Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 4, April 2016, pg.296 – 307

AN EFFICIENT SENTIMENT CLASSIFICATION BASED ON TWITTER DATASET

Ms.Ponmani.S[1], Ms.Lina Dinesh[2], Ms.Dharani.R[3]
[1]Department of CSE, Sri Eshwar College Of Engineering, Coimbatore, India
[2]Department of CSE, Sri Eshwar College Of Engineering, Coimbatore, India
[3]Member Technical Staff, Zoho corporation private Ltd, Chennai, India
[1] linadinesh@sece.ac.in; [2] ponmanigct@gmail.com; [3] dharani.r@zohocorp.com

Abstract— Focus on using Twitter, the most popular social blogging platform, for the task of sentiment analysis. To automatically collect a corpus for sentiment analysis and opinion mining purposes are shown. The POS tagger that is able to determine the positive and negative sentiment words. Finally NaiveBayes Classification introduces a computationally trivial approach for classifying a given Tweet's sentiment. If a Tweet contains an emoticon corresponding to a particular emotional sentiment, then it is simply assumed that the rational author actually felt and/or sought to convey the stated sentiment. Here, use classifier for sentiment analysis and efficient pre-processing are used. This proposed system discusses an approach where a publicized stream of tweets from the Twitter micro blogging site are pre-processed and classified based on their subjectivity word and semantic phrase content as positive, negative and irrelevant. Analyses the performance of various classifying algorithms based on their precision and recall in such cases. In this system, focus on using Twitter, the most popular micro-blogging platform, for the task of sentiment analysis. Show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. Using the corpus, we build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a twitter document. Experimental evaluations show that proposed techniques are efficient and perform better than previously proposed methods.

Keywords— clustering, information extraction and visualization

I. INTRODUCTION

This Opinion mining (or sentiment analysis) has attracted great interest in recent years, both in academia and industry due to its potential applications. One of the most promising applications is analysis of opinions in social networks. Lots of people write their opinions in forums, micro blogging or review websites. The data is very useful for business companies, governments, and individuals, who want to track automatically attitudes and feelings in those sites. Namely, there is a lot of data available that contains much useful information, so it can be

analyzed automatically. For instance, a customer who wants to buy a product usually searches the Web trying to find opinions of other customers or reviewers about this product. In fact, these kinds of reviews affect customer's decision. The booming micro-blog service, Twitter, attracts more people to post their feelings and opinions on various topics. The posting of sentiment contents can not only give an emotional snapshot of the online world but also have potential commercial, financial and sociological values. However, facing the massive sentiment tweets, it is hard for people to get overall impression without automatic sentiment classification and analysis. Therefore, there are emerging many sentiment classification works showing interests in tweets.
Topics discussed in Twitter are more diverse and unpredictable. Sentiment classifiers always dedicate themselves to a specific domain or topic named in the paper. Namely, a classifier trained on sentiment data from one topic often performs poorly on test data. One of the main reasons is that words and even language constructs used for expressing sentiments can be quite different on different topics. Taking a comment "read the book" as an example, it could be positive in a book review while negative in a movie review. In social media, a Twitter user may have different opinions on different topics. Thus, topic adaptation is needed for sentiment classification of tweets explicitly borrowed a bridge to connect a topic dependent feature to a known or common feature. Such bridges are built between product reviews by assuming that the parallel sentiment words exist for each pair of topics, such as books, DVDs, electronics and kitchen appliances. However, it is not necessarily applicable to topics in Twitter, especially the unpredictable ones. It is worth mentioning that detecting and tracking topics from tweets is another research topic. Ad-hoc micro blog search in Text Retrieval Conference (TREC) 2011 -2012 is hopefully a choice for people to query tweets on emerging topics, and sentiment classification can be conducted afterwards.

II. OVERVIEW DATA MINING

Data Mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data Mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods.

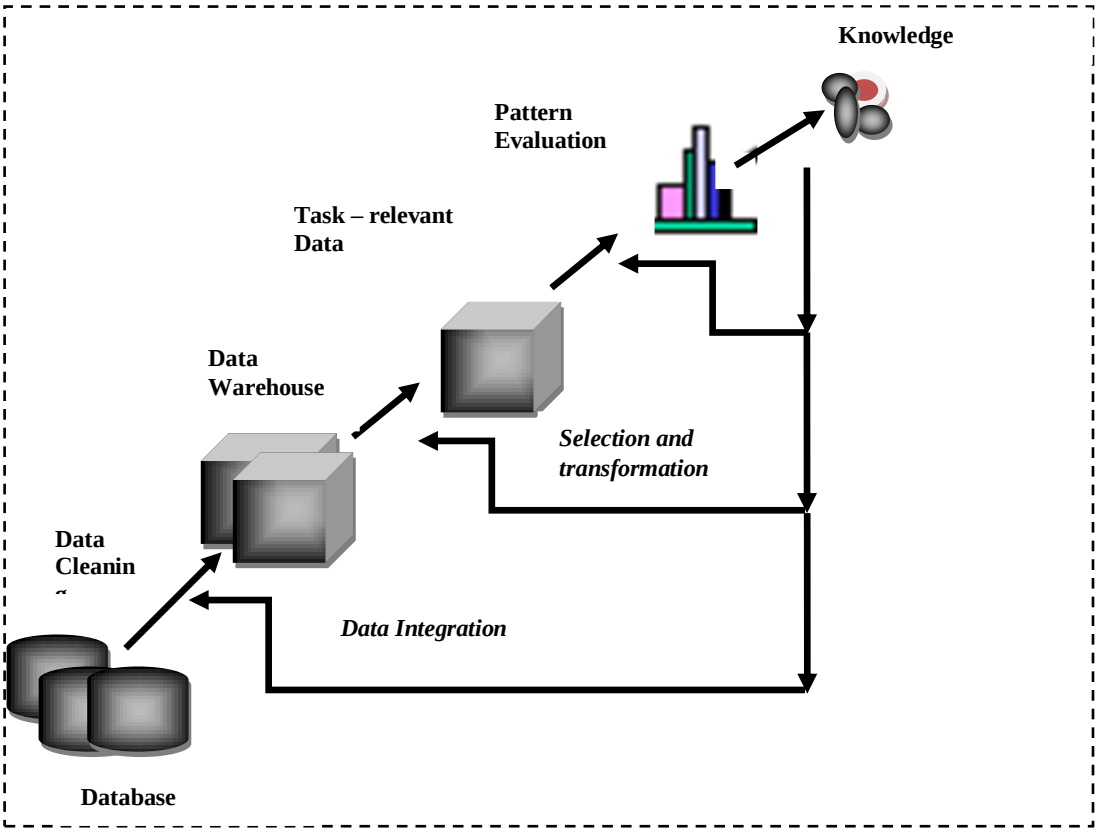

Figure 2.1 Data Mining Knowledge Discovery Process

Figure 2.1 show the overview of data mining process and discover techniques. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

Consequently, Data Mining consists of more than collecting and managing data, it also includes analysis and prediction. Data Mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis.

III. MOTIVATION WORK

A number of works have been done on the area of opinion mining especially for sentiment classification. This section mentions some of these works.

F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu says, extracting sentiment and topic lexicons is important for opinion mining. Previous works have showed that supervised learning methods are superior for this task. However, the performance of supervised methods highly relies on manually labeled training data. In this paper, we propose a domain adaptation framework for sentiment- and topic- lexicon co-extraction in a domain of interest where we do not require any labeled data, but have lots of labeled data in another related domain. The framework is twofold. In the first step, we generate a few high-confidence sentiment and topic seeds in the target domain. In the second step, we propose a novel Relational Adaptive bootstraPping (RAP) algorithm to expand the seeds in the target domain by exploiting the labeled source domain data and the relationships between topic and sentiment words. Experimental results show that our domain adaptation framework can extract precise lexicons in the target domain without any annotation.

K. Liu, L. Xu, and J. Zhao proposes a novel approach to extract opinion targets based on word-based translation model (WTM). At first, we apply WTM in a monolingual scenario to mine the associations between opinion targets and opinion words. Then, a graph-based algorithm is exploited to extract opinion targets, where candidate opinion relevance estimated from the mined associations, is incorporated with candidate importance to generate a global measure. By using WTM, our method can capture opinion relations more precisely, especially for long-span relations. In particular, compared with previous syntax-based methods, our method can effectively avoid noises from parsing errors when dealing with informal texts in large Web corpora. By using graph-based algorithm, opinion targets are extracted in a global process, which can effectively alleviate the problem of error propagation in traditional bootstrap-based methods, such as Double Propagation. The experimental results on three real world datasets in different sizes and languages show that our approach is more effective and robust than state-of-art methods.

G. Qiu, L. Bing, J. Bu, and C. Chen says, Analysis of opinions, known as opinion mining or sentiment analysis, has attracted a great deal of attention recently due to many practical applications and challenging research problems. In this article, we study two important problems, namely, opinion lexicon expansion and opinion target extraction. Opinion targets (targets, for short) are entities and their attributes on which opinions have been expressed. To perform the tasks, we found that there are several syntactic relations that link opinion words and targets. These relations can be identified using a dependency parser and then utilized to expand the initial opinion lexicon and to extract targets. This proposed method is based on bootstrapping. We call it double propagation as it propagates information between opinion words and targets. A key advantage of the proposed method is that it only needs an initial opinion lexicon to start the bootstrapping process. Thus, the method is semi-supervised due to the use of opinion word seeds. In evaluation, we compare the proposed method with several state-of-the-art methods using a standard product review test collection. The results show that our approach outperforms these existing methods significantly.

X. Ding, B. Liu, and P. S. Yu proposed one of the important types of information on the Web is the opinions expressed in the user generated content, e.g., customer reviews of products, forum posts, and blogs. In this paper, we focus on customer reviews of products. In particular, we study the problem of determining the semantic orientations (positive, negative or neutral) of opinions expressed on product features in reviews. This problem has many applications, e.g., opinion mining, summarization and search. Most existing techniques utilize a list of opinion (bearing) words (also called opinion lexicon) for the purpose. Opinion words are words that express desirable (e.g., great, amazing, etc.) or undesirable (e.g., bad, poor, etc) states. These approaches, however, all have some major shortcomings. In this paper, we propose a holistic lexicon-based approach to solving the problem by exploiting external evidences and linguistic conventions of natural language expressions. This approach allows the system to handle opinion words that are context dependent, which cause major difficulties for existing algorithms. It also deals with many special words, phrases and language constructs which have impacts on opinions based on their linguistic patterns. It also has an effective function for aggregating multiple

conflicting opinion words in a sentence. A system, called Opinion Observer, based on the proposed technique has been implemented. Experimental results using a benchmark product review data set and some additional reviews show that the proposed technique is highly effective. It outperforms existing methods significantly.

T. Ma and X. Wan says, news comments on the web express readers' attitudes or opinions about an event or object in the corresponding news article. And opinion target extraction from news comments is very important for many useful Web applications. However, many sentences in the comments are irregular and informal, and sometimes the opinion targets are implicit. Thus the task is very challenging and it has not been investigated yet. In this paper, we propose a new approach to uniformly extracting explicit and implicit opinion targets from news comments by using Centering Theory. The approach uses global information in news articles as well as contextual information in adjacent sentences of comments. Our experimental results verify the effectiveness of the proposed approach.

W. Jin and H. H. Huang proposes a model called merchants selling products on the Web often ask their customers to share their opinions and hands-on experiences on products they have purchased. As e-commerce is becoming more and more popular, the number of customer reviews a product receives grows rapidly. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. In this research, we aim to mine customer reviews of a product and extract highly specific product related entities on which reviewers express their opinions. Opinion expressions and sentences are also identified and opinion orientations for each recognized product entity are classified as positive or negative. Different from previous approaches that have mostly relied on natural language processing techniques or statistic information, we propose a novel machine learning framework using lexicalized HMMs. The approach naturally integrates linguistic features, such as part-of-speech and surrounding contextual clues of words into automatic learning. The experimental results demonstrate the effectiveness of the proposed approach in web opinion mining and extraction from product reviews.

Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai says, we define the problem of topic-sentiment analysis on Weblogs and propose a novel probabilistic model to capture the mixture of topics and sentiments simultaneously. The proposed Topic-Sentiment Mixture (TSM) model can reveal the latent topical facets in a Weblog collection, the subtopics in the results of an ad hoc query, and their associated sentiments. It could also provide general sentiment models that are applicable to any ad hoc topics. With a specifically designed HMM structure, the sentiment models and topic models estimated with TSM can be utilized to extract topic life cycles and sentiment dynamics. Empirical experiments on different Weblog datasets show that this approach is effective for modeling the topic facets and sentiments and extracting their dynamics from Weblog collections. The TSM model is quite general; it can be applied to any text collections with a mixture of topics and sentiments, thus has many potential applications, such as search result summarization, opinion tracking, and user behavior prediction.

W. X. Zhao, J. Jiang, H. Yan, and X. Li propose a model called discovering and summarizing opinions from online reviews is an important and challenging task. A commonly-adopted framework generates structured review summaries with aspects and opinions. Recently topic models have been used to identify meaningful review aspects, but existing topic models do not identify aspect-specific opinion words. In this paper, we propose a MaxEnt-LDA hybrid model to jointly discover both aspects and aspect-specific opinion words. We show that with a relatively small amount of training data, our model can effectively identify aspect and opinion words simultaneously. We also demonstrate the domain adaptability of our model.

Z. Liu, H. Wang, H. Wu, and S. Li says, mining opinion targets from online reviews is an important and challenging task in opinion mining. This paper proposes a novel approach to extract opinion targets by using partially-supervised word alignment model (PSWAM). At first, we apply PSWAM in a monolingual scenario to mine opinion relations in sentences and estimate the associations between words. Then, a graph-based algorithm is exploited to estimate the confidence of each candidate, and the candidates with higher confidence will be extracted as the opinion targets. Compared with existing syntax-based methods, PSWAM can effectively avoid parsing errors when dealing with informal sentences in online reviews. Compared with the methods using alignment model, PSWAM can capture opinion relations more precisely through partial supervision from partial alignment links. Moreover, when estimating candidate confidence, we make penalties on higher-degree vertices in our graph-based algorithm in order to decrease the probability of the random walk running into the unrelated regions in the graph. As a result, some errors can be avoided. The experimental results on three data sets with different sizes and languages show that our approach outperforms state-of-the-art methods.

K. Liu, H. L. Xu, Y. Liu, and J. Zhao says, the vast majority of existing approaches to opinion feature extraction rely on mining patterns only from a single review corpus, ignoring the nontrivial disparities in word distributional characteristics of opinion features across different corpora. In this paper, we propose a novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-specific corpus (i.e., the given review corpus) and one domain-independent corpus (i.e., the contrasting corpus). We capture this disparity via a measure called domain relevance (DR), which characterizes the relevance of a term to a text collection. We first extract a list of candidate opinion features from the domain review corpus by defining a set of syntactic dependence rules. For each extracted candidate feature, we then estimate its intrinsic-domain relevance (IDR) and extrinsic-domain relevance (EDR) scores on the domain-dependent and domain-independent corpora, respectively. Candidate features that are less generic (EDR score less than a threshold) and more domain-specific (IDR score greater than another threshold) are then confirmed as opinion features. We call this interval thresholding approach the intrinsic and extrinsic domain relevance (IEDR) criterion. Experimental results on two real-world review domains show the proposed IEDR approach to outperform several other well-established methods in identifying opinion features.

IV. RELATED WORK

Opinion target and opinion word extraction are not new tasks in opinion mining. There is significant effort focused on these tasks [1], [6], [12], [13], [14]. They can be divided into two categories: sentence-level extraction and corpus level extraction according to their extraction aims. In sentence-level extraction, the task of opinion target word extraction is to identify the opinion target mentions or opinion expressions in sentences. Thus, these tasks are usually regarded as sequence-labeling problems [13], [14], [15], [16]. Intuitively, contextual words are selected as the features to indicate opinion targets/words in sentences. Additionally, classical sequence labeling models are used to build the extractor, such as CRFs [13] and HMM [17]. Jin and Huang [17] proposed a lexicalized HMM model to perform opinion mining. Both [13] and [15] used CRFs to extract opinion targets from reviews.

However, these methods always need the labeled data to train the model. If the labeled training data are insufficient or come from the different domains than the current texts, they would have unsatisfied extraction performance. Although [2] proposed a method based on transfer learning to facilitate cross domain extraction of opinion targets/words, their method still needed the labeled data from out-domains and the extraction performance heavily depended on the relevance between in-domain and out-domain. In addition, much research focused on corpus-level extraction. They did not identify the opinion target/word mentions in sentences, but aimed to extract a list of opinion targets or generate a sentiment word lexicon from texts.

Most previous approaches adopted a collective unsupervised extraction framework. As mentioned in our first section, detecting opinion relations and calculating opinion associations among words are the key component of this type of method. Wang and Wang [8] adopted the co-occurrence frequency of opinion targets and opinion words to indicate their opinion associations. Hu and Liu [5] exploited nearest-neighbor rules to identify opinion relations among words. Next, frequent and explicit product features were extracted using a bootstrapping process. Only the use of co-occurrence information or nearest-neighbor rules to detect opinion relations among words could not obtain precise results. Thus, [6] exploited syntax information to extract opinion targets, and designed some syntactic patterns to capture the opinion relations among words. The experimental results showed that their method performed better than that of [5]. Moreover, [10] and [7] proposed a method, named as Double Propagation, which exploited syntactic relations among words to expand sentiment words and opinion targets iteratively. Their main limitation is that the patterns based on the dependency parsing tree could not cover all opinion relations. Therefore, Zhang et al. [3] extended the work by [7]. Besides the patterns used in [7], Zhang et al. further designed specific patterns to increase recall.

Moreover, they used an HITS [18] algorithm to compute opinion target confidences to improve precision. Liu et al. [4] focused on opinion target extraction based on the WAM. They used a completely unsupervised WAM to capture opinion relations in sentences. Next, opinion targets were extracted in a standard random walk framework. Liu's experimental results showed that the WAM was effective for extracting opinion targets. Nonetheless, they present no evidence to demonstrate the effectiveness of the WAM on opinion word extraction. Furthermore, a study employed topic modeling to identify implicit topics and sentiment words [19], [20], [21], [22]. The aims of these methods usually were not to extract an opinion target list or opinion word lexicon from reviews. Instead, they were to cluster for all words into corresponding aspects in reviews, which was different from the task in this paper.

In this process, we penalize high-degree vertices to weaken their impacts and decrease the probability of a random walk running into unrelated regions on the graph. Calculate the prior knowledge of candidates for indicating some noises and incorporating them into our ranking algorithm to make collaborated operations on candidate confidence estimations. Finally, candidates with higher confidence than a threshold are extracted. Compared to the previous methods based on the bootstrapping strategy, opinion targets/words are no longer extracted step by step. Instead, the confidence of each candidate is estimated in a global process with graph co-ranking. Intuitively, the error propagation is effectively alleviated. Impacts and decrease the probability of a random walk running. Graph-based co-ranking algorithm to estimate the confidence of each candidate. The algorithm does not end until no new optimal alignment is found. The opinion associations between opinion target candidates and opinion word candidates. This feature indicates the salience degree of a candidate in reviews.

4.1 LIMITATION

Structural correspondence learning(SCL) for domain adaptation. It employed the pivot features as the bridge to help cross-domain classification. Spectral feature alignment (SFA) algorithm to bridge the gap between the domains with domain independent words. Probabilistic topic model to bridge each pair of domains in a semantic level. All the above studies focus on the review data set. Document-level sentiment classification works, aspect-based sentiment analysis detects topic spans, and associations of topic aspects, sentiment words, and opinion holders in a document or even sentence. Distant supervised learning approach for automatically classifying the sentiment of tweets using emoticons as noisy labels for training data. Twitter data as a corpus for sentiment analysis and tracking the influence of a particular brand activity on the social network. Effective unsupervised learning algorithm, called semantic orientation, for classifying reviews. A web-kernel based measurement was proposed as PMI-IR to measure the weight of a sentiment word, which is independent to the corpus collection in hand. In previous methods, mining the opinion relations between opinion targets and opinion words was the key to collective extraction. To this end, the most adopted techniques have been nearest-neighbor rules and syntactic patterns. Several methods exploited syntactic information, in which the opinion relations among words are decided according to their dependency relations in the parsing tree. The collective extraction adopted by most previous methods was usually based on a boot strapping framework, which has the problem of error propagation.

V. PROPOSED SYSTEM

The motivation for building the sentiment detection and classification system described in this paper. Fully analyzing and classifying opinions involves tasks that relate to some fairly deep semantic and syntactic analysis of the text. These include not only recognizing that the text is subjective, but also determining who the holder of the opinion is, what the opinion is about, and which of many possible positions the holder of the opinion expresses regarding that subject. In this project, we are presenting three of the components of our opinion detection and organization subsystem, which have already been integrated into our larger question-answering system. These components deal with the initial tasks of classifying articles as mostly subjective or objective, finding opinion sentences in both kinds of articles, and determining, in general terms and without reference to a specific subject, if the opinions are positive or negative.

Micro blogging today has become a very popular communication tool among Internet users. Millions of users share opinion s on different aspects of life every day. Therefore micro blogging web-sites are rich sources of data for opinion mining and sentiment analysis. Because micro blogging has appeared relatively recently, there are a few research works that were devoted to this topic. In this project focus on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. It show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. It perform Navie bayes analysis of the collected corpus and explain discovered phenomena. It builds a sentiment classifier that is able to determine positive, negative and neutral sentiments for a real time collected tweets from twitter web site. Experimental evaluations show that our proposed techniques are efficient and perform better than previously proposed methods.

VI. SYSTEM ARCHITECTURE

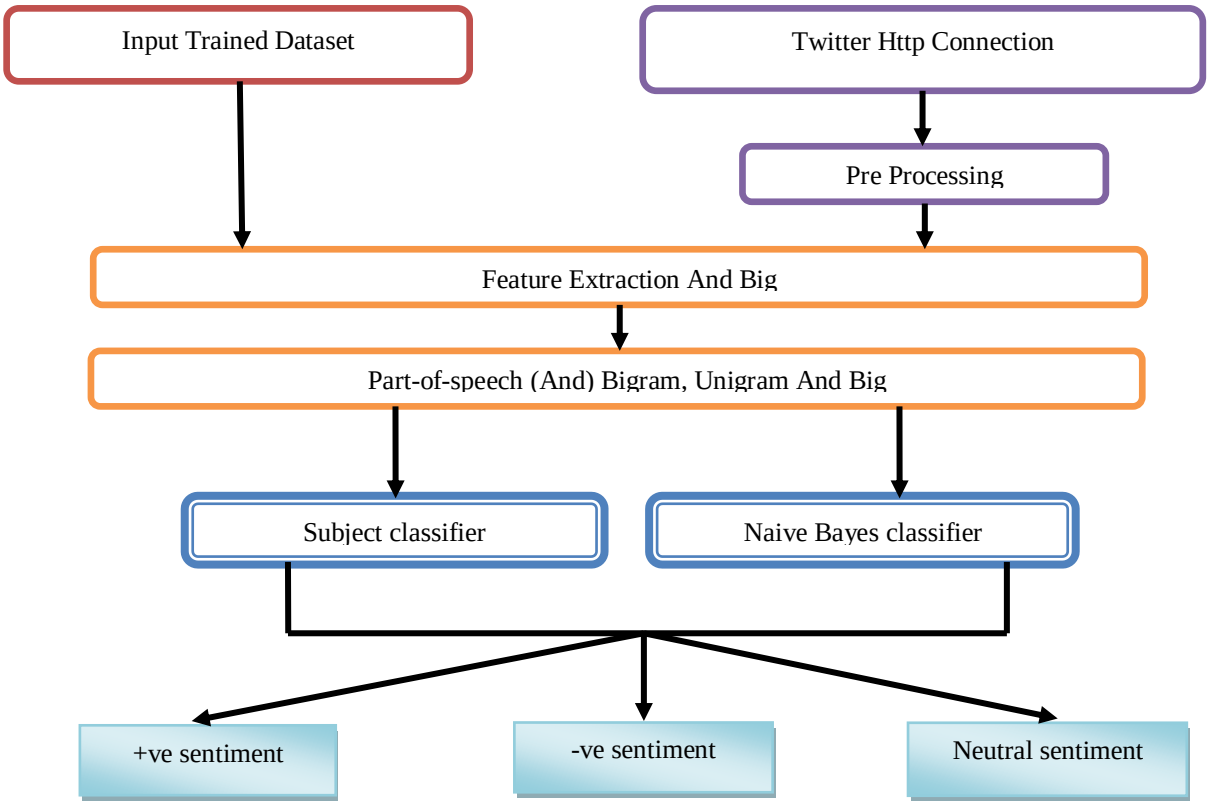

Figure 6.1 Architecture Diagram

6.1 SYSTEM MODULES

The proposed system contains four modules.

1. Twitter data collection
2. Preprocessing
3. Pos tagger feature extraction
4. Sentiment classification

6.1.1 TWITTER DATA COLLECTION

Twitter for two types of emoticons is queried: The two types of collected corpora will be used to train a classifier to recognize positive and negative sentiments. In order to collect a corpus of objective posts, text messages from Twitter accounts of popular newspapers and magazines are retrieved, such as "New York Times", "Washington Posts" etc. The queried accounts of 44 newspapers to collect training set of objective texts. Because each message cannot exceed 140 characters by the rules of the micro blogging platform, it is usually composed of a single sentence. Therefore, an emoticon within a message represents an emotion for the whole message and all the words of the message are related to this emotion is assumed.

6.1.2 PREPROCESSING

The corpus has considerable amount of metadata such as date, time, and identity number etc. Here is a sample of the raw data before data processing.

Sample Twits
_315,41199, AM 0:26:17, News Analysis: In Second Debate Obama Strikes Back (NY Times): Share With Friends: | | Top News - ... http://t.co/PXoW7wiD @NCAAlarms,-1,0,1__

1. *Spelling correction* - As Twitter users generally use informal language, there

Are often incorrect spellings in tweets. The Jazzy Open Source Spell Checker is used to detect incorrect spellings in the tweets and replace them with the closest word from the English dictionary.

2. *Filtering* - Tweets contained a lot of metadata and quite a bit of noise which were Removed. The following data was filtered, _ Identity numbers, date, time etc. of the tweets

1.Irrelevant tags

2.Hyperlinks

3.#tags e.g. #msnbc2012

4.Twitter handles e.g. @pavanred

5.punctuation, special characters and digits

After subjecting a tweet to this data processing process, the natural language content of the tweet is only left and the human annotated sentiment that is used by the classification algorithm in supervised learning. After data processing, the sample tweet that is considered earlier about would be transformed to, used for sentiment classification.

"news analysis second debate obama strikes back ny times share friends top news,-1 1 for positive, -1 for negative are the annotations"

6.1.3 POS TAGGER FEATURE EXTRACTION

In order to perform machine learning, it is necessary to extract certain clues from the text that may lead to an effective correct classification. Clues about the original data are usually stored in the form of a feature vector, F = (f1, f2…….. fn). Each coordinates of a feature vector represents one clue, also called a feature, "fi "of the original text.

Part Of Speech features -Parts of Speech information is most commonly exploited in all NLP tasks. One of the most important reasons is that they provide a crude form of word sense disambiguation. Since the language used in Twitter is generally informal, part of speech tagging isn't very accurate for tweets. Both NLTK Part Of Speech tagger and Open NLP Part Of Speech tagger are used along with a heuristic that adjectives and/or adverbs, JJ, JJR, JJS, RB, RBR and RBS in the Penn Tree bank target, are generally used to articulate opinions in natural language.

Performing grammatical tagging will indicate that "dogs" is a verb, and not the more common plural noun, since one of the words must be the main verb, and the noun reading is less likely following "sailor" (sailor !→ Dogs). Semantic analysis can then extrapolate that "sailor" and "barmaid" implicate "dogs" as 1) in the nautical context (sailor→<verb>←barmaid) and 2) an action applied to the object "barmaid" ([subject] dogs→barmaid). In this context, "dogs" is a nautical term meaning "fastens (a watertight barmaid) securely; applies a dog to".

6.1.4 SENTIMENT CLASSIFICATION

A conditional probably is a probably that event X will occur, given the evidence Y. That is normally written P(X | Y). The Bayes rule allows us to determine this probability with the probability of the opposite result and of the two components individually: P(X | Y) = P(X) P(Y | X) / P(Y). This restatement can be very helpful that are trying to estimate the probability of something based on examples of it occurring.

Formula looks like this.

P(sentiment | sentence)=P(sentiment)P(sentence|sentiment)/P(sentence).

So, the initial formula looks like this.

P(sentiment | sentence) = P(sentiment)P(sentence | sentiment) / P(sentence)

The dividing P(line) is dropped, as it's the same for both classes, and to rank them rather than calculate a precise probability. The use independence assumption to let, treat P (sentence | sentiment) as the product of P (token | sentiment) across all the tokens in the sentence.

So, the estimation P (token | sentiment) as

count(this token in class) + 1 / count(all tokens in class) + count(all tokens).

6.2 PERFORMANCE ANALYSIS

It suggests some social information can indeed help opinion retrieval in Twitter. The URL feature is the most effective feature, perhaps because most textual content in these tweets are objective introductions. Also, spammers usually post tweets including links and features dealing with links might help reduce spam. The effect

of URL, Statuses and Followers features for tweets ranking also supports our approach of using social information and structural information to generate "pseudo" objective tweets.

The computational accuracy of the classifier on the whole evaluation dataset, i.e.:

$$\text{accuracy} = \frac{\text{N(correct classifications)}}{\text{N(all classifications)}}$$

The measure of accuracy across the classifier's decision

$$\text{decision} = \frac{\text{N(retrieved documents)}}{\text{N(all documents)}}$$

The impact of the dataset size on the performance of the system is also examined. To measure the performance, The F-measure is used,

$$F = (1 + \beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{recall} + \text{recall}}$$

Classifiers	**Avg. Accuracy**	**Max. Accuracy**	**Avg F**
WAM	69.82	71.33	0.688
PSWAM	73.25	77.60	0.728
Naive Bayes Classification	91.86	94.82	0.920

TABLE 6.2 Comparison of Classification Algorithm

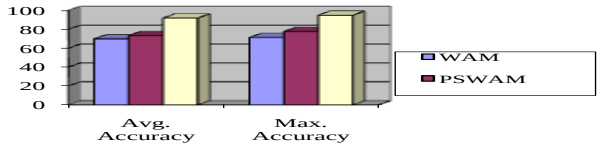

Fig 6.3 Comparison of classification results

Polarity	SVM			Naive Bayes Classification		
	Precision (%)	Recall (%)	F	Precision (%)	Recall (%)	F
Positive	75	74.3	75	92.1	72.3	92
negative	72.3	74	72.1	89.6	88.2	89.6
Natural	74	65	66.5	71	73	71

TABLE 6.2 Comparison Precisions and Recall

6.3 RESULTS

We used the performance metrics to validate the proposed algorithm with results obtained in this papers are shown in Following figure.

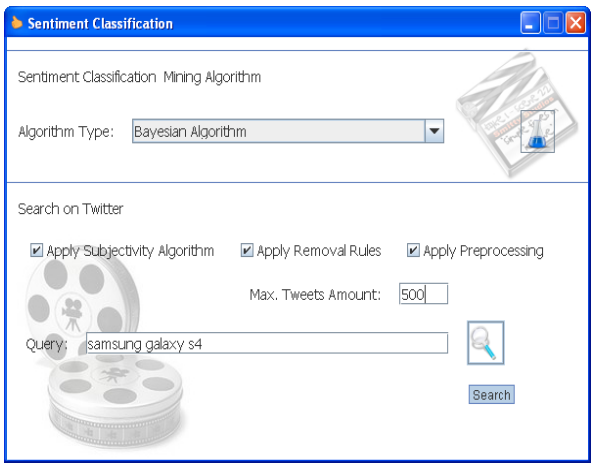

Figure 6.3.1 Main Form

Figure 6.3.2. Opinion Result

Figure 6.3. No of sentiment word graph

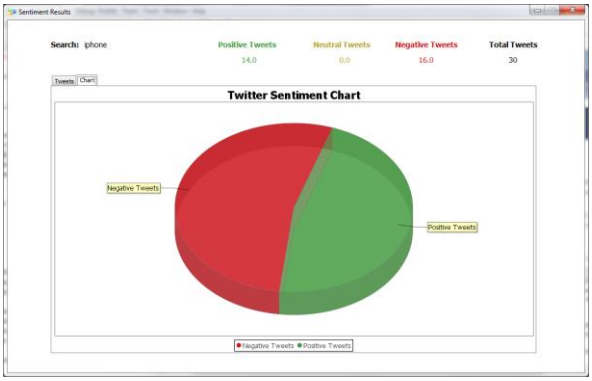

Figure 6.4. Sentiment analysis chart

Thus the proposed scheme is very significant and effective when comparing with existing methods.

VII. CONCLUSION

Twitter sentiment analysis has attracted much attention recently. In this proposed address real time sentiment classification sentiment classification of tweets. Our results show that a simple Naive Bayes classifier can be enhanced to match the classification accuracy of more complicated models for sentiment analysis by choosing the right type of features and removing noise by appropriate feature selection. Naive Bayes classifiers due to their conditional in dependence assumptions are extremely fast to train and can scale over large data sets. They are also robust to noise and less prone to over fitting. Ease of implementation is also a major advantage of Naive Bayes. They were thought to be less accurate than their more sophisticated counterparts like support vector machines and logistic regression but we have shown through this paper that a significantly high accuracy can be achieved. The ideas used in this paper can also be applied to the more general domain of twitter text classification.

VIII. FUTURE WORK

In future work, we will explore even richer linguistic analysis, for example, parsing, semantic analysis and efficient topic modeling.

1. It concentrates on the following aspects
2. Maximum Entropy (ME) classification method.
3. Tweets containing more no of other languages can be considered.
4. Genetic Algorithm for fast searching of tweets.

ACKNOWLEDGEMENT

I am thankful to Ms. Lina Dinesh for her guidance, support and supervision. And also providing the information and ready to help anytime in completion of this paper.

REFERENCES

[1] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, "Cross-domain coextractionof sentiment and topic lexicons," in Proc. 50th Annu.Meeting Assoc. Comput. Linguistics, Jeju, Korea, 2012, pp. 410–419.

[2] K. Liu, L. Xu, and J. Zhao, "Opinion target extraction using wordbasedtranslation model," in Proc. Joint Conf. Empirical MethodsNatural Lang. Process. Comput. Natural Lang. Learn., Jeju, Korea, Jul.2012, pp. 1346–1356.

[3] G. Qiu, L. Bing, J. Bu, and C. Chen, "Opinion word expansion andtarget extraction through double propagation," Comput. Linguistics,vol. 37, no. 1, pp. 9–27, 2011.

[4] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach toopinion mining," in Proc. Conf. Web Search Web Data Mining, 2008,pp. 231–240.

[5] T. Ma and X. Wan, "Opinion target extraction in chinese newscomments." in Proc. 23th Int. Conf. Comput. Linguistics, Beijing,China, 2010, pp. 782–790.

[6] W. Jin and H. H. Huang, "A novel lexicalized HMM-based learningframework for web opinion mining," in Proc. Int. Conf. Mach.Learn., Montreal, QC, Canada, 2009, pp. 465–472.

[7] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentimentmixture: modeling facets and opinions in weblogs," in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 171–180.

[8] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspectsand opinions with a MaxEnt-LDA hybrid," in Proc. Conf. EmpiricalMethods Natural Lang. Process., Cambridge, MA, USA, 2010.
[9] Z. Liu, H. Wang, H. Wu, and S. Li, "Collocation extraction usingmonolingual word alignment method," in Proc. Conf. EmpiricalMethods Natural Lang. Process., Singapore, 2009, pp. 487–495.
[10] K. Liu, H. L. Xu, Y. Liu, and J. Zhao, "Opinion target extractionusing partially-supervised word alignment model," in Proc. 23rdInt. Joint Conf. Artif. Intell., Beijing, China, 2013, pp. 2134–2140.pp. 56–65.
[11] Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, "Identifying featuresin opinion mining via intrinsic and extrinsic domain relevance,"IEEE Trans. Knowledge Data Eng., vol. 26, no. 3, p. 623–634, 2014. website.

