ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

Case Study on Online Reviews Sentiment Analysis Using Machine Learning Algorithms

I.Hemalatha[1], Dr. G.P.S.Varma[2], Dr. A.Govardhan[3]

Assistant Professor, Dept of IT, S.R.K.R. Engineering College, Bhimavaram, India[1]
Professor, HOD, Dept of IT, S.R.K.R. Engineering College, Bhimavaram, India[1]
Professor & Director of SIT, JNTU Hyderabad, India[3]

ABSTRACT: The main objective of the research paper is to prove the effectiveness of Analyzing social media data. Twitter is a valuable resource for data mining because of its prevalence and recognition by famous persons. In this paper we present a system which collects Tweets from social networking sites, we'll be able to do analysis using machine learning techniques on those Tweets and thus provide some prediction of business intelligence. Results of trend analysis will be display as tweets with different sections presenting positive, Negative and neutral.

KEYWORDS: Pre-processing, Sentiment analysis, Classification.

I. INTRODUCTION

Sentiment analysis has been an important topic for data mining, while the prevailing of social networking, more and more tweet analysis research focuses on social networking. Many people use Twitter as the media for sharing information, driven the wave of using Twitter as a communication tools, which makes sentiment analysis on Twitter become a valuable topic for further discussion. In this paper we introduce a sentiment analysis tool, it comprises three functions: sentiment analysis among Twitter tweets, finding positive, negative and neutral tweets from information resources. This tool focuses on analyzing tweets from those media sites, thus provide a way to find out technology trends in the future.

II. RELATED WORK

A. Social Network Analysis

Social network analysis is a methodology mainly developed by sociologists and researchers in social psychology. Social network analysis views social relationships in terms of network theory, while individual actor being seen as a node and relationship between each node are presented as an edge. Social network analysis has been define in [1] as an assumption of the importance of relationships among interacting units, and the relations defined by linkages among units are a fundamental component of network theories. Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, and sociolinguistics 1. In 1954, Barnes [2] started to use the term systematically to denote patterns of ties, encompassing concepts traditionally. Afterwards, there are many scholars expanded the use of systematic social network analysis. Due to the growth of online social networking site, online social networking analysis becomes a hot research topic recently.

B. Twitter

Twitter is an online social network used by millions of people around the world to be connected with their friends, family and colleagues through their computers and mobile phones [3]. The interface allows users to post short messages (up to 140 characters) that can be read by any other Twitter user. Users declare the people they are interested in following, in which case they get notified when that person has posted a new message. A user who is being followed by another user need not necessarily reciprocate by following them back, which renders the links of the network as directed. Twitter is categorized as a micro-blogging service. Micro-blogging is a form of blogging that allows users to send brief text updates or other media such as photographs or audio clips. Among variety of microblogging include Twitter, Plurk, Tumblr, Emote.in, Squeelr, Jaiku, identi.ca, and others, Twitter contains an enormous number of text

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

posts and grows quickly every day. Also, audience on Twitter varies from regular users to celebrities, company representatives, politicians [4], and even country presidents therefore provide a huge base for data mining. We choose Twitter as the source for trend analysis simply because of its popularity and data volume.

III. MACHINE LEARNING ALGORITHMS IN SENTIMENT ANALYSIS

The approach contains two major parts. Pre-processing and applying supervised learning algorithm. In pre-processing, we remove the data to increase data consistency then we get higher accurate results. The supervised learning algorithm classifiers are Naive Bayes and Support vector machine.

3.1 PREPROCESSING

The pre-processing is necessary because there are some words or expressions in the review don't return any meaning and by the presence of those words we cannot get the correct sentiment analysis. So by doing pre-processing we get higher accurate results.

In pre-processing we do the following steps.

Remove URLs We remove all the links from reviews. Because they don't have any meaning. So by removing the URLs we can get the result in less time. In reviews users include these URLs to give detailed information on which he get some idea. But for analysing we don't need to go through all this information so we remove these urls from our reviews.

Remove Repeated Letters The repeated letters in a word are also removed. For example we have good word in English, we don't have words like gooooooooood in English. In reviews these words are come because user like a product, then he give the review as gooooooooooood. In his point of view, he likes the product so much. So we remove the repeated letters and make it as good. And the word huuuuuuuungry make it as hungry, because in English we have a chance that a letter can come two times. So we remove the letters from a word which are occurring more than two times.

Remove Special Symbols We remove the symbols like : ; })] [({ etc. Because they don't have any meaning.

Remove Questions We remove the questions because by getting am answer only we can anlayse, so the questions are removed. For example how r u? is a review, it cannot get any sentimental meaning, so we remove that question.

By Pre-processing, we get the reviews which have a complete sentimental meaning so we can easily anlayse it.

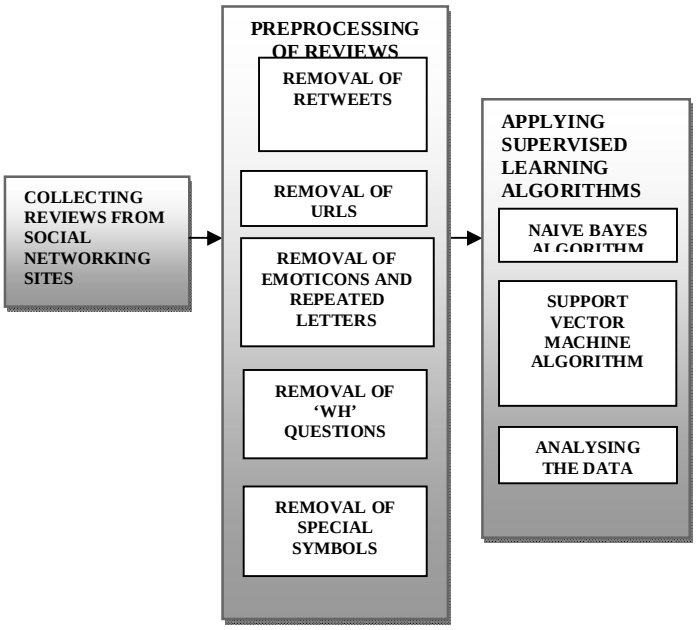


Fig 1: Architecture of Sentiment Analysis Using Machine Learning Algorithms

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

3.2 a)Naive Bayes algorithm

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

$C^* = \text{argma}_c P_{NB}(c|d)$

$P_{NB}(c|d) := (P(c) \sum^{m}_{i=1} P(f|c)^{n1(d)}/P(d)$……(1)

In the equation(1), f represents a feature and in(d) represents the count of feature fi found in tweet d. There are a total of m features. Parameters P(c) and P(fjc) are obtained through maximum likelihood estimates, and add-1 smoothing is utilized for unseen features.

• ***Proposed Naive bayes classifier***

Input: messages m= {m1, m2, m3.....mn},
Database: Naive Table NT
Output: Positive messages p= {p1, p2...},
Negative messages n= {n1, n2, n3....},
Neutral messages nu= {nu1, nu2, nu3...}
M= {m1,m2,m3.....}
Step: 1 Divide a message into words mi ={w1, w2, w3.....},i=1,2,....,n
Step 2: if wi NT Return +ve polarity and -ve polarity
Step 3: Calculate overall polarity of a word=log(+ve polarity)-log(-ve polarity)
Step 4: Repeat step 2 until end of words
Step 5: add the polarities of all words of a message i.e. total polarity of a message.
Step 6: Based on that polarity, message can be positive or negative or neutral.
Step 7: repeat step 1 until M NULL

b) Proposed Maximum Entropy classifier

Input: messages m= {m1, m2, m3....., mn},
Database: Naive Table NT
*Output***:** Positive messages p= {p1, p2...},
Negative messages n= {n1, n2, n3....}
Neutral messages nu= {nu1, nu2, nu3...}
M= {m1,m2,m3.....}
Step: 1 Divide a message into words
mi ={w1, w2, w3.....},i=1,2,....,n
Step 2: if wi NT return +ve polarity and -ve polarity
Step 3: Calculate overall polarity of a word=((+ve polarity)*log(1/+ve polarity))- ((-ve polarity)*log(1/-ve polarity))
Step 4: Repeat step 2 until end of words
Step 5: add the polarities of all words of a message i.e. total polarity of a message.
Step 6: Based on that polarity, message can be positive or negative or neutral.
Step 7: repeat step 1 until M NULL

c) *Proposed Support Vector Machine*

Input: messages m= {m_1, m_2, m_3.....m_n}

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

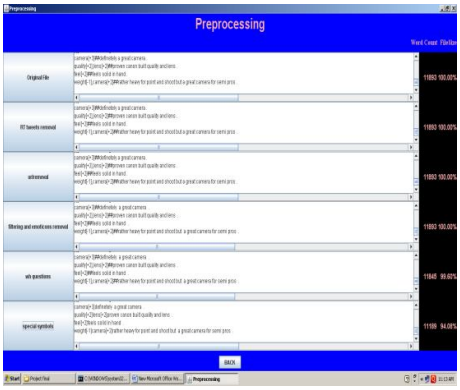
Vol. 2, Issue 2, February 2014

Database: sentistrength table S_T , negetorslist table NE_T , boosterslist table B_T .
Output: Positive messages p= {p_1, p_2...}
Negative messages n= {n_1, n_2, n_3....}
Neutral messages nu= {nu_1, nu_2, nu_3...}

M= {m_1, m_2, m_3.....}
Step: 1 Divide a message into words
m_i ={w_1, w_2, w_3.....},i=1,2,....,n
Step 2: if $w_i \in S_T$
return polarity p
Step 3: if $w_i \in NE_T$
return polarity and add it to p
Step 4: if $w_i \in B_T$
return polarity and add it to p
Step 5: Repeat step 2 until end of words
Step 6: add the polarities of all words of a message i.e. total polarity of a message.
Step 7: Based on that polarity, message can be positive or negative or neutral.
Step 8: repeat step 1 until M∈NULL

IV. DATA COLLECTION

The training data is an XML/Text file containing about 5400 tweets, each one labeled with sentiment polarity and the corresponding topics. The goal consists in providing automatic sentiment and topic classification for the test data, which is also available in Text file. In the pre-processing module we have to remove the unnecessary data of the message like RT tweets, removal of urls, filtering and emotion icons, removal of WHquestions, removal of special symbols. We measure the size after pre-processing. The results are as shown. classification algorithms have been used in sentiment and subjectivity classification: support vector machines (SVM), Maximum Entropy and Naive Bayes. These algorithms have a similar performance in sentiment classification at the text level. In the study presented here, experiments with classifiers were conducted using the same data with default settings in order to evaluate the impact of algorithm choice on sentence-level sentiment classification results in a controlled setting.

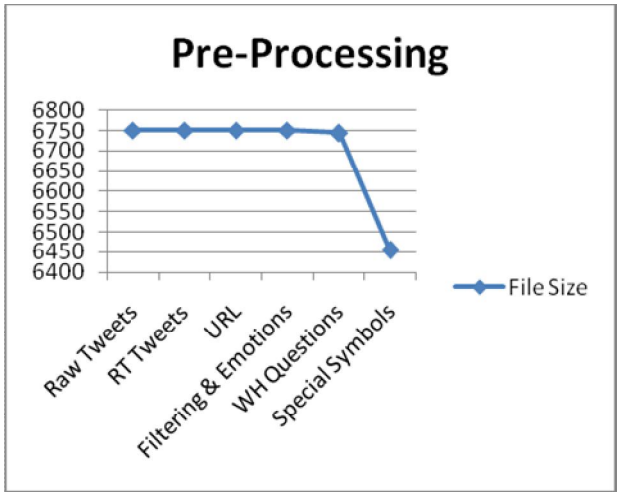


ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014



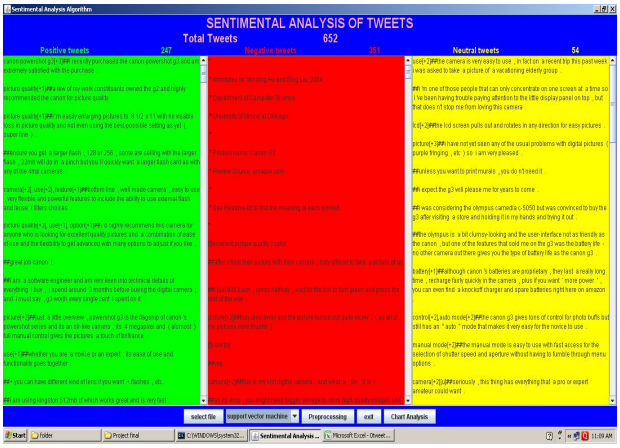


TABLE I

Algorithm	Classification		
	Positive	**Negative**	**Neutral**
SVM	247	351	54
Naive Bayes	356	176	120
Maximum Entropy	200	361	91

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

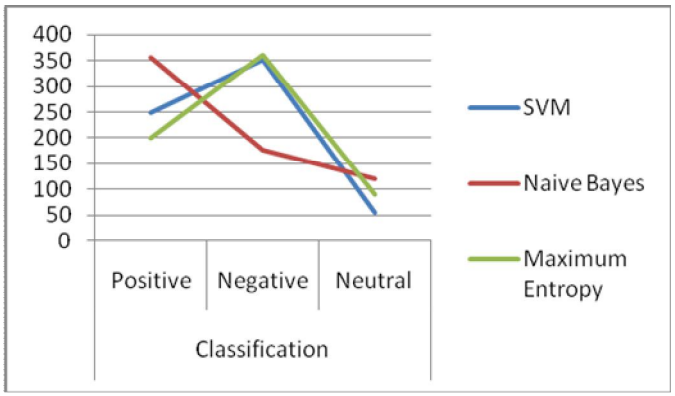
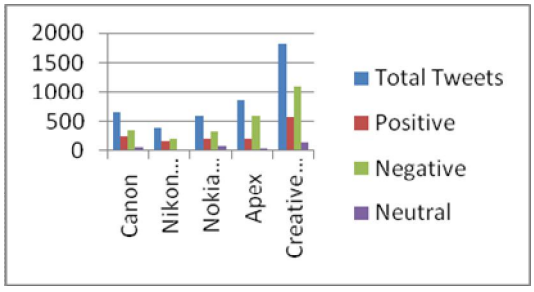


TABLE III: EXAMPLE FOR CLASSIFIER EVALUATION

	Total Tweets	Positive	Negative	Neutral
Canon	652	247	351	54
Nikon Coolpix	390	159	200	22
Nokia 6610	597	194	320	83
Apex	849	209	599	41
Creative Labs	1821	580	1103	138



V. CONCLUSION

In this research paper, we have proposed a novel method for extracting the user opinions of products. One of the distinctive features of the proposed methodology is to pre-process the tweets while attempting to discover user opinions as well as characterize the classifiers. After the preprocessing phase, the cleaned and refined data is stored in a database meant to be used for the Machine learning process.For this determination, we had to offer a detailed Machine learning method, which is devoted to sentiment analysis. The key benefit of our proposed method is to study the sentiment analysis with minimal support as a composite problem that can be solved by succeeding partitions. In the supervised learning, the objective was to calculate the sentiment score of product features by aggregating opinion polarities of opinion words around the product features. We considered several techniques that could improve the classification performance. Therefore, choosing the features that are directly related to sentiment analysis is important, because it can improve performance and time and space efficiency.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

REFERENCES

1. Hemalatha, I., GP Saradhi Varma, and A. Govardhan. "Sentiment Analysis Tool using Machine Learning Algorithms."
2. Hemalatha, I., A. Govardhan, and G. P. Varma. "Machine Learning Methods in Classification of Text by Sentiment Analysis of Social Networks." *International Journal of Advanced Research in Computer Science* 2.5 (2011).
3. Hemalatha, I., GP Saradhi Varma, and A. Govardhan. "Preprocessing the Informal Text for efficient Sentiment Analysis." *International Journal* (2012).
4. G.P.Saradhi Varma,A.Govardhan, I.Hemalatha. "Sentiment Analysis Tool Using Machine Learning Algorithms." *Elixir International Journal, Elixir Comp. Sci. & Engg.* 58 (2013): 14791-14794.
5. Hemalatha, I., A. Govardhan, and G. P. Varma. "Machine Learning Methods in Classification of Text by Sentiment Analysis of Social Networks." *International Journal of Advanced Research in Computer Science* 2.5 (2011).
6. B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1{135, 2008.
7. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79{86, 2002.
8. Twitter Sentiment Classification using Distant Supervision by Alec Go, Richa Bhayani, and Lei Huang.
9. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Association for Computational Linguistics, 2005.

BIOGRAPHY

I.Hemalatha received her M.Tech degree from Andhra University, pursuing Ph.D in computer Science Engineering. A member of CSI, Co-ordinator for Microsoft Student Education Academy, Member in Infosys Campus connect Programme. Working as Assistant Professor in S.R.K.R. Engineering College, China-Amiram, Bhimavaram.