Opinion Mining Of Social Media Data Using Machine Learning Techniques

Manu Krishna Bhardwaj
Computer Science and Engineering,
Manav Rachna College of Engineering,
Faridabad,India

Brajesh Kumar
Computer Science and Engineering,
Manav Rachna College of Engineering,
Faridabad,India

Abstract— **Data mining is to explore the large amount of information from various repositories. Different languages can be a barrier between monolingual communities, and the dynamics of language choice could explain the prosperity of local languages in an international setting. Also, with the revalidation of English as a lingua Hindi on Twitter, we reveal users of the native non-English language have higher influence than English users, and the language convergence pattern is consistent across the regions. Furthermore, we explore for which topics these users prefer to their native language rather than English. The interplay of language and network structure in diverse, multi-lingual societies, using Twitter. In our analysis, we are particularly interested in the role of multi-lingual. Concretely, we attempt to quantify the degree to which users are the bridge-builders between monolingual language groups, while monolingual users cluster together.**

Keywords—component, formatting, style, styling, insert (key words)

I. Introduction

There are number of micro blogging sites which are social media sites to which act as social media sites for users to make short and frequent posts. Twitter is one of the most popular micro blog social media site on which user can read and post messages which are 149 characters in length. The twitter posts and messages normally called as tweets. Nowadays, many Industries use sentiment analysis to know the opinion of customers about their product, this helps to improve the credentials of company, whose main aim is to consider customer satisfaction.

To obtained the twitter data known as twitter corpus that gives free information in the form of stream. This information has led to a variety of research, eg. prediction of movie reviews, recent product reviews, share market predictions, estimation of public sentiment during elections and recession. These survey convey very important role for online public reviews. This is helpful for their business performance monitoring from customer perspective instead of making customer surveys, which are expensive and time consuming [1]. Hence for automated classification of opinions, sentiment analysis tool using machine learning techniques into positive or negative aspects of public.

A. Overview of Sentiment:

Sentiment analysis involves study of opinion mining. Sentiments are defined as “an acquired and relatively permanent major neuropsychic disposition to react emotionally, cognitively, and co natively toward a certain object (or situation) in a certain stable fashion, with awareness of the object and the manner of reacting.” Gordon [3] has a similar definition; sentiments are “socially constructed patterns of sensations, expressive gestures, and cultural meanings organized around a relationship to a social object, usually another person or group such as a family.”

Examples of sentiments include romantic love, parental love, loyalty, friendship, patriotism, hate, as well as more transient, acute emotional responses, to social losses (sorrow, envy) and gains (pride, gratitude) [2].

B. Overview of Opinion:

In the case of opinions, not all words used in the sentence have significance. Some words are classified as noise because they are of no use in the process of classifying the polarity of the opinion.

According to Kim and Hovy [3], an opinion consists of the following four parts: topic, opinion holder, claim, and sentiment. That is, for each opinion, there is a holder who believes a claim about a topic and then associates a positive, negative, or neutral (neutral here does not mean absence, e.g., “The winter has arrived.” It is not good or bad, just saying sentiment with the claim).

Mary said the dress is beautiful

Fig 1. Example on an opinion [2][3]

II. DATA SOURCE OF OPINION MINING [4]

People and companies across disciplines exploit the rich and unique source of data for varied purposes. The major decisive factor for the improvement of the quality services rendered and enrichment of deliverables are the user consumers opinions. Review sites, blogs and micro blogs provide a good understanding of the reception level of products and services.

A. Review Sites[4]

Opinions are the major and actual data or more precise a decision for any user in making a purchase. The user generated reviews for products and services are mainly available on internet. The sentiment classification uses reviewer's data are gathered and composed from the websites like www.gsmarena.com (mobile reviews),www.amazon.com (product reviews).

B. Blogs [4]

The name associated to universe of all the blog sites is called blogosphere. People write about the topics they want to share with others on a blog. Blogging is a happening thing because of its ease and simplicity of creating blog posts, its free form and unedited nature. We find a large number of posts on virtually every topic of interest on blogosphere. Sources of opinion in many of the studies related to sentiment analysis, blogs are used[4][5].

C. Micr-Blogging [4]

A very accepted communication tool among Internet users is micro-blogging. We use this as one of the data source formed as a dataset of collected messages from Twitter. Twitter contains a very large number of very short messages created by the users, consumers of this micro blogging platform. Millions of messages appear daily in well-liked web-sites for micro-blogging such as Twitter, Tumblr, Facebook.

III. SENTIMENT CLASSIFICATION

A. Document level:

Document level sentiment classification is based on the sentiments executed on the overall sentiments expressed by authors. Documents classified according to the sentiments instead of topic. It is very useful in summarizing the whole document as positive or negative polarity about any object (camera, fridge, mobile, car, movie, and politician). Sentiment Classification Using Phrase Patterns in used Special tags opinion words. System constructed some phrase patterns and compute sentiment orientation using unsupervised learning algorithm. Proposed system achieved 86% accuracy. Investigated perspective from which a document was written. They build Naïve Bayes based model and test on Israeli-Palestinian conflict. Their corpus consists of articles published on the bitter lemons website. They used NB-B (full Bayesian inference) and NB-M (Maximum a posteriori).

B. Sentence level: [4]

Sentence level sentiment classification models is used for the extraction of the sentences contained in the opinionated terms, opinion holder and opinionated object. It is one level deep to document level and just concerns to the opinionated words but not the features. Total number of positive and negative words are counted from the extracted and classified sentences and if positive words are maximum then opinion about object is positive and if the negative words are more than opinion object is negative otherwise the opinion object will be neutral. To mine the customer reviews on a product proposed unsupervised algorithm is used and in this the algorithm find frequent features using Apriori algorithm. Chinese WordNet set classify opinion words in clauses (pos, neg or neutral) to summarize the comments. Sentence level opinion mining uses subjective and polarity (orientation) to find strength of opinions at the clause level. [4][6], all these are a notable work in this regard. To find the strength of opinions a new idea of syntactic clues is used. They use a wide range of features to find the strength of opinions. The system is about to provide tools and support for information analysts in government, commercial, and political domains, who want to be able to automatically track attitudes and feelings in the news and on-line forums. Opinion Analysis based on Lexical Clues and their Expansion to improve combination of rule-based algorithms and machine learning techniques. [4][7] proposed semi-supervised learning method based on highly precise seed rules. Subjectivity discovered at the sentence level. Polarity of the sentence defined as Positive, Negative or Neutral as well as opinion holders identified. The experimental results demonstrate that system achieved 45% Accuracy to extract opinionated sentences and 35% Accuracy to identify opinion holders.

C. Feature based level:

In customer reviews document, reviewer express positive, negative or both sentiments about the object and attributes. Document level and sentence level classification does not tell the likes and dislikes of consumer about particular attributes of object . When consumer comment on object (product, person, and topic, organization) he comment on the features of object[4][8].For example, if users commented on a Mobile Phone they basically comment on Camera result, LCD size, speaker, weight etc. On camera output 125 comments express the positive opinions and 25 comments may be negative. If a new customer is interested in camera quality of mobile he can take decision easily to purchase the product or not. To explore the detailed opinion on product or any topic, a detailed opinion mining study is required that is called feature based opinion mining [4][9]. Statistical Opinion Analyzer (SOA) extract the polarity of online customer reviews using Bayesian probability and frequency distribution. The proposed system helps the new customer to purchase the product and manufacturer to enhance the product's functionality. Reviews crawled, preprocess, tagged (GO tagger) and insert in SOA to find the positive and negative opinion probability and frequency distribution as well. The proposed system originated the very promising results. In a web based system SUMView crawled reviews from Amazone.com, decompose into sentences and tagged to find the nouns and noun phrases. Product features extracted using Hu and Liu (2004) method and top five extracted features were suggested to the users on the basis of frequency.

IV. TEXT MINING PROCESS AND METHODS[10]

Computer system endured the analytical thinking of text. By addressing different research areas we can define Text mining in number of ways which deals by view of each area [10].

- Information Retrieval (IR): One way that assumes text mining relation correlated to Information Retrieval i.e. retrieval of concept from text.
- Knowledge Organization Structure (KOS): for examining knowledge organization structure.

Text mining is the process of computation of extracting information from bulk quantity of data with the help of representing subordinate data to robust form. Text mining is the process of uncovering new unidentified information by retrieving entropy from numerous written and digital resources.

A. Process of Text Mining

1) Document Gathering:

In the first step, the different formats of text documents are collected. The document might be in form of .pdf, .doc,.docx, html doc, css etc.

2) Document Pre- Processing:

In this process, the given input document is processed for removing redundancies, inconsistencies, separate words, stemming and documents are prepared for next step, like as:

a) Tokenization:

The given document is considered as a string and identifying single word in document.

b) Removal of Stop word:

In this step the removal of usual words like a, an, but, and, of, the etc.

c) Stemming:

A stem is a natural group of words with equal (or very similar) meaning. This method describes the base of particular word.

3) Text Transformation:

A text document is collection of words (feature) and their occurrences.

4) Feature Selection (attribute selection):

This method results in giving low database space, minimal search technique by taking out irrelevant feature from input document. There are two methods in feature selection i.e. filtering and wrapping methods.

5) Data mining/Pattern Selection:

In this stage the conventional data mining process combines with text mining process.

6) Evaluate:

This stage Measures the outcome.

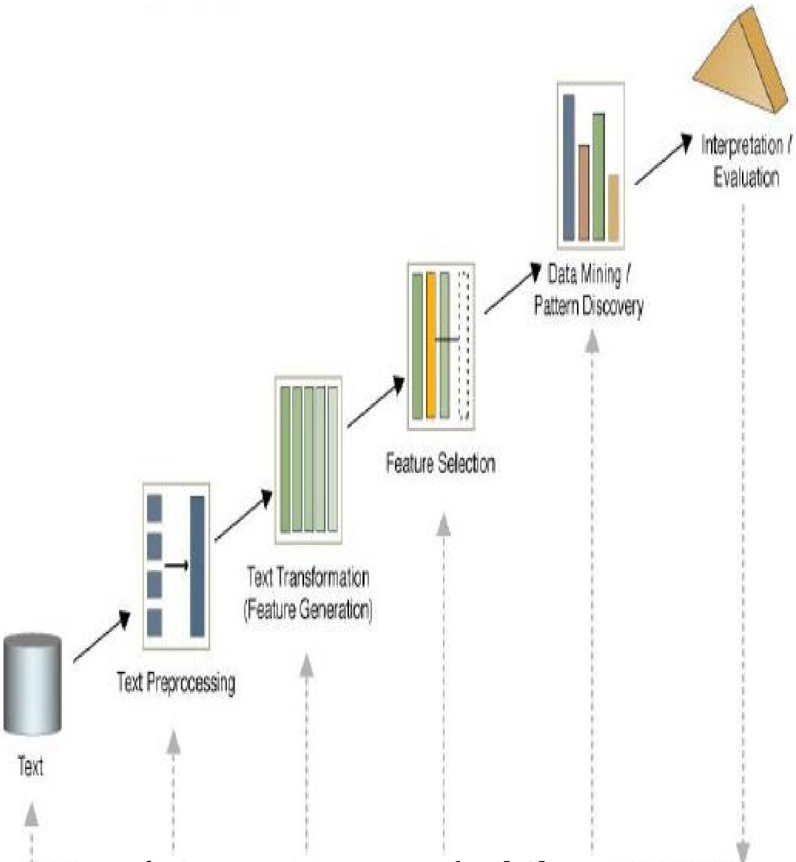

fig 2. Text Mining Process Flow [10]

B. Data Mining Methods for Text Mining

1) Information Extraction:

The bulk of data is available in an unstructured form. The drawback of conventional system is that it considers that the information to be mined is in the form of relational databases.

Information Extraction system addresses this problem of traditional system. The system looks for text stream available in the given document and determines the relationship and similarity between words and phrases.

As illustrated in fig. 2 The database drawn from given document by Information Extraction system is supplied to knowledge discovery databases for advanced mining processes. After mining knowledge from extracted data, Information Extraction system can predict information missed by the previous extraction using discovered rules DISCOTEX [10][12].

2) Topic Detection:

Conventional keyword search engines are restricted to a given data model and cannot easily adapt to unstructured, semi-structured or structured data [10][13]. Topic detection system overcomes this issue. Topic is an originative or event directly related along with all events and activities. Topic tracking mechanism helps to detect such an event which is related to target topic [10][14].

Topic Detection and Tracking (TDT) refers to automatic techniques for finding topically related material in streams of data (e.g., newswire and broadcast news). Work on TDT began about a year ago, is now expanding, and will be a regular feature at future Broadcast News workshops [10][14]. Topic Detection system uses two Topic Mining Models which are as follows:

1. Vector Space Model:

Despite of its simple data structure without using any explicit semantic information, the vector space model enables very efficient analysis of huge document collections. It was originally introduced for indexing and information retrieval [10][15] merely applies to document extraction scheme and also in text data mining technique.This model forms the matrices of high dimension for respective document. The compounding of document and term is nothing but text document[10][16] . There are many such documents that appears in the form of structured and semi-structured format. Text file like book, spreadsheets, telephone directory etc. Terms are words or group of word that are retrieved from document. Vector model represent each document as vector and each value in vector is represent word number at that value represent the number of occurrences of the word in document. Suppose collection of document $M=(m_1,m_2,m_3,\ldots\ldots.,m_N)$ and collection of terms $N=(n_1,n_2,n_3,\ldots,n_L)$ then we need to model vector Vi j = (vi,1, vi,2,……..,vi,L)in L dimension space. In Boolean expression if Vi,j=1, then term ni belongs to mi and if Vi,j=0 then ni does not belongs to mi .

2. Probabilistic Topic Model:

Machine learning researchers have developed probabilistic topic modelling, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information [10][17].

3. Natural Language Processing:

Text Mining is widely used in field of Natural Language Processing and Multilingual aspects. In NLP, Text Mining applications are also quite frequent and they are characterized by multilinguals. Use of Text Mining techniques to identify and analyse web pages published in different languages, is one of its example. The General Architecture for Text Engineering (GATE) is a framework for the development and deployment of language processing technology in large scale (Cunningham, Maynard, Bontcheva, & Tablan, 2002). GATE can be used to process documents in different formats including plain text, HTML, XML, RTF, and SGML.

4. *Clustering*:

Text mining becomes more challenging because of its characteristics such as volume, dimensionality, scarcity and complex semantics involved in it. These characteristics require clustering techniques to be scalable to large and high dimensional data, and be able to handle sparsity and semantics. Typical text clustering activity involves with the document representation, document similarity measure and clustering techniques. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity)[18].

V. Related Work[20]

The work described herein is related to the development of multilingual sentiment analysis systems and sentiment classification from tweets.

A. Methods for Multilingual Sentiment Analysis

In order to produce multilingual resources for subjectivity analysis, Banea et al. (Banea et al., 2008) apply bootstrapping to build a subjectivity lexicon for Romanian, starting with a set of 60 words which they translate and subsequently filter using a measure of similarity to the original words, based on Latent Semantic Analysis (LSA) (Deerwester et al., 1990) scores. Another approach to mapping subjectivity lexica to other languages is proposed by Wan (2009), who uses co-training to classify un-annotated Chinese reviews using a corpus of annotated English reviews. (Kim et al., 2010) create a number of systems consisting of different subsystems, each classifying the subjectivity of texts in a different language. They translate a corpus annotated for subjectivity analysis (MPQA), the subjectivity clues (Opinion Finder) lexicon and re-train a Naive Bayes classifier that is implemented in the Opinion Finder system using the newly generated resources for all the languages considered. (Banea et al., 2010) translate the MPQA corpus into five other languages (some with a similar ethimology, others with a very different structure). Subsequently, they expand the feature space used in a Naive Bayes classifier using the same data translated to 2 or 3 other languages. Finally, (Steinberger et al., 2011a; Steinberger et al., 2011b) create sentiment dictionaries in other languages using a method called “triangulation”. They translate the data, in parallel, from English and Spanish to other languages and obtain dictionaries from the intersection of these two translations.

B. *Sentiment Classification from Tweets*

One of the first studies on the classification of polarity in tweets was (Go et al., 2009). The authors conducted a supervised classification study on tweets in English, using the emoticons (e.g.“:)”, “:(”, etc.) as markers of positive and negative tweets. (Read, 2005) employed this method to generate a corpus of positive tweets, with positive emoticons “:)”, and negative tweets with negative emoticons “:(”. Subsequently, they employ different supervised approaches (SVM, Naive Bayes and Maximum Entropy) and various sets of features and conclude that the simple use of unigrams leads to good results, but it can be slightly improved by the combination of unigrams and bigrams. In the same line of thinking, (Pak and Paroubek, 2010) also generated a corpus of tweets for sentiment analysis, by selecting positive and negative tweets based on the presence of specific emoticons.
Subsequently, they compare different supervised approaches with n-gram features and obtain the best results using Naive Bayes with unigrams and part-of-speech tags. Another approach on sentiment analysis in tweet is that of (Zhang et al., 2011). Here, the authors employ a hybrid approach, combining super-vised learning with the knowledge on sentiment-bearing words, which they extract from the DAL sentiment dictionary (Whissell, 1989). Their preprocessing stage includes the removal of retweets, translation of abbreviations into original terms and deleting of links, a tokenization process, and part-of-speech tagging. They employ various supervised learning algorithms to classify tweets into positive and negative, using n-gram features with SVM and syntactic features with Partial Tree Kernels, combined with the knowledge on the polarity of the words appearing in the tweets. The authors conclude that the most important features are those corresponding to sentiment-bearing words. Finally, (Jiang et al., 2011) classify sentiment expressed on previously-given “targets” in tweets. They add information on the context of the tweet to its text (e.g. the event that it is related to). Subsequently, they employ SVM and General Inquirer and perform a three-way classification (positive, negative, neutral).

VI. Sentiment Analysis in Tweets[20]

Our sentiment analysis system is based on a hybrid approach, which employs supervised learning with the Weka (Weka Machine Learning Project, 2008) implementation of the Support Vector Machines Sequential Minimal Optimization (Platt, 1998) linear kernel, on unigram and bigram features, but exploiting as features sentiment dictionaries, emoticon lists, slang lists and other social media-specific features. We do not employ any specific language analysis software. The aim is to be able to apply, in a straightforward manner, the same approach to as many languages as possible. The approach can be extended to other languages by using similar dictionaries that have been created in our team. They were built using the same dictionaries we employ in this work and their corrected translation to Spanish. The new sentiment dictionaries were created by simultaneously translating from these two languages to a third one and considering the intersection of the translations as correct terms. Currently, new such dictionaries have been created for 15 other languages. The sentiment analysis process contains two stages: pre-processing and sentiment classification.

A. *Tweet Pre-processing*

The language employed in Social Media sites is different from the one found in mainstream media and the form of the words employed is sometimes not the one we may find in a dictionary. Further on, users of Social Media platforms employ a special “slang” (i.e. informal language, with special expressions, such as “lol”, “omg”), emoticons, and often emphasize words by repeating some of their letters. Additionally, the language employed in Twitter has specific characteristics, such as the markup of tweets that were reposted by other users with “RT”, the markup of topics using the “#” (hash sign) and of the users using the “@” sign. All these aspects must be considered at the time of processing tweets. As such, before applying supervised learning to classify the sentiment of the tweets, we preprocess them, to normalize the language they contain. The pre-processing stage contains the following steps: In the first step of the pre-processing, we detect repetitions of punctuation signs (“.”, “!” and “?”). Multiple consecutive punctuation signs are replaced with the labels “multi stop”, for the full stops, “multi exclamation” in the case of exclamation sign and “multi question” for the question mark and spaces before and after. In the second step of the pre-processing, we employ the annotated list of emoticons from SentiStrength2(Thelwall et al., 2010) and match the content of the tweets against this list. The emoticons found are replaced with their polarity (“positive” or “negative”) and the “neutral” ones are deleted.

B. *Tweet Pre-processing*

Subsequently, the tweets are lower cased and split into tokens, based on spaces and punctuation signs. The next step involves the normalization of the language employed. In order to be able to include the semantics of the expressions frequently used in Social Media, we employed the list of slang from a specialized site 3. At this stage, the tokens are compared to entries in Rogets Thesaurus. If no match is found, repeated letters are sequentially reduced to two or one until a match is found in the dictionary (e.g.

“perrrrrrrrrrrrrrrrrrfeeect” becomes “perrfeect”, “perfeect”, “perrfect” and subsequently “perfect”). The words used in this form are maked as “stressed”. Further on, the tokens in the tweet are matched against three different sentiment lexicons: GI, LIWC and MicroWNOp, which were previously split into four different categories (“positive”, “high positive”, “negative” and “high negative”). Matched words are replaced with their sentiment label - i.e. “positive”, “negative”, “hpositive” and “hnegative”. A version of the data without these replacements is also maintained, for comparison purposes. Similar to the previous step, we employ a list of expressions that negate, intensify or diminish the intensity of the sentiment expressed to detect such words in the tweets. If such a word is matched, it is replaced with “negator”, “intensifier” or “diminisher”, respectively. As in the case of affective words, a version of the data without these replacements is also maintained, for comparison purposes. Finally, the users mentioned in the tweet, which are marked with “@”, are replaced with “PERSON” and the topics which the tweet refers to (marked with “#”) are replaced with “TOPIC”.

C. Sentiment Classification of Tweets

Once the tweets are pre-processed, they are passed on to the sentiment classification module. We employed supervised learning using SVM SMO with a linear kernel, based on boolean features - the presence or absence of n-grams (unigrams, bigrams and unigrams plus bigrams) determined from the training data (tweets that were previously pre-processed as described above). Bigrams are used specifically to spot the influence of modifiers (negations, intensifiers, diminishers) on the polarity of the sentiment-bearing words.

D. Obtaining Multilingual Data for Sentiment Analysis in Tweets

Subsequent to the tweet normalization, we translate the Twitter data (the training and development data in the SemEval Task 2 campaign) using the Google machine translation system to four languages - Italian, Spanish, French and German. The reason for choosing the development dataset for testing is that this set is smaller and allows us to manually check and correct it, to obtain a Gold Standard (and ensure that performance results are not biased by the incorrect translation in both the training, as well as the development data). Further on, we extract the same features as in the case of the system working for English – unigrams and bigrams - from these obtained datasets. We employ the features to train an SVM SMO classifier, in the same manner as we did for English.

VII. CONCLUSION AND FUTURE WORK

The Sentiment analysis has been done on the feedback of a product on some multidisciplinary languages like Hindi, English from different social networking sites. The most of the comments on social media sites would be written in different languages for the ease of public .Thus, the impact of future work is to analyses the social networking data with Hindi words spelled in English language.

REFERENCES

[1] Eman M.G. Younis, “ Sentiment Analysis And Text Mining For Social Media Microblogs Using Open Source Tools: An Empirical Study”, International Journal Of Computer Applications(0975 – 8887), February 2015, Volume 112 – No. 5.

[2] AmrutaTarlekar, Prof. Kodmelwar M.K, “Sentiment Analysis of Twitter Data from Political Domain Using Machine Learning Techniques”, International Journal of Innovative Research in Computer and Communication Engineering.,ISO 3297: 2007 Certified Organization, June 2015,Vol. 3, Issue 6.

[3] S. Kim,E. Hovy, “Determining the sentiment of opinions”, In Proc. 20th Int. Conf. Comput.l Linguistics, PA, USA, 2004, pp. 1367–1363.

[4] Nidhi R. Sharma ,Vidya D. Chitre, “Opinion Mining, Analysis and its Challenges”, International Journal of Innovations & Advancement in Computer Science,ISSN 2347 – 8616,Volume 3, Issue 1,April 2014.

[5] Zhai Z, Liu B, Xu H, and Jia P, “Grouping Product Features Using Semi-supervised Learning with Soft-Constraints”, In Proceedings of COLING. 2010.

[6] ZhongchaoFei, Jian Liu, and Gengfeng Wu: “Sentiment Classification Using Phrase Patterns”, Proceedings of the Fourth International Conference on Computer and Information Technology in 2004.

[7] Wilson, T., Wiebe, J. and Hwa, R, “Just how mad are you? Finding strong and weak opinion clauses”, Proceeding of National Conference on Artificial Intelligence in 2004.

[8] Hiroshi, K., Tetsuya, N., and Hideo, W. “Deeper sentiment analysis using machine translation technology”. In Proceedings of the 20th international Conference on Computational Linguistics (Geneva, Switzerland) in 2004.

[9] B. Liu, “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers, April 22, 2012.

[10] Patil Monali S1, Kankal Sandip, “A Concise Survey on Text Data Mining”, International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 9, September 2014

[11] Vandana Korde and C. Namrata Mahender, ”Text Classification and Classifiers :A Survey”, International Journal of Artificial Intelligence & Application, Vol.3, No.2, March 2012.

[12] N. Kanya and S. Geetha, “Information Extraction: A Text Mining Approach”,International Conference on Information and Communication Technology in Electrical Sciences, IEEE (2007).

[13] Guoliang Li, Beng Chin Ooi, Jianhua Feng ,Jianyong Wang and Lizhu Zhou, ” EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data”.

[14] Charles L. Wayne, ”Topic Detection & Tracking (TDT) Overview & Perspective”.

[15] G.Salton, A. Wong, and C. S. Yang.” A vector space model for automatic Indexing”,Communications of the ACM, 18(11):613–620, 1975.

[16] Sushmita Mitra, Tinku Acharya “ Data Mining Multimedia, Soft Computing, and Bioinformatics”.

[17] Jonathan G. Fiscus and George R. Doddington, ” Topic Detection and Tracking Evaluation Overview”.[18] Charu C Aggrawal and Chengxiang Zhai,”Mining Text Data”.

[18] David M Blei, Princeton University, ”Introduction to Probabilistic Topic Models”.

[19] Vishal Gupta and Gurpreet S. Lehal, ”A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[20] Alexandra Balahur, Marco Turchi, “Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data”, Proceedings of Recent Advances in Natural Language Processing, pages 49–55,Hissar, Bulgaria, 7-13 September 2013.