Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques

Shweta Rana
ASET-IT, Amity University Uttar Pradesh
Noida, India
shweta11rana@gmail.com

Archana Singh
ASET-IT, Amity University Uttar Pradesh
Noida, India
archana.elina@gmail.com

***Abstract*— In the recent few years several efforts were dedicated for mining opinions and sentiment automatically from natural language in online networking messages, news and business product reviews. In this paper, we have explored sentiment orientation considering the positive and negative sentiments using film user reviews. We applied the technique Naive Bayes' classifier.). We have performed the sentiment analysis on the reviews using the algorithms like Naive Bayes, Linear SVM and Synthetic words. Our experimental results indicate that the Linear SVM has provided the best accuracy which is followed by the Synthetic words approach. The result also evaluate that the highest accuracy rate is of drama.**

Keywords: online reviews, Naïve Bayes, SVM, Opinion Mining, Sentimental analysis.

I. INTRODUCTION

The expanding web, social networking increases and people began to share data through various types of online networking. The expansive number of reviews makes it available for the makers to take responses of potential customers. They confront extra challenges in seeking after extensive variety of products, exchanged on online sites. It is valuable to make a framework to identify markers of execution of an item, and area particular measurements, to compress the sentiments got from the extensive measure of reviews, in a few positive and negative angles.

The objective of the certain data around a organization that a business provides for a customer. They all around have a positive or negative review associated with them which occasionally impacts the purchasers from the product's/organizations constant experience. The subjective points of view are a social occasion of opinion, review, proposals, remarks, appraisals and individual experience shared by various customers on the different sites, along with the valuable information. Sentimental Analysis is a kind of natural language processing for finding sentiments of people for a specific product or thing. Opinion mining, includes building a framework to gather and inspect sentiments about the product made in blog entries, remarks, surveys or reviews. Sentiment Analysis can be valuable in a few ways.

Sentimental Analysis is the opinion mining which is used for recognizing the content on the web. It is only to get the real voice of individuals for particular product, services, films, news, issues and so on. The aim of sentimental analysis is to decide the sentiment of a person regarding a few subjects. Sentimental Analysis should be possible at three different levels as, sentence level, record level and component or attribute level. The mentality might be a man's judgment for the specific product. Criticism or feedback is extremely important for customer and maker in light of the fact that most of the general population buy or deal the product on the web. Single buyers may need the suppositions of existing clients for the product some time recently obtaining it.

In this paper, we analyze the three steps.

Data Collection and Preprocessing: Identify the user reviews and detect the opinion. In this unnecessary words are omitted.

Mining: Two classification algorithms are used i.e. Linear SVM and Naïve Bayes on the movie dataset. A model is trained to evaluate the performance.

Result: The result is obtained and describes the accuracy rate of the different genre of the movie and display the end result.

In this paper, the several techniques are used to perform the sentimental analysis and sentimental orientation. The paper is distributed in different sections. Section II discuss about the related studies about the user reviews, techniques of mining and sentimental analysis. Section III describe about the methodology of the proposed work, dataset and techniques used. Section IV discuss about the experiment that performed in this paper and its subsequent result. Section V describe the conclusion and future scope of the paper.

II. RELATED WORK

Sentiment analysis is fundamentally used to express opinion of the distinctive individual. Current cutting edge in conclusion characterized classes into two class's positive, negative. This section describes the literature study regarding the sentimental analysis, techniques applied on user reviews.

Dhanashri Chafale and Amit Pimpalkar proposed that It gathers all the customer reviews for various items which contain the certainties and suppositions. The subjective sentences are ordered into three classes as positive, negative and neutral utilizing the machine learning technique based hierarchical clustering. The Plutchik's wheel of feeling is utilized to further characterize the positive and negative sentences into various emotions. For this machine learning based neural network is used. Association rule based approach is used to classify product feedbacks according to sentiments and the corpus is developed in hierarchical form. Finally the fuzzy logic is used to prediction and gives the best product. [7]
Ion SMEUREANU, Cristian BUCUR work on the technique for analysis of sentiments, reviews made by users after watching movies. Grouping of reviews in both negative and positive classes is done in light of a Naive Bayes algorithm. To enhance classification we first remove insignificant words and presented in classification of words (n-grams). For n = 2 groups we achieved substantial improvement in classification. [4]
Yessenov Kua has used the comments on articles from Digg as our content corpora. We assessed the fitness of various feature selection and learning algorithms (Supervised and unsupervised) on the classification of comments as indicated by their subjectivity (subject/objective) and their polarity (Positive/Negative). The outcomes demonstrate that simple bag of words model can perform relatively good, and it can be further refined by the decision choice of features based on syntactic and semantic information from the text. [1]
G.Vinodhini examined about the work it is apparent that neither one of the classifications demonstrate reliably beats the other, diverse sorts of components have unmistakable distributions. It is additionally found that distinctive sorts of components and arrangement calculations are consolidated in a productive route so as to defeat their individual downsides and advantage from each other's merits, lastly upgrade the sentiment classification execution. [9]

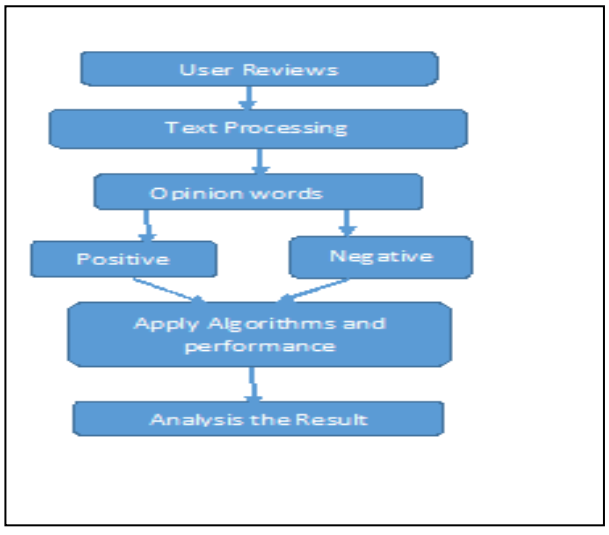

fig1 : Research Methodology

III. METHODOLOGY

For our tests, we worked with movies reviews. This area is experimentally advantageous in light of the fact that there are large on-line accumulations of such reviews Authors and Affiliations.

A. *Dataset*

For our paper, we worked with movie reviews. This domain is helpful because there are large on-line collection of user reviews and summarize there overall sentiment with positive and negative review. Our information source was the Internet Movie Database (IMDb) archive of the rec.arts.movies.review. For the work described in this paper, we focused just on positive and negative user reviews. We utilize dataset which contains 1000 negative and 1000 positive reviews. This dataset will be available on-line at ttp://www.cs.cornell.edu/individuals/pabo/-movie-review-data/ (the URL contains hyphens just around "review").

B. *Text Processing*

Text processing includes computer commands which conjure content, content changes, and cursor development, for instance to seek and supplant, position, produce a handled report of the content of, or filter a document or report of a content record. The content preparing of a consistent expression is a virtual altering machine, having a primitive programming dialect that has named registers (identifiers), and named positions in the arrangement of characters containing the content. Utilizing these the "Text processor" can, for instance, mark a region of text, and after that move it. The text processing a utility is a filter program, or filter. [10].

C. *Porter Algorithm*

The Porter stemming algorithm is a process for removing suffixes from words in English. Removing suffixes automatically is an operation which is particularly valuable in the field of Information Retrieval. In a typical IR environment a report is represented by a vector of words, or terms. Terms with a typical stem will for the most part have similar meaning, for examples
CONNECT
CONNECTED
CONNECTIONS
CONNECTIONS
Much of the time, the execution of an IR framework will be enhanced if the terms gatherings, for example, this are conflated into a solitary term. This might be finished by evacuation of the different additions –ED, - ING, - ION, IONS to leave the single term

CONNECT. What's more, the addition procedure will diminish the aggregate number of terms in the IR framework, and henceforth decrease the size and multifaceted nature of the information in the framework, which is dependably advantageous.[[6]

D. Linear Support Vector Machine

Linear Support vector Machine become one of the most prominent learning methods for solving the regression and classification problems. Linear Support Vector Machine works on large online dataset which are taken from the online sites and become popular because of its applications in text classification, word-sense disambiguation. Support Vector Machines (SVM) are another truthful learning strategy that can be seen as another method for get ready preparing classifiers considering polynomial capacities, extended premise capacities, neural systems. Support Vector Machines use a hyper-plane isolating plane to make a classifier. For issues that cannot be directly isolated in the space of information, this machine offers a probability to find an answer by making a improvement of the principal information space into a high dimensional feature space, where a perfect separating hyper plane can be found.

E. Naïve Bayes

The Naive Bayes classifier is a likelihood classifier, in light of Bayes' hypothesis. Bayes' hypothesis determines scientifically the connection between likelihood of two occasions An and B, P(A) and P (B) and contingent likelihood of occasion A molded by B and occasion B adapted by A, P (A | B) and P (B | A). Consequently Bayes' equation is :

$$P(A/B) = \frac{P(B/A)\ P(A)}{P(B)}$$

This hypothesis empowers us to decide a contingent likelihood having the likelihood of opposite occasion and autonomous probabilities of occasions. In this way, we can gauge the likelihood of an occasion taking into account the case of its event. Along these lines, we can evaluate the likelihood of an occasion in view of the case of its event. For this situation, we assess likelihood that a record is sure or negative, in a specific setting, or the probability that an occasion to happen in the event that it was foreordained to be certain or negative.[4]

F. Evaluate the execution of calculation

We utilize two particular measures for data recovery IR frameworks to assess the consequences of calculation utilized: (accuracy) and the review, both contrasting the outcomes and pertinence. To express these ideas will be utilized precision table.

TABLE 1. Table of correctly classified reviews

	Relevant	Irrelevant
Detected Opinion	True pos (tp)	Pred. pos (Pp)
Undetected Opinion	Pred. neg (Pt)	True neg (tp)

Precision is the proportion of the effectively arranged removed feelings and all extricated sentiments, the rate of accurately characterized assessments from ordered ones:

$$\text{Precision} = \frac{tp}{tp+Pp}$$

Recall communicates the proportion of effectively grouped removed conclusions and characterized suppositions in information source, the percent of accurately arranged feelings from all assessments in a class:

$$\text{Recall} = \frac{tp}{tp+Pn}$$

Another assessment measure for calculation might be precision, communicating the rate. Right made arrangements, a weighted consonant mean of accuracy and review:

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+Pp+Pn}$$

We compute exactness of classifier, the review and accuracy for the two classes, preparing the calculation on 1000 sentences for every class of pre-grouping test illustrations and applying it on whatever is left of the remaining cases:
An answer for enhance the nature of the calculation is to take out inconsequential words for characterization. Algorithm initially grouped words without lexical substance, so that other than things, verbs, pronouns and descriptive words, are considered articles, relational words and pronouns without semantic quality.

IV. EXPERIMENT AND RESULT

The proposed research is completely experimental with a practical implementation described in detail. This segment depicts and shows the execution of our technique on user reviews and obtained result. In our experiment, The tool used is Rapid Miner. Below figures display the dataset of positive and negative reviews which are processed in the rapid miner.

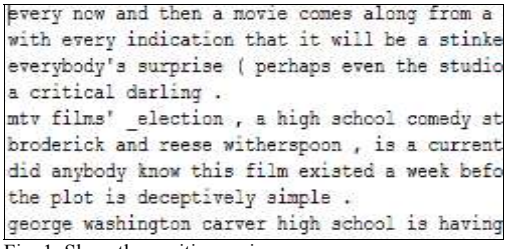

Fig. 1. Show the positive review

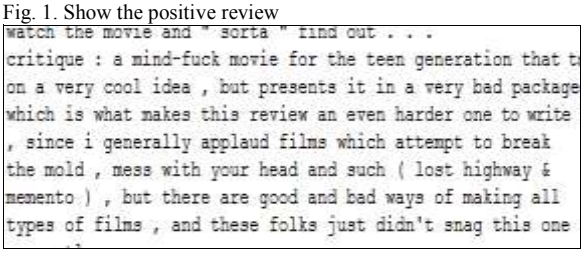

Fig. 2. Show the Negative Review

The below figure show that how these reviews are processed in the rapid miner by using process documents with files. Step to load the positive and negative data will be used in process documents from files and configure.

Fig. 3. Show the process documents from files.

The following below figure display the operators which are used in the process for further process such as tokenize, filter token, [3] stem(porter) and stopwords these are the text processing operators which are used in this step.

Tokenize splits the texts of a document into a series of words or tokens. We will use the 'non-letters' mode which will generate single word tokens but other options are available.

Filter Tokens we can filters tokens or words (in word mode) by length. We set the minimum to 4 and the maximum to 25. Obviously, depending on your context be careful not to exclude important words.

Stem (porter) stemming is a very important concept in natural language parsing. It allows one to reduce words to their base or stem. The aim of stemming is to reduce related forms of a word to a common base form. i.e fishing", "fished", "fish", and "fisher" to the base word, "fish". One of the most popular stemmer is the Porter stemmer.

Filter stopwords this operator removes common English words such as 'a' and 'the' etc. Word like these are very noisy unless removed.

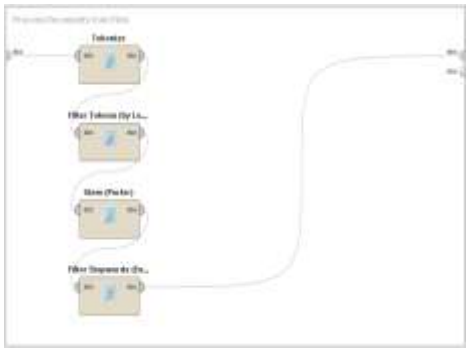

Fig. 4. Show the functions used in the process

The following figure displayed the new operator are discussed in this figure validation and store. Store simply output our word vector to a document and catalog of our picking. Acceptance quickly; cross-approval is a standard approach to survey the exactness and legitimacy of a factual model Our information set is separated into two sections, a preparation set and a test set. The model is prepared on the preparation set just and its exactness is assessed on the test set. This is rehashed n number of times. Stratified inspecting assembles irregular subsets and guarantees that the class dispersion in the subsets is the same as in the entire case set. Our selection is stratified.

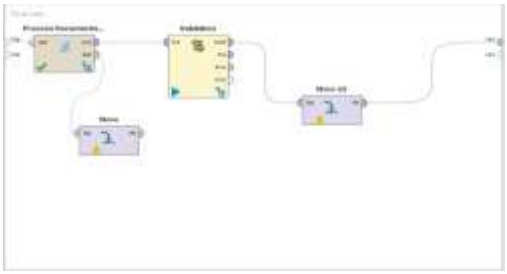

Fig. 5. Show the validation and store operators.

The below figures describe the use of linear SVM and Naïve Bayes. To quantify the exactness utilize the execution administrator to gauge the accuracy and review values.

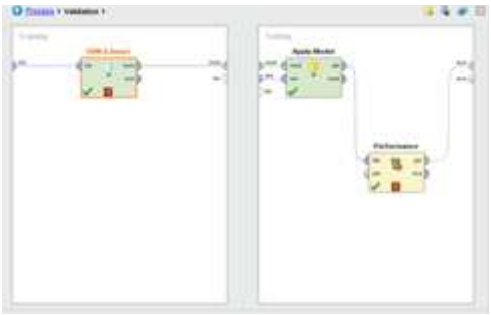

Fig. 6. Show the Linear SVM is used to analysis the result.

The following figures display the accuracy rate, precision and recall values by using linear SVM and Naïve Bayes of different genre of movies.

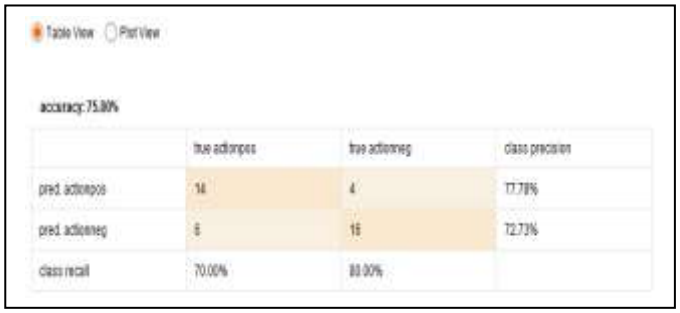

Fig. 7. Show accuracy rate, precision and recall value using SVM.

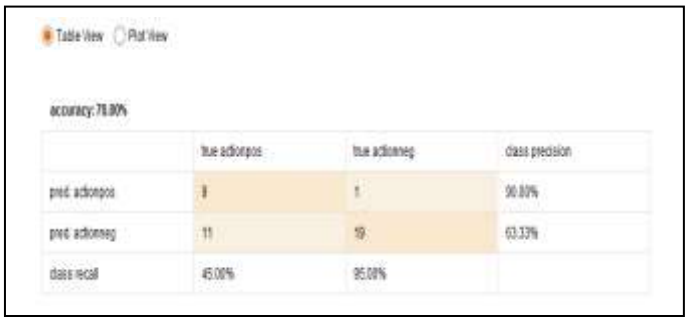

Fig. 8. Show accuracy rate, precision and recall value using Naïve Bayes.

Action Movies			
	true actionpos	true actionneg	class precision
Pred.actionpos	14	4	77.78%
Pred.actionneg	6	16	72.73%
class recall	70.00%	80.00%	
Adventure Movies			
	true adventurepos	true adventureneg	class precision
Pred.adventurepos	12	3	80.00%
Pred.adventureneg	8	17	68.00%
class recall	60.00%	85.00%	
Drama Movies			
	true dramapos	true dramaneg	class precision
Pred.dramapos	19	4	82.61%
Pred.dramaneg	1	16	94.12%
class recall	95.00%	80.00%	
Romantic movies			
	true romancepos	true romanceneg	class precision
Pred.romaancepos	18	6	75.00%
Pred.romanceneg	2	14	87.50%
class recall	90.00%	70.00%	

Fig. 9. Display the Precision and Recall value of movies using LSVM.

Action Movies			
	true actionpos	true actionneg	class precision
Pred.actionpos	9	1	90.00%
Pred.actionneg	11	19	63.33%
class recall	45.00%	95.00%	
Adventure Movies			
	true adventurepos	true adventureneg	class precision
Pred.adventurepos	14	3	66.67%
Pred.adventureneg	6	13	68.42%
class recall	70.00%	65.00%	
Drama Movies			
	true dramapos	true dramaneg	class precision
Pred.dramapos	16	4	80.00%
Pred.dramaneg	4	16	80.00%
class recall	80.00%	80.00%	
Romantic movies			
	true romancepos	true romanceneg	class precision
Pred.romaancepos	15	4	78.95%
Pred.romanceneg	5	16	76.19%
class recall	75.00%	80.00%	

Fig. 10. Display the Precision and Recall value using the Naive Bayes.

Table 2. Show the Accuracy Rate using LSVM.

Genre	Accuracy
Action	75.00
Adventure	72.50
Drama	87.50
Romantic	80.00

Table 3. Display the Accuracy Rate using Naive Bayes.

Genre	Accuracy
Action	70.00
Adventure	67.50
Drama	80.00
Romantic	77.50

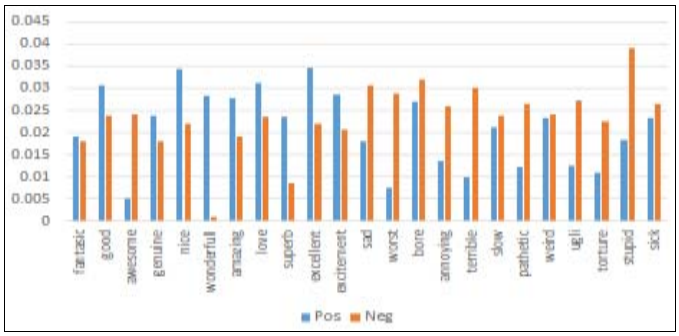

Fig. 11. Display the polarity of words.

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have analyzed the sentiment of user reviews about movies. We evaluated the linear SVM and Naïve Bayes algorithms on the movie user review dataset to check the accuracy of the different genre and opinions. On the basis of analysis the result demonstrated that the movie drama has the high accuracy rate among the different genre of the movies. The sentiment orientation describe that the user prefer to watch drama type of movie. The graph shows the polarity of the different words.

The future scope of the work is that we can explore our data to a wider genre of different products on social networking sites or e-commerce as day by day the user is moving online and they prefer buying stuff online so we can identify the accuracy rates of the products like books, games etc.

REFERENCES

[1]. Yessenov, Kuat, and Saša Misailovic. "Sentiment analysis of movie review comments." *Methodology* (2009): 1-17.

[2]. Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[3]. Verma, Tanu, and Deepti Gaur Renu. "Tokenization and Filtering Process in RapidMiner." International Journal of Applied Information Systems (IJAIS)–ISSN (2014): 2249-0868.

[4]. Smeureanu, Ion, and Cristian Bucur. "Applying Supervised Opinion Mining Techniques on Online User eviews." InformaticaEconomica 16.2 (2012): 81-91.

[5]. Kechaou, Zied, Mohamed Ben Ammar, and Adel M. Alimi. "Improving e-learning with sentiment analysis of users' opinions." Global Engineering Education Conference (EDUCON), 2011 IEEE. IEEE, 2011.

[6]. www2.bui.haw-hamburg.de/pers/ursula.schulz/astep/porter.pdf

[7]. Chafale, Dhanashri, and Amit Pimpalkar. "Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic. "International Journal of Computer Sciences and Engineering (IJCSE) 2 (2014): 14-18.

[8]. Wang, Weiping, and Yuanzhuang Zhou. "E-Business websites evaluation based on opinion mining." Electronic Commerce and Business Intelligence, 2009. ECBI 2009. International Conference on. IEEE, 2009.

[9]. Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." International Journal 2.6 (2012).

[10]. https://en.wikipedia.org/wiki/Text_processing