A sentiment analysis of customer product reviews using machine learning

Student Name:

Student ID:

Table of Contents

Table of Figure

Chapter 1: INTRODUCTION AND OVERVIEW

1.1 Research Background

It is known that commercial online platform is associated with the trading of products with the help of e-commerce applications and websites. People are more inclined towards digitalization as the world becomes digitalized in the selling of products. In order to buy a product's customer, go through the range of reviews that are provided by previous customers. According to a study, there are 88% of reviews on products of amazon are genuine. As the customer of online products is increasing, the reviews are also increasing. For performing better sales and marketing there is a requirement of understanding customer needs that can be developed by machine learning algorithms. It is known that machine learning works for the efficient and smoother workflow of complex tasks. Sentiment analysis is an approach of natural language processing (NLP) that identifies the emotions of the customer towards a product and service of the company. There is a range of machine learning tasks included in performing the research such as data mining, data visualization, text mining, and algorithms (Haque et al., 2018, pp. 1-6).

Sentiment analysis is helpful for the design of strategy with the help of insights that are generated by online sources. The data that is generated by the online resources is present in an unstructured format. The unstructured data is generated by tickets, blogs, emails, social media, web chats, social media comments, and customer surfing. The research on sentiment analysis for the product review includes a study on NLP, text analysis, and computation. In order to perform the product review, there is a need for the extraction of evaluations, attitudes, judgments, opinions, and thoughts. Sentiment analysis methods can be distinguished into two types which include the machine-learning-based and lexicon-based methods. The paper includes the algorithms of machine learning

for performing sentiment analysis. The sensitive analysis is classified into three parts which are negative, positive, and neutral. Mainly, sensitive analysis is performed on the textual data. The reason behind the sensitivity analysis is to understand the behaviour of customers for changing the strategy of the product. The dataset is generated as the part of web crawling of the Flipkart website. The dataset contains the text and numeric data that will be classified with the help of the naïve Bayes algorithm. A web crawler is defined as SpiderBot or spider that is mainly used for browsing the data from the world wide web. The research is performed on the basis of the research onion framework that consists of different sections. Sentiment analysis is useful for companies for establishing communication. Artificial neural networks play a vital role in sentiment analysis. Data which is used in sentiment analysis is present in the unstructured format. Apart from Machine learning, there is a requirement for deep learning (Onan, 2021, p.e5909).

1.2 Research Problem

As the world moves towards digitalization thus it is necessary that companies should incorporate advanced technology in their marketing and sales processes. Sentiment analysis can be beneficial for Flipkart to increase its business area and sales. For increasing the number of sales then there is the requirement to identify the circumstances and conditions in which consumer is motivated for buying the product. It is also necessary that Flipkart should focus on its service thus engagement of the customer can be established with the platform. To analyze the preference of customers on an eCommerce website there is a requirement for analysis of previous activity which is known as sentiment analysis. To perform the proper analysis there is a requirement of proper selection of an algorithm otherwise it leads to inaccuracy in a mathematical expression. The research incorporated various mathematical computations for performing the machine learning analysis.

1.3 Research Scope

The research covers a range of areas such as the Flipkart business process, sentiment analysis, Artificial intelligence (AI), machine learning, and deep learning uses. Machine learning and deep learning are integrated with the range of potentials that help the industry in various terms. Social media provides wider reachability to companies. The research is able to provide an overview of the use of sentiment analysis in the eCommerce industry. The research can be used for monitoring product and brand sentiment on the basis of feedback from customers. It is also able to develop an understanding of customer needs with the help of advanced technologies. The research considers the development of a model on the basis of extracting the information from the webpage thus it can be stated that research is able to provide an overview of web crawling and web scraping. Machine learning covers a range of areas such as statistics, computer programming, and data evaluation. So, this research is able to provide knowledge about these areas. The system with sentiment analysis on the basis of the data enhances the decision-making process of the organization. This research is able to contribute to the new entrants of the market because they are able to understand the market demands.

1.4 Research Justification

AI application enables a range of business areas that can be used for business growth. Machine learning is associated with the range of algorithms that can be used for the performance of various tasks in the industry. Nowadays, sentiment analysis is an important aspect of the development of a strategy for an e-commerce company. The research is performed on the basis of analysis of a dataset that is collected by the Kaggle. Machine learning is the best method for the development of research in order to analyse product reviews. The dataset contains a range of entries thus there

is a requirement for advanced technology that can be performed efficiently which is the reason behind the selection of machine learning algorithms for sentiment analysis in research. The main issue that is faced during the study is the selection of appropriate algorithms on the basis of the characteristics of the dataset. If a wrong algorithm is chosen for the dataset then it will lead to inaccuracy with the machine learning model thus there is a requirement for a proper understanding of the dataset. In order to perform the data operation there is a need for proper cleaning of data so appropriate outcomes can be achieved. The research is associated with the new analysis therefore it can be used for further research. It will work as the series for the new development of analysis in the eCommerce segment.

1.5 Aims and objectives

The main aim of the research is:

To perform and analyse the sentiment analysis on the basis of Flipkart product review using machine learning.

Objective

- To identify the algorithm of machine learning that is used for sentiment analysis on the basis of the dataset.
- To illustrate the mathematical expression for sentiment analysis.
- To develop the sentiment analysis using a machine learning algorithm for Flipkart product review.

1.6 Research Questions

Q1 How sentiment analysis works for analysing the product review of an e-commerce company?

Q2 How does a machine learning algorithm work on an available dataset for sentiment analysis?

Q3 What is the procedure for illustrating the exploratory and predictive data analysis?

Q4 How does mathematical expression work for sentiment analysis?

Q5 What is the procedure for the development of a machine-learning model for sentiment analysis?

1.7 Rationale

The research incorporates machine learning algorithms for sentiment analysis for a product review of Flipkart. There is a range of customer who provides review on a single product. If the product range increases then it will become difficult to analyse the review of the product by a human being. For reducing the complexity of the task there is a need for advanced technologies that can be performed a complex task in an efficient way and reduces human involvement. The research consists of data from previous research. For the purpose of the sentiment analysis, the Flipkart product data was already taken by the Flipkart website as a part of web crawling. The research includes machine learning that involves a range of algorithms which is an important part of the research. There is a range of research that takes place for performing the sentiment analysis of various products. This research will provide an overview of those research in the literature review section. The research will be distinguished into the three domains such as positive, negative, and neutral that helps to understand the demands of the customers. It also enables the exact review of the customer apart from the classification.

Chapter 2: Literature Review

2.1 Sentiment Analysis

According to Dang et al. (2020, p.483), Sentiment analysis is defined as a process that is used for the extraction of information from an entity. Sentiment analysis automatically identifies the entity on the basis of the objective. The main aim of the sentiment analysis is to determine the sentiment of users by differentiating their views into three parts named Negative, positive, and neutral opinions. There is three level of extraction that takes place as part of sentiment analysis which is listed below.

1. Feature-level extraction
2. Sentence-level extraction
3. Document-level extraction

Yang et al. (2020, pp.23522-23530), defined sentiment analysis as the collection of a number of reviews by the user on online shopping platforms. As a result of rapid development, the number of users increases the interest in eCommerce platforms. There are two machine learning methods that can be used for sentiment analysis which are the classification by SVM and other ne is feature extraction on the basis of the Gini index. The study also defines the SJSM (Supervised Joint emotional model), It is used for the identification of sentiments on the basis of data that is generated by the comments. There are four machine learning algorithms are used for the development of SJSM. These algorithms are the J48, naïve bayes, OneR, and BFTree. There are three approaches are used for sentiment analysis problems which are described below.

2.1.1 Lexicon Based technique

It is the first method that is used for sentiment analysis. It is distinguished into two parts named corpus-based and dictionary-based methods. To perform the dictionary-based sentiment analysis there is a requirement for the classification of sentiments on the basis of terms of the dictionary. Whereas corpus-based sentiment analysis does not have a dependency on the predefined dictionary. It uses sentiment analysis for performing the content of documents with the help of KNN (K-nearest neighbor), HMM (hidden Markov models), and CRM (conditional random field) techniques.

2.1.2 Machine learning-based techniques

There are two types of machine learning methods used for sentiment analysis which are the traditional method or model and deep-learning-based models. The traditional methods of machine learning consist of a range of techniques such as the SVM (support vector machine), maximum entropy classifier, and naïve Bayes classifier. To perform the task with these algorithms there is a requirement the process the input with the help of the sentiment lexicon-based features, lexical features, adjectives, adverbs, and parts of speech. The accuracy of the traditional method has a dependency on these features. Whereas the deep learning model provides better accuracy than the traditional model. There is a range of models are used in deep learning models such as the RNN, DNN, and CNN. These techniques are useful for performing the classification on the basis of the aspect level, sentence level, and document level.

Yang et al. (2020, pp.23522-23530) define a deep-learning-based model that consists of the co-occurrence features in statistical form and semantic features. The author defines the use of the n-gram feature in producing the desired output with the help of CNN. The model consists of the analysis of semantic polarity with the tweets data. The name of this model is a Target-dependent

convolutional neural network. This model works for defining the relationship between the surrounding words and target words. The mechanism uses the LSTM network to perform the analysis.

The expression for sentiment analysis:

$$\text{senti}\,(w_i) = \begin{cases} sw_i, & w_i \in SD \\ 1, & w_i \notin SD \end{cases}$$

2.1.3 Hybrid approach

The hybrid approach is more useful because it integrates both machine learning and lexicon-based approaches. The sentiment lexicon plays a vital role in identifying the strategies. There are different tasks that can be performed with this approach. These tasks may be subjective or objective classification, aspect or feature-based sentiment analysis, and polarity sentiment detection. Aspects-based sentiment analysis refers to the analysis that contains the various entities like service, cleanliness, location, and value. The subjectivity of the phrases and words has a dependency on the object and context document by containing the subjective sentences. There are two components are considered for the sentiment analysis which is the intensity and polarity with the help of the score sentiment analysis. Intensity is used as an indicator that is helpful for identifying the strength of sentiment whereas polarity works for defining the sentiment type as positive, negative, or neutral.

2.2 Procedure for ML algorithms

As per the view of Haque et al. (2018, pp. 1-6), Sentiment analysis is also known as opinion mining that uses machine learning algorithms for analysing reviews on products. The business model

works for the detection of the emotion of customers on the basis of their information such as their names, gender, and reviews, this model also enables the detection of fake reviews. To perform the machine learning algorithm there is a requirement for the scripting language. The most commonly used language for machine learning algorithms is python and R. To perform the sentiment analysis there is a need for classifiers because there is a need for classification among the words, phrases, and documents. The commonly used classifiers are the support vector machine and Multinomial naïve Bayesian. The author uses the supervised learning algorithms for the prediction of product rating on the basis of the given reviews and uses the numerical scale for the text classification. To perform the machine learning algorithm, data is split into two parts. The data is split on the basis of cross-validation which uses 70% of data for training and other data for testing. The research also contains natural language processing for performing sentiment analysis.

2.2.1 Data Acquisition

It refers to the first step of machine learning model building. Ecommerce platforms have a large number of reviews so manual labelling for the data cannot be performed. To perform the labelling on the dataset there is a requirement for an active learner for the label. Active learning is part of semi-supervised ML algorithms. Active learning reduces the time complexity of the task with the help of pull-based active learning. To perform the labelling there is a need for some part of data for the learning purpose. After that algorithm learns itself for labelling of the data. The accuracy of data is calculated on the basis of ratio. If the probability of the data is greater than 90% then it is considered that the data is labelled properly. After that pre-labelled and labelled data is combined for further ML steps.

2.2.2 Data Pre-Processing

Tokenization: It separates the sequence of phrases or strings in the individual words, keywords, and symbols. These elements are known as the token, in the process punctuation is used and marked as discarded. The token is considered as the input for the text mining process.

Removal of stop words: Stop words refers to the objects of the sentence that does not impact the analysis. This is the reason behind the removal of this type of word. It also leads to increasing the accuracy and model performance.

POS tagging: This process contains the part of speech and assigns them tagging. There are different grammar aspects are considered such as verbs, nouns, adjectives, adverbs, conjunction, and pronoun.

2.2.3 Feature extraction

Bag of words: It is a process of feature extraction by presenting simplified data or text. It is mainly used in information retrieval and natural language processing. In this method, a document or text is defined as a bag of numerous sets of words. After that sentiment analysis contains the useful words in it. After the processing of the bag, POS tagging is applied to it and separates the bag on the basis of the adjectives and nouns.

Chi-Square: Chi-square is a mathematical method that is used for identifying errors in model building. It determines the difference between the expected data and observed data. There is a need for a pipeline method for performing the chi-square task. After the pre-processing of data, the dataset is divided into two parts which are the testing and training datasets (Bahassine et al., 2020, pp.225-231).

2.2.4 Evaluation

For the evaluation of the classification model, there is a need for various parameters. Accuracy is one of the most among them. Accuracy is defined as the percentage of correct observations among the selected observations for the training of the model. The evaluation of the model is performed on the basis of the evaluation matrix that contains the measures (Kou et al., 2020, p.105836).

- **True positive (TP)**
- **False positive (FP)**
- **False negative (FN)**
- **True negative (TN)**

1. **Precision:** It is the measure of classifiers for identifying the number of correct documents. Higher precision leads to less number of false positives.

$$P = \frac{TP}{TP+FP}$$

2. **Recall:** It is used for the calculate the sensitivity of the classifier, it is the measure of the output of positive data. Higher recall refers to fewer false negatives.

$$R = \frac{TP}{TP+FN}$$

3. **F-Measure:** It combines the recall and precision that refers to the single metrics. The weighted harmonic mean for define as follows.

$$F = \frac{2P.R}{P+R}$$

4. **Accuracy:** Accuracy is the measure for the prediction of the correct decision of the classifier.

$$Accuracy = \frac{Correct\ Prediction}{Total\ data\ points}$$

2.3 EDA (Exploratory Data Analysis) and Predictive data analysis Using ML

2.3.1 Exploratory data analysis

EDA is used for the investigation and analysis of datasets for identifying insights from them. It is helpful for discovering patterns, testing hypotheses, spotting anomalies, and checking assumptions. It also works for establishing the relation between the different dataset variables (Saldaña 2018). EDA is associated with statistical computation and it is widely used for data discovery. EDA makes the assumptions on the basis of the dataset and identifies the pattern and error from the data by the detection of outliers. The main aim of the EDA for business is to achieve the desired outcome for the business decision-making process. It uses various mathematical computations and parameters such the categorical variables, standard deviations, and confidence intervals. EDS can be distinguished into four primary types which are described below.

- **Univariate non-graphical:** It consists of only one variable for data analysis. It is not inclined with the relationships and is caused because of a single variable. It is mainly used for the identification of patterns.

- **Univariate graphical:** It is a graphical method of data analysis that includes a range of graphs such as Histograms, box plots, and stem-and-leaf plots.
- **Multivariate nongraphical:** It contains more than one variable and it is mainly used for establishing the relationship among them.

Multivariate graphical: It displays the information in a graphical form. Bar chart and grouped bar plot is the most commonly used method of representation. Multivariate graphics uses a range of plots such as the run chart, scatter plot, multivariate chart, heat map, and bubble chart (IBM, 2020).

According to AlQahtani (2021), EDA is a process of interpretation and visualization of information on the basis of the hidden columns and row formats. It uses a number of charts for maximizing the insights from the available dataset. Python and R is the most useful tool for performing data analysis. Data analysis contains a range of tasks with the above-mentioned task such as data visualization, data cleaning, data pre-processing, and data modelling with the help of machine learning algorithms. The author performs the EDA on amazon products for the process of sentiment analysis. They examine the review of mobile phones, the author presents the data in a graphical format. The visualization helps with sentiment analysis by the development of understanding of data.

2.3.2 Predictive data analysis

From the perspective of Tuladhar et al. (2018, pp. 279-289), Predictive analysis is defined as an advanced technique of analysis that is used for the prediction of future events by analyzing historic data. The predictive analysis contains a range of techniques such as statistical algorithms, machine learning algorithms, data mining, and artificial intelligence. For performing the sentiment analysis

by considering the sentiment of the customer there is the requirement of some algorithms that are described below.

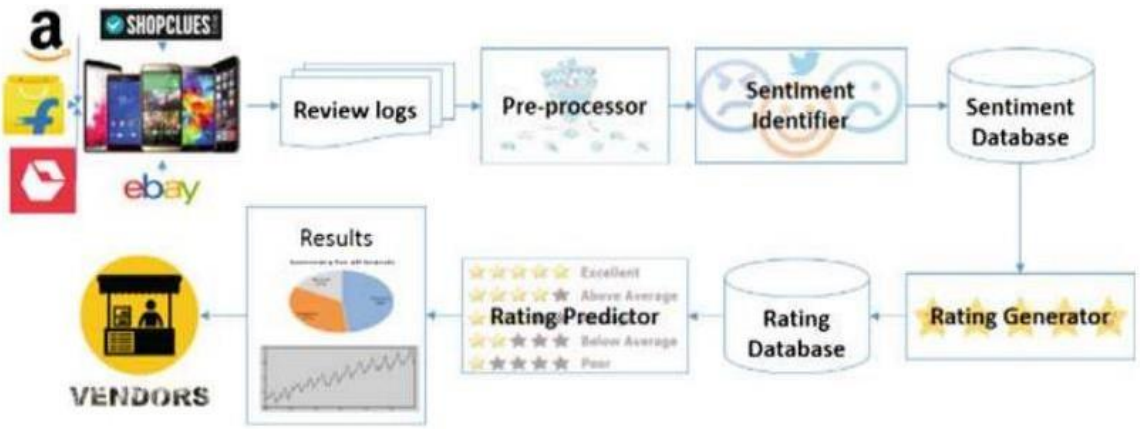

Figure 1: Steps for predictive analysis

(Source: Tuladhar et al., 2018, pp. 279-289)

The first step refers to the collection of data which is collected with the help of the python-based web scrapper. The predictive analysis mainly contains four steps which are the pre-processor, sentiment analysis, rating generation, and rating prediction. Pre-processor contains the concept of part of speech and exhibits the reviews on the basis of parts of speech. The sentiment analysis provides the review in three terms which is negative, positive, and neutral. The rating generator contains the data of the user and provides the rating on the basis of user sentiment. In the final stage rating predictor provides the output as a prediction for the success and failure of the product one eCommerce platform.

2.3.3 Algorithm for pre-processor

The main process for the pre-processor is tokenization that uses the (Natural language tool kit) NLTK algorithms. NLTK algorithm filters the review on the basis of the suffix. The working procedure of NLTK algorithms is depicted in the diagram.

Fig. 2 Preprocessor

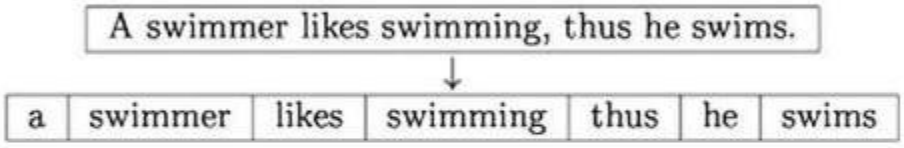

Fig. 3 Example of tokenization

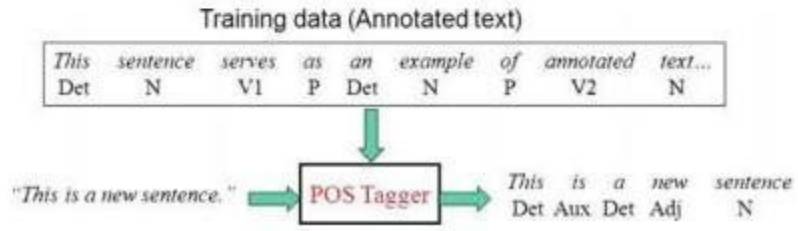

Fig. 4 Example of tagging

Figure 2: Pre-processor algorithmic steps

(Source: Tuladhar et al., 2018, pp. 279-289)

```
Step 1: Start
Step 2: for each review Ri from the review log RL
          use word_tokenize to generate the list of
                                              tokens T
          tag Part-Of-Speech to each ti ϵ T {(t_i,
                    POS_i), (t_j, POS_j), (t_k, POS_k), …}
          convert to base words using stemming
                                              process
          filter the words based on Part-Of-Speech
          remove stop words in the review
        end for
Step 3: Output the pre-processed review PR
Step 4: End
```

Figure 3: Algorithm 1 for pre-processor

(Source: Tuladhar et al., 2018, pp. 279-289)

2.3.4 Sentiment analysis algorithm

The sentiment analysis uses the Pre-processor output as the input in the form of a list. The list contains the bag of words by considering the naïve Bayes algorithms. The classifier classifies the review into three categories which is the neutral, negative, and positive on the basis of the Naïve Bayes algorithm.

```
Step 1: Start
Step 2: Intialize a bag of words in sentiment
        database
Step 3: for each PRi ϵ PR
            match PR[i] in PR with the bag of words
            if match is not found
                discard the item
            else
                sentiment_value[] = Extract the value
                                    from the bag of words
            end if
            calculate ∑sentiment_value[]
        end for
Step 4: if ∑sentiment_value > 0
            PR is "positive"
        else if ∑sentiment_value < 0
            PR is "negative"
        else
            PR is "neutral"
        end if
Step 5: Store PR in Sentiment Database
Step 6: End
```

Figure 4: Algorithm 2 for sentiment analysis

(Source: Tuladhar et al., 2018, pp. 279-289)

2.3.5 Rating Generator algorithm

```
Step 1: Start
Step 2: Accept pre-processed PR review along with the
        sentiment value.
Step 3: for each PR_i ϵ PR
          sum= ∑sentiment value {PR_1, PR_2, PR_3, …}
          average = sum / total no. of reviews with
                          sentiment value in that
                          time period
        end for
Step 4: Rate the product according the average
Step 5: Output product rating of every review
Step 6: End
```

Figure 5: Algorithm 3 for rating generation by a classifier

(Source: Tuladhar et al., 2018, pp. 279-289)

2.3.5 Rating prediction

Rating prediction contains the rating of the product by considering the forecasting technique of time series. The prediction result occurs in the form of the failure and success of the product.

```
Algorithm 4: Rating Generator
  Step 1: Start
  Step 2: Accept the product rating
  Step 3: for each product rating at time Δt
             if product rating Δt-1<=product rating Δt
                Product will be a success
             else
                Product will fail
             end if
          end for
  Step 4: End
```

Figure 6: Algorithm 4 for prediction of the success and failure of a product

(Source: Tuladhar et al., 2018, pp. 279-289)

2.4 Mathematical Intuition Behind Sentiment analysis

2.4.1 Naïve Bayes classification

As per the view of Shin et al. (2020), The naïve Bayes algorithm is a type of supervised learning algorithm that is designed on the basis of the Bayes theorem. It is designed for the solution of classification problems. The main aim of the algorithm is to provide high-dimensional training for the dataset for the task of text classification. This algorithm works on the probability and predicts the output. There are four types of probabilistic functions are used in the Naïve Bayes which are described below. Naïve Bayes is used when the data is present in the form of a discrete variable.

Conditional probability: It is the computation of the probability of a specific event with respect to another event that is occurring simultaneously.

Joint probability: It defines the likelihood of the occurrence of two or more events at the same time.

Proportionality: It refers to the two events that multiplicatively have a connection on the basis of simple terms and constants.

Bayes Theorem: It defines the probability of a phenomenon on the basis of various conditions.

The naïve Bayes classification theorem can be defined as follows (Berrar, 2018, 403):-

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

When the input variables are taken into the form of x and y variables then the expression can be written as follows.

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

As the naïve state that variables are independently present in the available class then the equation can be written in the following form.

$$P(X|y) = P(x_1|y) * P(x_2|y) * \ldots * P(x_n|y)$$

From the above equation, it is clear that P(X) is constant so this can be removed by the equation, and remained equation leads to the proportionality.

$$P(y|X) \propto P(X|y) * P(y)$$

-or-

$$P(y|X) \propto P(y) * \prod_{i=1}^{n} P(x_i|y)$$

Now the above equation shows the last state of naïve Bayes, it refers to the selection of y with the maximum probability. Argmax is a function that performs the operation in finding the maximum probability. This function finds the maximum value of the target function.

$$y = argmax_y[P(y) * \prod_{i=1}^{n} P(x_i|y)]$$

Li and Liu (2019, pp.150630-150641) defines the Steps of naïve Bayes algorithms.

Input – P_0 = Individual set

E = Function for determination of quality

Output – Best variable individually

Stage 1 – It ensures the current population meets with S, Step 5 if Satisfy, Stage 2 if not;

Stage 2 – Computation of E of an individual event in the current population;

Stage 3 – Execution of genetic operators for the generation of a new population;

Stage 4 – Execution of step 1;

Stage 5 – Termination of algorithm;

2.4.2 SVM for text classification

According to Goudjil et al. (2018, pp.290-298), Support vector machine classifiers are part of the supervised machine learning method. SVM works for finding the optimal hyperplane by the separation of two classes. On the basis of the difference between the hyperplane, it develops the model. The main aim of the SVM is to obtain high accuracy for classification tasks such as text categorization and pattern recognition. SVM is distinguished into two parts which are the linear and non-linear SVM. Linear SVM works for finding the boundaries between the two classes on a hyperplane. Whereas nonlinear cases use the Kernel function for mapping the high-dimensional features. In the transformed space a discriminant function that is associated with the hyperplane is defined as follows.

$$f(x) = w\Phi(x) + b.$$

There is the need for only one optimal hyperplane of separation of classes that have the maximum distance among the classes. The optimal hyperplane can be obtained by the following cost function.

$$\Psi(w, \xi) = \frac{1}{2}w^2 + c\sum_{i=1}^{N} \xi_i.$$

For the representation of active learning, there is a need for the following parameters.

1) C: Supervised classifier.
2) S: Supervisor who has the ability to assign the label of true class to unlabelled sample.
3) Q: Query function that used informative samples (unlabelled) of U.
4) T: Training set (Labelled).
5) U: Pool of samples (unlabelled).

Algorithm for active learning:

1. Selection of samples (Unlabelled) by the pool (randomly in a small set) and assigning a class for every sample. It is considered as a training set (T) initially.
2. Training of classifier (C) with T that was constructed in the first step.

Repeat

3. Query from a sample by the pool U with the help of the query function Q.
4. S – assign a label for class to T.
5. Addition of new samples (labelled) to T.
6. Retrain classifier.

These steps are run until the desired outcome is achieved.

2.4.3 Logistic regression for Sentiment analysis

Logistics regression is a type of statistical model that is mainly used for prediction and classification. Logistics regression works for predicting the probability of the occurrence of an event. The outcome of the dependent variable in terms of probability is bounded between 0 to 1. In order to perform the logistic regression it uses the logit transformation for distinguishing the probability of failure and the probability of success. The following equation defines the formula of logistic regression (: IBM 2022).

$$Logit(pi) = 1/(1+ exp(-pi))$$

$$ln(pi/(1-pi)) = Beta\_0 + Beta\_1*X\_1 + ... + B\_k*K\_k$$

Figure 7: Logistics regression equation

(Source: IBM 2022)

For performing the sentiment analysis on the social media data, logistics regression is considered the best method because it consists of the facility of frequency extraction that helps in analyzing the text data. It analyses the positive and negative reviews and provides the probability of the dependent variable in the range of 0 to 1 (Mashalkar 2020).

2.5 Model development

Step 1) **Data collection:** It is the most important step for performing the analysis because there is the requirement of appropriate data for finding the desired outcomes. For gathering the data there is a need for an understanding of the problem thus the appropriate method can be selected. The step of data collection requires some tasks such as the

identification of data sources, collection of data, and integration of data that is obtained by the different sources.

Step 2) Library selection: for the development of the ML model there is a need for various python libraries for the data processing and model development task. The use of a library reduces the complexity of the program. Python library enables a range of functions that can be used like a simple English language.

Step 3) Text/Data processing: After the collection of data there is a need for the preparation of data so appropriate action can be performed on the dataset. It requires a range of tasks such as data exploration, cleaning, and pre-processing. Data exploration defines the nature of data by considering the various aspects such as the format, characteristics, and quality of data. Data cleaning is also a necessary task because there is a range of null values and improper data is present that leads to the inaccuracy of the model thus it is necessary to the consideration of complete steps. It is also known as data wrangling which deals with missing values, invalid data, duplicate data, and noise. Mohammad (2018), states that appropriate classification is necessary because there is the presence of irrelevant content in the reviews section. The author describes the n-gram method for the classification of the clearing of text data. The data cleaning stage works for removing the rare words that can not be used for any type of analysis. Apart from them it also removes the blacklisted works that are marked as abusive language (Doo et al. 2021, pp.37-45).

Step 4) Data visualization and analysis: As the prepared data is present thus it is necessary that an understanding of the data should be developed with the help of visualization and analysis. Data analysis contains a range of analytical techniques, model

development, and evaluation. Machine learning enables a range of analytical techniques which are regression, classification, association, and cluster analysis. The text-based sentiment analysis process contains three major steps, in which the first one is to define the identity of the sentiment on the basis of the characteristics of sentiments. The other step refers to the selection of a feature that is useful for the analysis. Then it leads to the classification on the basis of machine learning algorithms. It produces the result into three parts which are negative, positive, and neutral (Zad et al., 2021, pp. 0285-0291).

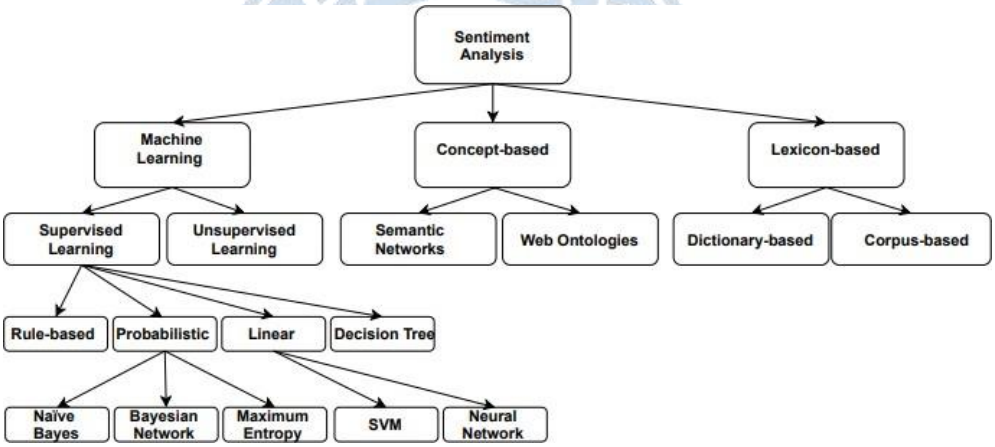

Figure 8: Sentiment classification techniques

(Source: Zad et al., 2021, pp. 0285-0291)

Step 5) Model training and testing: After the selection of the model, there is a requirement for training the model with the help of the dataset. In order to train the model, there is a need for machine learning algorithms. The model uses most of the part of dataset for the learning purpose and the remaining part is used for the testing of the model. Testing

of the model also provides the accuracy of the model. The model uses text data for the training of the model. If the unsupervised type learning method is selected then the model has the ability to work with the self-learning capacity.

Step 6) Model deployment: When the desired outcome is achieved by the model then the model is deployed in the real-world system. Before the deployment of the model, there is a need for the evaluation of the speed performance of the model.

2.6 Hypotheses

It is predicted that a machine learning model for sentiment analysis of Flipkart product reviews will be designed on the basis of the NLTK network. NLTK is a library of python that is used for the computation of statistics. It will also consider the concept of deep learning such as the convolutional neural network model. In order to perform the classification task, a Naïve Bayes classification will be used because there will be a requirement for separation of class on the basis of the product review. It will also perform the regression task with the help of the logistic regression functions. Sentiment analysis of Flipkart product reviews follows some stages of the machine learning life cycle of model development. The machine learning model uses the statistical concept for providing the output. The machine learning model will provide the output in three parts which are Positive, negative, and neutral.

3 Design and Methodology

In order to perform the research, it consists the research onion framework as the part of the research methodology. The research onion model refers for the different stages for attempting the research. It can also define as the method for performing the research. The research onion model was developed by the Saunders, Thornhill, and Lewis. The framework contains the six layers for performing the research that are described in the context of the research (Melnikovas, 2018, pp.29-44).

Research Philosophy: It is considered the first layer of the research onion framework, which consists of the principles of research. In order to perform the research, the researcher used the positivism approach. The reason behind the selection of positivism is that it contains the scientific results on the basis of the analysis. The research uses the public dataset of Flipkart company from the data science community Kaggle (Bojer and Meldgaard, 2021, pp.587-603). The analysis requires the scientific method because it uses the computation method of machine learning. The dataset contains the values in numeric and string format. In order to attempt the literature review there are various machine learning method and their mathematical intuition are included. The literature review chapter includes the different methods for developing the understanding of the appropriate method on the basis of the characteristics of the dataset. The research analyses the views of different authors that have published their research for sentiment analysis on the basis of the review of the products. The research also contains the hypotheses that define the future use of scientific methods. the method will be performed on the basis of the python programming language. The research also includes the statistical method that is used as part of mathematic intuition in order to perform the machine learning algorithm. Epistemology is considered for the research because it refers to the collection of information from an authentic source so the validity

of the research can be maintained. It uses resources from journal articles, books, and authenticated website that refers to the originality of the data.

Research Approaches: After the selection of the appropriate philosophy, there is a requirement of the section on the suitable research approach that also refers to the second layer of the model. The research uses the inductive approach because it refers to the creation of new theories on the basis of existing research and development. The research collects the data from the articles and also collects the data in a structured format in order to perform the analysis. The inductive approach also refers to the search for patterns and trends on the basis of available observations or theories. In order to identify the trends, the research will use data visualization in the next chapter. It also consists of the theoretical understanding of the visualization from the different research that is performed by machine learning-based researchers. The main aim behind the use of the inductive approach is to integrate the finding from the different data sources. In order to perform the research, it contains the data in both structured and text format. So, it integrates the theoretical and python-based analysis data for getting the results. Inductive research can also be performed for hypothesis testing.

Research Strategy: The research strategies suggest the inclusion of experimental research, action research, surveys, case study, and interviews. Experimental research involves the relationship between different variables. The literature review contains the different theories for the evaluation. The research contains the algorithmic procedure for performing the sentiment analysis. The research strategy aims to focus on a single subject by considering in-depth knowledge. The research uses the literature review section for the evaluation of the different theories. The research provides a deep study of sentiment analysis and defines the sentiment analysis of product reviews as opinion mining. The research defines sentiment analysis as the natural language processing that

works for developing the understanding of users' emotions. The sentiment analysis helps the organization to decide the suitability of the product. The research strategy is able to define the decision-making process of Flipkart so the company is able to choose the requirement of products. The research strategy also defines that the sentiment analysis provides the result in the three-part which are negative, positive, and neutral. It is considered the third layer of the research onion framework.

Research Choice: The research used the quantitative method for performing the study. The quantitative method refers for the analysis of the quantitative data. The research uses the python programming language for performing the quantitative research. The quantitative method includes observation, experiments, and surveys. The aim of the use of quantitative data is to achieve a better understanding. The quantitative method is able to develop new and innovative solutions for existing problems. The research contains the different open-source python-based library that is used for different computation. The mainly used libraries are NumPy and Pandas. These libraries are used for the processing and visualization of data (Stančin and Jović, 2019, pp. 977-982). The research uses the statics regression and naïve Bayes classification. Regression classification is the two different techniques that are used for different tasks. Regression is used for establishing the linear relationship between the independent and dependent variables and providing predictive analysis. The classification separates the two classes, it is mainly used on the categorical variable and delivers the result and predicts the future outcome. Both techniques will be used for sentiment analysis and provide predictive results. These approaches used the concept of statistics for performing sentiment analysis.

Time horizon: As part of the time horizon, the research uses longitudinal data because the research will be designed on the basis of observation. The research collects data from Kaggle. The data contains a range of variables so it is easy to perform the operation. If the dataset contains a range of variables then it is easy to establish the relationship in the various events. The dataset is collected for a specific time period and it also considers the variable of date so it refers to the longitudinal data because it contains the data from different years, months, days, and quarters. The time horizon is considered as the fifth layer of the framework. The dataset that contains the regular values also refers for the time series data so it easy to analyze the data. The longitudinal study repeatedly detects the event over a specific time period. The research will use two different methods in a duration of a short period of time. The longitudinal research can also be defined as the statistics term correlation research because it does not influence the other variable. It adds the specification in order to develop the research.

Data collection technique and procedure: It is considered the final layer of the framework. It defines the different procedures and techniques that are used in the research. In order to perform the study, we used a secondary data source for collecting the data. The data is collected from a research articles, journals, reviewed articles, official websites, and Kaggle. The data on Kaggle contains a range of variables which are the Product ID, Product title, rating of the product, summary, review, location, date, upvotes, and downvotes. The source of metadata is Flipkart's official website. This data is collected with the help of the Beuatifulsoup library of python (Uzun, 2018, pp.87-92). The library works for web scraping and collects the data. The reason behind the selection of secondary data is to develop a theoretical understanding of machine learning and statistical methods. The review of this data from the perspective of different authors is mentioned

in the chapter second which will help in the analysis part. The analysis of the dataset requires a range of machine learning and deep learning-based open-source libraries.

Ethical consideration: The research maintains the following values that increase the integrity of the research.

- **Scientific validity:** The research uses a range of scientific methods in terms of statistics, machine learning, and deep learning thus it is necessary to maintain the scientific value. The research does not hamper the originality of the method. The research uses appropriate principles of statistics that can be perfectly suited to sentiment analysis.
- **Social value:** It is known that advanced technology plays a vital role in for improvement of the society and workflow of the organization so the research maintains social value. The research has the scientific understanding for the improvement of the caring, treating, and preventing for people.
- **Fair selection of subject:** In order to choose the subject of the research, we maintain the current need of society. We also analyzed the various research that are performed already. After the analysis, we have chosen the sentiment analysis for the Flipkart product review so the research is able to contribute to the new startups for building the strategy.

4 Results, evaluation, and discussion

The evaluation of the research and generation of results is performed with the help of the python programming language. There are various tools can be used for performing the python script such as the Python IDE, Spyder, Jetbrains Pycharm, Anaconda, visual studio code, Jupyter, and online interpreter (Bercowsky, 2022, pp. 29-57). Python is considered as a dynamic language that can be used in various dimensions such as application development, GUI development, website, development, machine learning model building, and artificial intelligence-based task (Python, 2022). The research uses the python scripting language for performing the sentiment analysis on the Flipkart product reviews. Machine learning can be performed with the help of the python and R languages. Python enables a range of open-source libraries that are beneficial for the processing, visualization, and modelling of data. The sentiment analysis can be performed with the different algorithms of machine learning. We perform the sentiment analysis with the help of classification and regression. In order to perform the classification, we have used the naïve Bayes classification whereas we adopt the logistic regression for performing the regression analysis. The algorithm of naïve Bayes consider as the classification of probability or it can also define as the probabilistic classifier. The naïve Bayes classification designed a model on the basis of the probability and considers the independence assumptions for predicting the probability value (Dey et al., 2020, pp. 217-220). Logistic regression also refers to the estimation of the occurrence of a probabilistic event by considering the independent and dependent variables. Logistic regression follows the concept of linear regression. The logistic regression model also works on various assumptions.

- Sentiment Analysis is a process that is used to extract information from an entity. Entity is identified on the basis of conditions. Sentiment Analysis is used mainly to whether one's view is positive, neutral or negative.

- Sentiment analysis is collection of multiple reviews of a user on an online platform. Describes How sentiment analysis is performed by SVM and feature extraction, also what are the algorithms like, J48, BFTree and so.
- Lexicon Based Technique is the first method used for sentiment analysis. It used 2 concepts called Dictionary-based analysis and Corpus-based analysis.
- There are 2 types of ML methods used for sentiment analysis which are Traditional Methods and Deep-Learning-Based Models. Traditional Methods process input with help of Lexicon Technique and uses range of techniques like SVM, Maximum Entropy Classifier and Naïve Bayes Classifier. Whereas deep learning model provides more accuracy than traditional methods and uses models such as RNN, DNN, and CNN.
- Hybrid approach is an integration of ML and Lexicon-based approaches which use both statistical methods and knowledge-based methods to detect biased views.
- To perform ML algorithm in sentiment analysis we use certain procedures,

i. Data Acquisition is the first to build an ML model. Companies have many reviews, and the data should be labelled properly.

ii. Data Pre-Processing refers to the manipulation of data in order to get enhanced performance out of it. It uses keywords to keep the useful data in and trash the rest out.

iii. Feature Extraction is used to extract the useful information out of the selected review and convert it into a refined text.

iv. Evaluation, using the metrics of Precision, Accuracy, F-score and recall one can search for the result accordingly.

- Exploratory Data Analysis is an important step used to refine the datasets for better insights. EDA can be defined into 4 types, Univariate non-graphical, univariate graphical, multivariate non-graphical and multivariate graphical.
- Predictive Data Analysis deals with manipulating data from existing datasets and identification of new trends from those datasets.
- The main purpose of a pre-processor is to use NLTK algorithms and filter out the reviews on the basis of suffix.
- The Sentiment Analysis uses the pre-processor (output) as the input in form of a list. The list then goes through classifiers like Naïve Bayes Algorithms and categories the reviews into positive, neutral and negative it uses rating generator algorithm for the same and rating prediction to predict weather the product will be a failure or a success.
- There is some mathematical intuition behind sentimental analysis,
  i. Naïve Bayes Classification is a supervised learning algorithm that is designed on Bayes theorem. This is used to provide high-dimensional training for the dataset.
  ii. SVM, is used to find an optimal hyperplane by separating 2 class and obtain high accuracy for classification tasks.
  iii. Logistic Regression is a type of statistical model which is used for classification and prediction. It works with predicting the probability of the occurrence of an event.
- There are particular steps followed for the deployment of models,
  i. Data Collection: It is most important step, because to move further ahead there is need of accurate data.
  ii. Library Selection: Using a correct library is important to reduce the complexity of program to do that we can use liabraries like NumPy, Matlib and so.

iii. Text/Data Processing: It is step with various aspect such as, data exploration, Cleaning and pre-processing to keep the necessary data and trash out the improper data.

iv. Data Analysis: The data is examined by various statistical algorithms and helps identify important data to plot.

v. Model training and testing: The refined data is fined tuned according to the needs of the model and conditions and is tested repeatedly to get accurate output.

vi. Model Deployment: When the desired outcome is achieved then the model is deployed in the real world, and before deployment, there is a need for speed performance of the model.

- It can be predicted that the machine learning model for sentiment analysis of Flipkart product reviews will be designed on the basis of the NLTK network, which is deep component of neural network model. It will perform regression tasks with the help of logistic regression functions and help provide the results in positive, neutral and negative.

In order to perform the research we have used the following libraries for the efficient workflow of the machine learning algorithms.

- **NumPy:** NumPy has defined as the python language-based library that is used for numerical computation by enabling multi-dimension matrices and arrays, statistical calculation, and a range of mathematical formulas. It is considered as efficient as the efficient library because python does not have the in-built array but with the help of the NumPy, the programmer is able to use the array in the NumPy. The various complex mathematical formula can be performed in a single command with the help of the Open-source NumPy library. It also enables the core mathematical function such as the linear routine of algebra, Fourier transform, and random number occurrence (Harris et al., 2020, pp.357-362).

- **PANDAS:** Pandas is considered the software library that is used by the python programming language for the manipulation, processing, and visualization of data. It is known that ML works on historic data so it provides efficiency in handling the dataset. The main aim behind the development of pandas is to provide efficiency for data science-based tasks in the industry. It is the most widely used library for performing exploratory data analysis that provides various statistics results (Lemenkova, et al., pp.13-32).
- **Matplotlib:** It is also a python-based library that is defined as the upgraded and graphical version of the NumPy library. Matplotlib consists of a range of packages that can be used in different tasks. The main aim behind the development of the matplolib is to plot the relation in between the available variables of the dataset. So, it can also define as the visualization library and numerical extension of the NumPy library. It is mainly used for generating the embedding plots of object-oriented API. The purpose of the library is to integrate with efficient tolls such as Qt, Tkinter, and GUI (Lemenkova, et al., pp.13-32).
- **NLTK:** It stands for the natural language toolkit that is mainly used for the processing of NLP. It is associated with the different packages. The NLTK is used for processing of the words as it is known that dataset contains the reviews of the user in the string format. The main aim of the NLP is to processing of words. So, is used in the analysis part. It provides the analysis with the same efficiency as machine machine learning model (Schimtt et al., 2020, pp. 338-343).
- **Seaborn:** Seaborn have the same functionality as matplotlib. It is a visualization library for the high-level interface so an attractive design can be generated. Matplotlib is associated with some limitations because it is able to plot some basic graphs. Seaborn enables the development of various complex relationships such as the correlation matrix and heatmap (Lemenkova, et al., pp.13-32). Seaborn, is a library used for visualizing the statistical graphs plotting in Python.

Default styles and palettes of color define plots as more attractive and beautiful. All these plots are built in a Matplotlib Library and also data is closely integrated from Pandas. Sns in python tells the code to give an alias of sns by seaborn.

- **Plotly:** Plotly is python based library that provides 40 charts in a different contexts such as financial, statistical, scientific, and geographic. It is mainly used for the tasks of data and business analytics with the collaboration of different statistical tools. It can be integrated with the R, Python, MATLAB, Arduino, and REST (Van Der Donckt et a., 2022).
- **Sklearn:** It is also named Scikit-learn that is associated with the various packages for the python programming language. It includes a range of features such as regression, clustering, and classification. The library Sklearn is widely used for performing the support vector machine algorithms (Komar et al., 2019, pp. 97-111).

The programming is performed with the help of the Jupyter tool which is integrated with the visual studio code. The python file requires the '.ipynb' extension for the compatibility with the Jupyter notebook. After importing the library there is a requirement of the processing of the data with the help of the pandas library. The pandas provide the command read_csv that process the complete dataset in the csv format.

Histogram of rating score by distribution

Plotting the x, y parameters by defining the distribution of rating, where the title is rating score by distribution and variable x is ratting and the variable y is count, and giving palettes blue provide a different color maps of values. The output represents the histogram graph, and distribution of scores by rating.

Concatenating

Concatenating means combining series together, objects in panels, and various kinds of data frames. This function does all the lifting and performs operations horizontal or vertical axis while performing the union or intersection set of the logic of the given index named rating in the above code.

Next is to return the series in ascending order of a defined row rating in the data frame.

Histogram of distribution of rating

Distributing the rating on the basis of rating and count in the given data frame, where the x variable is rating and the y variable is count and the result is in the form of a histogram displaying rating scores and giving palettes blue provides a different color map of values. The output represents the histogram graph and distribution of scores by rating.

Screening out review rating up votes and down votes

Screening out the columns named review rating and upvotes and downvotes from the given data set by using the function stated in the codes. Then we used the info command to get a concise summary of the given data frame. It gives easy access to analyze the data.

Creating a regular expression

To extract all regular emoji we use this operation called extract emoji in python, we used this function in our data frame to extract emoticons, symbols & pictographs, map symbols & transport, flags, and some Chinese characters.

Stemming and Stop words

We used snowball stemmer to reduce a word in English to its base word or stem in such a way that words similar kind would lie under it. There are also other stemming algorithms like Porter stemmer and Lancaster stemmer.

The Stop words command is used to remove nonrequired English terms used in the dataset, it is the pre-processed thing that is done before the acquisition of another command to make understood by the computer.

Text Analysis

Receiving a review and cleaning it by using the methods which follow certain steps which include removing all nonwords and then transforming the review in lower case which removes all stop words after all these steps are performed stemming for test analysis.

Then return command is used to end the function. It is used to supplicate a function so that the statement which is passed can be executed.

Insertion of rating values

Filtering the n values from the product using python. The goal is to filter the product that has n value reviews in each rating. Plotting pie from the python library is to inspect the set of values exhibited in rating. By this, we could show the percentage to understand the well because of its different colorful sections.

Plotting word cloud

Visualizing the n set of values of the rating represents the text data indicating its frequency and importance of the data set. The word cloud was generated using a CSV file in the dataset. Some advantages of word clouds

- Keeping a record of customer
- Analyzing feedback driven by customer and employee
- Identify targeting keywords

Some parameters are generated containing an argument of text data. These arguments define the size and height of the cloud image. The default background is black. Bilinear is used to display smoother images.

Sentiment analysis by VEDOR

It is the process of determining writing computationally whether it demonstrates positive negative or neutral. A sentiment lexicon contains features that are generally labeled.

On the basis of rating, upvotes, and downvotes the review is classified as positive, negative, or neutral. if X= positive , Y= negative , Z= neutral

X>(Y, Z), it is positive.

Y>(X, Z), it is negative.

Z>(X, Y), it is neutral.

Thus, the sentiment score is neutral with a value of 772.819.

- Mapping the ratting, 0 to negative and 1 to positive, for the value count of sentiment score is:

  0-2020

  1-1000

- Plotting the distribution of sentiment labeling negative and positive in the pie chart with positive 33.11% and negative 66.89%.
- Word cloud was plotted for both negative and positive reviews by using a word cloud.

5. Conclusion

By analyzing the changing demands of customers, Flipkart choose to assimilate with tools like Sentiment Analysis. Sentiment Analysis incorporates technologies i.e., Machine Learning, Artificial Intelligence, NLP, and Data Mining and uses them to determine whether the response given by the user for a product, or a brand is Positive, Neutral or Negative. This research focuses on central questions like working of sentiment analysis for product review; working of machine learning on a dataset; illustrating for data analysis; mathematical expression in sentiment analysis; and procedure for developing a machine-learning model for sentiment analysis are answered. These questions are answered and discussed with Literature Reviews of the experts (from secondary data sources) and algorithms for data pre-processing, sentiment analysis of the data, rating generated analysis via classifier, lastly predicting success or failure of the product. The code uses a great degree of mathematical intuition to perform sentiment analysis, Naïve Bayes Classification is used which is a collection of multiple classification algorithms which are based on Bayes' Theorem, which is used to find the probability of an event to occur while knowing the probability of another event which has already occurred. While Logistic Regression or Logit Model is a statistical model used for classification and predictive analytics, logit used independent variables to estimate the probability of an event occurring. Since the data used in the code is from Kaggle, which has data in form of independent variables, Logit Model is best suited to get near-perfect answers. Python Programming is used to execute the code and generate the results for the same since python language is dynamic in nature and has multiple use cases like GUI, Web Development, Machine Learning, AI integration, Modelling and so. Python can embed a range of open-source libraries like NumPy, Pandas, Matplotlib, NLTK, Seaborn, Plotly, and Sklearn to perform activities related to machine learning with ease.

5.1 Recommendation

Sentiment Analysis can be used to predict the success or failure of a product in this case, Flipkart is a firm dealing with sales of products,

- It is recommended, that Flipkart should use the data-driven approach for decision-making.
- It is recommended, that Flipkart should look for advanced methods instead of traditional methods. Traditional Methods are obsolete and have high risk due to the involvement of human resources, whereas Advance Methods use technology to get the outcomes for the problems and are considered more accurate and efficient.
- It is observed that after some companies adopted sentiment analysis as their decision-making approach, those companies have shown positive results. Therefore, it is recommended to study their behaviour and strategies.
- It is recommended, to improve the performance of data by creating new version of the dataset, to get missing values according to the need of organizations.
- It is recommended, to improve the performance of algorithms by creating refined data representations and well-performing algorithms according to need of the company and so proper algorithm tuning for better performance and smooth algorithms for the datasets, by setting more defined parameters it's easy to find the optimum value for a more accurate model.
- It is recommended, that Flipkart should go with Ensemble Methods and create multiple well-performing models to get a combined prediction than a highly tuned one, by use of Ensemble Methods we can get more accuracy in the model for accurate customer sentiments.

5.2 Future work

Sentiment Analysis is a very useful tool for businesses and brands to find out their standing among their customers. To date, brands use sentiment analysis to understand their position among their competitors and the data and reaction of customers grow over the internet the more will sentiment analysis grow. This research offers multiple future scopes,

- To those who are looking to understand the concept of Sentiment Analysis, and to implement it in their businesses either Large Capital or Small Capital.
- To educational institutions who wish to teach the concept of sentiment analysis to their students and implement it in their institute for inner growth.
- To Market Research Analyst, who wish to understand the customer's buying habits and spot upcoming trends.
- For Businesses who wish to create customer satisfaction strategies by analysis their feedbacks.
- Further studies on sentiment analysis like spam detection, improved identification algorithms, handle biased sentiments, and so on.
- For researchers, who wish to cite this research for further studies.

References

AlQahtani, A.S., (2021). 'Product sentiment analysis for amazon reviews'. International Journal of Computer Science & Information Technology (IJCSIT) Vol, 13.

Bahassine, S., Madani, A., Al-Sarem, M. and Kissi, M., (2020). 'Feature selection using an improved Chi-square for Arabic text classification'. *Journal of King Saud University-Computer and Information Sciences*, *32*(2), pp.225-231.

Bercowsky Rama, A., (2022). 'Python: Data handling, analysis and plotting. In Bioimage Data Analysis Workflows–Advanced Components and Methods' (pp. 29-57). Springer, Cham.

Berrar, D., (2018). 'Bayes' theorem and naive Bayes classifier'. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics.

Bojer, C.S. and Meldgaard, J.P., (2021). 'Kaggle forecasting competitions: An overlooked learning opportunity'. *International Journal of Forecasting*, *37*(2), pp.587-603.

Dang, N.C., Moreno-García, M.N. and De la Prieta, F., (2020). 'Sentiment analysis based on deep learning: A comparative study'. *Electronics*, *9*(3), p.483.

Dey, S., Wasif, S., Tonmoy, D.S., Sultana, S., Sarkar, J. and Dey, M., 2020, February. A comparative study of support vector machine 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.

Doo, I.C., Shin, H.D. and Park, M.H., (2021). 'Automated product review collection and opinion analysis methods for efficient business analysis'. *International Journal of Computing and Digital Systems*, *10*(1), pp.37-45.

Goudjil, M., Koudil, M., Bedda, M. and Ghoggali, N., (2018). 'A novel active learning method using SVM for text classification'. *International Journal of Automation and Computing*, *15*(3), pp.290-298.

Haque, T.U., Saber, N.N. and Shah, F.M., (2018, May). 'Sentiment analysis on large scale Amazon product reviews'. In *2018 IEEE international conference on innovative research and development (ICIRD)* (pp. 1-6). IEEE.

Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J. and Kern, R., (2020). 'Array programming with NumPy'. *Nature*, *585*(7825), pp.357-362.

IBM. (2020). '*What is Exploratory Data Analysis?'* [online] Available at: https://www.ibm.com/in-en/cloud/learn/exploratory-data-analysis [Accessed 10 Nov. 2022].

IBM. (2022). '*What is Logistic regression?*. IBM.com. [online] Available at: https://www.ibm.com/in-en/topics/logistic-regression [Accessed 26 Nov. 2022].

Komer, B., Bergstra, J. and Eliasmith, C., (2019). 'Hyperopt-sklearn'. In Automated Machine Learning (pp. 97-111). Springer, Cham.

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y. and Alsaadi, F.E., (2020). 'Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods'. *Applied Soft Computing*, *86*, p.105836.

Lemenkova, P., (2020). 'Python Libraries Matplotlib, Seaborn and Pandas for Visualization Geo-spatial Datasets Generated by QGIS'. Analele stiintifice ale Universitatii" Alexandru Ioan Cuza" din Iasi-seria Geografie, 64(1), pp.13-32.

Li, M. and Liu, K., (2019). 'Causality-based attribute weighting via information flow and genetic algorithm for naive Bayes classifier'. *IEEE Access, 7,* pp.150630-150641.

Mashalkar. A., (2020). '*Sentiment Analysis using Logistic Regression and Naive Bayes'*. [online] Medium. Available at: https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b [Accessed 26 Nov. 2022].

Melnikovas, A., (2018). 'Towards an explicit research methodology: Adapting research onion model for futures studies'. *Journal of Futures Studies, 23*(2), pp.29-44.

Mohammad, F., (2018). 'Is preprocessing of text really worth your time for online comment classification'?. *arXiv preprint arXiv:1806.02908*.

Onan, A., (2021). 'Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks'. *Concurrency and Computation: Practice and Experience*, *33*(23), p.e5909.

Python. (2022). 'Applications for Python'. [online] Available at: https://www.python.org/about/apps/ [Accessed 12 Dec. 2022].

Saldaña, Z.W., (2018). 'Sentiment Analysis for Exploratory Data Analysis'. *Programming Historian*.

Schmitt, X., Kubler, S., Robert, J., Papadakis, M. and LeTraon, Y., (2019, October). 'A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate'. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 338-343). IEEE.

Shin, T. (2020). '*A Mathematical Explanation of Naive Bayes in 5 Minutes'*. [online] Medium. Available at: https://towardsdatascience.com/a-mathematical-explanation-of-naive-bayes-in-5-minutes-44adebcdb5f8 [Accessed 10 Nov. 2022].

Stančin, I. and Jović, A., (2019), May. 'An overview and comparison of free Python libraries for data mining and big data analysis'. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 977-982). IEEE.

Tuladhar, J.G., Gupta, A., Shrestha, S., Bania, U.M. and Bhargavi, K., (2018). 'Predictive analysis of e-commerce products'. In *Intelligent Computing and Information and Communication* (pp. 279-289). Springer, Singapore.

Uzun, E., Yerlikaya, T. and Kırat, O., (2018). 'Comparison of python libraries used for web data extraction'. *Journal of the Technical University-Sofia Plovdiv Branch, Bulgaria*, *24*, pp.87-92.

Van Der Donckt, J., Van Der Donckt, J., Deprost, E. and Van Hoecke, S., (2022). 'Plotly-resampler: Effective visual analytics for large time series'. *arXiv preprint arXiv:2206.08703*.

Yang, L., Li, Y., Wang, J. and Sherratt, R.S., (2020). 'Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning'. *IEEE access*, *8*, pp.23522-23530.

Zad, S., Heidari, M., Jones, J.H. and Uzuner, O., (2021, May). 'A survey on concept-level sentiment analysis techniques of textual data'. In *2021 IEEE World AI IoT Congress (AIIoT)* (pp. 0285-0291). IEEE.