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. 3, Issue 5, May 2015

Multi Aspect Based Document Level Sentiment Analysis for Educational Institute Analysis

Kalpana Razdan, Abhinav Raj, Vaidehi Dastapure, Parth Srivatava, Mrunal Shinde, Uma Nagaraj

Department of Computer, University of Pune, Pune, Maharashtra, India

ABSTRACT*:* Sentiment Analysis is used to determine the attitude of a writer with respect to some topic or the overall contextual polarity of a document. The objective of our project is to build an interactive portal wherein the comparative analysis of various colleges can be visualized. In doing so ,instead of the basic factual information, analysis will be done based on the feedback and reviews acquired from various sources. In this approach , document level sentiment analysis will be done considering every aspect of the same using the techniques of natural language processing.

KEYWORDS: Sentiment Analysis, Bag of Words, POS tagging, Entity Extraction**.**

I. INTRODUCTION

Sentiment Analyses concentrates on classifying documents according to their opinion and emotions expressed by their authors. Judging a document's orientation as positive or negative is a common two-class problem in sentiment analysis, which is also known as sentiment orientation analysis in text classification. [1] Text classification has been found very useful in many areas. With the expansion of internet, users are encouraged to leave feedback and comments and based on those, a lot of business trends are coming up. Many approaches have already been implemented for the same. One of the major approaches is bag of words and bag of nouns which uses the underlying noun-adjective clustering thus forming proper structures and extracting features. (Dim vector approach is also used for text classification.) In our project we have used the bagging and boosting technique to cluster features and the rest of the algorithm lies on that. [2]

II. RELATED WORK

There are two basic procedures to detect sentiments from text. They are Symbolic techniques and Machine Learning techniques. The next two sections deal with these techniques.

A. Symbolic Techniques

Much of the research in unsupervised sentiment classification using symbolic techniques makes use of available lexical resources. Turney used bag-of-words approach for sentiment analysis. In that approach, relationships between the individual words are not considered and a document is represented as a mere collection of words. To determine the overall sentiment, sentiments of every word is determined and those values are combined with some aggregation functions. He found the polarity of a review based on the average semantic orientation of tuples extracted from the review where tuples are phrases having adjectives or adverbs. He found the semantic orientation of tuples using the search engine Altavista. Kamps et al. used the lexical database WordNet to determine the emotional content of a word along different dimensions. They developed a distance metric on WordNet and determined the semantic orientation of adjectives. WordNet database consists of words connected by synonym relations. Baroni et al. developed a system using word space model formalism that overcomes the difficulty in lexical substitution task [1]. It represents the local context of a word along with its overall distribution. Balahur et al. introduced EmotiNet, a conceptual representation of text that stores the structure and the semantics of real events for a specific domain. Emotinet used the concept of Finite State Automata to identify the emotional responses triggered by actions. One of the participants of SemEval 2007 Task No. 14 used coarse grained and fine grained approaches to identify sentiments in news headlines. In coarse grained

DOI: 10.15680/ijircce.2015.0305109

approach, they performed binary classification of emotions and in fine grained approach they classified emotions into different levels. Knowledge base approach is found to be difficult due to the requirement of a huge lexical database. Since social network generates huge amount of data every second, sometimes larger than the size of available lexical database, sentiment analysis became tedious and erroneous. [3]

B. Machine Learning Techniques

Machine Learning techniques use a training set and a test set for classification. Training set contains input feature vectors and their corresponding class labels. Using this training set, a classification model is developed which tries to classify the input feature vectors into corresponding class labels. Then a test set is used to validate the model by predicting the class labels of unseen feature vectors. A number of machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews. Some of the features that can be used for sentiment classification are Term Presence, Term Frequency, negation, n-grams and Part-of-Speech. These features can be used to find out the semantic orientation of words, phrases, sentences and that of documents. Semantic orientation is the polarity which may be either positive or negative. Domingos et al. found that Naive Bayes works well for certain problems with highly dependent features. This is surprising as the basic assumption of Naive Bayes is that the features are independent. Zhen Niu et al. introduced a new model in which efficient approaches are used for feature selection, weight computation and classification. The new model is based on Bayesian algorithm. Here weights of the classifier are adjusted by making use of representative feature and unique feature. ’Representative feature’ is the information that represents a class and ’Unique feature’ is the information that helps in distinguishing classes. Using those weights, they calculated the probability of each classification and thus improved the Bayesian algorithm. Barbosa et al.designed a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy training set to reduce the labeling effort in developing classifiers. Firstly, they classified tweets into subjective and objective tweets. After that, subjective tweets are classified as positive and negative tweets. Celikyilmaz et al. developed a pronunciation based word clustering method for normalizing noisy tweets. In pronunciation based word clustering, words having similar pronunciation are clustered and assigned common tokens. They also used text processing techniques like assigning similar tokens for numbers, html links, user identifiers, and target organization names for normalization. After doing normalization, they used probabilistic models to identify polarity lexicons. They performed classification using the BoosTexter classifier with these polarity lexicons as features and obtained a reduced error rate. Wu et al. proposed a influence probability model for twitter sentiment analysis. If @username is found in the body of a tweet, it is influencing action and it contributes to influencing probability. Any tweet that begins with @username is a retweet that represents an influenced action and it contributes to influence probability. They observed that there is a strong correlation between these probabilities. Pak et al. created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features. In that method, there is a chance of error since emotions of tweets in training set are labeled solely based on the polarity of emoticons. The training set is also less efficient since it contains only tweets having emoticons. [3]

Xia et al. used an ensemble framework for sentiment classification. Ensemble framework is obtained by combining various feature sets and classification techniques. In that work, they used two types of feature sets and three base classifiers to form the ensemble framework. Two types of feature sets are created using Part-of-speech information and Word-relations. Naive Bayes, Maximum Entropy and Support Vector Machines are selected as base classifiers. They applied different ensemble methods like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy. Certain attempts are made by some researches to identify the public opinion about movies, news etc from the twitter posts. V.M. Kiran et al. utilized the information from other publicly available databases like IMDB and Blippr after proper modifications to aid twitter sentiment analysis in movie domain.[2]

III. METHODOLOGY

The proper working of our web portal is according to the Diagram which is shown as following down under.

The steps which are followed in the processing of the data analysis is given as follows:-

1. DATA SOURCE

The data source is mainly the feedbacks which are collected from the user has given for a particular college. The users can give there reviews or feedback on any aspect of a college. These feedbacks or data is passed through the Entity Database in which the feedbacks are saved according to the aspects that also acts as an entity in the database.

2. DATA PRE-PROCESSING

When the feedback given by any random user on any aspect of a college, so our job is to thoroughly analyse the data and then show a proper result which can add to the sentiment score of an aspect of a college. For the analysis of the feedbacks we use some processing techniques which are mainly, Sentiment Extraction, Stopword Elimination, and Term Stemming.

i. Sentiment Extraction: Sentiment extraction is used to extract the positivity or negativity of a feedback given by the user. In this we are naturally summing up the number of positive and negative words in the reviews through which we can calculate the sentiment score of a feedback.

ii. Stopword Elimination : Some of the words in the feedback which do not contribute to the sentiment score of any aspect are called stopwords. These word are mainly, the, are, as, is, etc. These words are to be eliminated from the feddback so as to get the proper sentiment score of that feedback.

iii. Term Stemming: Term Stemming is the process of stemming of the words which are used in the feedback. This can help in getting the proper meaning of the word used in the feedback. For example, if a word “getting” is used is the feedback, so in this the “-ing” part of the word will be stemmed out during the processing if the feedback.

3. FEATURE SELECTION

After the processing on the data has been done, the feedback is now ready for the feature extraction. Suppose a user came to your web portal and gives a feedback on any of the one college. The user may give review on all the aspect of the college in one feedback only. So to extract the sentiment score on all the aspects, clustering is being done, where all the aspects are clustered to give a valid sentiment score. Through this method the score on every aspect is collected.

4. OPINION ORIENTATION IDENTIFICATION

When all the features of the college is clustered separately through the Feature Selection. Now we just have to calculate the sentiment score on different aspect of the college. The number of positive word and negative are calculated in order to find out the polarity of the feedback towards any aspect. The sentiment score is calculated according term frequency of the words used in the review. This helps in getting the proper sentiment analysis of the a review.

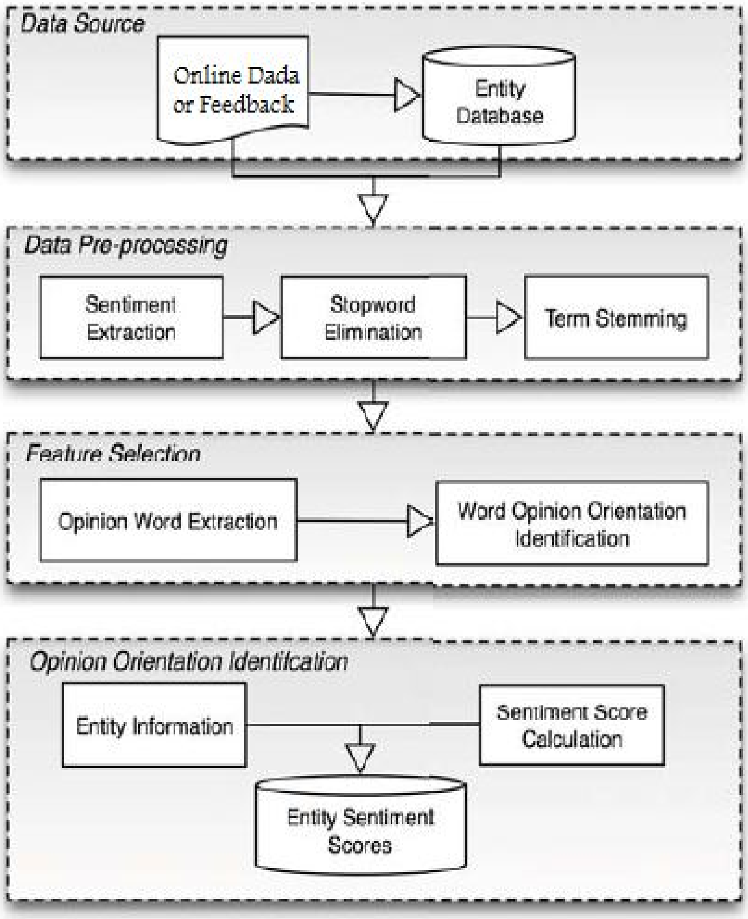


Figure 2: Methodology Diagram

IV. IMPLEMENTATION

The whole implementation structure can be divided into 3 layers namely :-

i) Presentation layer- It compromises of user interface elements designed using JSP,HTML and CSS it basically deals with appearance of the system to external users.

ii) Application layer- It comprises functionality related to NLP like stopword removal , stemming, pos tagging, tokenizer, sentiment word extractor or feature extractor this is done writing code and logic involved in each step in java servlet, It is the core part of implementation.

iii) Database layer- All the data base handling job is done here it has various components like Sentiment Word list with Sentiment Score, Various tables consisting of information related to users and their feedbacks, Also the output after each NLP steps is stored in database for further reference, Database handling is done using MYSQL technology it is the backend of the system

V. THE EXPERIMENT AND RESULTS

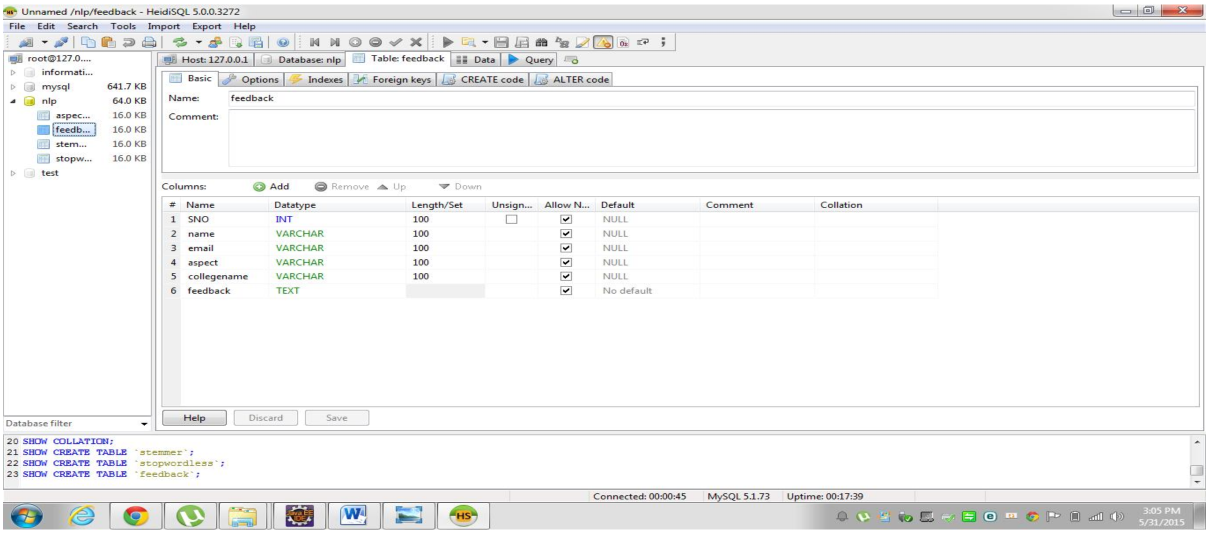
The feedback obtained from various sources is analyzed with the help of opinion mining and a comparative chart is drawn based on pre-defined aspect .

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. 3, Issue 5, May 2015



Name: feedback

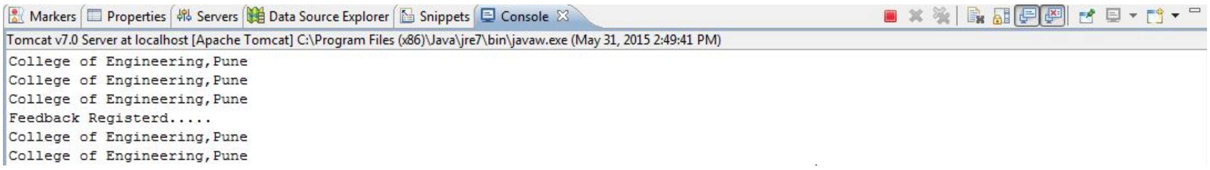
Comment:

Columns:

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Default	Comment	Collation
1	SNO	INT	100	□	☑	NULL		
2	name	VARCHAR	100		☑	NULL		
3	email	VARCHAR	100		☑	NULL		
4	aspect	VARCHAR	100		☑	NULL		
5	collegename	VARCHAR	100		☑	NULL		
6	feedback	TEXT			☑	No default		

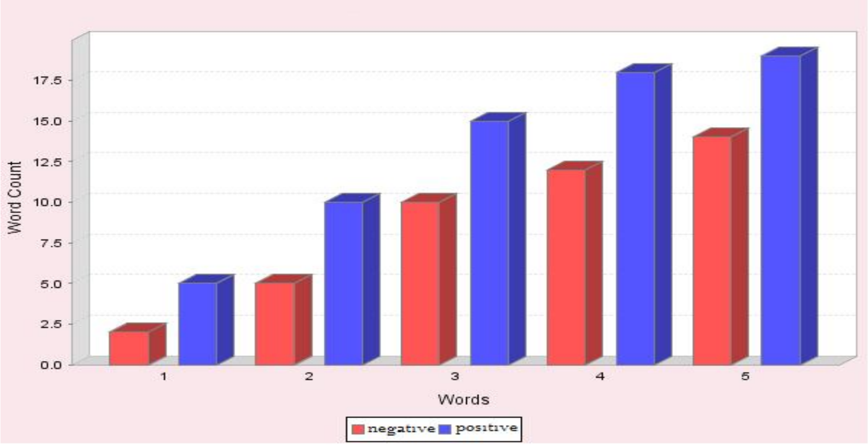
```
SHOW COLLATION;
SHOW CREATE TABLE `stemmer`;
SHOW CREATE TABLE `stopwordless`;
SHOW CREATE TABLE `feedback`;
```

After running Mysql in backend feedback table is created .This table contains information about the feedback form.Table ,it has Name , Datatype and all the comments. After Stemming and Stopword removing process Stemmer and Stopward table is created at the backend .



```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files (x86)\Java\jre7\bin\javaw.exe (May 31, 2015 2:49:41 PM)
College of Engineering,Pune
College of Engineering,Pune
College of Engineering,Pune
Feedback Registerd.....
College of Engineering,Pune
College of Engineering,Pune
```

In console we can see actual result “Feedback is registered”.
And number of Positive and negative words also there in result.



This is graphical representation of scoreboard of Sentiment Analysis of Educational Data Catering Process.

DOI: 10.15680/ijircce.2015.0305109

VI. CONCLUSION

In this paper document level sentiment analysis is evaluated by by obtaining aspect level sentiment score and the corresponding weightages given to the elements. It is noted that when we compute aspect based document level sentiment analysis accuracy is high as compared to sentiment analysis at direct document level. Rather of giving information in terms of only positive and negative class our approach gives a multiaspect sentiment analysis providing close-grained view of sentiments. Also negation handling is very important for calculating accurate sentiments otherwise it can mislead to inconsistent information.

REFERENCES

1. N. D. Valakunde and Dr. M. S. Patwardhan 2013“Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process”. Book By Han and Kamber. Data Mining.
2. Janxiong Wang and Andy Dong 2010“A Comparison of Two Text Representations for Sentiment Analysis”. ”SentiMeter-Br: a New Social Web Analysis Metric to Discover Consumers Sentiment”
3. Renata Lopes Rosa, Demstenes Zegarra Rodrguez.,2013 IEEE 17th International Symposium on Department of Computer Engineering, MIT AOE 37
4. ”Sentiment Analysis on Tweets for Social Events” Xujuan Zhou and Xiaohui Tao, Jianming Yong.,Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design.
5. ”Sentiment Analysis in Twitter using Machine Learning Techniques” Neethu M S and Rajasree R.,IEEE – 31661
6. ”Twitter Sentiment Analysis: A Bootstrap Ensemble Framework” Ammar Hassan* and Ahmed Abbasi+ and Daniel Zeng.,SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013
7. ”Text Sentiment Analysis Algorithm Optimization Platform Development in Social etwork” Yiming Zhao, Kai Niu, Zhiqiang He, Jiaru Lin, and Xinyu Wang.,2013 Sixth
8. International Symposium on Computational Intelligence and Design. Sentiment Analysis: A Combined Approach Rudy Prabowo, Mike Thelwall.