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. 6, June 2016, pg.11 – 22

Improved Optimized Sentiment Classification On Dynamic Tweets

Mrs. Elakkiya.R[1], Mrs. Jayasudha.M[2], Mr. Sivanesh Waran[3]

[1]PG Scholar, Department of CSE, SECE, Coimbatore & Anna university, India
[2]Assistant professor, Department of CSE, SECE, Coimbatore & Anna university, India
[3]Software Solution Engineer, Meteonic Innovation Private Limited, Banglore, India
[1] r.elakkiya4@gmail.com; [2] jayasudha.m@sece.ac.in; [3] siva@meteonic.com

Abstract— Real time Sentiment analysis is a subfield of Natural Language Processing concerned with the determination of opinion and subjectivity in a text, which has many applications. In this paper, classifiers for sentiment analysis of user opinion towards through comments and tweets sing Support Vector Machine (SVM) is described. The goal is to develop a classifier that performs sentiment analysis, by labeling the users comment to positive or negative. The extremely sparse text of tweets also brings down the performance of a sentiment classifier. In this paper, we propose a semi-supervised topic-adaptive sentiment classification (TASC) model, which starts with a classifier, built on common features and mixed labeled data from various topics. It minimizes the hinge loss to adapt to unlabeled data and features including topic-related sentiment words, authors' sentiments and sentiment connections derived from "@" mentions of tweets, named as topic-adaptive features. Text and non-text features are extracted and naturally split into two views for co-training. The TASC learning algorithm updates topic-adaptive features based on the collaborative selection of unlabeled data, which in turn helps to select more reliable tweets to boost the performance. We also design the adapting model along a timeline (TASC-t) for dynamic tweets. It also beats those semi-supervised learning methods without feature adaption. Finally, with timeline visualization of "river" graph, people can intuitively grasp the ups and downs of sentiments' evolvement, and the intensity by color gradation.

Keywords— Natural Language Processing,Support Vector Machine (SVM), Topic-Adaptive Sentiment Classification (TASC) model, and Adapting.

I. INTRODUCTION

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.

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."A different method for determining sentiment is the use of a scaling system whereby words commonly associated with having a negative, neutral or positive sentiment with them are given an associated number on a -10 to +10 scale (most negative up to most positive) and when a piece of unstructured text is analyzed

using natural language processing, the subsequent concepts are analyzed for an understanding of these words and how they relate to the concept.

Each concept is then given a score based on the way sentiment words relate to the concept, and their associated score. This allows movement to a more sophisticated understanding of sentiment based on an 11 point scale. Alternatively, texts can be given a positive and negative sentiment strength score if the goal is to determine the sentiment in a text rather than the overall polarity and strength of the text.

II. LITERATURE SURVEY

More and more people express their opinions on social media such as Facebook and Twitter. Predictive analysis on social media time-series allows the stake-holders to leverage this immediate, accessible and vast reachable communication channel to react and proact against the public opinion. In particular, understanding and predicting the sentiment change of the public opinions will allow business and government agencies to react against negative sentiment and design strategies such as dispelling rumors and post balanced messages to revert the public opinion.

In this paper, we present a strategy of building statistical models from the social media dynamics to predict collective sentiment dynamics. We model the collective sentiment change without delving into micro analysis of individual tweets or users and their corresponding low level network structures. Experiments on large-scale Twitter data show that the model can achieve above 85% accuracy on directional sentiment prediction.

In this paper, we develop a statistical model to predict the sentiment change in the social media and to address the following questions:How long back to the tweet history is most appropriate to learn a sentiment prediction model?How long does it take for the social media to demonstrate its response (sentiment change) after certain dynamics/events/activities occur?How long does the response on social media last? Additionally, we introduce three parameters: history window size, prediction bandwidth, and response time, and discover how they would the sentiment prediction quality. Comprehensive experiments are conducted to evaluate our sentiment prediction model on large-scale twitter data.

Sentiment analysis and opinion mining is the field of study that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. For the first time in human history, we now have a huge volume of opinionated data recorded in digital form for analysis.

Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world. For this reason,

when we need to make a decision we often seek out the opinions of others. This is true not only for individuals but also for organizations.

In summary, determining review helpfulness is an important research topic. It is especially useful for products and services that have a large of number reviews. To help the reader get quality opinions quickly, review sites should provide good review rankings. However, I would also like to add some cautionary notes. First, as we discussed in the chapter about opinion search and retrieval, we argued that the review ranking (rankings) must reflect the natural distribution of positive and negative opinions. It is not a good idea to rank all positive (or all negative) reviews at the top simply because they have high quality scores.

Various semi-supervised learning methods have been proposed recently to solve the long-standing shortage problem of manually labeled data in sentiment classification. However, most existing studies assume the balance between negative and positive samples in both the labeled and unlabeled data, which may not be true in reality. In this paper, we investigate a more common case of semi-supervised learning for imbalanced sentiment classification. In particular, various random subspaces are dynamically generated to deal with the imbalanced class distribution problem. Evaluation across four domains shows the effectiveness of our approach.

In this paper, we address semi-supervised learning for imbalanced sentiment classification. We first adopt under sampling to generate multiple sets of balanced initial training data and then propose a novel semi-supervised learning method based on random subspace generation which dynamically generates various subspaces in the iteration process to guarantee enough variation among the involved classifiers. Evaluation shows that our semi-supervised method can successfully make use of the unlabeled data and that dynamic subspace generation significantly outperforms traditional static subspace generation. To the best of our knowledge, this is the first work that systematically addresses the imbalanced class distribution problem in sentiment classification, especially under semi-supervised learning.

We investigated major SSL methods for identifying opinionated sentences in three domains. For movie review data, SSL methods attained state-of the-art results with a small number of labeled sentences. Even without a natural feature split, different co-training strategies increased the baseline SL performance and outperformed other SSL methods. Due to the nature of the movie review data, we suspect that opinion detection on movie reviews is an ‘easy’ problem because it relies, strictly speaking, on distinguishing movie reviews from plot summaries, which also involves genre classification. For other manually created data sets that are expected to reflect real opinion characteristics, the SSL approach was impeded by low baseline precision and showed limited improvement. With the addition of out-of-domain labeled data, however, self-training exceeded full SL. This constitutes a successful new approach to domain adaptation.

The rapid development of Web technology has resulted in an increasing number of hotel customers sharing their opinions on the hotel services. Effective visual analysis of online customer opinions is needed, as it has a significant impact on buildinga successful business. In this paper, we present OpinionSeer, an interactive visualization system that could visually analyze a large collection of online hotel customer reviews. The system is built on a new

visualization-centric opinion mining technique that considers uncertainty for faithfully modeling and analyzing customer opinions. A new visual representation is developed to convey customer opinions by augmenting well-established scatter plots and radial visualization.

To provide multiple-level exploration, we introduce subjective logic to handle and organize subjective opinions with degrees of uncertainty. Several case studies illustrate the effectiveness and usefulness of OpinionSeer on analyzing relationships among multiple data dimensions and comparing opinions of different groups.

III. EXISTING SYSTEM

Cross-domain sentiment classification is challenging and many works proposed their solutions. The existing approach called structural correspondence learning (SCL) for domain adaptation. It employed the pivot features as the bridge to help cross-domain classification. Another existing a spectral feature alignment (SFA). The supervised learning approach for automatically classifying the sentiment of tweets using emoticons as noisy labels for training data. The existing hash tags in tweets to build training data and demonstrated that part-of-speech features might not be useful for sentiment analysis of tweets.

The Twitter data as a corpus for sentiment analysis and tracking the influence of a particular brand activity on the social network. We focus on sentiment classification problem for tweets. Their lack sentiment labels in tweets for supervised learning of a sentiment classifier. The graph of hash tag co-occurrence for sentiment classification on hash-tag-level.The another algorithm developed a semi-supervised factor graph model which incorporates both the single image features and the image correlations to better predict the emotional impact. In social networks, showed that users generated the images which embedded users' emotional impact and influenced the emotional impact of each other. The predicted users' emotions in a social network, based on a dynamic continuous factor graph model which modeled the users' historic emotion logs and their social network.

The social relationships could be used to improve user-level sentiment analysis. Their approach was based on that the users who connected with each other might be more likely to hold similar opinions. However, our work focuses on the tweet-level sentiment classification across diverse topics, considering rich non-text features, such as user's sentiment, and network of "@" mentions. Visualizing themes and dynamics of complex data even in the area of text mining has been studied. However, there are only a few studies dealt with the sentiment visualization. As far as we know, proposed the opinion triangle and ring to visualize the hotel reviews on different polarities, and the periodic pattern is not applicable to visualize the sentiment evolvement of event series.

An intuitive sentiment visual with Opinion Blocks on different aspects of a product, which was an analysis of overall reviews. The used pixel cell-based sentiment calendars and high density geo maps for visualization. The presented simple directed paths to show the temporal relations between sentiment events. Nevertheless, those visualizations cannot show the dynamics and trend of sentiment along a timeline of a topic. In our sentiment visualization, we use a "river" graph to intuitively show the sentiment classification results and its dynamic process on a tweet topic.

IV. PROPOSED SYSTEM

The proposed an effective unsupervised learning algorithm, called polarity semantic orientation, for classifying reviews. A web-kernel based measurement is proposed as point wise mutual information measure the weight of a sentiment word, which is independent to the corpus collection in hand. But such a measurement is suitable for common sentiment words which always have a stable weight and fixed sentiment polarity, i.e., negative or positive in contexts of diverse topics. Besides, the text content of tweets is too sparse to extract plenty of salient sentiment words. Except text feature of sentiment classification based on support vector machine algorithm, there have been attracted many studies which considered other features to improve the classification result. Polarity Classification over Twitter offers different organizations a fast and effective way to monitor the feelings/emotions of general public towards their brand, business, politicians etc.

A wide range of features for training polarity classifiers for Twitter datasets have been researched in recent years with varying results. In this paper, we introduce a novel approach for automatically classifying and adding semantics as additional features the polarity of Twitter messages. These messages are classified as positive or negative or neutral with respect to a query term. The propose system focuses on addressing polarity classification for product features in product reviews by building semantic association between product features and polarity words. The results show that our method is encouraging.

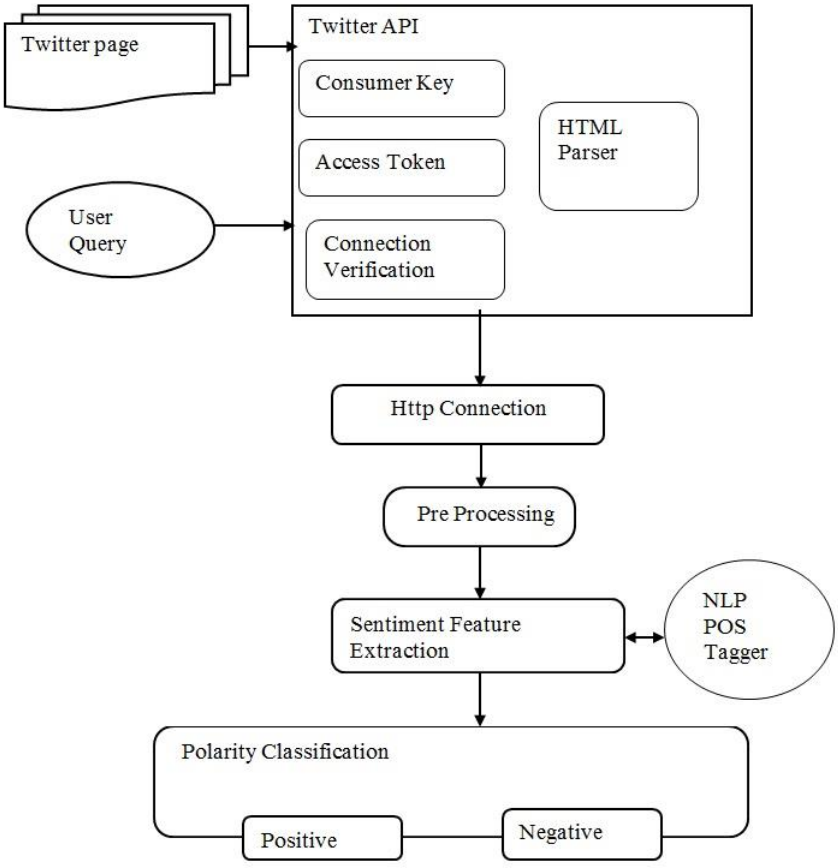

Figure.1 Architecture Diagram

4.1 ADVANTAGE OF PROPOSED SYSTEM

POS tagging for tweets on a topic and removing the common sentiment words

- Sentiment classification using unlabelled dataset.
- avoid bringing much noise review analysis
- The weight of a topic-adaptive sentiment word
- The iterations between optimization and adapting to unlabeled data agree with the co-training framework.
- TASC-t is designed to adapt along a timeline for the dynamics of tweets.
- The proposed algorithm designed visualization graph is demonstrated in the experiments, showing its effectiveness of visualizing the sentiment trends and intensities on dynamic tweet.

4.2 SYSTEM MODULES

The proposed system contains three modules.

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

4.2.1 TWITTER DATA COLLECTION

Using Twitter API a corpus of text posts are collected and formed a dataset of three classes: positive sentiments, negative sentiments, and a set of objective texts (no sentiments). To collect negative and positive sentiments, the same procedure is followed. 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. In the research, English language is used. However, the method can be adapted easily to other languages since Twitter API allows specifying the language of the retrieved posts.

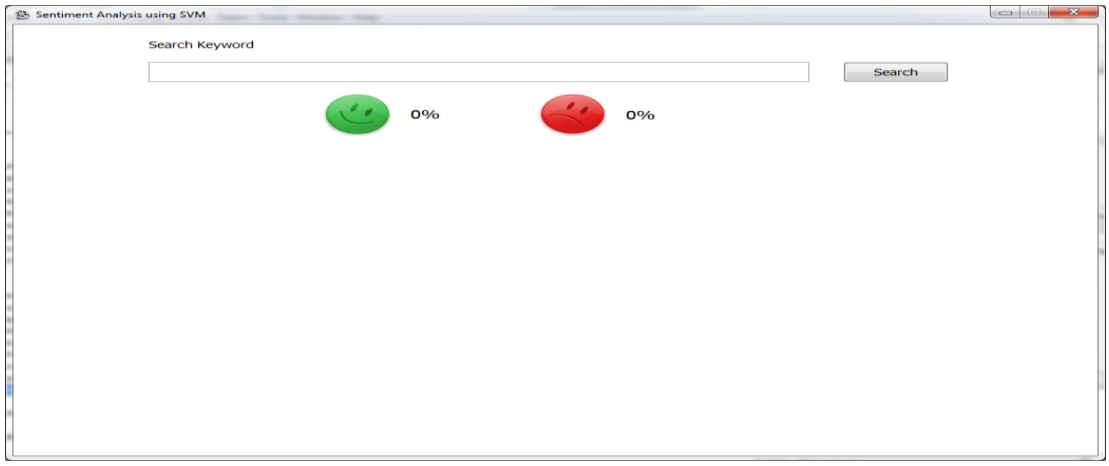

Figure.2 Main Form of Twitter Data

The above figure.2 to show main form of twitter sentiment analysis using some preprocessing techniques and classification algorithm and user submit query.

4.2.2 PREPROCESSING

The corpus has considerable amount of metadata such as date, time, and identity number etc. and to extract natural language content the data to a series of processing steps before the data can be used to extract features and train a classifier are need to be subjected. Here is a sample of the raw data before data processing.

- **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.

- **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
    - Irrelevant tags
    - Hyperlinks
    - #tags e.g. #msnbc2012
    - Twitter handles e.g. @pavanred
    - 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.

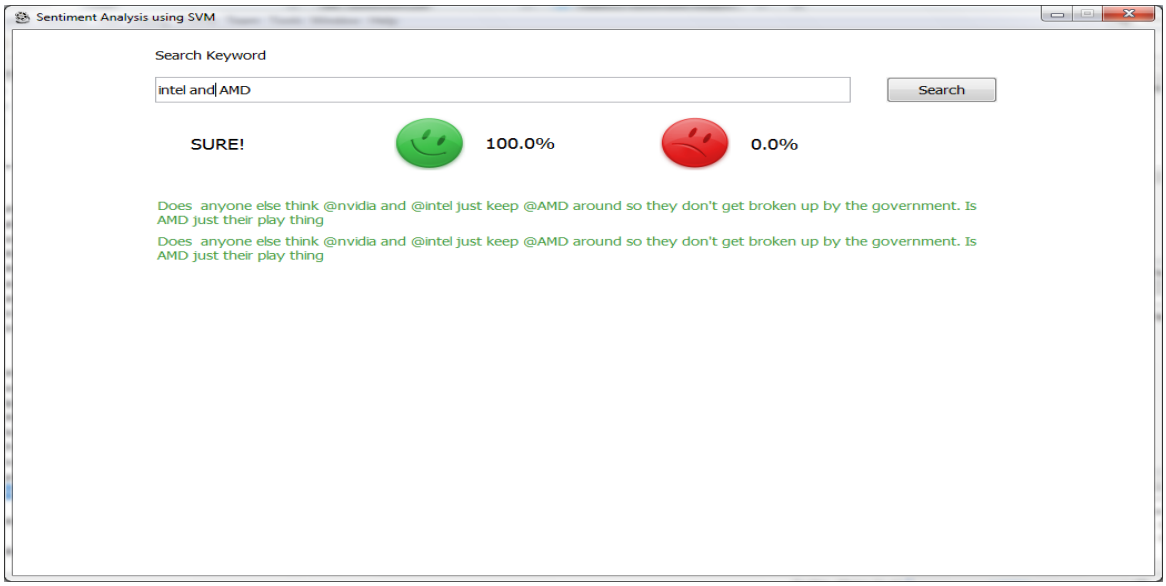

Figure.3 PreprocessingOpinion Result

The above figure.3 to show to select the Bayesian algorithm to perform the classification the total retrieved tweets is 500 and positive tweet is 216 and negative tweet is 101 neutral tweets 182.

4.2.3 POS TAGGER FEATURE EXTRACTION

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.

On setting out to classify a document, starts generally with depicting a very large number of words that need to be considered, even though very few of the words in the corpus are actually expressing sentiment. These extra features have two clear drawbacks that need to be eliminated. The first is that they show down the process of document classification, since there are far more words than needed. The second is that they can actually reduce accuracy, since the classifier is obliged to consider these words when classifying a document.

Clearly, there is an advantage in using fewer features; so in order to remove some of the unnecessary features, we resort to feature selection. As the name suggests, feature selection is a process through which we run across the corpus before the classifier has been trained and remove any features that seem unnecessary. This allows the classifier to fit a model to the problem more quickly as there will be less information to consider, thus allowing it to classify items faster.

4.2.4 SENTIMENT CLASSIFICATION

Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization, generally outperforming Naive Bayes. They are large-margin, rather than probabilistic, classifiers, in contrast to Naive Bayes and Maximum Entropy. In the two-category case, the basic idea behind the training procedure is to find a maximum margin hyper plane, represented by vector w, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible.

Support vector machines basically attempt to find the best possible surface to separate positive and negative training samples. Support Vector Machines (SVMs) are supervised learning methods used for classification. In this project, SVM is used for sentiment classification. First module is sentiment analysis and Support vector machines perform sentiment classification task on review data. The goal of a Support Vector Machine (SVM) classifier is to find a linear hyper plane (decision boundary) that separates the data in such a way that the margin is maximized. Look at a two class separation problem in two dimensions observe that there are many possible boundary lines to separate the two classes. Each boundary has an associated margin. The idea for SVM is to find a boundary (known as a hyper plane) or boundaries that separate class of data.

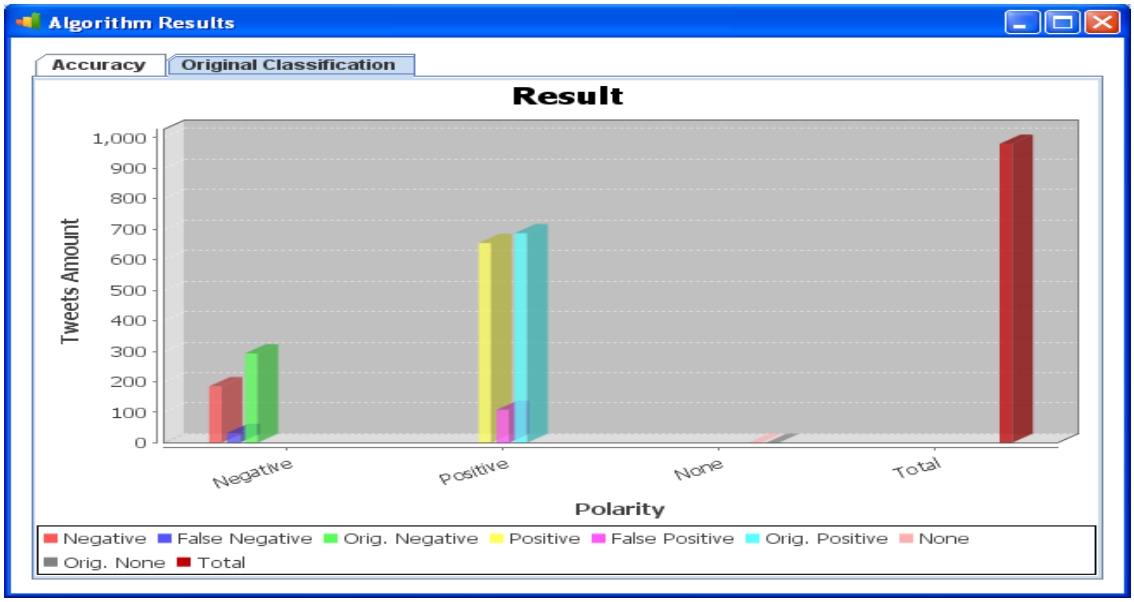

Figure.4 No of Sentiment Word Graph

The above figure.4 to show subjectivity algorithm based classification total no of tweets to parse total no of objective, subjective, none and total to perform the chat.

V. PERFORMANCE ANALYSIS

Performance analysis 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.

TABLE .1 Comparison of Classification Algorithm

Classifiers	**Avg. Accuracy**	**Max. Accuracy**	**Avg F**
SCL	55.82	53.11	0.52
TASC	63.11	80.12	0.85
Polarity Classification	92.43	95.32	0.99

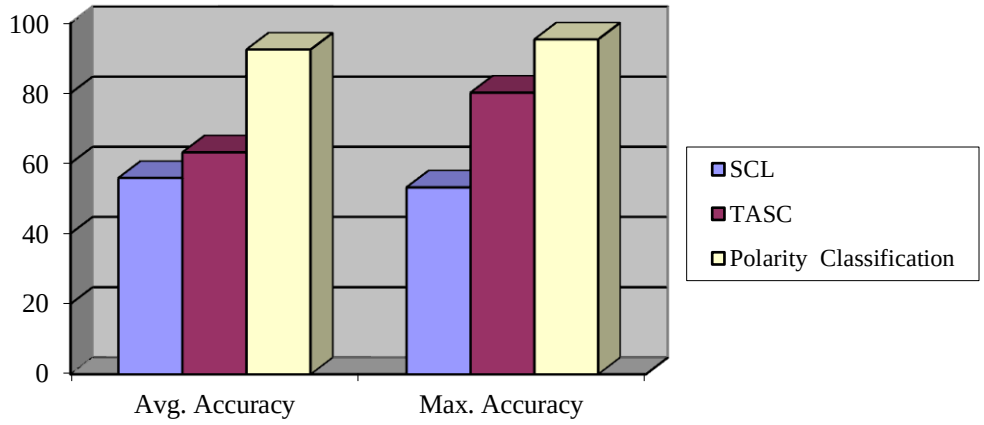

Figure.5 Comparison of classification results

TABLE .2 Comparison Precisions and Recall

Polarity	TASC			Polarity Classification		
	Precision (%)	Recall (%)	F	Precision (%)	Recall (%)	F
Positive	75	74.3	75	94.2	81.8	92
negative	72.3	74	72.1	91.3	89.2	91.6
Natural	74	65	66.5	73	75	78

VI. CONCLUSION

The proposed Sentiment analysis is a subfield of NLP concerned with the determination of opinion and subjectivity in a text, which has many applications. In this paper we will be studying a bout classifiers for sentiment analysis of user opinion towards political candidates through comments and tweets sing Support Vector Machine (SVM) the goal is to develop a classifier that performs sentiment analysis, by labeling the us ers comment to positive or negative. The proposed a semi-supervised sentiment classification method which focuses on using ranked opinion words to build a semi-supervised sentiment classifier based on the co-training framework. The method itself doesn't rely on other language resources. Our method only needs a small number of labeled training instances and some unlabeled instances. We find that ranked opinion words are helpful for improving the final sentiment classification accuracy.

ACKNOWLEDGEMENT

I am thankful to Mrs.M.Jayasudha Assistant Professor Of CSE Department for her guidance, support and supervision. And also for providing the information and ready to help any time in completion of this paper.

REFERENCES

[1] L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from time-series social media," in Proc. 1st Int. Workshop Issues Sentiment Discovery Opinion Mining, 2012, p. 6.
[2] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lect. Human Lang. Technol., vol. 5, no. 1, pp. 1–167, 2012.
[3] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, "Cross-domain co-extraction of sentiment and topic lexicons," in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics: Long Papers, 2012, pp. 410–419.
[4] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 751–760.
[5] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification," in Proc. 22nd Int. Joint Conf. Artif. Intell., 2011, pp. 1826–1831.
[6] N. Yu and S. Kubler, "Filling the gap: Semi-supervised learning for opinion detection across domains," in Proc. 15th Conf. Comput. Natural Language Learn., 2011, pp. 200–209.
[7] S. Gao and H. Li, "A cross-domain adaptation method for sentiment classification using probabilistic latent analysis," in Proc. 20th ACM Int. Conf. Inform. Knowl.Manage., 2011, pp. 1047–1052.
[8] S. Liu, F. Li, F. Li, X. Cheng, and H. Shen, "Adaptive co-training SVM for sentiment classification on tweets," in Proc. 22Nd ACM Int. Conf. Inform. & Knowl.Manage., 2013, pp. 2079–2088.
[9] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: A graph-based hash tag sentiment classification approach," in Proc. 20th ACM Int. Conf. Inform. Knowl.Manage., 2011, pp. 1031–1040.
[10] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu, "OpinionSeer: Interactive visualization of hotel customer feedback," IEEE Trans. Visualization Comput. Graph., vol. 16, no. 6, pp. 1109–1118, Nov. 2010.
[11] M. Thelwall, K. Buckley, and G. Paltoglou, (2011) "Sentiment in twitter events," J. Am. Soc. Inform. Sci.Technol., vol. 62, no. 2, pp. 406–418.
[12] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, (2011) "Sentiment analysis of twitter data," in Proc. Workshop Lang. Soc. Media, pp. 30–38.
[13] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, (2011) "User-level sentiment analysis incorporating social networks," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 1397–1405.
[14] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, (2011)"Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach," in Proc. 20th ACM Int. Conf. Inform. Knowl.Manage., pp. 1031–1040.

