International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 2, Issue 2, March – April 2013 ISSN 2278-6856

Sentiment Analysis Tool using Machine Learning Algorithms

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

[1]Research Scholar JNT University Kakinada, Kakinada, A.P., INDIA

[2]Professor & Head, Dept., of Information Technology, S.R.K.R. Engineering College, Bhimavaram, A.P., INDIA.

[3]Director of Evaluation JNT University, Hyderabad, A.P.,India.

Abstract: ***As the increase of social networking, people started to share information through different kinds of social media. Among all varieties of social media, 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 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: component; Sentiment analysis; Twitter; social media

1. 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.

2. RELATED WORK

2.1. 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.

2.2. 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 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.

2.3. Social Network Analysis on Twitter

A social networking service is an online service that focuses on building social network among people who are willing to share interests, activities, information, or real-life connections. As the fast-growing popularity on the Internet, social network service platform therefore provide adequate information for social network analysis. In [5], Ahn, Han, Kwak, Moon, and Jeong analysis whether online relationships and their growth patterns are as same as in real-life social networks by comparing the

structures of three online social networking services: Cyworld, MySpace, and orkut.

Among all kinds of social networking service, Twitter, as a micro-blogging service is the second popular social networking site [6]. With its special limitation that only 140 characters can be entered in each tweet, Twitter therefore provide a good position for social network analysis. Many researches has focus on social network analysis on Twitter. Longueville, Smith, and Luraschi [7] focus on how Twitter can be used as a source of spatio-temporal information; Sakaki, Okazaki, and Matsuo [8] present an investigation of the real-time nature of Twitter and proposes an event notification system that monitors tweets and delivers notification promptly; Pak and Paroubek [9] used Twitter as a source of opinion mining and sentiment analysis tasks.

3. SYSTEM FRAMEWORK

We present a model which collects tweets from social networking sites and thus provide a view of business intelligence. In our framework, there are two layers in the sentiment analysis tool, the data processing layer and sentiment analysis layer. Data processing layer deals with data collection and data mining, while sentiment analysis layer use a application to present the result of data mining. More details will be introduced in the following sections.

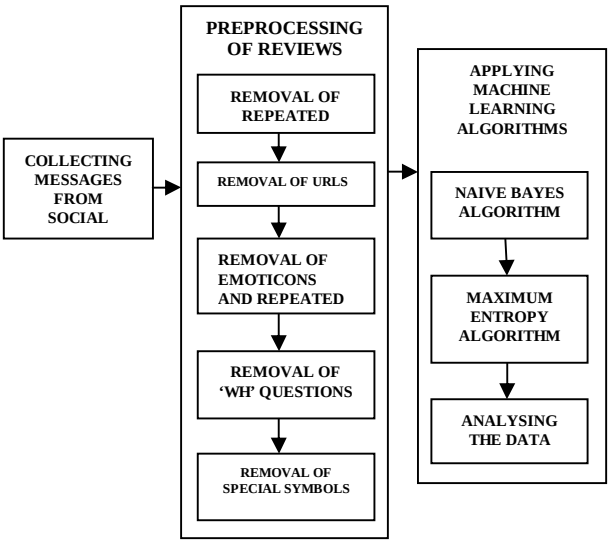


Figure 1: Architecture for SAT using Machine learning algorithms.

3.1. Data Collection and Preprocessing

As now, we have set up the list of tweets or comments on different products manually. We then go through the website of social network sites to collect tweets. All data collected will be stored in a database for further analysis. During the analysis process, words and their polarities are taken into considerations. Combining with social semantic analysis and natural language processing, tweets about daily gossips or unrelated contents will be discarded, and thus relative contents are accurately extracted.

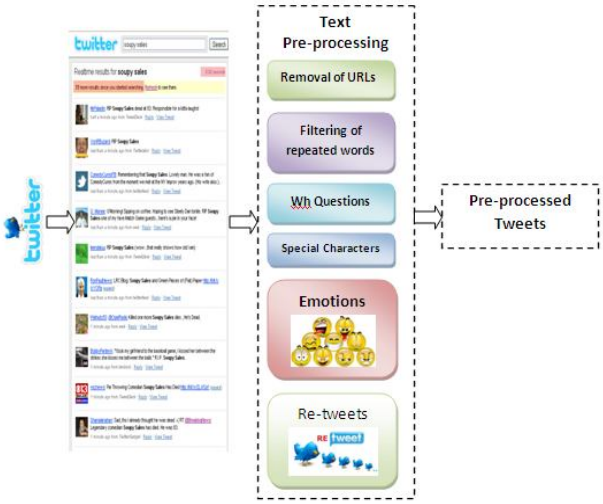


Figure 2: Overview of Data Collection and Preprocessing Process

The system architecture consists of three parts i.e. collecting messages from social networking sites, preprocessing and applying algorithms. First we have to select the messages from text file or excel file to preprocess the messages. In preprocessing, they remove the unnecessary data like repeated messages (tweets), repeated letters, urls, emotion icons, WHQuestions, special symbols. We have to select an algorithm after preprocess the messages by classifying to get the reviews on any product like cinemas, phones, iPods, electronic media etc…so this project contains two modules.

3.2. Machine Learning Algorithms

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too large to be covered by the set of observed examples (training data). Hence the learner must generalize from the given examples, so as to be able to produce a useful output in new cases.

3.3 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.

$$c* = argmax_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

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= {m_1, m_2, m_3.....m_n},
Database: Naive Table N_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 $w_i \in N_T$
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) Maximum Entropy algorithm

Maximum entropy (ME) models, variously known as log-linear, Gibbs, exponential and multinomial logic models, provide a general purpose machine learning technique for classification and prediction which has been successfully applied to fields as diverse as computer vision and econometrics. In natural language processing, recent years have seen ME techniques used for sentence boundary detection, part of speech tagging, parse selection and ambiguity resolution, and stochastic attribute-value grammars, to name just a few applications. A leading advantage of ME models is their flexibility: they allow stochastic rule systems to be augmented with additional syntactic, semantic, and pragmatic features. However, the richness of their presentations is not without cost. Even modest ME models can require considerable computational resources and very large quantities of annotated training data in order to accurately estimate the model's parameters. While parameter estimation for ME models is conceptually straightforward, in practice ME models for typical natural language tasks are usually quite large, and frequently contain hundreds of thousands of free parameters. Estimation of such large models is not only expensive, but also, due to sparsely distributed features, sensitive to round-off errors. Thus, highly efficient, accurate, scalable methods are required for estimating the parameters of practical models. In this paper, we consider a number of algorithms for estimating the parameters of ME models, including Generalized Iterative Scaling and Improved Iterative Scaling, as well as general purpose optimization techniques such as gradient ascent, conjugate gradient, and variable metric methods. Surprisingly, the widely used iterative scaling algorithms perform quite poorly, and for all of the test problems, a limited memory variable metric algorithm outperformed the other choices.

$$P_{ME}(c|d,\lambda) = \frac{\exp[\Sigma_i \lambda_i f_i(c,d)]}{\Sigma_{c'} \exp[\Sigma_i \lambda_i f_i(c,d)]} \quad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

In the equation(2), c is the class, d is the tweet, and ¸ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the lambdas so as to maximize the conditional probability. We use the Stanford Classifiers to perform MaxEnt classification. For training the weights we used conjugate gradient ascent and added smoothing (L2 regularization). Theoretically, MaxEnt performs better than Naive Bayes because it handles feature overlap better. However, in practice, Naive Bayes can still perform well on a variety of problems.

Proposed Maximum Entropy classifier

Input: messages m= {m_1, m_2, m_3....., m_n},
Database: Naive Table N_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 $w_i \in N_T$ 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

4. RUNTIME EXECUTION

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 preprocessing. The results are as follows

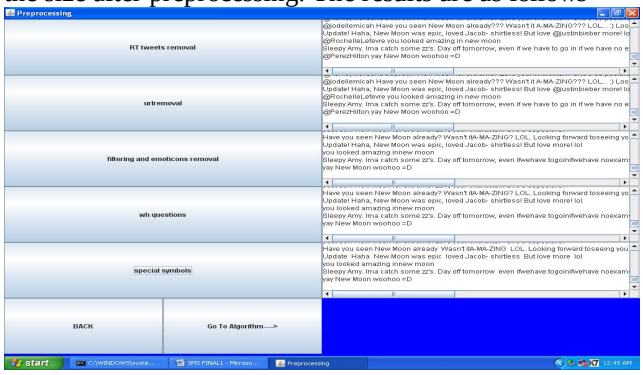


Figure 3: Preprocess the messages.

Table 1: Compare the file size with Pre-processing techniques.

PRE-PROCESSING	FILE SIZE(KB)	TOTAL %
Before Preprocessing	82.9	100%
After removal of RT tweets	80.4	96.9%
After removal of url	80.0	96.5%
After Filtering and removal of Emotion icons	77.8	93.8%
Ater removal of WHQuestions	76.7	92.5%
After removal of Special symbols	76.7	92.5%

In the preprocessing the original message size is 100% so, after removing the RTtweets the file size is gradually decreased to 96.9%. after url removal the file size is decreased as well as in filtering technique and emotion icons, removal of WHquestions, removal of special symbols also decreases the file size of the messages. so, after pre-preprocessing the messages it gradually decreases to 92.5%.

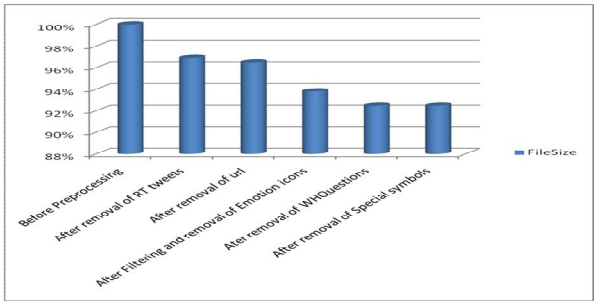
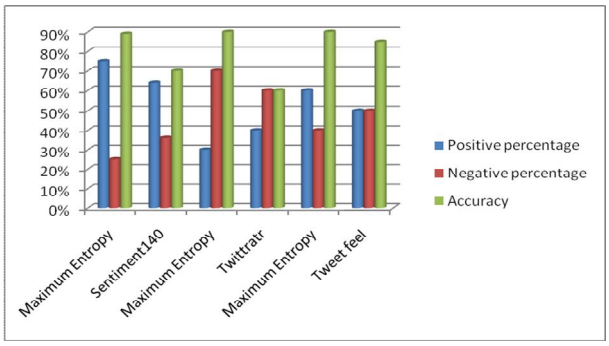


Figure 4: Comparison of file size with pre-processing techniques.



- *Accuracy:*

Accuracy is the ratio of number of correctly classified documents and total number of documents.

$$Accuracy = \frac{Number\ of\ correctly\ classified\ documents}{Total\ number\ of\ documents} \quad ....(3)$$

The maximum entropy algorithm is compared with another tools of sentimental analysis like sentiment140, twittratr,tweetfeel. We give same messages to both maximum entropy and sentimental tools. These messages are classified as positive, negative and neutral. Itgives high accuracy and it is given same messages to both maximum entropy and sentimental analysis tools.it gives high accurate results than sentimental analysis tools.

Figure 5. A Snapshot of the FrontPage including the messages classified as positive ,negative and neutral.

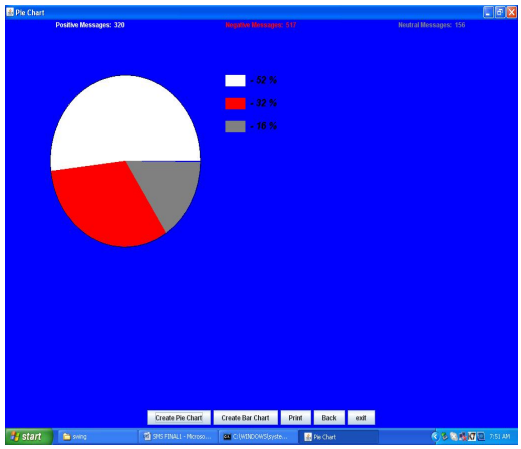


Figure 6. A Snapshot of the Sentiment Analysis of classified Messages.

5. CONCLUSION

We show that using emoticons as noisy labels for training data is an effective way to perform distant supervised learning. Machine learning algorithms (Naive Bayes, maximum entropy classification) can achieve high accuracy for classifying sentiment when using this method. Although Twitter messages have unique characteristics compared to other corpora, machine learning algorithms are shown to classify tweet sentiment with similar performance.

References

[1]. B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1{135, 2008.

[2]. 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.

[3]. Twitter Sentiment Classification using Distant Supervision by Alec Go, Richa Bhayani, and Lei Huang.

[4]. J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Association for Computational Linguistics, 2005.

[5]. K. Nigam, J. Lafferty, and A. Mccallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61{67, 1999.

[6]. [Mikheev, 1999] Andrei Mikheev. Feature lattics and maximum entropy models. *Machine Learning*, 1999.

[7]. [Nigam *et al.*, 1999] Kamal Nigam, Andrew McCallum,Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 1999.

[8]. [Csisz_ar, 1996] I. Csisz_ar. Maxent, mathematics, and information theory. In K. Hanson and R. Silver, editors, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 1996.

[9]. [Rosenfeld, 1994] Ronald Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD thesis, Carnegie Mellon University, 1994.

AUTHOR

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.