ISSN:0975-9646

Social Media Analytics for E-commerce Organisations

Monish Narwani [1],Sanjay Lulla [2],Vivek Bhatia [3],Rishi Hemwani [4],Prof. Gresha Bhatia [5]
Dept. of Computer Science, Vivekanand Institute of Technology, Chembur

***Abstract*—Social media analytics plays an important role in e-commerce for retrieving the useful information of a product or service. Sentiment Analysis has become the key function of social media analytics. Opinion Mining is to used to analyze the polarity of sentiment expressed in data .Taking the data flood from online social media, in its many forms, and transforming it into useful knowledge for strategic decision making is the backbone of this paper. The paper proposes a model for doing customer review analytics on social media using big data for improving target advertising and improved business decision making.**

***Keywords*- big data, social media, sentiment analysis, clustering**

I. INTRODUCTION

The term "Big data" is used for huge volume of data sets whose size is so large that a normal software tool cannot collect, arrange and process it within a certain time limit.[1] 3Vs (volume, variety and velocity) are three basic blocks of big data. Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the frequency of data processing. An Opinion is a judgment or belief a majority of people. Sentiment analysis, is a natural language processing tool to find public mood about a product or topic. The tool used for Opinion Mining processes a collection of search results for a given product,generating product attributes (quality, features etc.) and aggregating opinion[12].OM is automatic extraction of knowledge from others opinions on a particular topic/problem. Opinion Mining is beneficial for strategizing organization's marketing campaigns by studying the purchasing patterns of the people of particular region which helps the organization to get the insights of trending products.[2]

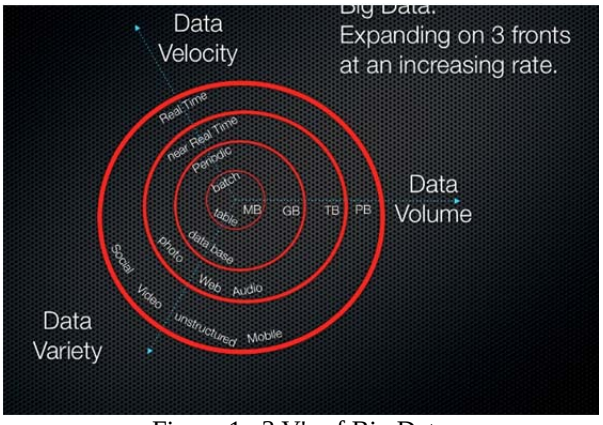

Figure 1. 3 V's of Big Data

II. LITERATURE STUDY

A. *Sentiment Analysis using product review data for Amazon[5]*

This paper tries to solve a major problem of sentiment analysis,i.e. categorization of polarity. The paper proposes an algorithm which uses a mathematical approach in order to compute score for a particular review. The approach used in the review and searches for the negative word or phrase and base on frequency of the negativity assigns a corresponding score to the review. The author performs two experiments on the proposed algorithm i.e. sentence and review level sentiment categorization.

B. *How Big Data helped increase Walmart sales turnover [6]*

Walmart was the world's largest retailer in 2014 in terms of revenue. Big Data became popular in the industry. Walmart uses data mining to discover patterns in point of sales data. Data mining helps were bought together or which products were bought before the purchase of a particular product. A familiar example of effective data mining through association rule learning technique at Walmart .

C. *Sentiment Analysis of Flipkart reviews using Naiive Bayes and Decision Tree Algorithm[7]*

In this paper, the authors have presented their study on the reviews collected from Flipkart and applied sentiment analysis algorithms on the reviews. The paper presents the entire process for sentiment analysis starting with the collection of reviews, preprocessing them using NLP, tokenizing the review and extracting the stopwords using stemming, transforming the tokens into target words using Wordnet, applying Naiive Bayes Algorithm to classify the target words as positive or negative and finally evaluating the review using Decision Tree Algorithm.

D. *Real Time Analysis for Measuring User's Influence on Twitter[8]*

Through this paper, the authors have done real time analysis on the tweets collected from the Twitter and presented the ways for measuring influence of user on the social media platform. The paper discusses the use of the Twitter by people in order to express their thoughts and their concerns about the society. The paper tries to explore the use of Twitter by the online marketeers to express their reviews regarding the products they buy. This can be helpful to the e-commerce organizations as well as the users who wish to purchase the same in the near future.

III. PROBLEM STATEMENT

In the last couple of years the social medium Twitter has become more and more popular. Since Twitter is the most used microblogging website with about 500 million users and 340 million tweets a day, it is an interesting source of information.Because Twitter is widely adopted through all strata, it can be seen as a good reflection of what is happening around the world. Among all that happens, the latest trends are most interesting for companies. The latest trends can be analyzed and when identified, reacted to. From a marketing point of view, these latest trends can be used to respond with appropriate activities, like product advertisements. Analyzing tweets can therefore be a goldmine for companies to create an advantage to competitors.Because Twitter is widely adopted through all strata, it can be seen as a good reflection of what is happening around the world. Among all that happens, the latest trends are most interesting for companies. The latest trends can be analyzed and when identified, reacted to. From a marketing point of view, these latest trends can be used to respond with appropriate activities, like product advertisements. Analyzing tweets can therefore be a goldmine for companies to create an advantage to competitors[19]. Our research work investigates below problem:

How social media sites can be used for the business advantage?

How organization can increase its product sales by making the use of social networking websites?

To obtain insights about the current trends in the market, what is the particular customer taste and preferences?

Apart from the organization's point of view, our research work investigates the user's dilemma while purchasing a product from the social media websites.[1]

Challenges:

1. Huge amount of data which needs to analyzed.
2. Presence of unstructured data which needs to be structured.
3. Absence of single tool for capturing and analyzing all types of data available on different social media platforms.
4. Sentiment Classification for people living in different parts of the world.

What British Say	What British Mean	What Other Understand
With the greatest respect..	I think you are an idiot	He is listening to me
That is very brave proposal	You are insane	He thinks I have courage
I'm sure it's my fault	It's your fault	Why do they think it's their fault
I almost agree	I don't agree at all	He agrees
I hear what you say	I don't want to discuss further	He accepts my point of view

Figure 2. Sentiment Analysis between British and EU people[8]

IV. PROPOSED WORK

The proposed architecture of the system comprises of four basic modules namely, Data Retrieval, Data Aggregation, Simulator Manager and User Interface. The Data Retrieval module focuses on retrieving the data sets using various APIs like Twitter, Facebook, Amazon, Pinterest, etc. These datasets are ingested into Data Aggregation module which collects the data in raw format and converts it into target words and stores them in data warehouse. The target words along with the proprietary data collected from the user interface are then passed on to analytics model to perform the sentimental analysis. The output of the analysis is stored in separate database for monitoring. The Simulator Manager on request of information from the client API extracts the data from database for monitoring and produce the data in the form of reports and graphical diagram which is helded by reporting component.

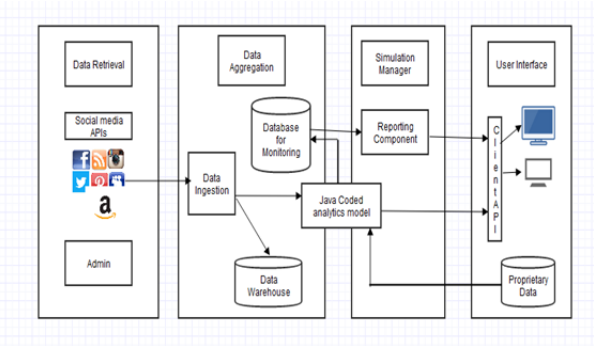

Figure 3. Proposed System Architecture

Further each module consist of sub modules which illustrate the entire working of the system in order to perform the sentimental analysis of the data collected from the social media APIs.

A. INFORMATION / DATA RETRIEVAL:

Information or Data Retrieval is responsible from extracting the information from the social media APIs and online datasets. The Data Retrieval module takes single product at a time and extracts information regarding the product from all the available social media and then with the help of regular expression performs web url scraping. [17]The url scraping produces the data in XML format while the web APIs and online datasets produce the data in textual format. Both the format being different from each other need to be brought into a single format which is informal textual reviews.

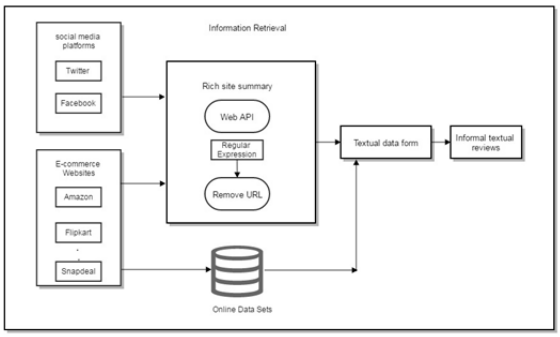

Figure 4. Data Retrieval

B. PROCESSING AND CLEANING

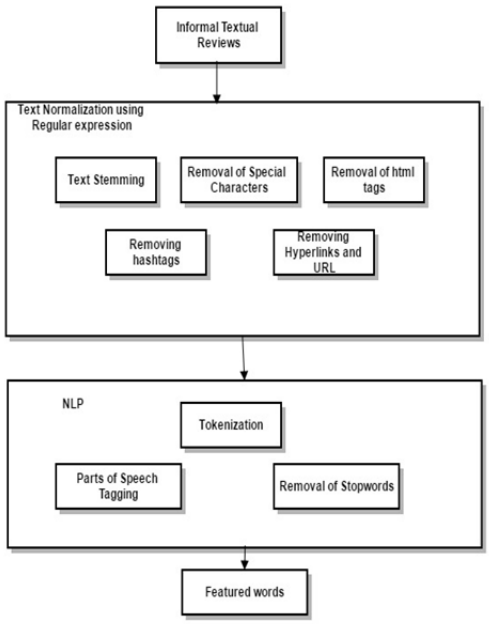

Figure 5. Processing and Cleaning

The informal textual reviews are processed further in order to remove the unwanted data with the help of Natural Language Processing (NLP). The NLP performs the text filtering by removing the html tags, removing the urls and hyperlinks, removing the special characters and splitting the sentence into different parts of speech[18]. Each important part of speech like noun, adjective, adverb, etc are counted and based on the word frequency, a set of feature words is prepared.

C. TARGET WORDS EXTRACTION

Using semi supervised learning, which is a Machine Learning Approach, the feature words are converted into target words with the help of WordNet which uses lexicon based data dictionary approach.[19]

Figure 6. Target Word Extraction

D. CLASSIFICATION OF WORDS

The target words which are obtained from the Word Net are further used to classify them into positive, negative and neutral words. Using Naiive Bayes Algorithm, the positive words are placed in positive class and negative words are placed in negative class which are augmented to prepare a sentence which is stored in knowledge base for future use. The augmented sentence is further provided to Naiive Bayes Algorithm.

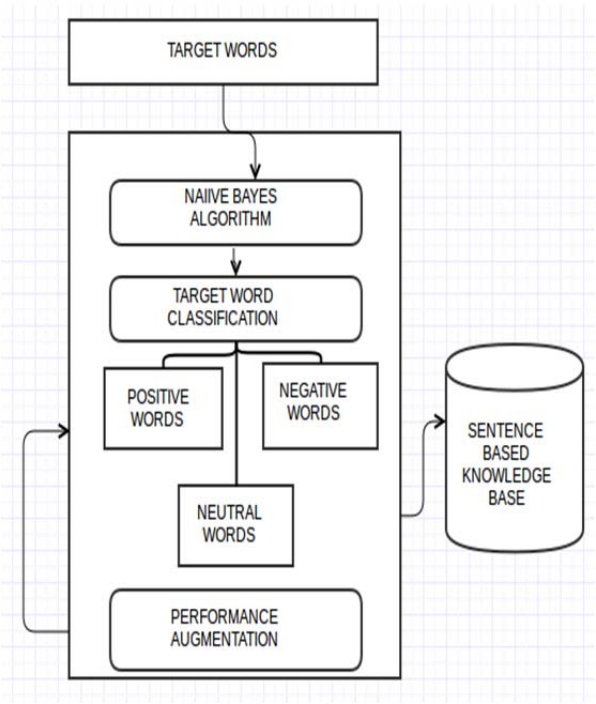

Figure 7. Classification of words

E. AGGREGATION AND EVALUATION

The sentence knowledge base is used to extract the sentiments for reviewing and classify the target word's polarity for the review. The result of both processes is stored in the sentence level knowledge base. This entire process completes the aggregation part. In the evaluation part, the polarities of the words are used to represent the data in statistical and graphical form using evaluation algorithm.

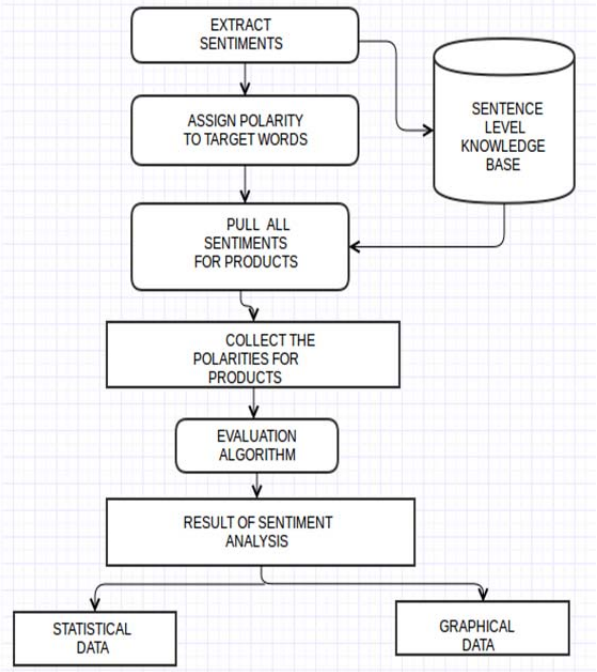

Figure 8. Aggregation and Evaluation

V. METHODOLOGY

A. PRELIMINARIES

Forming a group of people with similar opinion is called clustering. Clustering is a method through which the grouping of users datasets with similar data can be clustered into small sets. In this paper, we discuss two clustering algorithms: k-means and DBSCAN

1) k-means for opinion mining: K-means gets its name from k and means, where k is the number of clusters to group data into, and means because it effectively assigns the centroid of each cluster to the mean of the data-points in the cluster.[10]

2) DBSCAN for opinion mining: Density based spatial clustering of applications with noise (DBSCAN) is a generic clustering algorithm.[9] The algorithm works in two steps: Initially, it forms clusters using the points from all the possible neighbours. Further, it grows each cluster by adding nearby neighbouring points and iterates this till all the points are covered. In DBSCAN algorithm, the number of nodes need not be specified.

B. APPROACH

Corpus Based Approach:

A corpus based approach is based on assigning the emotional affinity to each word and then find the probabilistic score of each of them from the huge corpus. The corpus based approach is used to check the happiness factor of the words which helps to decide the positiveness or negativeness of the tweets, blogs or posts. The method is to pick up each tweet,post or blog from the knowledge base and assign a corresponding positivity factor based on the frequency of the positive or negative words and then decide the overall points to the entire tweet.

Dictionary Based Approach:

Dictionary based approach uses Sentiwordnet to obtain the proper synonym of the feature words for classification. The approach to this is picking up each adjective, adverb or verb and try to find the most closest meaning using the synsets. The Senti Wordnet contains the synonyms as well as the hyponyms of the words from the past history. The Dictionary based method assigns a number to each word according to the meaning which is done using unsupervised learning approach.

Feature based opinion mining:

Using Sentiment Analysis, a user review is graded at three different levels- document, sentence and feature levels. Grading a review at first level i.e. document, the entire text is classified positive or negative based on sentiments expressed in that text. At second level i.e sentence, each and every sentence from tweet, post or blog is classified as positive or negative by comparing the overall positivity factor obtained from the first level grading.The final level i.e. feature level summarizes the overall text as positive or negative. The major tasks of feature based opinion mining are - (1) identify features of products in review, (2) classify the review as positive or negative (3) provide the summary of the entire information and also store it for future references.

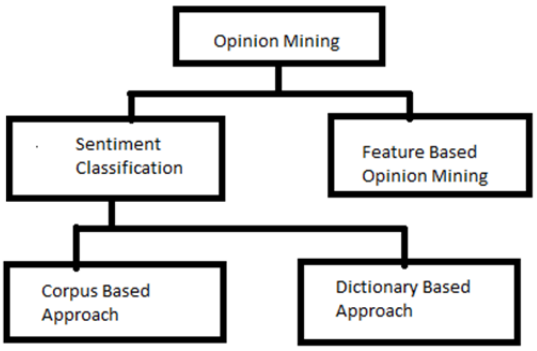

Figure 9. Classification of Methods

VI. CONCLUSION

This paper has focused on the process of opinion mining using text analytics and discussed briefly about the challenges in opinion mining.[3] It is much needed for a company to know the opinions of users about its products. Sentimental Analysis is used to analyse useful content in the text known as dataset.[4]By applying k-means and DBSCAN algorithm to the Twitter dataset, split the data into different categories by taking sentiment as a consideration factor, after classifying the opinions as clusters, an analysis of different products versus different sentiments process has been explained.

REFERENCES

[1] Aditya Patel , Hardik Gheewala , Lalit Nagla, ”Using Social Big Media for Customer Analytics”,978-1-4799-3064-7/14/$31. 00©20 14 IEEE

[2] Jeevanandam Jotheeswaran, Dr. S. Koteeswaran, “Sentiment Analysis: A Survey of Current Research and Techniques”, International Journal of Innovative Research in Computer and Communication Engineering,Vol. 3, Issue 5, May 2015

[3] S. Surya Kumari , G. Anjan Babu, “Sentiment on Social Interactions Using Linear and Nonlinear Clustering”, International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEI CB16)

[4] Xing Fang, Justin Zhan, “Sentiment analysis using product review data”,Journal of Big Data (2015) 2:5 DOI 10.1186/s40537-015-0015-2

[5] D.MALI,M.ABHYANKAR,P. BHAVARTHI, K. GAIDHAR,M.BANGARE,”SENTIMENT ANALYSIS OF PRODUCT REVIEWS FOR E-COMMERCE-RECOMMENDATION”,Proceedings of 44th IRF International Conference, 29th November 2015, Pune, India, ISBN: 978-93-85832-59-8

[6] How Big Data Analysis helped increase Walmart's-Sales-turnover?, https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

[7] Gurneet Kaur, Abhinash Singla, “Sentiment Analysis of Flipkart reviews using Naiive Bayes and Decision Tree Algorithm”, International Journal of Advanced Research in Computer Engineering & Technology, Volume 5, Issue 1,January 2016.

[8] Velissarios Zamparas, Andreas Kanavos and Christos Makris,”Real Time Analytics for Measuring User Influence on Twitter”,2015 IEEE 27th International Conference on Tools with Artificial Intelligence.

[9] D. Godfrey et al. “A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets”. In:(2014). arXiv: stat.ML/1408.5427.

[10] M. P. S. Bhatia and D. Khurana. “Experimental study of Data clustering using k-Means and modified algorithms”. In: International Journal of Data Mining and Knowledge Management Process 3.3 (2013), pp. 17–30.

[11] Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets & classification algorithms for sentiment classification. ISciences, 181(6), 1138-1152.
[12] Pak, A., Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC.
[13] Khan, K., Baharudin, B. B., & Khan, A. (2009, June). Mining opinion from text documents: A survey. In Digital Ecosystems and Technologies, 2009. DEST 09. 3rd IEEE International Conference on (pp. 217-222). IEEE.
[14] Java, A., Song, X., Finin, T. and Tseng, B. (2007).Why_We_Twitter:Understanding Microblogging Usage and Communities. Joint WEBKDD and SNA-KDD Workshop on Web Mining and Social Network Analysis,pp. 56-65.
[15] Java, A., Song, X., Finin, T. and Tseng, B. (2009). Why We Twitter:An Analysis of a Microblogging Community. Advances in Web Mining and Web Usage Analysis,Volume 5439, pp. 118-138.
[16] Ricardo Baeza-Yates, Berthier Ribeiro-Neto,"Modern Information Retrieval", ISBN-13: 978-0201398298,2013
[17] Haruna Isah, Paul Trundle, Daniel Neagu "Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis"2014 IEEE.
[18] 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.
[19] [Thesis] R. de Groot,"Data Mining for Tweet Sentiment Classification", Utrecht University, Faculty of Science, Department of Information and Computing Sciences, Netherlands.