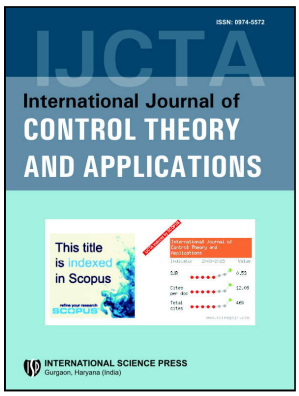

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 19 • 2017

Opinion Analysis of Users' Review for Educational Institutes

Deepali Londhe[1], Emmanuel M.[2] and Aruna Kumari[3]

[1] Ph.D. Scholar, Department of Computer science and Engineering K.L.University, Vijayawada, Andhra Pradesh, India, Email: deep_londhe@yahoo.com
[2] Department of Information Technology PICT, Pune, Maharastra, India, Email: emman2001@gmail.com
[3] Department of Electronics and Computer Engineering K.L.University, Vijayawada,Andhra Pradesh, India, Email: aruna_d@kluniversity.in

Abstract: The growing interest in the field of opinion mining and its applications in numerous areas of information and social science has triggered many researchers to take a deep look at the field. As an education pattern, Educational Institutes are becoming more and more important. But it is relevant to have gainful insight of users' view about a particular institute and lead to an opinion accordingly. Users can decide about selecting the Institute with the help of Predictive analytics on collected reviews. Opinions in the form of reviews, extracted from websites and social media will go through automatic sentiment analysis to identify the sentiment of opinions. The system has considered the problem of classifying documents not by topic, but by overall sentiment. This paper uses a known supervised learning algorithm 'Naive Bayes' to create a prediction model.

Keywords: *Sentiment Analysis, machine learning, Social Media, Educational Data Mining, Opinion mining*

1. INTRODUCTION

Sentiment analysis or opinion mining [7, 8] is an application of language processing, and text analysis. Sentiment Analysis can be used to identify the attitude, judgment, evaluation or emotional communication of the author or a speaker of the document.

The Increasing use and interactivity of the Internet and the growth in expressed opinion on the social platform such as review sites, social Media [6], on line forums and web blogs, collectively called the user generated content, provide many others with an opportunity to understand the opinions of many people, at a scale which was not possible before. Analyzing direction-based texts, i.e. texts having views and sentiments is vital sentiment analysis technique [1]. However, as these expressed opinions are generally huge in numbers, making it difficult, for users to read all of the reviews and create an overall view about the Institutes and also identity their orientations [15]. Therefore, the inkling behind this experiment lies in providing such data by forming a system where the reviews will be categorized into positive or negative, referencing to the expressed sentiment within or via those reviews.

In this paper, to help User to know the opinions of candidates on Educational Institutes and to help them improve the services, the method of Opinion mining is applied. An opinion mining system for reviews on the Educational Institutes has been developed using R language [22]. The purpose of this system is to extract and summarize the opinions and reviews, and determine whether these reviews and opinions are positive or negative.

2. RELATED WORK

In recent years large amount of data is available on line in the form of reviews on blogs, forums, social media sites, etc. which is way beyond the visual capacity of any person. Hence, It lead to an urgent need to come up with an autonomous innovative techniques that can automatically collect and analyses the attitudes of users expressed in their reviews.

As such, performing an automatic sentiment classification of blogs and projecting the overall reviews of specific institutes as positive or negative would definitely be useful to users. So far, sentiment classification has been conducted on a wide range of domains such as feature film reviews, product reviews, travel reviews and e-learning reviews. But for Educational Reviews, huge scope is available for further work. Many scientists have so far have used training machine learning algorithms [18, 24] to classify reviews.

In this section, earlier approaches, pros and cons to the opinion mining problem are discussed. Opinion mining has two subtasks: the extraction of information and the analysis of sentiment. The information extraction task embarks on the ways of extracting elements which constitute opinions. Subsequently in sentiment analysis task, the focus is on classifying of documents according to sentiment orientation such as positive vs. Negative. The previous finding states that there are 2 most efficient ways in which experiments of sentiment analysis can be carried out: Naïve Bayes, and Support Vector Machine algorithms.

Pang [16] have nicely stated a meaningful work based on topic classification techniques. The author aims to verify whether a certain selected bunch of machine learning algorithms can positively provide good result when sentiment analysis is perceived as document topic analysis bearing two topics: positive and negative. Actually, [16] shows results for experiments by means of: Naïve Bayes, Maximum Entropy and Support Vector Machine algorithms. Interestingly the performed experiments have exhibited results comparable to other solutions ranging from 71 to 85% depending on the method and test data sets .Similarly [3] has used these same classifiers to classify educational data in terms of student feedback comments about faculty performance. In Direct document level sentiment analysis SVM gives 67.23% accuracy whereas NB 65.10 accuracy, Inspect based document level sentiment analysis SVM gives 81 % accuracy whereas NB gives 72.80% accuracy.

Kechaou [1] has analyzed compared various single and hybrid classifiers like HMM, SVM and NB on a particular set. Each classifier gives various levels of accuracy. Although SVM is most accurate NB is considered faster and more reliable for small dataset. Raut, Londhe[4,5] presented machine learning and SentiWordNet based approach for opinion mining and opinion summarization. The paper has used various methods such as NB, SVM, Decision Tree, SentiWordNet [4, 12] method which gives different accuracy %.

NB classifier gives highest accuracy of 88% whereas decision tree method gives lowest accuracy of 78%. [4] has performed opinion mining on user's opinion about different hotels where the reviews are collected from web. This paper states that the NB classifier gives highest accuracy of 88%.

This paper will put a bright light on a new area of study by implementing opinion mining in the context of Educational Institutes. Opinion mining is an important area that has been recently explored in the business and computing domains. In order to analyze the data gathered through this process, we need to apply adequate sentiment classification techniques to the domain of mining reviews gathered from blogs and forums.

3. METHODOLOGY

Sentiment classification aims to assign a category/class to a data or document from a predefined set of categories/ classes. The predefined category/classes set mostly made up of some sentiment classes, e.g. 'positive' or 'negative'. In supervised machine learning, a trained statistical classifier is used for sentiment classification. The trained classifier projects the orientation of sentiment based on input documents. Subsequently the machine learning algorithms cannot work directly with the input documents. Testing of algorithm is necessary in case of supervised machine learning algorithm. That is why a preliminary stage, also called pre-processing, is required. Firstly, textual data should be pre-processed prior classifying it. Pre-processing transforms document data into a relevant format which will make it to the classification phase. The pre-treated documents, go through the phase of removal of punctuation, numbers, stop words and finally stemming. Both training and test sets of documents are pre-processed in the same way.

The proposed methodology in this paper is shown in Fig. 1. Sentiment analysis process classified broadly into four stages. In the first stage the data is extracted from social media and websites. The reviews collected are then preprocessed by applying various filtering techniques such as stemming, stop word removal etc. Numerical representation of reviews is done in the next stage to identify relevancy of terms. Sentiment classification is important stage where reviews are classified into specified sentiment classes. Many algorithms are available for

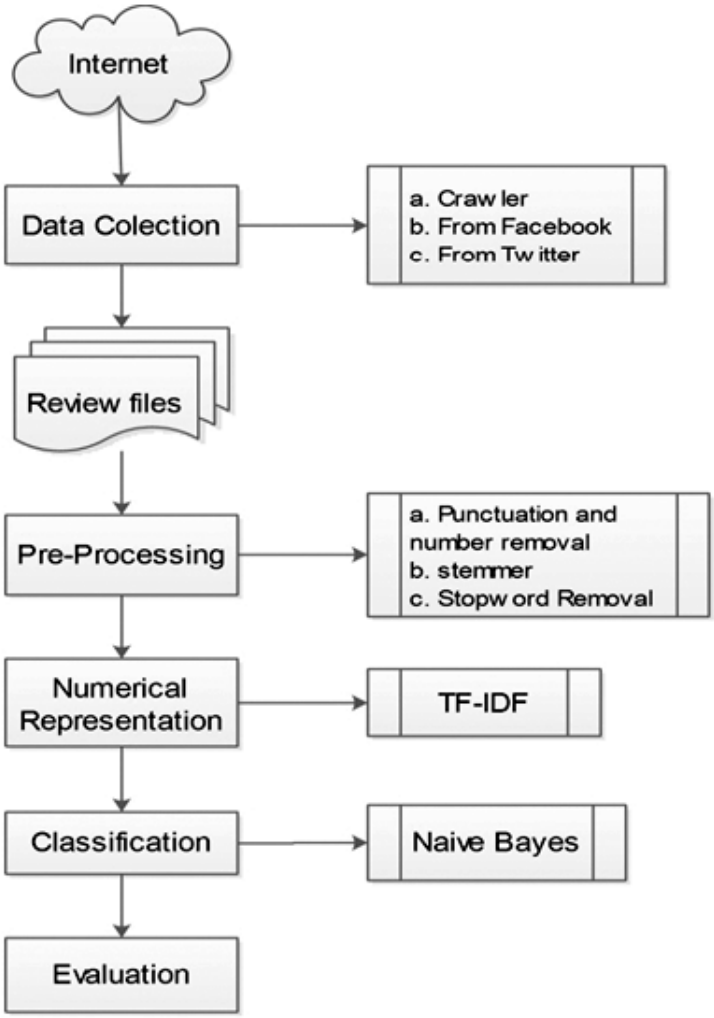

Figure 1: Sentiment analysis Methodology

sentiment classification such as Naive Bayes, Support vector machine, Decision Trees, etc[4, 6, 16]. Naïve Bayes classifier is selected for this study.

3.1. Data Collection

For Opinion mining on educational institutes, data in the form of reviews are collected from Social Sites, Twitter and Facebook. The sites where people submit reviews of colleges like www.Shiksha.com, www.career360.com, the reviews are extracted. The extraction of these reviews need design of crawler specific to the these websites, which will help to get specific review for the required college. For extraction of reviews from Twitter and Facebook, the respective APIs are used.

The inventory of seed URLs helps any Internet crawler. The basic formula takes it as input and executes the subsequent steps in loop. Generally an address from the address list is extracted and verified against the science address of its host name. Transfer of the corresponding document and extraction of links are performed [19]. For each and every of the extracted links its Associate absolute address is verified. Later it is added to the list of URLs to get transferred and provided. . General Algorithm steps are given below:

i. A seed URL as input.

ii. Create ontology tree and find out the knowledge path.

iii. Downloads all the URL's that are associated with input URL.

iv. Extract all links present in downloaded web page and insert into URL frontier.

v. extract the reviews which are present in the web-page downloaded with help of parser using its html tags

vi. Repeat these steps until to get more relevant result

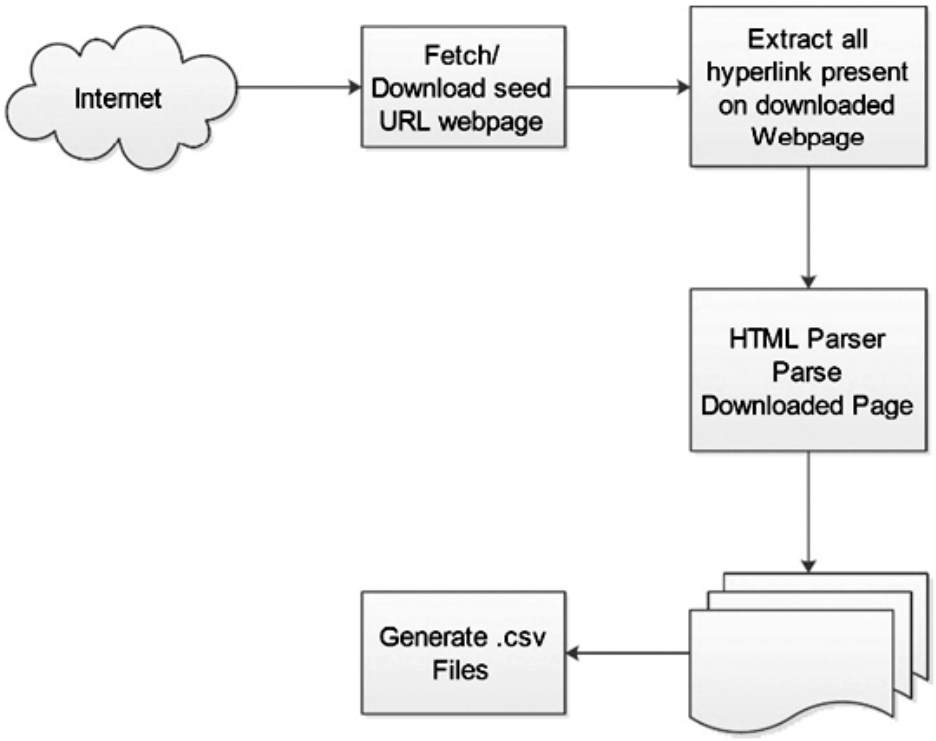

Figure 2: Working of Crawler

One of the source of collecting college information is from twitter in real time . For getting tweets easily, an twitter API is used[14, 20]. The system has been cooperated with R and twitter API using "*OauthFactory"* function of "*twitteR*" package [26].

To receive tweets by specifying keyword a filter stream function in "*streamR*" package can be used. One can acquire recent tweets specifying keyword for setting time long. Subsequently with "*filterStream*" function, you can get set of tweets ,user name, contribution time. However this function can output data by .csv file, but this method uses set of tweet information directly on R. In this paper educational reviews are acquired from Facebook [17] using "R" package known as "*Rfacebook*" [25].

3.2. Pre-Processing

Unwanted symbols, words, numbers have to be removed from collected reviews as these things are not useful in the process of Sentiment Analysis. Stemmer can be used for bringing the words to the root form. Stopword removal can be achieved using the list of stopwords.

Punctuations like ",; : ' { }) (" and numbers, if present in data after pre-processing , they might reduce the accuracy of classifier. So this step is very important in data pre-processing. After collecting final data, some unwanted punctuations and numbers are removed from each sentence in data set. Text mining package in "R" called "tm" [24] helps to remove unwanted numbers and punctuations form text.

Stemming [1, 3] is the term used in information retrieval to describe the process for reducing inflected (or sometimes derived) words to their root form. The stems are preferred over words serves two beneficial role in sentiment analysis. First, the sparseness in the data is decreased, since there are fewer distinct stems compared to distinct words. Second, semantic information captured and gathered in a better way. For example, words like "Teaching", "Understanding", etc., after stemming is converted into "teach", "Understand", etc. Hence, with the usage of stem (root) word, sentence words are nicely compared with number of positive/negative words making it easy for analysis.

Stop words [1, 3], generally known to be noise words or most common words, which don't solve any significant purpose in the field of sentiment classification. Stop words are the most common words in a any language, but there is no specific, single, explicit universal list of them. They are used by all natural language processing tools, and indeed not all tools even use such a list. With this context, we have established a list of words containing mainly English pronouns, particles, special characters and numbers e.g. prepositions, pronouns, and some adverbs. For example "The Institute is good" is a positive opinion sentence. After performing the step of stop words removal, the output derived is as "Institute good". This is an efficient and easier way to identify sentiment bearing words and entities for classifying the document.

3.3. Numerical Representation

In this paper, the textual data is represented into numerical format for computation. Computation of TF-IDF [1] value of each and every feature (sentiment bearing words) from the set of training documents is performed. TF-IDF is one of the most widely used representations. TF-IDF methods consider both term frequencies in a document as well as the relevance of a term in the entire collection of documents. The formula for calculating the TF-IDF can be written as follows:

$$TF=Log\ (f\ (t,\ d)\ +1) \tag{1}$$

TF stands for term frequency i.e. the number of times the term appears in the document; IDF is the Inversed Document Frequency given by the below equation:

$$IDF=Log\ (N/n) \tag{2}$$

N is the total number of training documents and n is the number of documents the term appears in. IDF is useful in minimizing the weight of term with low discriminative value.

3.4. Classification

Naïve Bayes [4, 5, 9, 17, 20, 21, 23] is a probabilistic learning approach that assumes, the terms in documents occur independently [3].

One approach to text classification is to assign to a given document d the class. If *D* is dataset of training data instances *X = (X1, X2 ... Xn)* having n attributes and m classes *C1, C2 ... Cn*. The classifier predicts text *X* belongs to class having higher probability values for given conditions. This is shown in equation (3)

$$P(C_i/X) > P(C_j/X) fpr\ 1 \le j \le m, j \ne i \tag{3}$$

Where is calculated using Bayes theorem,

$$P(C_i/X) = \frac{P(X/C_i)P(C_i)}{P(X)} \tag{4}$$

Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization still tends to perform surprisingly well.

4. EXPERIMENT

For the experimentation student feedback, comments about Educational institutes is done is collected. This dataset is collected from various sites such as "Facebook, Twitter, stupidsid.com, carrier360.com" which has 1500 positive and negative Reviews about Educational institutes. We labeled each and every comment for classification purpose as positive and negative. For training NB and SVM[3, 4, 5, 9, 11, 17, 23]classifier we randomly chose 1000 reviews, having 500 positive and 500 negative reviews and remaining 500 comments are used for testing. We tested different machine learning classification algorithms NB and SVM for opinion classification of comments/reviews using RStudio. We evaluated our approach of opinion classification based on parameters such as Precision(P), Recall(R), Fmeasure(F) [2, 10].

Table 1 illustrates the confusion matrix used for the validation stage.

Table 1
Confusion Matrix

Original Class	*Positive*	*Negative*
Positive	N P,P	N P,N
Negative	N N,P	N N,N

Precision (P) is the ratio of correct cases through the system outputs.

$$\text{Precision(Positive)} = \frac{N_{P,P}}{N_{N,P} + N_{P,P}} \tag{5}$$

$$\text{Precision(Negative)} = \frac{N_{N,N}}{N_{N,N} + N_{P,N}} \tag{6}$$

Recall (R) is the ratio of correct cases that the system assigned as compared to the base of all cases in which a human analyst is manually associated to a text class.

$$\text{Recall(Positive)} = \frac{N_{P,P}}{N_{P,P} + N_{P,N}} \tag{7}$$

$$\text{Recall(Negative)} = \frac{N_{N,N}}{N_{N,N} + N_{N,P}} \tag{8}$$

To combine these two measures in a single value, the F-measure is often used. The F-measure reflects the relative importance of recall versus precision.

$$\text{F-Measure(Positive)} = \frac{2.\,P(Positive).\,R(Positive)}{P(Positive) + R(Positive)} \tag{9}$$

$$\text{F-Measure(Negative)} = \frac{2.\,P(Negative).\,R(Negative)}{P(Negative) + R(Negative)} \tag{10}$$

Table 2
Classification Results

Classifier	*NB*	*SVM*
CorrectlyClassified	436	421
Accuracy	87.2%	84.2%
Precision	0.872	0.842
Recall	0.864	0.840
F-measure	0.868	0.841

5. CONCLUSION AND FUTURE WORK

In this paper, Method of opinion mining to know users' opinions about engineering colleges has been applied. The goal is to extract and summarize the opinions and reviews, and determine whether these reviews and opinions are positive or negative. For opinion classification of colleges reviews, we have used supervised machine learning algorithm Naive Bayes for implementation along with R as Development tool.

The results of this experiment show that the approach is helpful and provide high accuracy rate. By integrating and enhancing the capabilities of these essential technologies and ideas, we hope to introduce Sentimental analysis based model which will help different users to formulate an opinion about colleges.

The future scope will be improving the performance of the system and integrating it to with major Sites and Institutes to get more and precise data for further analysis. Also the scope and data of this experiment can be extended further as for all over India or even over all over the world.

REFERENCES

[1] ZiedKechaou, Mohamed Ben Ammar, Adel. M Alimi, "Improving e-learning with sentiment analysis of users' opinions", 2011 IEEE Global Engineering Education Conference (EDUCON)

[2] Dan Song, Hongfei Lin, Zhihao Yang, "Opinion Mining in e- Learning System", 2007 IFIP International Conference on Network and Parallel Computing – Workshop

[3] N. D. Valakunde, Dr. M. S. Patwardhan,"Multi-Aspect and Multi- Class Based Document Sentiment Analysis of Educational

Data Catering Accreditation Process", 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies

[4] Vijay B. Raut, D.D. Londhe, "Opinion Mining and Summarization of Hotel Reviews", IEEE 2014 Sixth International Conference on Computational Intelligence and Communication Networks.

[5] Vijay B. Raut, D.D. Londhe, "Survey on Opinion Mining and Summarization ofUser Reviews on Web", Vijay B. Raut et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1026-1030

[6] ImeneGuellil, KamelBoukhalfa , "Social Big Data Mining : A Survey Focused on Opinion Mining and Sentiments Analysis", 2013 International Conference on Cloud & Ubiquitous Computing

[7] Hung-Yu Kao and Zi-Yu Lin, "A Categorized Sentiment Analysis of Chinese Reviews by Mining Dependency in Product Features and Opinions from Blogs", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

[8] Ivan Khozyainov, EvgenyPyshkin, VitalyKlyuev, "Spelling Out Opinions: Difficult Cases of Sentiment Analysis", 2013

[9] RanjeetaRana,Mrs. VaishaliKolhe,"Analysis of Students Emotion for Twitter Data using Naïve Bayes and Non-Linear Support Vector Machine Approachs",2015 International Journal on Recent and Innovation Trends in Computing and Communication

[10] Chakrit Pong-inwong, WararatSongpan (Rungworawut),"TeachingSenti-Lexicon for Automated Sentiment Polarity Definition in Teaching Evaluation"

[11] HuosongXia,MinTao,YiWang,"Sentiment Text Classification of Customers Reviews on the Web Based on SVM", 2010 Sixth International Conference on Natural Computation (ICNC 2010)

[12] Stefano Baccianella, Andrea Esuli, and FabrizioSebastiani ,"SENTI WORD NET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining" 2008

[13] Jinan Fiaidhi, Osama Mohammed,Sabah Mohammed Simon Fong, Tai hoon Kim,"Opinion Mining over Twitterspace: Classifying Tweets Programmatically using the R Approach",2013

[14] Kushal Dave, Steve Lawrence,David M. Pennock,"Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews",2006

[15] ŽeljkaPožgaj,BlaženkaKneževiü,"E-learning:survey on students opinion",2007

[16] Bo Pang and Lillian Lee,ShivakumarVaithyanathan,"Thumbs up? Sentiment Classification using Machine Learning Techniques", 2005

[17] Shankar Setty, RajendraJadit, Sabya Shaik, ChandanMattikalli, VmaMudenagudi," Classification of Facebook News Feeds and Sentiment Analysis",2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

[18] Rajdeep Singh, Roshan Bagla, HarkiranKaur,"Text Analytics of Web Posts' Comments Using Sentiment Analysis", 2015 IEEE

[19] Gunjan H. Agre, Nikita V.Mahajan,"Keyword Focused Web Crawler", 2015 IEEE sponsored 2'nd International Conference on Electronics and communications.

[20] Xin Chen, MihaelaVorvoreanu, and Krishna Madhavan,"Mining Social Media Data for Understanding Students' Learning Experiences", 2013 IEEE

[21] Bingwei Liu , Erik Blasch , Yu Chen , Dan Shen, and GensheChen,"Scalable Sentiment Classification for Big Data Analysis Using Na ýve Bayes Classifier", 2013 intl conf. On Big Data

[22] Bogdan Batrinca• Philip C. Treleaven , "Social media analytics: a survey of techniques, tools and platforms", AI &Soc (2015)

[23] WalaaMedhat, Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal (2014) 5, 1093–1113

[24] Package 'tm' – CRAN [Online]. Available: https://cran.r-project.org/web/packages/tm/

[25] Package 'Rfacebook' – CRAN [Online]. Available: https://cran.r-project.org/web/packages/Rfacebook/

[26] Package 'twitteR' – CRAN [Online]. Available: https://cran.r-project.org/web/packages/ twitteR