ISSN No:-2456-2165

A Comparative Study on Sentiment Analysis

Anbarasi M.
Asst. Professor, School of Computer science and Engineering
VIT Vellore, India

Ayush Patel
School of Computer Science and Engineering
VIT Vellore, India

Sarthak Panigrahi
School of Computer Science and Engineering
VIT Vellore, India

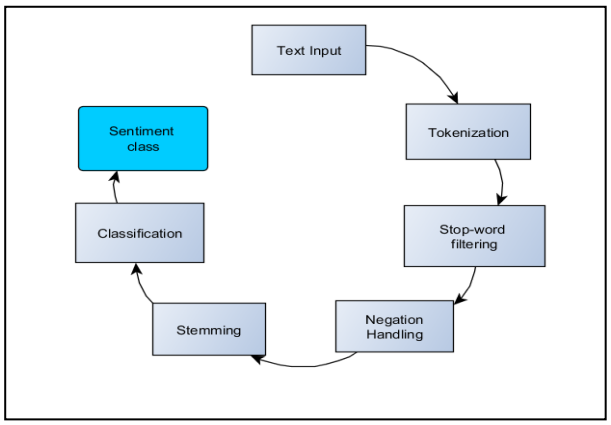
Abstract:- Collecting data from manifold sources of compiled consumer reviews, a framework can be constituted to compare and analyze customer opinions. Sentiment analysis is the process of identifying emotions and attitudes of a writer towards a particular topic, product, service etc. Customers place their reliance on the text reviews in form of experiences and opinions regarding any product available with an enterprise. The main process behind sentiment analysis is categorizing texts in order to find the writer's emotions. Sentiment analysis has a huge number of applications in a variety of fields. This technique can be very helpful to determine the popularity of a product in the online selling business and can play a major role in making decisions for companies such as amazon, flipkart, e-bay etc. It can also be used in social media sites like twitter, Facebook to identify the attitude of people towards different celebrities. This paper performs a comparative study on the different approaches used for sentiment analysis.

Keywords:- *k-means,sentiment analysis,SVM, Big data, Bayes classifier.*

I. INTRODUCTION

The market is mostly moved by people's attitude towards a certain product or an idea. Predicting how the market will behave can have a huge impact on the decisions made by the giant companies in the market which in turn will have effect on lives of the people.

Sentiment Analysis is the process by which we can identify the behavior of people towards something by extracting information from the things that they write about. For example, reviews of a certain product in online selling



sites. They have a huge impact in the buying behavior of the consumers. It is impossible to perform sentiment analysis by analyzing each text by human participation. The available content on the cyber space is basically unstructured data and we use opinion mining to refine such data. It aids the means of both users (consumers) as well as dealing organization -by hindering users to go through innumerable reviews manually

and by identifying new marketing opportunities, predicting sales tab and managing online reputation of any dealer [3].

Opinion mining is the computational study to process public expression in order to generate information upon user discretion. There are different approaches by which we can perform sentiment analysis with little or no human participation.

Sentiment analysis involves steps as follows

- Text Input-The input can be taken from a variety of platforms such as tweets, facebook posts, comments, blogs, product repots, news articles etc.
- Tokenization-It is the process by which the input text is differentiated into words, special characters, brackets etc.
- Stop word filtering- There are a lot of common words that occur frequently in the text such as on, at, when etc. which needs to be filtered to improve the performance
- Negation Handling-Negation words such as no, not, hardly, less, de-, dis- etc. can change the polarity of the opinionated words and can change the sentiment of the text. Hence it is very important to identify these words to determine the correct sentiment of the writer.
- Stemming-It is, in a general sense stripping of the suffix of the words to collapse distinct word forms to reduce the vocabulary size and improve efficiency.
- Classification-The basic function of sentiment analysis is to classify the different texts as positive or negative or neutral by using various algorithms which is going to be discussed in this paper.
- Sentiment Class-The text is classified into different classes known as the sentiment class which represents the sentiments of the writer which can be positive, negative or neutral.

Trust-worthiness of any seller or dealer highly determines their purchase estimate; trust is built with the feedback given by the user regarding any particular product or general service. Any enterprise makes use of this analysis to find out data about their products, services, their competitors and their general reputation put forth before the people. The neutral comments and reviews in the sentiment are often filtered out or ignored to focus on binary nature of classes involved, but

however using three classifiers improve the accuracy of the task at hand [4].

II. RELATED WORK AND BACKGROUND

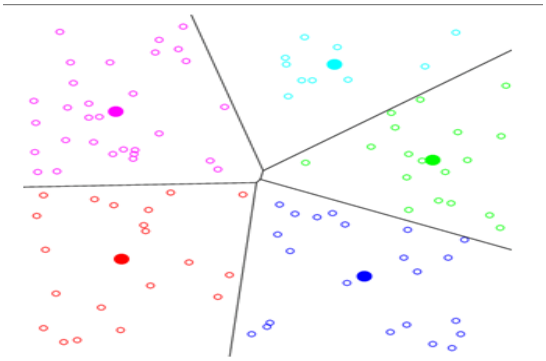
There are certain challenges faced in the procedure like firstly, reviews are gathered for varied resources over different domains so it is increasingly cumbersome to maintain a single dataset of nouns to satisfy all the test cases; hence needing a dedicated web-crawler to manage the issue. Secondly, feature extraction must be done to filter out comments or words that are irrelevant to the context and put our focus on object identification along with features related to the object or service; it requires a suitable algorithm to extract features correctly. Thirdly, we need to adjust for the irregularities occurring in polarities and public viewpoint from time to time and recognize the cause for such changing opinions. We also need to account for the misspelled words in reviews along with finding an accurate way to quantify the sentiment strength. The reviews with negations need to be dealt properly so as to not mislead the user.

Sentiment analysis proves to be a very promising approach to tackle smart city project. Research is carried out for sentimental analysis along with networking protocols which helps us realise parallelism, using multiple processors [1].However the complexity increased with this. We also need to prioritize customer reviews along with pre-existing parameters like pricing, ratings and other offer some other techniques employed are Word Alignment model(unsupervised) and lexicon-based model. We can combine the above along with lexical database like WordNet to form a semi-supervised model.

III. SENTIMENT ANALYSIS METHODS

Sentiment analysis plays a major role in the decision-making process of various companies. There needs to be a discussion on the different approaches that can be used to achieve this process so we can find which process can give accurate results efficiently.

A. K-means clustering

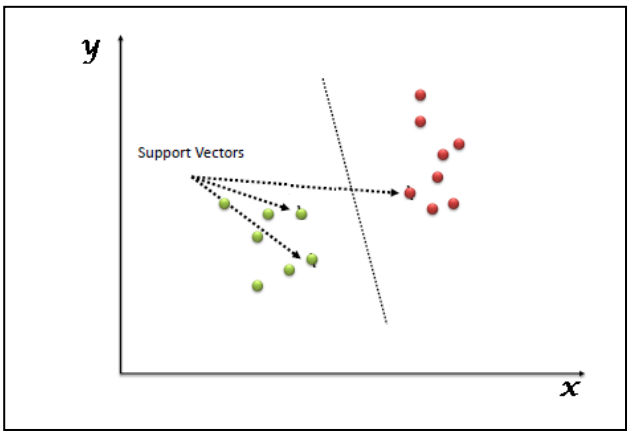


K means clustering method starts with a random number of cluster taken as initial means. We find the distances of each records with the cluster means. Assign each record to the cluster having the closest mean. We repeat this process until we reach a stable solution.

B. SVM- Support Vector Machines

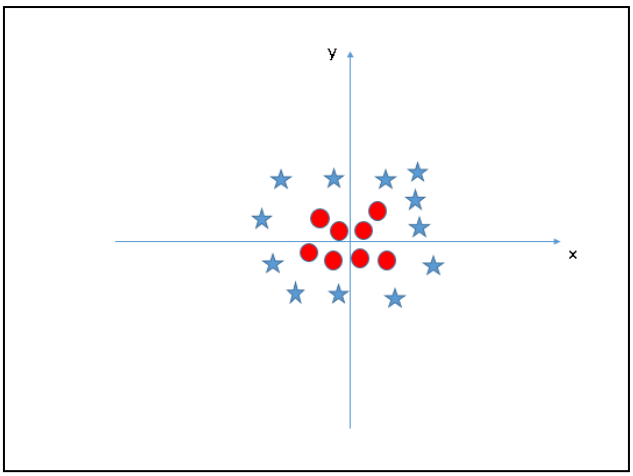
Support Vector Machines is a supervised machine learning algorithm which is mostly used for classification. In this, the data points are plotted in a n dimensional space known as support vectors. The support vectors are then differentiated by finding a hyper-plane which perform the classification the support vectors.

A significant advantage of SVM is that when we don't have a linear hyper-plane that divides the support vectors.



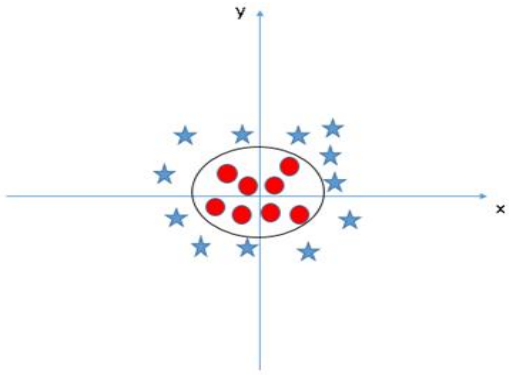
SVM can solve this by adding additional features by converting the lower dimensional input space to a higher dimensional input space. For example, in the above situation it uses a circle equation to differentiate the data points

$$z^2 = x^2 + y^2$$



In the original input space the hyper-plane would look like this Support Vector Machine works very well to handle large features and classify texts as positive or negative records with the cluster means. Assign each record to the cluster having the closest mean. We repeat this process until we reach a stable solution. or heads unless they are unavoidable.

ISSN No:-2456-2165



C. Naïve Bayesian classifier

The system gives the user the required information regarding any product of their discretion and it also aids them in decision making for any product. The initial steps are to retrieve or acquire a proper information base-unstructured data that matches the query of user; and then pre-processing the same datasets into more comprehensible format, using sophisticated methods hence demanding equal importance as classification [2].

We apply machine learning methods such as semantic orientation scheme of extracting relevant n-grams of text to be classified under bad or good response. The approach is based on semi supervised learning using WordNet as dataset dictionary for transforming features into target words. Thereafter, classification is carried out using training dataset. We can consider deterministic classifiers like nearest centroid classifier, SVM and k-nearest neighbour classifier, but for the give scenario, it is more suitable to use probabilistic classification like Naïve Bayes Algorithm, which is one of the fastest. It is one of the basic approaches for research in decision-making domain. We make use of multiclass classifier, where classes belong to different dealer sites.

$$P(ci/x) = \frac{P(x/ci)P(ci)}{P(x)}$$

where,

ci ∈ C = {c1,..,cn} (n is no. of e-commerce websites)

x ∈ X =(a1, a2,...,am) (m is number of target words)

P(x⁄ci) = probability or likelihood of target word x over class ci.

To make it more effective, it is backed by using Artificial Neural Networks and Back Propagation Network. A relative probabilistic model has to be considered along with randomness of variable to comply with the sample space. Aggregation (polarity of each review) and evaluation is carried out at a sentence-level to find the polarity of comments and draw conclusive result graphically and statistically.

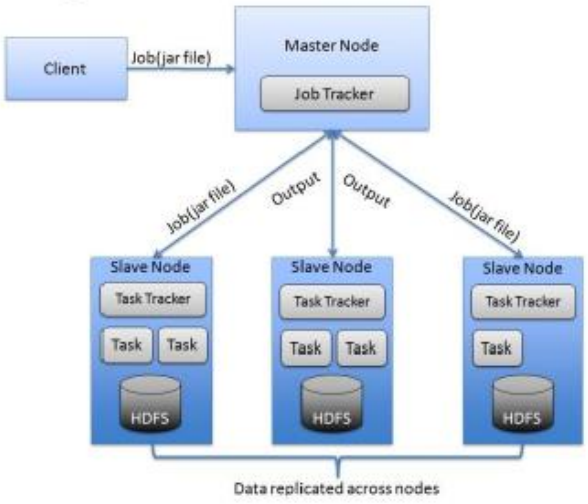
D. MapReduce in HDFS

In this method, for data collection and pre-processing we use the Apache flume framework. Any event from the source website/blog flows through the channel with the host process called agent. The data payload reaches a terminal sink from the client interface, from which it is accessed by the HDFS. The Hadoop architecture runs on a distributed system i.e. a master/slave system, on a cluster of manifold machines enabling it to analyze humongous amount of data. The master node controls access of client over the various duplicated data nodes. The analysis of classification is done using Map Reduce technique. The statistical visualization of data is done using the R language, where data is fetched from the HDFS. But, R has a restriction on physical memory size making the process long or cause to terminate, when working on very large datasets-hence we make the use of Rhadoop for such cases [5].

We need to evaluate the sentiment classification by separately parsing the tokenized data payload against the available trained dataset. We can analyze the same using a scale of custom range values (say, from -10 to 10) and a formula,

$$S = P_v + N_v - Ne_v$$

Where, sentiment value is sum of positive and negative values subtracted by the neutral value.



E. POS tagging

The motive is to decide upon the choice of specifically particularizing the product and the most suitable enterprise for the purchase. Initial phases include data collection from websites like amazon, e-bay, etc. and pre-processing of the feedback comments to filter out the stop words (that are frequently used and don't emphasize any feature of product) and focus on the more important terms. Then we have to segment the sentence in accordance with the different Parts of Speech like nouns (N), adjectives(A), verbs(V), prepositions(PREP), conjunction(CONJ), etc. Owing to ambiguity in certain words in POS tagging, we can incorporate fine-grained POS tags like Noun-plural, Pronoun-first person, etc. This method is essential to classify words into tags or classes like nouns, verbs and adverbs as well as has primary use text and opinion mining. The next task is to identify the feature from the opinion given; the features of the product like price, quality, durability, etc. is done from the adjectives along with the adverb, using the adjective to determine the sentiment value.

ISSN No:-2456-2165

Finally, the comments are classified under positive, negative or neutral by using the feature identification of words and checking against the set of words in the pre-defined dataset for the said case. Such NLP techniques are used along with text mining to determine the trust score by rating individual products of different enterprises.

F. Random Forest

Random forest is one of the best classification techniques. It can be used for classification as well as regression. In Random Forest the combined prediction value is used for classification from the different decision trees. In general, it is a mix of different decision trees put together to give a reliable solution.

The Random Forest is good for classification because it brings extra randomness in the model while creating the RF. In contrast with decision tree it does not select the best feature while separating attributes, it selects best features from a random subset of features to bring more randomness in the RF. This process creates a diverse model so that do not learn irregular pattern and over fit their training sets.

IV. CONCLUSION

All the approaches for sentiment analysis have several advantages and disadvantages. There cannot be a one approach that is better than every approach possible. The selection for classification should also be based on the dataset used for the training model and the purpose it is used for.

V. ACKNOWLEDGMENT

The paper undertaken in Vellore Institute of Technology, Vellore.

REFERENCES

[1]. Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K., “A tutorial review on Text Mining Algorithms”, International Journal of Advanced Research in Computer and Communication Engineering, Pages: 223-233, Vol. 1, Issue 4, June 2012.

[2]. I.Hemalatha, Dr. G. P Saradhi Varma, Dr. A.Govardhan“Preprocessing the Informal Text for efficient Sentiment Analysis” proceeding in International Journal of Emerging Trends & Technology in Computer Science 2012.

[3]. A. Ghose and P.G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining Text and reviewer characteristics," IEEE Trans. Knowl. Data Eng., vol.23, no. 10, pp. 1498-1512. Sept.2010.K. Elissa, “Title of paper if known,” unpublished.

[4]. M. Koppel and 1. SchIer (2005) "The Importance of Neutral Examples for Learning Sentiment". In IJCAI

[5]. Kumar, S., Singh, P., & Rani, S. (2016, September). Sentimental analysis of social media using R language and Hadoop: Rhadoop. In Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016 5th International Conference on (pp. 207-213). IEEE.

[6]. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[7]. Zainuddin, N., & Selamat, A. (2014, September). Sentiment analysis using support vector machine. In Computer, Communications, and Control Technology (I4CT), 2014 International Conference on (pp. 333-337). IEEE.

[8]. Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. Procedia Computer Science, 87, 44-49.

[9]. Riaz, S., Fatima, M., Kamran, M., & Nisar, M. W. (2017). Opinion mining on large scale data using sentiment analysis and k-means clustering. Cluster Computing, 1-16.

[10]. D.Mali, M.Abhyankar, P.Bhavarthi, K.Gaidhar, M.Bangare "Sentiment analysis of product reviews for ecommercerecommendation";Proceedings of 44th IRF International Conference, 29th November 2015, Pune, India, ISBN: 978-93-85832-59-8.

[11]. Ramanujam, R. S., Nancyamala, R., Nivedha, J., & Kokila, J. (2015, April). "Sentiment analysis using big data". In Computation of Power, Energy Information and Commuincation (ICCPEIC), 2015 International Conference on (pp. 0480-0484). IEEE.

[12]. Wajgi, R. D., & Bagul, S. J. (2016, February). Design feedback analysis system for E-commerce organization. In Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on (pp. 1-4). IEEE.

[13]. Hegde, Y., & Padma, S. K. (2017, January). Sentiment Analysis Using Random Forest Ensemble for Mobile Product Reviews in Kannada. In Advance Computing Conference (IACC), 2017 IEEE 7th International (pp. 777-782). IEEE.