International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 521-525 www.ijaem.net ISSN: 2395-5252

Product Data Review using SVM

P.Haribabu[1], J.Sai Mohith[2], K.Rose Rachel[3], M.Suswanth[4], Sai Prasad[5]

Assistant professor[1], Student[2,3,4,5] B. Tech (CSE), COMPUTER SCIENCE ENGINEERING

Submitted: 01-06-2022 Revised: 05-06-2022 Accepted: 08-06-2022

ABSTRACT: Sentiment analysis or opinion mining is one of the major tasks of NLP (Natural Language Processing). Sentiment analysis has gained much attention in recent years. In this paper, we aim to tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. A general process for sentiment polarity categorization is proposed with detailed process descriptions.

I. INTRODUCTION

Sentiment is an attitude, thought or judgmentprompted byfeeling.Sentiment analysis, which is also known as opinion mining, studies people's sentiments towards certain entities. Internet is a resourceful place with respect to sentiment information. From a user's perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites. From a researcher's perspective, many social media sites release their application programming interfaces

Data used in this study are online product reviews collected from Amazon.com. Experiments for both sentence-level categorization and review-level categorization are performed with promising outcomes. At last, we also give insight into our future work on sentiment analysis.Sentiment analysis is defined as the process of mining of data, view, review or sentence to predict the emotion of the sentence through natural language processing (NLP). The sentiment analysis involves classification of text into three phase "Positive", "Negative" or "Neutral".

Sentiment analysis is text-based analysis, but there are certain challenges to find the accurate polarity of the sentence. This states that there is need to find the better solution toget much better results than the previous approach or technique used to find polarity of sentence. Therefore, to find polarity or sentiment of, user or customer there is a demand for automated data analysis techniques. In this paper, a detailed survey of different techniques or approach is used in sentiment analysis and a new technique which is proposed in this paper.

Sentiment analysis or opinion mining is the computational study of people's emotions, opinions, sentiments by considering their reviews in the form of text in recent years. This is the most active research area in the field of natural language processing and text mining. Since it is based on the opinions and as all the humans make decisions dependent on other opinions its popularity is increasing day by day. In this paper, our ultimate goal is to tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis.

Therefore, this project deals with analysis of Products data getting from the UCI Depositories, DataWorld and Jmcauley where we need products reviews and their customers names.

(APIs), prompting data collection and analysis by researchers and developers.

For instance, Twitter currently has three different versions of APIs available, namely the REST API, the Search API, and the Streaming API. With the REST API, developers are able to gather status data and user information; the Search API allows developers to query specific Twitter content, whereas the Streaming API is able .

II. LITERATURE SURVEY-

A.Modeling Sentiment Terminologies: Target Based Polarity Phenomena In [5], the researchers presented a subject sensitive sentiment analysis approach, which includes the context of tweets. According to authors the text cleansing techniques for input data before classification process can improve the results. Text cleansing includes normalization and vector representation of input data. They have pointed out that the subject aware classification brings the better results as compare to subject un-aware classification. The results can be further improved, if uni-gram approach is used instead of bi-gram or n-gram approach. A twitter dataset about word "Obama" was selected first. Features from tweets of selected

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 521-525 www.ijaem.net ISSN: 2395-5252

dataset were extracted through Alchemy API, Tweet NLP and NTLK. From dataset, 30% of the data was used for training purpose and the rest of 70% as the test data.

Keyword_Bundle- in conjunction with their specific topics to retain the target and context of the tweets. This technique further helped for the development of input matrix for SVM to classify the tweet with improved accuracy. Then two more datasets were selected "Movie Review", and "Apple" to have a comparative analysis. 85.00 %, 84.00%, and 88.00% accuracies were achieved of "Obama", "Movie Review" and "Apple" datasets respectively, making a cumulative accuracy of 85.60%

B.Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process

In [6], authors presented an approach that classified the documents into multiple categories by keeping in view the multiple aspects. The existing problems of document level sentiment analysis such as entity identification, subjectivity detection and negation were also taken into consideration in this study. The proposed framework was used for educational data mining. The faculty performance was evaluated using the (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018 186 | P a g e www.ijacsa.thesai.org comments provided by the students as feedback. The dataset contained 5000 comments about the faculty. The objective reviews which had no polarity inclination were filtered out, such as social comments, replies and questions. Java string tokenizer was used to divide the reviews into two token groups. After this, stopwords removal algorithms were used to remove special characters and some pronouns which would hold no significant value in the actual classification.

They used TF-IDF to represent the acquired data in a numerical form, which is further used by the classifiers. Two machine learning classifiers, i.e. Naïve Bayes and Support Vector Machine were applied on the pre-processed dataset. 81.00% and 72.80% accuracy were achieved by the SVM and Naïve Bayes, respectively for aspect based document level sentiment analysis.

C. Tweeple's Microblogs on Illegal Immigration in USAAuthors in [7] presented a process for opinion mining of tweeps (People who use Twitter). The topic that was specifically chosen in comparison with some other political topics was "Illegal Immigration" as it has been under discussion for decades in the US. The dataset used in this research was collected after the US Republican Presidential election debate on Oct 28, 2015. Three major categories of the topic were selected i.e. "reform/give citizenship to illegal immigrants", "deport all immigrants" or "deport only the criminal illegal immigrants". Binary classification of first two and multinomial classification of all three categories was done using the Random Forest, Multinomial Naïve Bayes, Linear SVM and Logistic Regression classifiers. The results obtained for all the four classifiers were promising with 82% of overall average. Linear SVM and ensemble based approach using Random Forest classifiers depicted optimal results and accuracy with the mean score of 90% and 84%, respectively for binomial and multinomial classification, for individual classes with lower error rate.

III. METHODOLOGY & IMPLEMENTATION

3.1. Proposed Work

The flow of proposed work is shown in Figure 2. The proposed work contains the following key points:

•The proposed work discusses about applying sentiment analysis and machine learning algorithm to investigate the relationships among the online reviews for smart phone products and the revenue of performance.

• The algorithm is to be applied on product reviews and predict collection of the product based on the reviews and analyses how much effect the reviews have on the collection. The product collection for the next day is predicted based on online reviews of the present day.

• A prediction of high or low collection is also predicted. From the website, the detailed information containing the values for the following: brands, product date, rank of sale, user's reviews, etc. of a smart phone were obtained.

•Part of Speech (POS) model in which a sentiment or textual review is represented as a vector, whose entries correspond to individual terms of a vocabulary. Part-of-speech information is supposed to be a significant indicator of sentiment expression.

•The score of each sentence in the dataset is calculated by sum of weight of each term in the corresponding sentences. Clustering of the review data based on the TF-IDF measure has been performed.

• Finally, the proposed work achieves high accuracy, the reviews are taken as appropriate and the success or failure of the smart phone product is predicted based on the reviews by using Support Vector Machine classification algorithm.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 521-525 www.ijaem.net ISSN: 2395-5252

3.2 Simulation Work

The simulation work is performed by using Java programming language, WAMP Server and MySQL. The complete simulation is accomplished by using the following modules:

- Text pre-processing
- Transformation
- Clustering
- SVM classification
- Evaluation

3.2.1 Text Pre-Processing

Text pre-processing techniques are divided into two sub-categories which are POS tagging and stop words removal. In POS, textual data comprises block of characters called tokens. The input reviews are separated as tokens and Figure 2. Flow of proposed work. Sentiment Analysis of Product Reviews using Support Vector Machine Learning Algorithm Vol 10 (35) | September 2017 | www.indjst.org Indian Journal of Science and Technology 4 start the pre-processing. A stop-list is the name commonly given to a set or list of stop words. Some of the more frequently used stop words for English include ”a”, ”of ”, ”the”, ”I”, ”it”, ”you”, and these are generally regarded as ’functional words’ which do not carry meaning. Hence remove those words that support no information for the task.

3.2.2 Transformation

In the transformation process, the score for each sentence is calculated in the document. For that, first the weight of each term is calculated by the product of term frequency and inverse document frequency.

3.2.3 Clustering Clustering

Of the document review is based on the TF-IDF measurement. Thus, points on the edge of a cluster, maybe in the cluster to a lesser degree than points in the center of cluster. It chooses the number of clusters and it finds centroid.

3.2.4 SVM Classification

After the removal of the outliers based on the clustering, the improved feature sets were used for sentiment classification.SVM is mainly used for the sentiment classification. It classifies the positive and negative reviews.

3.3 PROBLEM STATEMENT-

The main objective of this project is to go about an extra mile to provide the users with an output that is the analysis of thousands and thousands of reviews

3.4 PROBLEM SOLUTION-

To save time by analysing thousands of reviews in short period and if those reviews were anmanually it may take up to decades. The overall output is achieved by several steps like:

Step 1: Gathering the raw data from different repositories.

Step 2: Pre-processing the raw data in a usable format.

Step 3: Fetching the data set file using R language.

Step 4: Identifying the sentiment words in the reviews.

Step 5: Maintaining a count of different type of words.

Step 6: Data visualization in a bar graph.

3.5 WORKFLOW ALGORITHM-

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning**.**

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine

3.6 STEPS FOR THE PROJECT-

1.Data Pre-processing: After the data has been selected, it needs to be pre-processed

2. Tokenization: The process of breaking a stream of text up into phrases, words, symbols, or other meaningful elements called tokens. The goal of the tokenization is the exploration of the words in a sentence.

3. Stop-word Elimination: The most common words that unlikely to help text mining such as prepositions, articles, and pro-nouns can be considered as stop words. Since every text document deals with these words which are not necessary for application of text mining.

4.Bag-of-words Model: The bag-of-words model is one of the simplest language models used in NLP. It makes a unigram model of the text by keeping track of the number of occurrences of each word. This can later be used as features for Text Classifiers.

5. Training the classifier: We are training the classifier using the Features Extracted using the Bag-of Words Model.

The Features of both the training and test dataset are

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 521-525 www.ijaem.net ISSN: 2395-5252

com- pared. And this is giving to the classifier to give the predictions on the test data.
6. Sentimental Analysis: For sentimental analysis we are using the Decision treeclassifier and Naive Bayes and comparing the results. We also see which classifier has the most accuracy

3.7 DESIGN OF PROJECT-

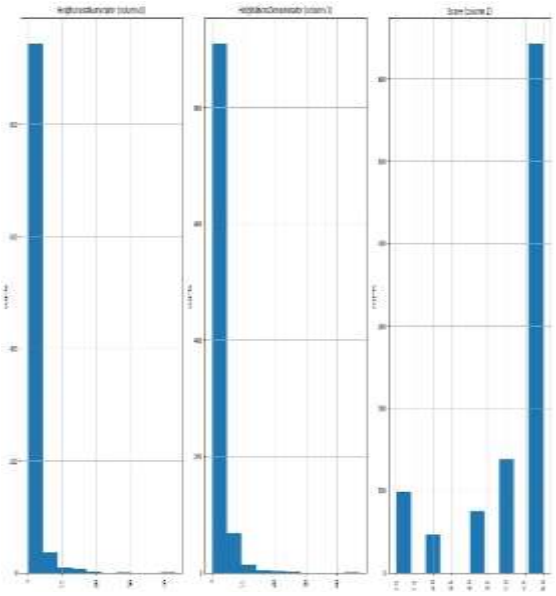
3.7.1Filtering thereviews:

There are many numbers of reviews given for a particular product so in that reviews we need only the adjective words and the remaining nouns and pronouns are to be ignored so to ignore those words we use Stop words. Stop words is a variable consists of all the noun and pronoun words. By removing stop words from the reviews, we get only the adjective words from filtering. The remaining words are taken by having spaces between them.

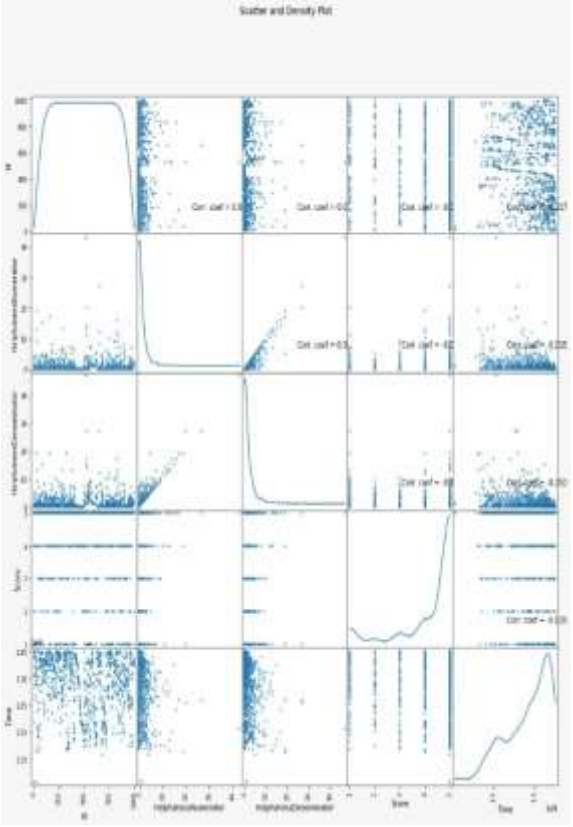
3.7.2Finding the count of sentiment words:

As our output is a bar-graph that outputs the emotions of customers based on the words used in their reviews, our first task is to find out count of those words. We used several packages to help with the count. In the previous step we removed all stop words and reviews are filtered so that can be used in this step.

IV. RESULT& DISCUSSION-

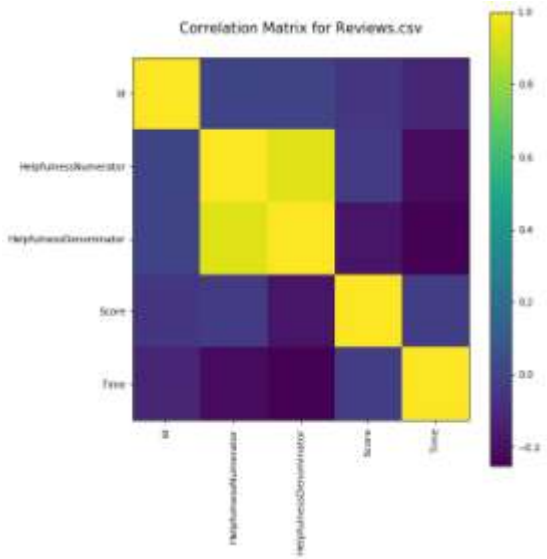


Bar graph



Heatmap

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 521-525 www.ijaem.net ISSN: 2395-5252



Scatterplot

Train Accuracy is 88%
Test Accuracy is 86%

V. CONCLUSION& FUTURE WORK-

This paper presents an experimental study along with a proposed model through SVM algorithm on different datasets of Product reviews to measure the polarity of reviews whether positive or negative and words related to the products such as good, bad, excellent, super hit. The performance resulting models are tested to measure accuracy of Support Vector Machine learning algorithm. Finally, the Support Vector Machine classification algorithm is achieved high accuracy and found better one than others 1.The ML approaches proffered the good outcomes to categorize product reviews. SVN got 98.170% accuracy and NB got 93.54% accuracy for Camera—related Reviews.

2.The approach utilizes the SVM, which encompasses several key parameters that are required to be set properly

3. Thus, the SVM renders BEST accuracy in theclassification.

4.The key aim is to analyze a large amount review by using amazon dataset which are already labeled.

REFERENCES –

[1]. TangFeilong,LuoyiFu,YaoBin,WenchaoXu.Aspectbasedfine-grainedsentimentanalysis for online reviews. Inf Sci.2019;488:190–204.

[2]. JagdaleRS,ShirsatVS,DeshmukhSN.SentimentanalysisonproductreviewsCognitive Informatics and Soft Computing, Springer, Singapore, pp. 639–647,2019.usingmachine learning techniques.

[3]. Vanaja S, Belwal M. Aspect-level sentiment analysis on e-commerce data. In: international conference on inventive research in computing applications (ICIRCA), IEEE, pp. 1275–1279, 2018.

[4]. Sentiment Analysis and Opinion Mining by Bing Liu (http://www.cs.uic.edu/~liub/FBS/Sentiment Analysisand-OpinionMining.html)

[5]. Y. Khaliq and M. Khaleeq, "Modeling Sentiment Terminologies : Target Based Polarity Phenomena," pp. 700–705, 2016.

[6]. N. D. Valakunde and M. S. Patwardhan, "Multi-aspect and multi-class based document sentiment analysis of educational data catering accreditation process," Proc. - 2013 Int. Conf. Cloud Ubiquitous Comput. Emerg. Technol. CUBE 2013, pp. 188–192, 2013.

[7]. S. M. Altarrazi and S. Sasi, "Tweeple's microblogs on illegal immigration in USA," Int. Conf. Electr. Electron. Optim.Tech. ICEEOT 2016, pp. 2011–2018, 2016