**SENTIMENT ANALYSIS OF BOOK REVIEW**

**Savitha O P – 1, Dharani P -2** Student, MSC (Decision and computing sciences), Coimbatore Institute of Technology, Coimbatore, India.

RamyaRajendran, Assistant professor, Computing department, Coimbatore institute of technology.

**ABSTRACT:**

The word data is heard atleast once a day. There are many applications, social mediaCollect data from comment, review of any movie or online shopping. About 30% of the data is structured and rest of them are unstructured. To process the unstructured data like audio, video, text natural language processing has been used. In NLPsentiment analysis is used to process the text data. Sentiment analysis is used to find the sentiment of the text that is positive or negative. Today, there are enormous amount of people were engaged on online shopping Here sentiment analysis is performed on a review of a book think and grow fast.People were buying goods on the basis of product reviews. People see reviews from various social media platforms like twitter, instagram, facebook, youtubeetc and then they decide to buy a product or not.

. Sentiment analysis is an interesting topic to identify the comment as positive or negative sentiment. Sentiment analysis is also known as opinion mining. There are many algorithms in sentiment analysis to identify comment, text as positive or negative sentiment. Naive bayes and support Vector Machine to measure sentiment analysis from user’s opinions. Naïve bayes gives good results provided with less training dataset. Whereas support vector machine provides good accuracy.Here accuracy between the test dataset of flipkart and amazon is compared.

Sentiment analysis helps amazon, flipkart to improve their product quality. Data used in this is collected from online Book review from Amazon and flipkart.

Keywords-Support vector machine,NaiveBayes,Sentiment Analysis

**1.INTRODUCTION:**

In starting of the 20th century people buy products from traditional stores. Now a days people show more interest to buy products on online shopping sites like amazon, flipkart.  Unstructured data makes up 80% and more of enterprise data, and is growing at the rate of 55% and 65% per year [1]. To process the unstructured data Natural language processing is used .Sentiment analysis is one of the application of Natural language processing. Sentiment analysis studies opinion text,there was almost no research into this topic from either the linguistic community or the NLP community before the year 2000,in part because almost no opinionated text was recorded in digital forms before then.With the explosive growth of the web and social media in the past twenty years, we now have a constant flow of opinionated data recorded in digital forms[2]. Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. A sentiment analysis system for text analysis combines natural language processing ([NLP](https://www.lexalytics.com/lexablog/what-is-natural-language-processing)) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase[3].Sentiment analysis is the process to identify the sentiment of the text data . Sentiment analysis is used to determine whether the text that is comment, review is positive or negative sentiment.It is a text based analysis method. It is also known as emotional extraction or opinion mining. It is an robust process to understand what customer think’s about the book.

Sentiment analysis is very popular field of research in predictive analytics. It analyse people sentiments, opinion, attitude,emotions towards elements such as products, individual’s etc. Nowadays, online shopping has grown tremendous these years and people share their opinions through ratings and comments.In this paper Sentiment Analysis is used on a book review to determine whether the book is reachedto the people in a positive or negative way. It helps to identify the overall review of the book.Most researches related to sentiment analysis uses support vector machine to improve the accuracy.

Here sentimental analysis is done in two methods namely naïve baye’s and support vector machine. The advantage of naïve baye’salgorithm is that it can work with less dataset to train the model. On the other hand support vector machine has good accuracy comparing to other classification algorithms.

Naive bayes and Support Vector Machine Classifier areclassification algorithm used for sentiment analysis. With the help of this method we are able to know the sentiment of the book. Accuracy of Amazon and flipkart dataset is compared. resultsDataset consists of review and sentiment column that is positive or negative. Dataset is collected from amazon and flipkart.The training dataset for both amazon and flipkart is 100 rows each and the test dataset for both of amazon and flipkart is 50.

Section 2 explains data preprocessing procedure so that text data is transformed into machine learning algorithm .Section 3 explains methodology here naïve bayes and support vector machine is explained and this part covers how the implementation is done. Section 4 explains the results which compares accuracy for train and test dataset for amazon and flipkart. Section 5 explains conclusion.

**2. DATA PREPROCESSING:**

Data preprocessing is an important tool for Data Mining (DM) algorithm. Twitter data is an unstructured data set it is a collection of information from people entered his/her feelings, opinion, attitudes, products review, emotions, etc. This type of information is growing day by day in the internet. May companies want to analyze customers opinions which like the product and the services[11].Data is preprocessed so that the text data that review is transformed into numeric data so that it can be feed into machine learning algorithm to determine the sentiment of the book. Sentiment analysis involved identifying a given text of content by first preprocessing it to detecting stop words and symbols, etc. and then checking the subjectivity contents. The getting the opinion content polarity is determined either on machine learning methods and lexical based methods[12].

There are various steps in data preprocessing procedure they are tokenization, stop words removal, lemmatization, Bag of words creation and positive and negative score calculation.

**TOKENIZATION:**

The process of tokenization is to split the sequence of strings into words. In Sentiment analysis, it is used to convert the sentences into words. Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words[13].

**STOP WORDS:**

Words that do not contain enough significance, and that can be removed from the review. After tokenization stop word is performed to remove the unwanted words from the tokenized words. Stop words are those words in the text which does not add any meaning to the sentence and their removal will not affect the processing of text for the defined purpose. They are removed from the vocabulary to reduce noise and to reduce the dimension of the feature set[14].

**LEMMITIZATION:**

The Process of lemmatization is used to group the words to their root word depending upon the context. After the removal of stop words lemmatization is performed to group the words to their base word depending upon the context**.**

**BAG OF WORDS:**

This method is used for feature extraction. Bag of Words (BOW) is a method to extract features from text documents. These features can be used for training machine learning algorithms. It creates a vocabulary of all the unique words occurring in all the documents in the training set[15].In this paper this method is used .This method gives the vector list for reviews.[4]

For example :

|  |
| --- |
| **Take a sentence :”It's an amazing book..”** |

* First the sentences is removed from special characters .
  + There is no special characters.
* Next the sentences is tokenized .Tokenization splits the sentence into words and discard comma and full stop.
  + List = [‘it’s’, ‘an’, ‘amazing‘, ‘book’]
* Then the stop word is removed from the list
  + Vocabulary = [ ‘amazing’, ’book’]
  + Here the stop words were it’s, an.
* Then the vector is formed. The vector is of length vocabulary. If the words in the list is found in the vocabulary then the vector is marked 1 else 0.
  + Vector =[ 1,1]
* Then the positive and negative score for each review is calculated and is added with the corresponding vector list.
  + Vector=[1,1,positive\_score,negative\_score]

Data

Data Preprocessing

Result

The first step is to find the vocabulary list. First the each comment is removed from special characters. Then each comment that is sequence of string is split into words and if the split words is found in the stop word it is removed and the words that are not in stop words is lemmatizied and then to the vocabulary list. Then the vocabulary list is passed to set function to remove the repeated words and then the vocabulary list is sorted.

The second step to vectorize the comment using the vocabulary list. The length of the vector is of size vocabulary list. Here each comment is lower cased and each vocabulary is traversed to check if any comment word is found in vocabulary list then one will be marked else zero is marked.

After the completion of bag of word method this step to find the positive and negative score of the comment. Each comment is split into words and the words senti\_synsets() function and then the positive and negative score for the word is found. Likewise do for all the word in a comment and add the scores to the final positive and negative score

Then using merge\_featuresfunction merge the vector and the positive,negative score.

Now the data is ready to be feed into naïve bayes and support vector machine Classification algorithm.

**3. METHODOLOGY:**

In this section the machine learning classification algorithm that is used in this paperis discussed. Here two classification algorithm is performed. Naive bayes algorithm preform good with less train dataset. Whereas support vector machine give good accuracy. In sentiment analysis determination of positive or negative sentiment is done by both of the classification algorithm.

**TWO METHODS:**

* + NAÏVE BAYES
  + SVM(SupportVectorMachine)

**3.1 NAIVE BAYES:**

[6]Naive bayes is one of the traditional machine learning method. Naive bayes classifier is a supervised learning algorithm used for classification.Naivebayes is one of the best method for sentiment analysis. It does not require much training dataset .In naïve bayesalgorithm,every feature is independent of each other. Naive bayes algorithm is not just an algorithm it’s a collection of algorithm that works on same concepts based on biased theorem. Naïve bayes classifier has three types namely

* + - * Gaussian
      * Multinomial
      * Bernoulli

[10]The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

Here in this paper multinomial has been used. MultinomialNB() .fit() function is used to train the model and predict function is used to predict the test dataset.

In MultinomialNV.Fit()function passing train features and training labels to the fit function. Training features contain the count vectorisation and positive or negative value.Training label contain the numerical value of the sentiments thatis 1 for positive and -1 for negative.Now using the predict function predict the test features.

**3.2 SUPPORT VECTOR MACHINE:**

. [5]Support vector machine is used to segregate the data points . Support Vector Machine is a supervised learning algorithm .It is used for classification and regression problems. SVM create a best line or decision boundary that can segregate n dimensional space into classes so that in future we can put new data point in a correct category.

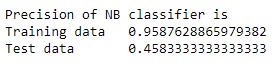
[9]SVC with parameter kernel=’linear’, but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples.This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme.

In SVM,LinearSVC().Fit()method to train the dataset and use predict function to test the dataset.

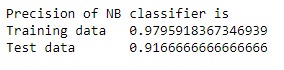
In LinearSVC().Fit() function passing train features and training labels. Training features contain the count vector and positive or negative value.Training label contain the numerical value of the sentiments thatis 1 for positive and -1 for negative.Now using the predict function to predict the test features.

**4. RESULT:**

The output for the Multinomial naïve bayesfor Amazon :



The output for the Multinomialnaïve bayes for Flipkart:



The output for the Linear support vector machinefor Amazon :

amma2.jpeg

The output for Linear support vector machine Flipkart:

flip2.jpeg

Comparing the accuracy produced by the amazon and flipkart dataset for the book think and grow fast.Flipkart dataset shows the good accuracy for both naïve bayes and support vector machine algorithms.The accuracy produced by the test dataset of flipkart for both the algorithms shows the same results. The accuracy produced by the test dataset of Amazon for both the algorithms shows that naïve bayes algorithm produce 7%more accuracy than SVM algorithm .Though the accuracy for both the models is less than 50%.

**5.CONCLUSION:**

In this paper an attempt to analyze the unstructured data. To analyze the unstructured data natural processing language has been used. In NLP sentiment analysis has been used to process the unstructured data. Sentiment analysis has been used to classify the text as positive or negative sentiment us. Sentiment analysis is done on review for a book think and grow rich. It helps the people whether to buy a book. It enables people to buy products all the time based on positive feedback. It helps amazon and flipkartto improve their service for customers by removing the most negative sentiment product. Sentiment analysis or opinion mining is a field of study that analyzes people’s sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Online product reviews from Amazon.com and flipkart.com are selected as data used for this study. A sentiment polarity categorization process has been proposed along with detailed descriptions of each step. Experiments for both sentence-level categorization and review-level categorization have been performed[16-20].

**6. REFERENCES:**

**1.**[**https://www.datamation.com/big-data/structured-vs-unstructured-data/#:~:text=On%20top%20of%20this%2C%20there,%25%20and%2065%25%20per%20year**](https://www.datamation.com/big-data/structured-vs-unstructured-data/#:~:text=On%20top%20of%20this%2C%20there,%25%20and%2065%25%20per%20year)

**2.**[**https://www.google.co.in/books/edition/Sentiment\_Analysis/PdX7DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover**](https://www.google.co.in/books/edition/Sentiment_Analysis/PdX7DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover)

**3.**[**https://www.lexalytics.com/technology/sentiment-analysis**](https://www.lexalytics.com/technology/sentiment-analysis)

4.[**https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/**](https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/)

5.[**https://www.goeduhub.com/3487/demonstrate-and-implement-support-vector-machine?show=3487#q3487**](https://www.goeduhub.com/3487/demonstrate-and-implement-support-vector-machine?show=3487#q3487)

6.[**https://www.goeduhub.com/3188/demonstrate-implement-bayesian-classifier-classification#q3188**](https://www.goeduhub.com/3188/demonstrate-implement-bayesian-classifier-classification#q3188)

7.[**https://www.sciencedirect.com/science/article/pii/S187705091630463X**](https://www.sciencedirect.com/science/article/pii/S187705091630463X)

8.[**https://www.irjet.net/archives/V4/i4/IRJET-V4I4598.pdf**](https://www.irjet.net/archives/V4/i4/IRJET-V4I4598.pdf)

**9.https://scikitlearn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC**

**10.https://scikitlearn.org/stable/modules/generated/sklearn.naive\_bayes.MultinomialN.html#sklearn.naive\_bayes.MultinomialNB**

**11.** [**https://www.researchgate.net/publication/334670363\_DATA\_PREPROCESSING\_IN\_SENTIMENT\_ANALYSIS\_USING\_TWITTER\_DATA/link/5d393d3692851cd046844411/download**](https://www.researchgate.net/publication/334670363_DATA_PREPROCESSING_IN_SENTIMENT_ANALYSIS_USING_TWITTER_DATA/link/5d393d3692851cd046844411/download)

**12.**[**https://www.researchgate.net/publication/334670363\_DATA\_PREPROCESSING\_IN\_SENTIMENT\_ANALYSIS\_USING\_TWITTER\_DATA/link/5d393d3692851cd046844411/download**](https://www.researchgate.net/publication/334670363_DATA_PREPROCESSING_IN_SENTIMENT_ANALYSIS_USING_TWITTER_DATA/link/5d393d3692851cd046844411/download)

**13.**[**https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4**](https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4)

**14.**[**https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4**](https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4)

**15.** [**https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/**](https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/)

**16-20:** [**https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2**](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2)