**International Journal of Civil Engineering and Technology (IJCIET)**
Volume 8, Issue 10, October 2017, pp. 1829–1842, Article ID: IJCIET_08_10_184
Available online at http://http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=8&IType=10
ISSN Print: 0976-6308 and ISSN Online: 0976-6316



# ANALYZING CUSTOMER SENTIMENTS USING MACHINE LEARNING TECHNIQUES

**Magesh G**
School of Information Technology and Engineering, VIT University, Vellore, India

**Dr. P. Swarnalatha**
School of Computer Science and Engineering, VIT University, Vellore, India

**ABSTRACT**

*Nowadays in this digital world we see huge amount of data being created every day, Amazon is one of the leading e-commerce companies which possess such kind of data and Twitter is a famous micro blogging service where its users express their opinion on various topics as "tweets". We analyze these customer review data to help the customer to come to a conclusion for their purchases. The purpose of this paper is to help users who are trying to buy a new book by providing public opinion based on the Amazon user reviews by constructing an algorithm that can accurately classify sentiments in reviews and also to classify the tweets about those books. Main idea is that we can obtain this high accuracy on classifying sentiments in reviews using natural language processing and machine learning techniques such as bag-of- words, n-gram and Naive Bayes Classifier etc.,. Amazon review data for books for the past decade is itself more than 9GB it's more than billions on reviews from user around the globe to analyze it and return the most spoken feature about the product we are implementing hadoop technology to make it quick and feasible. This paper may also help the authors, publishers and researchers who want to know the public opinion of the book. The user sentiments will be broadly classified into three categories positive, negative and neutral. Top features of the book (product) will be used to make a user attractive word cloud.*

**Key words:** Text mining, sentiment classification, summarization, reviews.

**Cite this Article:** Magesh G and Dr. P. Swarnalatha, Analyzing Customer Sentiments Using Machine Learning Techniques. *International Journal of Civil Engineering and Technology*, 8(10), 2017, pp. 1829–1842.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=10

## 1. INTRODUCTION

The age of Internet has changed the way individuals express their perspectives. It is currently done through blog entries, online exchange gatherings, item audit sites and so on. Individuals rely on this client created substance all things considered. When somebody needs to purchase an item, they will look into its audits online before taking a choice. Online surveys are regularly our first port of call while considering items and buys on the web. Twitter a small

scale blogging website has turned into a selective apparatus for each updates over the world. It is a place where individuals accumulate and present their interests. This high variation information exhibit in locales has expanded the possibility of forecasts about particular results, without presenting the entire market machines. Programmed nostalgic investigation is exceptionally valuable to recognize and foresee present and future patterns. The measure of client created substance is too huge for a typical client to examine. So to computerize this, different feeling investigation strategies are utilized. The point of this venture is to concentrate client feeling from amazon surveys. Extricating such information and breaking down it is an issue in present time. The intense ascent and sudden impact of online networking as of late has put weight on associations to execute web-based social networking over their business.

There has been a few tasks and undertakings on nostalgic examination on amazon audit information and twitter tweets like Jansen have analyzed microblogging and its effect on different brands. In any case, we lac framework where numerous informational collections have been utilized to discover the client conclusions of an item. Most broadly utilized approach is utilizing twitter information in view of its programming interface usefulness. A portion of the activities likewise exist with amazon audit information yet very few incorporate numerous datasets from different areas. Likewise there aren't numerous wistful examination ventures which incorporates hadoop structures in it. This venture is done in way that expansive datasets are taken care by hadoop module and littler datasets are broke down by typical techniques.

The approach includes utilization of accumulation of item based dataset from various E-trade locales like amazon.com, twitter.com and so on. The surveys are gathered on items like books, fuel books and so on. The target of the work is to break down and anticipate item based audits by ordering them as positive, negative and impartial by utilizing calculations like credulous bayes and pack of words. Since information is about item audits that are unstructured, we perform pre-handling, removes highlights on to which remarks are made.

## 2. RELATED WORKS

Numerous analysts have worked in the field of supposition examination, every one proposing better approach for showing signs of improvement proficiency from machine learning approaches. A LSA to distinguish item highlight supposition words which are required to pick adjust sentences to wind up noticeably an outline of survey, with enabling just chose elements to demonstrate the final products, accordingly, decreasing genuine size of rundown [10]. In [11] creator discusses the particular issues inside conclusion investigation field which incorporates; record level, sentence level, highlight level, similar feeling and supposition dictionary issue. Bo string [1] considers arranging records not by subject, but rather by general feeling, finishing up whether an audit is sure or negative.

Surveys are changed over to straightforward choice by making utilization of methodologies, for example, gullible bayes, support vector machine by at first checking the quantity of positive and negative words in an archive. Since assessments are not generally coordinate e.g. "the nokia telephone is great" additionally it can be a near sentiment like "nokia telephone has preferred battery life over samsung". There exists three levels at which assessments are arranged: sentence level, record level, and highlight level [12]. At sentence level, subjective and target feelings exist, at archive level, a report is characterized in view of general assessment communicated by supposition holder. In any case, at highlight level, properties of items are thought about, which gives order inside and out.

Supposition Analysis [1] is the most unmistakable branch of normal dialect preparing. It manages the content arrangement keeping in mind the end goal to decide the expectation of the creator of the content. The goal can be of profound respect (positive) or feedback (Negative) sort. This paper shows a correlation of results got by applying Naive Bayes (NB) and Support Vector Machine (SVM) grouping calculation. These calculations are utilized to group a nostalgic survey having either a positive audit or negative audit. The dataset considered for preparing and testing of model in this work is marked in light of extremity motion picture dataset and a correlation with comes about accessible in existing writing has been made for basic examination.

Another strategy for estimation examination in Facebook that, beginning from messages composed by clients, underpins: (i) to separate data about the clients' assessment extremity (positive, impartial or negative), as transmitted in the messages they compose; and (ii) to display the clients' standard slant extremity and to identify critical passionate changes [2]. We have actualized this strategy in SentBuk, a Facebook application additionally exhibited in this paper. SentBuk recovers messages composed by clients in Facebook and orders them as indicated by their extremity, demonstrating the outcomes to the clients through an intelligent interface. It additionally bolsters passionate change identification, companion's feeling discovering, client arrangement as per their messages, and insights, among others. The characterization strategy actualized in SentBuk takes after a half and half approach: it consolidates lexical-based and machine-learning systems. The outcomes got through this approach demonstrate that it is attainable to perform feeling investigation in Facebook with high exactness (83.27%). With regards to e-learning, it is extremely helpful to have data about the clients' suppositions accessible. On one hand, this data can be utilized by versatile e-learning frameworks to help customized learning, by considering the client's enthusiastic state while prescribing him/her the most reasonable exercises to be handled at each time. Then again, the understudies' suppositions towards a course can fill in as criticism for educators, particularly on account of web based realizing, where eye to eye contact is less continuous. The handiness of this work with regards to e-learning, both for educators and for versatile frameworks, is depicted as well.

Sentiment Analysis (SA) is a continuous field of research in content mining field. SA is the computational treatment of feelings, estimations and subjectivity of content. This review paper handles a complete diagram of the last refresh in this field. Many as of late proposed calculations' improvements and different SA applications are examined and exhibited quickly in this review. These articles are classified by their commitments in the different SA procedures. The related fields to SA (exchange learning, feeling discovery, and building assets) that pulled in scientists as of late are talked about. The fundamental focus of this study is to give almost full picture of SA [3] strategies and the related fields with brief points of interest. The principle commitments of this paper incorporate the advanced arrangements of a substantial number of late articles and the representation of the current pattern of research in the opinion investigation and its related ranges.

In this paper [4], we propose a regulated term weighting plan in view of two fundamental components: Importance of a term in a report (ITD) and significance of a term for communicating slant (ITS), to enhance the execution of investigation. For ITD, we investigate three definitions in view of term recurrence. At that point, seven measurable capacities are utilized to take in the ITS of each term from preparing archives with class names. Contrasted and the past unsupervised term weighting plans began from data recovery, our plan can make full utilization of the accessible naming data to appoint suitable weights to terms. We have tentatively assessed the proposed technique against the best in class strategy. The test comes

about demonstrate that our technique beats the strategy and deliver the best precision on two of three informational collections.

The rise of Web 2.0 has definitely modified the way clients see the Internet, by enhancing data sharing, coordinated effort and interoperability [5]. Smaller scale blogging is a standout amongst the most well-known Web 2.0 applications and related administrations, similar to Twitter, have advanced into a functional means for imparting insights on all parts of regular daily existence. Therefore, smaller scale blogging sites have since turned out to be rich information hotspots for supposition mining and slant examination. Towards this course, content based assessment classifiers frequently demonstrate wasteful, since tweets regularly don't comprise of agent and linguistically reliable words, because of the forced character restrict. This paper proposes the arrangement of unique cosmology based strategies towards a more effective feeling investigation of Twitter posts. The curiosity of the proposed approach is that posts are not just described by a slant score, similar to the case with machine learning-based classifiers; however rather get an assessment review for each unmistakable idea in the post. Generally speaking, our proposed design brings about a more point by point examination of post conclusions with respect to a particular subject.

Because of the sheer volume of assessment rich web assets, for example, talk gathering, survey locales, online journals and news corpora accessible in computerized frame, a great part of the ebb and flow inquire about is concentrating on the territory of feeling investigation. Individuals are planned to build up a framework that can distinguish and characterize assessment or slant as spoke to in an electronic content. A precise strategy for foreseeing suppositions could empower us, to separate sentiments from the web and anticipate online client's inclinations, which could demonstrate profitable for monetary or advertising research [6]. Till now, there are couple of various issues prevailing in this examination group, to be specific, assessment order, include based characterization and taking care of invalidations. This paper shows a study covering the procedures and techniques in notion examination and difficulties show up in the field.

For instance, an article about a particular infection regularly comprises of various features, for example, manifestation, treatment, cause, analysis, visualization, and counteractive action [7]. In this manner, archives may have distinctive relations in light of various features. Capable inquiry instruments have been produced to enable clients to find arrangements of individual records that are most identified with particular catchphrases. Be that as it may, there is an absence of powerful investigation instruments that uncover the multifaceted relations of reports inside or cross the archive groups. In this paper, we show FacetAtlas, a multifaceted perception procedure for outwardly breaking down rich content corpora. FacetAtlas joins look innovation with cutting edge visual expository apparatuses to pass on both worldwide and nearby examples at the same time. We depict a few remarkable parts of FacetAtlas, including (1) hub factions and multifaceted edges, (2) an advanced thickness guide, and (3) computerized mistiness design upgrade for featuring visual examples, (4) intelligent setting switch between aspects. Furthermore, we exhibit the energy of FacetAtlas through a contextual analysis that objectives understanding training in the human services area. Our assessment demonstrates the advantages of this work, particularly in help of complex multifaceted information investigation.

We present a novel approach for consequently ordering the slant of Twitter [8] messages. These messages are named either positive or negative regarding a question term. This is helpful for customers who need to investigate the supposition of items before buy, or organizations that need to screen people in general slant of their brands. There is no past

research on grouping conclusion of messages on microblogging administrations like Twitter. We exhibit the aftereffects of machine learning calculations for characterizing the slant of Twitter messages utilizing inaccessible supervision. Our preparation information comprises of Twitter messages with emojis, which are utilized as uproarious marks. This kind of preparing information is plentifully accessible and can be gotten through computerized implies. We demonstrate that machine learning calculations (Naive Bayes, Maximum Entropy, and SVM) have precision over 80% when prepared with emoji information. This paper likewise depicts the preprocessing steps required keeping in mind the end goal to accomplish high precision. The fundamental commitment of this paper is utilizing tweets with emojis for far off administered learning.

The fast development in Internet applications in tourism has prompt a huge measure of individual surveys for travel-related data on the Web. These audits can show up in various structures like BBS, online journals, Wiki or discussion sites. All the more significantly, the data in these surveys is profitable to the two voyagers and professionals for different comprehension and arranging forms [9]. A characteristic issue of the mind-boggling data on the Internet, in any case, is data over-burdening as clients are just unfit to peruse all the accessible data. Inquiry works in web search tools like Yahoo and Google can enable clients to discover a portion of the surveys that they required about particular goals. The returned pages from these web indexes are still past the visual limit of people. In this examination, opinion characterization strategies were joined into the area of mining audits from travel online journals. In particular, we looked at three regulated machine learning calculations of Naïve Bayes, SVM and the character based N-gram demonstrate for feeling arrangement of the audits on travel web journals for seven well known travel goals in the US and Europe. Exact discoveries showed that the SVM and N-gram approaches outflanked the Naïve Bayes approach, and that when preparing datasets had an extensive number of surveys, every one of the three methodologies achieved exactness's of no less than 80%.

Sentiment mining [10] plans to utilize mechanized devices to recognize subjective data, for example, assessments, mentalities, and emotions communicated in content. This paper proposes a novel probabilistic displaying structure in light of Latent Dirichlet Allocation (LDA), called joint supposition/point demonstrate (JST), which identifies assumption and subject all the while from content. Not at all like other machine learning ways to deal with feeling grouping which frequently require named corpora for classifier preparing, the proposed JST demonstrate is completely unsupervised. The model has been assessed on the motion picture survey dataset to group the audit opinion extremity and least earlier data have additionally been investigated to additionally enhance the slant arrangement precision. Preparatory tests have indicated promising outcomes accomplished by JST.

Keeping in mind the end goal to cure this inadequacy, this paper introduces an exact investigation of notion arrangement on Chinese archives. Four component determination techniques (MI, IG, CHI and DF) and five learning strategies (centroid classifier, K-closest neighbor, winnow classifier, Naïve Bayes and SVM) are researched on a Chinese supposition corpus with a size of 1021 records. The exploratory outcomes show that IG plays out the best for nostalgic terms choice and SVM [11] displays the best execution for notion order. Moreover, we found that feeling classifiers are extremely reliant on spaces or themes.

We show another technique for conclusion order in view of removing and investigating evaluation gatherings, for example, ``very great'' or ``not horrendously clever''. An examination amass is spoken to as an arrangement of property estimations in a few undertaking free semantic scientific classifications, in light of Appraisal Theory. Semi-

computerized [12] strategies were utilized to fabricate a vocabulary of assessing descriptive words and their modifiers. We order motion picture surveys utilizing highlights in light of these scientific classifications consolidated with standard ``bag-of-words'' elements, and report cutting edge exactness of 90.2%. Also, we locate that a few sorts of examination have all the earmarks of being huger for assessment arrangement than others.

We exhibit that it is conceivable to perform programmed opinion characterization in the extremely loud area of client input information. We demonstrate that by utilizing vast component vectors in mix with include decrease, we can prepare straight help vector machines that accomplish high characterization precision [13] on information that present grouping challenges notwithstanding for a human annotator. We additionally demonstrate that, shockingly, the option of profound semantic investigation elements to an arrangement of surface level word n-gram highlights contributes reliably to grouping exactness in this area.

This paper acquaints an approach with notion investigation which utilizes bolster vector machines (SVMs) [9-14] to unite assorted wellsprings of conceivably relevant data, including se veral positivity measures for expressions and descriptive words and, where accessible, information of the theme of the content. Models utilizing the elements presented are additionally joined with unigram models which have been appeared to be successful previously (Pang et al., 2002) and lemmatized variants of the unigram models. Tests on film survey information from Epinions.com exhibit that cross breed SVMs which consolidate unigram-style include construct SVMs with those situated in light of genuine esteemed idealness measures get predominant execution, delivering the best outcomes yet distributed utilizing this information. Additionally tests utilizing a list of capabilities advanced with point data on a littler dataset of music audits hand-commented on for subject are likewise detailed, the consequences of which recommend that consolidating theme data into such models may likewise yield change.

Dealers offering items on the Web regularly request that their clients survey the items that they have obtained and the related administrations. As web based business is winding up increasingly famous, the quantity of client surveys that an item gets develops quickly. For a well-known item, the quantity of audits can be in hundreds or even thousands. This makes it troublesome for a potential client to peruse them to settle on an educated choice on whether to buy the item. It additionally makes it troublesome for the maker of the item to follow along and to oversee client feelings. For the maker, there are extra troubles in light of the fact that numerous dealer destinations may offer a similar item and the producer regularly creates numerous sorts of items. In this examination, we mean to mine and to abridge all the client surveys of an item. This rundown assignment is not quite the same as customary content outline since we just mine the components of the item on which the clients have communicated their assessments and whether the sentiments [3-15] are sure or negative. We don't condense the surveys by choosing a subset or revise a portion of the first sentences from the audits to catch the principle focuses as in the exemplary content outline. Our assignment is performed in three stages: (1) mining item highlights that have been remarked on by clients; (2) distinguishing feeling sentences in each survey and choosing whether every assessment sentence is certain or negative; (3) outlining the outcomes. This paper proposes a few novel procedures to play out these assignments. Our exploratory outcomes utilizing audits of various items sold online show the viability of the procedures.

Bo Pang et al consider the issue of ordering records not by theme, but rather by general supposition, e.g., deciding if a survey is certain or negative. Utilizing motion picture surveys as information, we find that standard machine learning methods definitively beat human-

delivered baselines. Be that as it may, the three machine learning techniques we utilized [16] (Naive Bayes, most extreme entropy classification, and support vector machines) don't execute also on conclusion classification as on conventional theme based order. We finish up by inspecting components that make the supposition classification issue additionally difficult.

Method for assumption examination utilizing hadoop which will prepare the tremendous measure of information on a hadoop bunch speedier progressively, Accuracy is observed to be 72.27 %. Utilization of Hadoop [17] guarantees the dispersed handling and it additionally brings down the get to time. Snide remarks are the ones which are extremely hard to distinguish. Tweets containing mocking remarks give precisely inverse outcomes inferable from the outlook of the creator. These are practically difficult to track. Likewise relying upon the setting in which a word is utilized, the understanding changes. For ex: "unusual" in "flighty plot" in setting of a land plot is negative while "capricious plot" in setting of a film's plot is certain.

Minqing Hu et al [18] discusses a recurrence based way to deal with recognizing the components in item audits. They arrange the thing phrases by recurrence and after that have diverse physically characterized settings to discover the elements (like lower cutoff, upper cutoff and so forth. Despite the fact that they can accomplish a decent workable framework with these strategies, their presumption that a component would dependably be a thing is not generally genuine. There can be multi word highlights like "optical zoom", "hot shoe streak" where one of the words is a modifier. They adopt a more all-encompassing strategy to the issue and utilize the assessment (notion) words to discover occasional elements and they played out the undertaking in three stages: (1) mining item highlights that have been remarked on by clients; (2) recognizing feeling sentences in each survey and choosing whether every supposition sentence is sure or negative; (3) abridging the outcomes

Alexander Pak et al discusses their attention on utilizing Twitter, the most well-known microblogging [19] stage, for the assignment of notion investigation. They demonstrate to consequently gather a corpus for assessment investigation and feeling mining purposes. They perform phonetic investigation of the gathered corpus and clarify found wonders. Utilizing the corpus, they construct a conclusion classifier that can decide positive, negative and nonpartisan opinions for an archive. Test assessments demonstrate that their proposed systems are effective and performs superior to already proposed techniques. In their exploration, they worked with English, in any case, the proposed procedure can be utilized with some other dialect

Efthymios Kouloumpis et al examines about the utility of etymological elements for distinguishing the slant of Twitter messages. They assess the convenience of existing lexical assets and additionally highlights that catch data about the casual and imaginative dialect utilized as a part of microblogging. [19-20]They adopt a managed strategy to the issue, yet use existing hashtags in the Twitter information for building preparing information. In this paper, they investigate one technique for building such information: utilizing Twitter hashtags (e.g., #bestfeeling, #epicfail, #news) to recognize positive, negative, and impartial tweets to use for preparing three-way assessment classifiers.

# 3. METHOD

## 3.1. Data Collection

The first step of this method is to collect the data on which sentimental analysis is performed to help the user, here we scrap data from amazon.com which can be said as datasets. These dataset will contain the metadata and review data in form of json data and review text is what we need.

***Sample Review:***

{

"reviewerID": "A2SUAM1J3GNN3B",

"asin": "0000013714",

"reviewerName": "J. McDonald",

"helpful": [2, 3],

"reviewText": "I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!",

"overall": 5.0,

"summary": "Heavenly Highway Hymns",

"unixReviewTime": 1252800000,

"reviewTime": "09 13, 2009"

}

where

reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B

asin - ID of the product, e.g. 0000013714

reviewerName - name of the reviewer

helpful - helpfulness rating of the review, e.g. 2/3

reviewText - text of the review

overall - rating of the product

summary - summary of the review

unixReviewTime - time of the review (unix time)

reviewTime - time of the review (raw)

Product description structure :

– asin - ID of the product, e.g. 0000031852

– title - name of the product

– price - price in US dollars (at time of crawl)

– imUrl - url of the product image

– related - related products (also bought, also viewed, bought together, buy after viewing)

– salesRank - sales rank information

– brand - brand name – categories - list of categories the product belongs to

## 3.2. Training the System

The bag-of-words makes a unigram model of the text by counting the each occurrence of the word and saving it for future use as features for text classifiers. After which we need to find

the subjectivity score of each word which is added upon to calculate the total subjectivity score of each text. This helps in the finding the sentiment i.e. positive word or negative word. To determine this subjectivity we need to determine the class probability of each word present in the bag-of-words.

Panda Dataframe can be used to find this value by utilizing it as data container (word in rows and class in columns). By simply dividing all elements of each rows by the total elements of that respective row we will get a Dataframe containing relative occurrences of each word in each class which is nothing but class probability of each word. Let's assign class 1 to be of negative and class 5 to be of positive. The word which occur only once will have 100% class probability. Therefore it's better to determine some cut off value i.e. words which occur less than the cut of value will not be included for calculation. 4 and 5 star reviews are labeled to be positive while1 and 2 are negative and 3 as neutral. With this classification we can determine with bag-of-words model whether a review is positive or negative with 60% accuracy (expected).

But unigram doesn't take grammar, position and context into account which will reduce the accuracy of the classification. Thus n-gram features is being included along where a features may have 2 or 3 words. This may increase the combinations of the words exponentially but still not all the combination makes sense so a small set of words which may alter the context of the feature id defined as dataset. This dataset will be helpful in forming the n-gram features. With this bag-of-words we can find the class of each word in the document and by adding the score of each word we can classify whether the given word is positive or negative.

By using Naïve Bayes model we take all the words present in the training dataset, the word which has not been appeared in the training set Laplacian smoothing can be applied. The training is done by iterating through all the training documents, a hash table can be built with the relative occurrence of each word per class is constructed.

## 3.3. Algorithm Implemented

### *Fetching Real-time Tweets*

- Create an Twitter app for utilizing its API
- Create access token and access token secret in the Twitter app dashboard
- Authenticate using consumer key, consumer token secret, access token and access token secret
- Provide the query to search and respective tweets count limit to search
- Receive and store the tweets
- Parse the tweets and remove links and other unwanted symbols form it
- Classify the tweets using tweety polarity check for individual words
- Calculate the tweet percentage for positive, negative, neutral classes
- Print the results of it along with 5 positive and negative tweets

### *Parsing Amazon Data*

- Remove unwanted Json items
- Store Rating and review posted by the user
- Training using Amazon Dataset

- Strip the '\n' character (new line) at the end of each review
- Add a (space) ' ' before and after :.,()[]:; - in order to avoid being wrongly interpreted
- Split the sentences into tokens using spaces
- Remove unwanted tokens like space, empty strings, punctuations, etc.
- Remove Stop words (the, of, this, etc.)
- Append the tokens for all reviews per class C
- Increment N, number of words in the current training set
- Store the relative occurrences of each word per class C as Vc in a hash table
- Classify the new document (test set) by using Vc
- For each reviews in the test set
- For each token in the review
- Find the class which has highest probability for the token from Vc
- Calculate the sentiment for each review
- Calculate the sentiment for entire test set

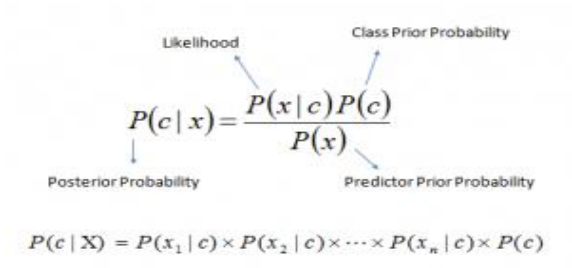


**Figure 1** Naïve Bayes..."

### 3.4. Visualization of Extracted Sentiments

The user has to select the product for which the sentimental analysis has to be done. After calculating the results it will be displayed to the user in various formats such as word cloud, charts and graphs etc. In this paper the features identified is used to create a wordcloud using PIL, Tkinter python libraries.

## 4. RESULTS

Once the tweets for the desired hashtag are collected and analyzed share of positive tweets, negative tweets and neutral tweets square measure calculated and displayed. 5 reviews of every category is additionally exhibited to the user. Once amazon knowledge is parsed and Naïve mathematician classification algorithmic rule is enforced higher than eightieth of accuracy over a category is measured. High option is additionally collected from the amazon user review knowledge by victimization hadoop streaming framework that is useful in formulating word cloud for the user.

Sentiment analysis is to classify the polarity of text in document or sentence whether or not the opinion expressed is positive, negative, or neutral. The most advantage of victimization Naïve mathematician is that it's simple to implement. We have a tendency to see here that Naïve mathematician is found to relinquish accuracy that's around eighty.5% result for this approach severally on the merchandise review dataset has been done. We have a

tendency to see that for text files that square measure overlarge in size take far more computation time. And conjointly word cloud is made with the options that has been extracted victimization these knowledge. We have a tendency to see that for text files that square measure overlarge in size take far more computation time. Automatic sentimental analysis is incredibly helpful to spot and predict current and future trends. Until currently opinion at feature level has been concerned however several limitations still exist which may be additional concerned.

**Table 1** Results

| Dataset | Class Neg – F1 Score(1000 test reviews) | Positive Tweets percentage | Negative Tweets Percentage | Neutral Tweets Percentage |
|---|---|---|---|---|
| The Martian | 0.83 | 9.85 | 0 | 90.14 |
| Fifty Shades of Grey | 0.94 | 33.33 | 66.66 | 0 |
| Hunger games | 0.43 | 28.95 | 13.15 | 57.89 |
| Goldfinch | 0.88 | 21.43 | 11.42 | 67.14 |
| Gone Girl | 0.90 | 3.03 | 0 | 96.97 |

## 5. CONCLUSIONS

The future scope of improvement are reviewing product primarily based opinions in multiple languages, Coping with downside of mapping slangs, Coping with sardonic opinions. Distinguishing comparative opinions and finding that among 2 product compared is best one and Coping with anaphora resolution like what's really being observed within the opinion

## REFERENCES

[1] Tripathy, A. Agrawal, And S. K. Rath, "Classification Of Sentimental Reviews Using Machine Learning Techniques," Procedia Comput. Sci., Vol. 57, Pp. 821–829, 2015.

[2] Z. H. Deng, K. H. Luo, And H. L. Yu, "A Study Of Supervised Term Weighting Scheme For Sentiment Analysis," Expert Syst. Appl., Vol. 41, No. 7, Pp. 3506–3513, 2014.

[3] W. Medhat, A. Hassan, And H. Korashy, "Sentiment Analysis Algorithms And Applications: A Survey," Ain Shams Eng. J., Vol. 5, No. 4, Pp. 1093–1113, 2014.

[4] Ortigosa, J. M. Mart??N, And R. M. Carro, "Sentiment Analysis In Facebook And Its Application To E-Learning," Comput. Human Behav., Vol. 31, No. 1, Pp. 527–541, 2014.

[5] Kontopoulos, C. Berberidis, T. Dergiades, And N. Bassiliades, "Ontology-Based Sentiment Analysis Of Twitter Posts," Expert Syst. Appl., Vol. 40, No. 10, Pp. 4065–4074, 2013.

[6] Vinodhini And R. Chandrasekaran, "Sentiment Analysis And Opinion Mining: A Survey," Int. J. Adv. Res. Comput. Sci. Softw. Eng., Vol. 2, No. 6, Pp. 282–292, 2012.

[7] N. Cao, J. Sun, Y. R. Lin, D. Gotz, S. Liu, And H. Qu, "Facetatlas: Multifaceted Visualization For Rich Text Corpora," Ieee Trans. Vis. Comput. Graph., Vol. 16, No. 6, Pp. 1172–1181, 2010.

[8] Go, R. Bhayani, And L. Huang, "Twitter Sentiment Classification Using Distant Supervision," Processing, Vol. 150, No. 12, Pp. 1–6, 2009.

[9] Lin And Y. He, "Joint Sentiment/Topic Model For Sentiment Analysis," Proc. 18th Acm Conf. …, Pp. 375–384, 2009.

[10] Q. Ye, Z. Zhang, And R. Law, “Sentiment Classification Of Online Reviews To Travel Destinations By Supervised Machine Learning Approaches,” Expert Syst. Appl., Vol. 36, No. 3 Part 2, Pp. 6527–6535, 2009.

[11] S. Tan And J. Zhang, “An Empirical Study Of Sentiment Analysis For Chinese Documents,” Expert Syst. Appl., Vol. 34, No. 4, Pp. 2622–2629, 2008.

[12] Whitelaw, N. Garg, And S. Argamon, “Using Appraisal Groups For Sentiment Analysis,” Proc. 14th Acm Int. Conf. Inf. Knowl. Manag. - Cikm ’05, P. 625, 2005.

[13] M. Gamon, “Sentiment Classification On Customer Feedback Data: Noisy Data, Large Feature Vectors, And The Role Of Linguistic Analysis,” Proc. 20th Int. Conf. Comput. Linguist., Pp. 841–847, 2004.

[14] M. Hu And B. Liu, “Mining And Summarizing Customer Reviews,” Proc. 2004 Acm Sigkdd Int. Conf. Knowl. Discov. Data Min. Kdd 04, Vol. 4, P. 168, 2004.

[15] T. Mullen And N. Collier, “Sentiment Analysis Using Support Vector Machines With Diverse Information Sources,” Conf. Empir. Methods Nat. Lang. Process., Pp. 412–418, 2004.

[16] B.Pang, L.Lee, S. Vaithiyanathan, “Thumbs Up? Sentiment Classification Using Machine Learning Techniques”, In: Proceedings Of Conference On Empirical Methods In Natural Language Processing.

[17] Javier Conejero, Peter Burnap, Omer Rana, Jeffrey Morgan, "Scaling Archived Social Media Data Analysis Using A Hadoop Cloud", 2013 Ieee Sixth International Conference On Cloud Computing, Vol. 00, No. , Pp. 685-692, 2013.

[18] Minqing Hu And Bing Liu, "Mining And Summarizing Customer Reviews". Sigkdd, 2004.

[19] Alexander Pak And Patrick Paroubek, “Twitter As A Corpus For Sentiment Analysis And Opinion Mining”, Lrec. Vol. 10. No. 2010. 2010.

[20] Kouloumpis, Efthymios, Theresa Wilson, And Johanna D. Moore. "Twitter Sentiment Analysis: The Good The Bad And The Omg!." Icwsm 11.538-541 (2011): 164.

[21] Xing Fang, Justin Zhan, “Sentiment Analysis Using Product Review Data”, Journal Of Big Data2015.

[22] Liu, Bing, “Sentimental Analysis And Opinion Mining”, Synthesis Lectures On Human Language Technologies 5.1(2012): 1-167

[23] Neethu M S, Rajasree R, “Sentiment Analysis In Twitter Using Machine Learning Techniques”, 4th Icccnt 2013.

[24] Maria Soledad Elli, Yi-Fan Wang Amazon Reviews, Business Analytics.

[25] Liu, Chien-Liang, Et Al. "Movie Rating And Review Summarization Inmobile Environment." Systems, Man, And Cybernetics, Part C: Ieee Transactions On Applications And Reviews,Volume 42 Issue 3, Pp.397-407,2012

[26] Feldman, Ronen. "Techniques And Applications For Sentiment Analysis."Communications Of The Acm 5,Volume 56 Issue4, Pp.82-89,2013.

[27] Wang, Min, And Hanxiao Shi. "Research On Sentiment Analysis Technology And Polarity Computation Of Sentiment Words." Progress Ininformatics And Computing (Pic), 2010 Ieee International Conference On. Vol. 1. Ieee, 2010

[28] Transl. J. Magn. Japan, Vol. 2, Pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, P. 301, 1982].

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

[30] Thirumalai, C.S., Senthil Kumar, M. Secured E-Mail System Using Base 128 Encoding Scheme (2016) International Journal Of Pharmacy And Technology, 8 (4), Pp. 21797-21806.

[31] Thirumalai, C.S., Senthil Kumar, M.Spanning Tree Approach For Error Detection And Correction (2016) International Journal Of Pharmacy And Technology, 8 (4), Pp. 5009-5020.

[32] Senthilkumar, M., Manikandan, N., Senthilkumaran, U., Samy, R. Weather Data Analysis Using Hadoop (2016) International Journal Of Pharmacy And Technology, 8 (4), Pp. 21827-21834.

[33] Chandrasegar, T., Senthilkumar, M., Silambarasan, R., Westphall, C.B. Analyzing The Strength Of Pell's Rsa (2016) International Journal Of Pharmacy And Technology, 8 (4), Pp. 21869-21874.

[34] Thirumalai, C., Senthilkumar, M., Vaishnavi, B. Physicians Medicament Using Linear Public Key Crypto System (2016) International Conference On Electrical, Electronics, And Optimization Techniques, Iceeot 2016, Art. No. 7755025, Pp. 1936-1939.

[35] Senthilkumaran, U., Manikandan, N., Senthilkumar, M. Role Of Data Mining On Pharmaceutical Industry-A Survey (2016) International Journal Of Pharmacy And Technology, 8 (3), Pp. 16100-16106

[36] Senthilkumar, M., Ilango, P. Analysis Of Dna Data Using Hadoop Distributed File System (2016) Research Journal Of Pharmaceutical, Biological And Chemical Sciences, 7 (3), Pp. 796-803.

[37] Senthikumar, M., Ilango, P. Big Data Optimization For Social Networking Tweet (2016) International Journal Of Soft Computing, 11 (5), Pp. 305-311.

[38] Senthilkumar, M., Ilango, P. A Survey On Job Scheduling In Big Data (2016) Cybernetics And Information Technologies, 16 (3), Pp. 35-51

[39] Senthil Kumaran, U., Nallakaruppan, M.K., Senthil Kumar, M. Review Of Asymmetric Key Cryptography In Wireless Sensor Networks (2016) International Journal Of Engineering And Technology, 8 (2), Pp. 859-862

[40] Senthilkumar, M., Nallakaruppan, M.K., Chandrasegar, T., Prasanna, S. A Modified And Efficient Genetic Algorithm To Address A Travelling Salesman Problem (2014) International Journal Of Applied Engineering Research, 9 (10), Pp. 1279-1288.

[41] Nallakaruppan, M.K., Senthil Kumar, M., Chandrasegar, T., Suraj, K.A., Magesh, G. Accident Avoidance In Railway Tracks Using Adhoc Wireless Networks (2014) International Journal Of Applied Engineering Research, 9 (21), Pp. 9551-9556.

[42] Sania Rahman, Magesh G: International Journal Of Professional Engineering Studies Volume Viii /Issue 4 / May 2017 Effective Ways Of Using Iot In Medical And Smart Health Care.

[43] Magesh G, And P Swarnalatha : Big Data And Its Applications: A Survey. Research Journal of Pharmaceutical, Biological and Chemical Sciences 03/2017; 8(2).

[44] Deepa, P, Magesh, G, Sanjana, N, Chowdhury, Ahana Roy: Framework for Reliable Re-Encryption in Cloud. Journal Of Chemical And Pharmaceutical Sciences 02/2017;

[45] Padmakumari P and Umamakeswari.A, Hybrid Statistical and Machine Learning Methods for Failure Prediction in Cloud, International Journal of Mechanical Engineering and Technology 8(8), 2017, p p. 714–719.

[46] Er. Harpal, Dr. Gaurav Tejpal and Dr. Sonal Sharma , Machine Learning Techniques for Wormhole Attack Detection Techniques in Wireless Sensor Networks, International Journal of Mechanical Engineering and Technology 8(9), 2017, pp. 337–348.

[47] Taran Singh Bharati and R. Kumar. Intrusion Detection System for Manet Using Machine Learning and State Transition Analysis. International Journal of Computer Engineering and Technology , 6 (1 2 ), 2015, pp. 01 - 08

[48] C.R. Cyril Anthoni, Dr. A. Christy, Integration Of Feature Sets With Machine Learning Techniques For Spam Filtering. International Journal of Computer Engineering and Technology, 2 (1), 2011, pp. 47-52

[49] Goverdhan Reddy Jidiga and Dr. P Sammulal, Machine Learning Approach To Anomaly Detection In Cyber Security With A Case Study Of Spamming Attack. International Journal of Computer Engineering and Technology, 4 (3), 2013, pp. 113-122