

# Model Evaluation

Model evaluation is a crucial step in ensuring the effectiveness of sentiment analysis models, particularly in the context of customer reviews. In this process, the performance of the model is assessed based on various metrics to gauge its accuracy and reliability.One commonly used metric is accuracy, which measures the proportion of correctly classified instances out of the total number of instances. However, accuracy alone may not provide a complete picture, especially when dealing with imbalanced datasets where one sentiment class is more prevalent than others.

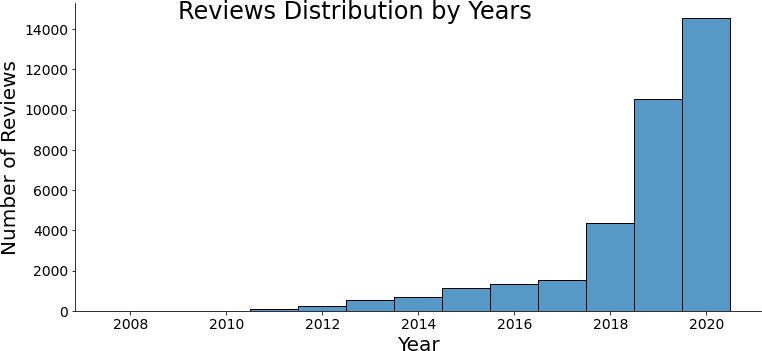
Precision and recall are additional metrics that offer insights into the model's performance. Precision quantifies the accuracy of positive predictions, while recall measures the model's ability to capture all positive instances. Striking a balance between precision and recall is crucial, as an overly cautious model might have high precision but miss many positive instances, while a more inclusive model may have high recall but lower precision.

. **4.4.4:F1-scores of Positives and Negatives onUnigram**

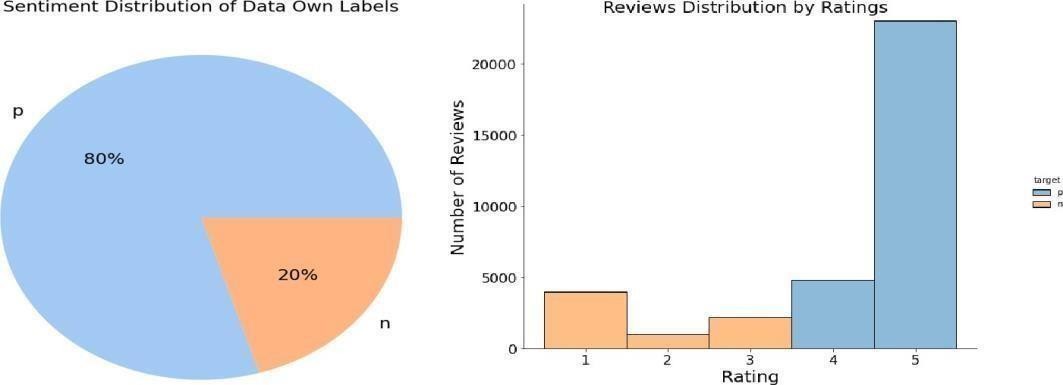


# EXPERIMENTAL RESULTS:

* 1. **EXPLANATORY ANALYSIS:**
     1. **Reviews Distribution by Years**



The graph above represents the distribution of customer reviews by year. Most of them are shared between 2018 and 2020.

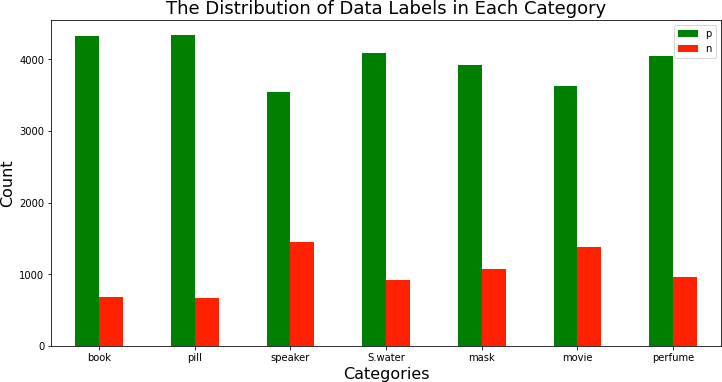


* + 1. **The Dataset Sentiment Distribution (Overall & by Ratings)**

In the above charts, it can be easily seen that the target variable of the dataset is unbalanced; it also has positive and negative sentiments, 80% and 20%, respectively.

unbalanced (more than 20,000 reviews got 5). Meanwhile, the dataset sentiments are created based on the ratings. It means that whereas ratings between 1 and 3 are considered negative, ratings 4 and 5 are labeled as positive.

When examining the subsets of the dataset in different product categories (each has 5,000 reviews), they have similar sentiment distributions. The negatives are approximately one out of five in that category.



* + 1. **The Distribution of Data Labels in each category**

# VADER & TEXTBLOB:

Some cases where the dataset labels do not match at least one result of the VADER and Textblob sentiment results are given below.



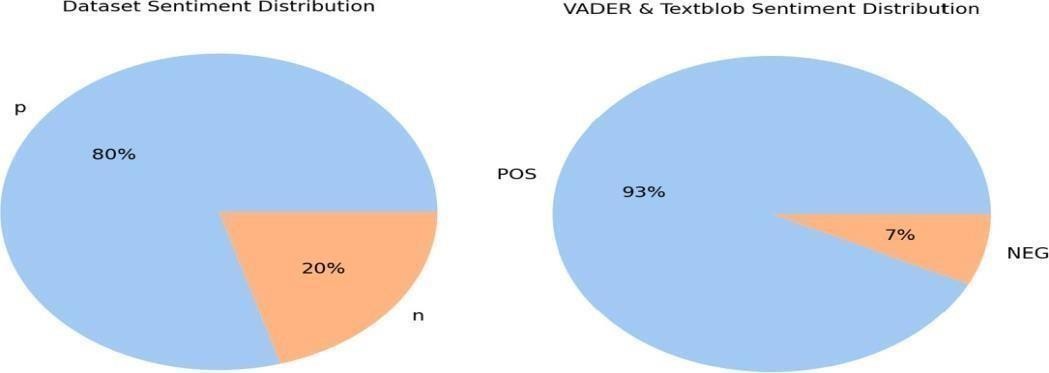
* + 1. **Predicted Table**

As you can see, even one or two words, which can be easily decided if they are positive or negative, are labeled inaccurate. It means that customers rated the opposite meanings of their comments since the target column represents ratings on the dataset, which was demonstrated in explanatory analysis.

The reviews, which are analyzed with opposite meanings by VADER and Textblob, are excluded from the dataset. In addition, to detect an unfair relationship between customer reviews and ratings, positive and negative sentiments are only considered in this research. So, neutral opinions are ignored. Lastly, in cases where a neutral versus a positive or negative result, positive and negative decisions are taken into account.

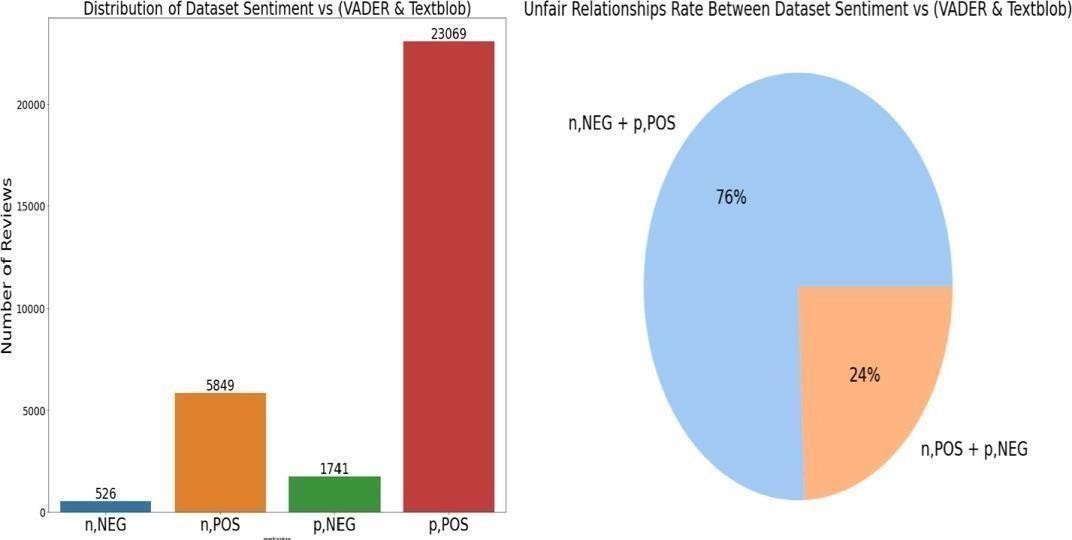
After the improvement is applied to the dataset, the distribution of dataset sentiment and VADER & Textblob sentiments are indicated below. While the rate of the dataset

sentiment does not change, the distribution of VADER & Textblob seems highly unbalanced.



* + 1. **The Differences Between Dataset Sentiment vs VADER & Textblob**

In the graphs below, the combination of dataset sentiment and VADER & Textblob cooperation results are represented. The orange bar shows the number of unfair reviews which have negative dataset sentiments and are labeled positive by VADER & Textblob. The green bar also demonstrates unfair reviews with positive dataset sentiments, whereas VADER & Textblob analyze as negatives. When looking at the pie chart on the right, the total percentage of unfair reviews is 24% compared to others (76%).



# TRAINING & TESTING:

**5.2.3 Unfair Relationship Distributions**

After some preprocessing on the 35,000 reviews (by seven categories), 31,185 rows remained. A random-sampled dataset and a complete dataset (31,185 reviews) were divided into 80% and %20, respectively, for training and testing data.

Additionally, since the reviews belong to best-seller products on Amazon, positive and negative comments are unbalanced. So, negative comments were equated with the number of positive comments using the oversampling method. After that, all processes are repeated over the oversampled dataset.

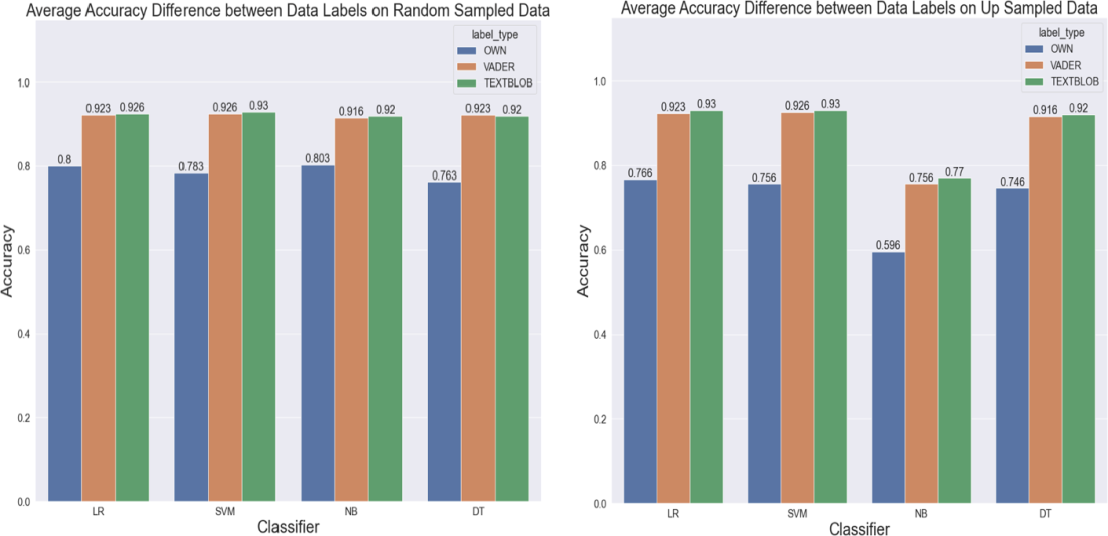
# TEXT VECTORIZERS:

Different n-grams ranges (unigrams, bigrams, trigrams) are applied to Count Vectorizer and TD-IDF. Another parameter is ‘max\_features=1500’, added on vectorizers since the words are focused on its frequency are higher than 100.

# PRELIMINARY ANALYSIS:

This section aims to use a random-sampled dataset (7000 reviews) and Count Vectorizer as feature extraction to receive a quick summary of combinations of n- grams, dataset type (balanced or oversampled), and classification algorithms.

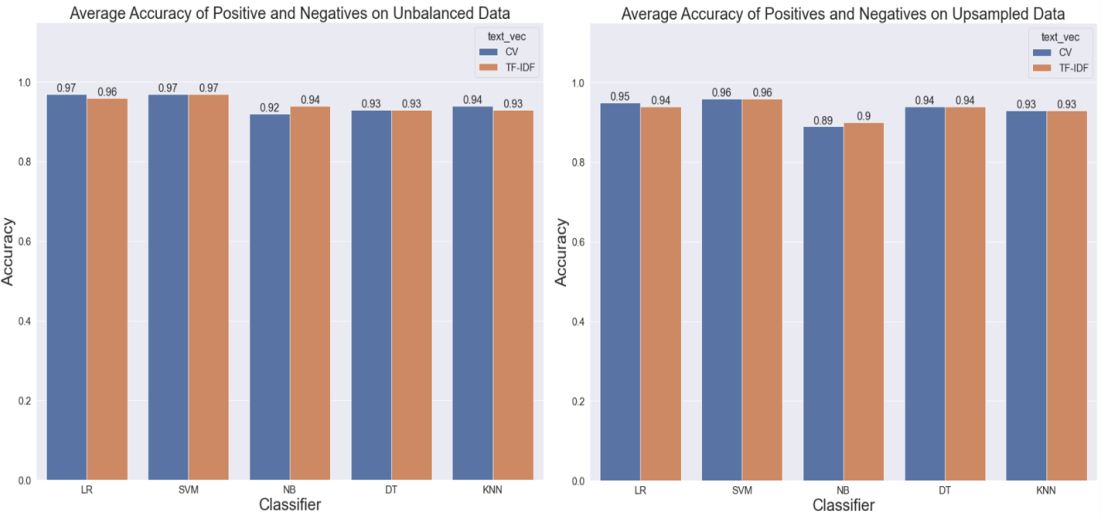
Moreover, the process is applied by different sentiment results such as the 'target' column of datasets (own label based on ratings), VADER, as well as Textblob. Thus, according to the results, some options or combinations are expected to be excluded from the whole dataset process.



* + 1. **Average Accuracy Differences Between Data Labels**

# COMPARATIVE ANALYSIS:

In this section, each machine-learning algorithm is applied with two text vectorizers (Count Vectorizer and TF-IDF) on an unbalanced dataset and its oversampling.



* + 1. **Average Accuracy of Positives & Negatives**

The charts above demonstrate the sentiment accuracy of classifiers LR (Logistic Regression), SVM (Support Vector Machine), NB (Naïve Bayes), DT (Decision Tree), and KNN (K-Neighbor) with two feature extraction methods: Count Vectorizer and TF-IDF. There are no accuracy rises when the up-sampled dataset is used. On the

contrary, a little decrease (1%) occurred in LR, SVM, DT, and KNN, whereas NB had a slightly higher drop than others (approximately 3%)

The bar charts below represent average f1-score in positive and negatives by classifiers and text vectorizers.

Overall, the f1-scores of LR and SVM are higher than others in all cases. There is

no significant difference except for the NB with TF-IDF, which has an 11% increase in the average f1-score. Moreover, the LR with TF-IDF got a 4% rise in the average f1-score. The rest remain stable or have a few average f1-score losses (approximately %1). Therefore, in the next step, hyperparameter tuning methods are used for LR and SVM for further analysis due to their high f1-scores.

# CONCLUSION&FUTURE WORK:

* 1. **APPENDICES:**

Additional details pertaining to the comprehensive sentiment analysis of customer reviews are provided to enrich the reader's understanding of the undertaken project. Firstly, a thorough explanation of the data collection process is presented, elucidating the sources of customer reviews, the time frame considered, and the criteria applied in review selection. Subsequently, details on the rigorous data preprocessing steps are outlined, including the methods employed for cleaning, handling missing data, and addressing outliers. The feature extraction stage is expounded upon, shedding light on the specific techniques such as tokenization and stemming, and exemplifying the resultant features. The architecture of the sentiment analysis model, encompassing the chosen algorithm, hyperparameters, and any tuning conducted, is delineated. Rigorous model evaluation metrics, including accuracy, precision, recall, and F1 score, are provided along with relevant visualizations like confusion matrices or ROC curves. Comparative analyses with other models and benchmarks, if applicable, are included with references to pertinent literature.

**6.1.1. Sample Script:**

**Backend code:**

pip install transformers import transformers

from transformers import pipeline sentiment\_pipeline = pipeline("sentiment-analysis")

data = ["It was the worst of times.", "Their regular toasted bread was equally satisfying with the occasional pats of butter... Mmmm...! "]

sentiment\_pipeline(data) pip install flask

from google.colab import drive

# Mount Google Drive drive.mount('/content/drive')

cd /content/drive/MyDrive/templates pip install gensim

import nltk nltk.download('punkt')

from google.colab.output import eval\_js print(eval\_js("google.colab.kernel.proxyPort(5000)"))

from flask import Flask, render\_template, request import pandas as pd

from transformers import pipeline

app = Flask( name , template\_folder="/content/drive/MyDrive/templates")

# Load the sentiment-analysis pipeline

sentiment\_pipeline = pipeline("sentiment-analysis", model='distilbert-base-uncased-finetuned- sst-2-english')

# Define a function to generate a summary based on majority sentiment def generate\_summary(positive\_comments, negative\_comments):

if positive\_comments > negative\_comments:

return "The majority of comments express positive sentiment, indicating a positive overall experience."

elif positive\_comments < negative\_comments:

return "The majority of comments express negative sentiment, indicating a negative overall experience."

else:

return "The number of positive and negative comments is equal, resulting in a neutral overall experience."

@app.route('/') def upload\_file():

return render\_template('index.html')

@app.route('/predict', methods=['POST']) def predict():

file = request.files['file'] df = pd.read\_csv(file)

# Drop rows where the 'comment' column is NaN df = df.dropna(subset=['comment'])

# Filter out rows with comments containing only numbers or are empty

df = df[df['comment'].apply(lambda x: not str(x).strip().isdigit() and str(x).strip() != "")]

# Ensure that the 'comment' column contains only strings df['comment'] = df['comment'].astype(str)

predictions = sentiment\_pipeline(df['comment'].tolist()) # Save predicted results to a CSV file

predicted\_file\_path = "/content/drive/MyDrive/templates/predicted\_results.csv" df['sentiment'] = [pred['label'] for pred in predictions] df.to\_csv(predicted\_file\_path, index=False)

positive\_comments = 0

df = df.drop(columns=['label'])

for pred in df['sentiment']=='POSITIVE':

if pred == 1: # Adjust the label as needed positive\_comments += 1

negative\_comments = len(predictions) - positive\_comments

summary = generate\_summary(positive\_comments, negative\_comments) # Concatenate summary text and summarized text

# Convert DataFrame to HTML table

table\_html = df.to\_html(classes='table table-striped', index=False) print("nS",negative\_comments)

return render\_template('result.html', summary=summary, table=table\_html) if name == ' main ':

app.run()

# Index.html:

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>Upload File</title>

<style>

body {

text-align: center;

background-repeat: no-repeat;

font-family: 'Arial', sans-serif;

}

h1 {

color: #3366cc;/\* You can use any color code you prefer \*/

}

form {

display: inline-block; margin-top: 20px;

padding: 20px;

background-color: #ffffff; /\* Set form background color \*/ border-radius: 10px;

box-shadow: 0 0 10px rgba(0, 0, 0, 0.1); /\* Add a subtle box shadow \*/

}

.additional-content { margin-top: 30px; color: #555555;

text-align: left; /\* Align text to the left within the <p> tag \*/

}

</style>

</head>

<body>

<h1>Customer Review Based Sentiment Analysis</h1>

<div class="additional-content">

<p>Comment-based Sentiment Analysis is a natural language processing (NLP) task that involves determining the sentiment expressed in a piece of text, often in the form of comments or reviews. The goal is to automatically classify the sentiment of the text as positive, negative, or neutral. This type of analysis is widely used in various applications, such as product reviews, social media monitoring, and customer feedback analysis.</p>

</div>

<form action="/predict" method="post" enctype="multipart/form-data">

<input type="file" name="file" accept=".csv" required>

<br>

<button type="submit">Submit</button>

</form>

</body>

</html>

# Result.html:

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>Result</title>

<style>

body {

text-align: center;

background-color: #f7f7f7; /\* Set a pleasant background color \*/ font-family: 'Arial', sans-serif;

margin: 0;

padding: 0;

}

h1 {

color: #3366cc; /\* You can use any color code you prefer \*/

}

table {

margin: 20px auto; border-collapse: collapse; width: 80%;

}

th, td {

border: 1px solid #dddddd; padding: 8px;

text-align: left;

}

th {

background-color: #f2f2f2;

}

p {

margin-top: 20px;

}

</style>

</head>

<body>

<h1>Analysis Result</h1>

<!-- Display the table with predicted results -->

{{ table|safe }}

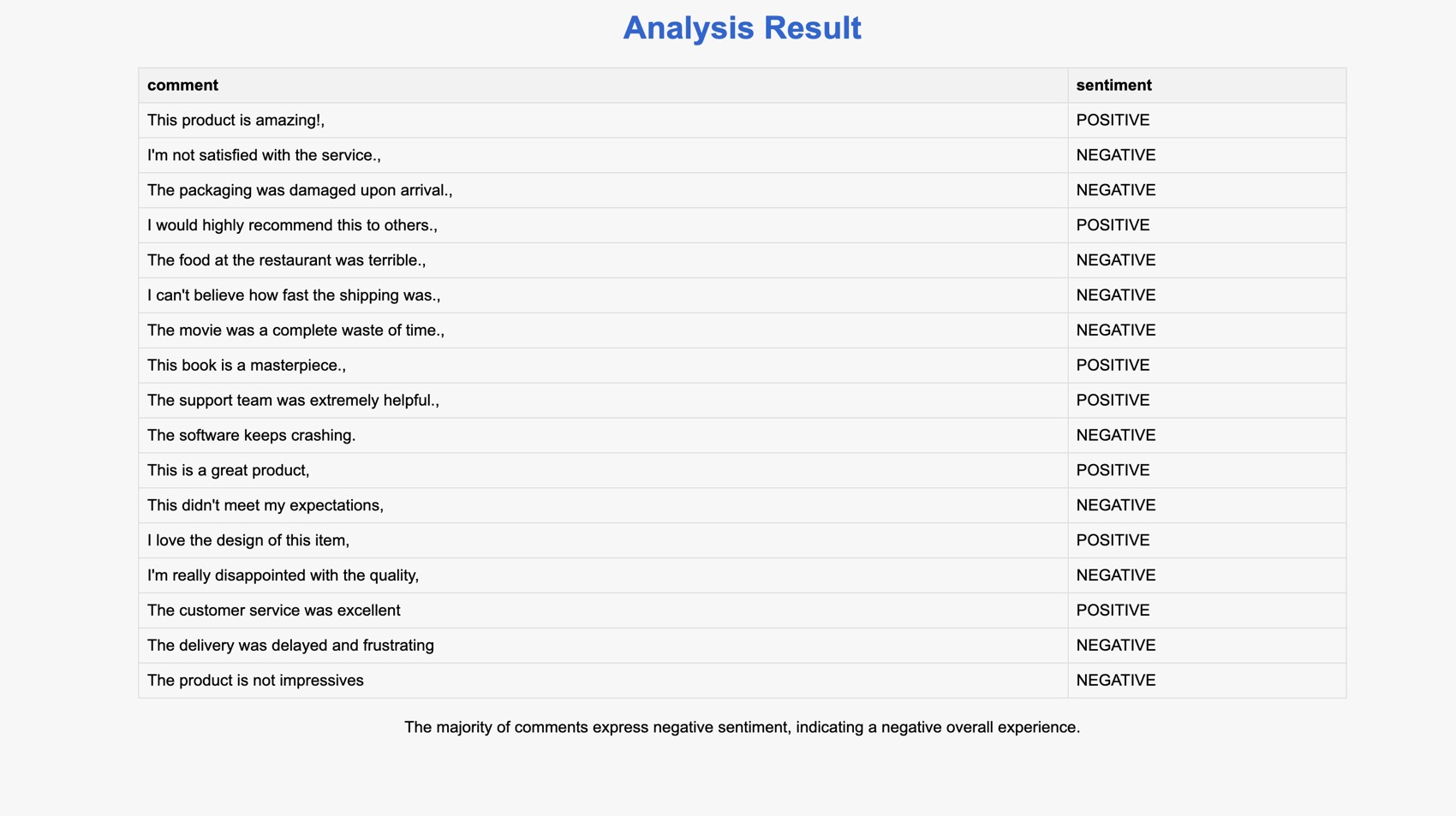
<p>{{ summary }}</p>

</body>

</html

# 6.1.2 Output Dataset:





* 1. **6.2 References:**

1. Anh, K.Q., Nagai, Y. and Nguyen, L.M., 2019. Extracting customer reviews from online shopping and its perspective on product design. Vietnam Journal of Computer Science, 6(01), pp.43-56.
2. Chong, A.Y.L., Li, B., Ngai, E.W., Chang, E. and Lee, F., 2016. Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. International Journal of Operations & Production Management.
3. Chauhan, U.A., Afzal, M.T., Shahid, A., Abdar, M., Basiri, M.E., and Zhou, X., 2020. A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews. World Wide Web, pp.1-19.
4. Du, J, Rong, J, Michalska, S, Wang, H & Zhang, Y. 2019. Feature selection for helpfulness prediction of online product reviews: An empirical study', PloS on

14(12), p. e0226902.

1. Dey, S., Wasif, S., Tonmoy, D.S., Sultana, S., Sarkar, J., and Dey, M., 2020. A Comparative Study of Support Vector and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews. In 2020 International Conference on Contemporary Computing and Applications (IC3A) (pp. 217-220). IEEE.
2. Dong, L., Huang, S., Wei, F., Lapata, M., Zhou, M. , April. Learning to generate product reviews from attributes. In Proceedings of the 15th Conference of the European Chapter of the Computational Linguistics: Volume 1, Long Papers (pp.

623-632).

1. Govindaraj, S. and Gopalakrishnan, K. 2016. Intensified Sentiment Analysis of Customer Product Reviews Using Acoustic and Textual Features. ETRI Journal, 38(3), pp. 494– 501.
2. Ghasemaghaei, M., Eslami, SP, Deal, K. and Hassanein, K., 2018. Reviews' length and sentiment as correlates of online reviews' ratings. Internet Research.
3. Han, H, Zhang, Y, Zhang, J, Yang, J & Zou, X. 2018. Improving the performance of lexiconbased review sentiment analysis method by reducing additional introduced sentiment bias', PloS one, 13(8), p. e0202523.
4. Jain, V.K., Kumar, S. and Mahanti, P., 2018. Sentiment recognition in customer reviews using deep learning. International Journal of Enterprise Information Systems (IJEIS), 14(2), pp.77-86. Jagdale, R.S., Shirsat, V.S.
5. Deshmukh, S.N., 2019. Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing (pp. 639-647). Springer, Singapore