Abhishek Tiwari, Kshitij Pathak, Upasana tiwari, Rupam das / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 1, Jan-Feb 2012, pp. 412-416

Opinion Polarity Detection in Blog Comments from Blog Rss Feed by Modified TF-IDF Algorithm

Abhishek Tiwari*, Kshitij Pathak, Upasana tiwari***,Rupam das******

*(I.T. Dept. , M.I.T. Ujjain)
** (I.T. Dept. , M.I.T. Ujjain)
***(I.T. Dept. , M.I.T. Ujjain)
**** (Head, Technical Operations, IICS, India)

ABSTRACT

Blogs are most common medium over web where user posts their opinion. It is considered to be a web space of the users where they share their views, beliefs and other philosophy. Blogs posted across the web can be extracted from their rss feed. Once a blog is posted, several readers leaves their comment on the blogs. Analyzing these comments can help in finding the opinion of people for the blog or about the topic in general. A general pattern for these comments are, they are short and sometimes not very grammatically accurate. Many a times the comments are generalized like an appreciative statement for the author or about the post. There are several opinion polarity mining techniques which are mainly eccentric around the theory of training a natural language processing machine with known opinionistic blogs and train a classifier based on this. The classifier further classifies the blogs based on their closeness with the trained datasets. A machine learning in natural language processing requires huge training data to build the decision rule and therefore classification time also increases naturally. Therefore this work proposes a unique technique of opinion polarity mining from comments of the blogs dynamically, through the RSS feed with a unsupervised classifier. The proposed technique is based on modified TF-IDF algorithm for first extracting the relevance of a comment with the topic and thereafter uses a scoring mechanism to identify the opinion based on occurance of opinionisitic terms and their order in comments. The algorithm is tested with various real time blog site like digitalinspiration.com, techmafia.org, integratedideas.co.in, kerryseo.co.in and so on. RSS of blogger,wordpress and blogspot powered blogs are tested for testing the efficiency of the detection. 100 posts in total are analyzed and are verified by the author of the posts about effectiveness of the opinion being detected. Result shows an overall accuracy of 81% in classifying the opinion.

Index Terms— ***Opinion Polarity Mining, Blog Sentiment Detection , TFID***

I. INTRODUCTION

A blog is a type of website or part of a website. Blogs are usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video. Entries are commonly displayed in reverse-chronological order. *Blog* can also be used as a verb, meaning *to maintain or add content to a blog*. In common web usage blogs are referred as user posts or users opinion over certain issues.

Consider following two random blogs.

"I feel that for past several years government has not undertaken much of new development work. The prime minister had many contributions towards economy but don't know why he is becoming ineffective."
"Other Government's development activities were good. I wish the current set of central ministers continued the same work."
The two sentences are reflecting the similar sentiments without any direct word wise similarity amongst the sentences. The second sentence mentions about "Current set of Ministers" instead of government which syntactically reflects the same thing. Moreover these reflections are over the similar subject matter. Hence extracting a subject associated with a blog and extracting the polarity or the blogger's opinion about the subject matter is a challenging aspect.
The objective of the work is to develop an engine for detecting the blogs containing user opinion about a particular subject and further extract the three opinion scenarios : positive, negative and neutral from the blogs. The technique takes the help of both syntactic and semantic analysis to mine the opinion and the polarity.
Due to more and more users posting their opinions and views as blogs, it becomes important that automated tools are developed to analyze such post to draw users or overall publics opinion on a certain issue. In the following paragraph we discuss about the opinion polarity mining and their importance.
Opinion Detection is one of the most exciting and challenging application of text analysis today. It is the ability of recognizing and classifying opinionated text within the

documents This ability is desirable for various tasks, including filtering advertisements, separating the arguments in online debate or discussions, ranking web documents cited as authorities on contentious topics, etc. In *Opinion Detection*, one has to check whether a given text has a factual nature (i.e. describes a given situation/event without giving any opinion about it) or expresses an opinion on its subject matter. This task can be performed on different levels of granularity, i.e. on word level, sentence level or on document level. As a conclusion of this task a given word, sentence or document can be declared as of opinionated nature (or subjective) or of factual nature (objective). Text with opinionated nature can further be analyzed for having negative or positive polarity of opinion and this subtask is called *Opinion Polarity Detection*. The objective of the work is to detect opinion polarity on a given subject amongst set of blog documents featuring the subject.

II. RELATED WORK

In this paper [1] author used regular expression based algorithm and it give the polarity of opinion for a particular news item but results are restricted to yahoo' sites only so this is the drawback of this work.

The approach described in this paper exploits Senti-WordNet [2] as lexical resource for opinion mining. The authors introduces a lexicon based method of analyzing the opinion even without any training data in this direction.

Khurshid et al[3] have developed a method for identifying the words that may surprise a native speaker by comparing the distribution of all the words in a collection of randomly sampled financial texts with that of the same words in a reference collection of texts. More prolific keywords in financial texts, the chances are that such a word will be less prolific in general language texts. Once it identified keywords, based on a statistical criteria in our training collection of texts, then the system look at the neighborhood of these keywords; and, then look at the neighborhood of the two word pair and so on.
This neighborhood, established on strict statistical criteria, yields information bearing sentences in the financial domain and, it turns out, sentences that typically carry sentiment information. These patterns are then used to build a finite state automaton. This automaton is then tested on an unseen set of texts – and the results vis-à-vis sentiment analysis are quite good.
[4] analyzes the various techniques for sentiment analysis in online data. Further they present a simplistic algorithm for the same which contains following.

Evaluation (positive/negative), Potency (powerful/unpowerful), Proximity (near/far), Specificity (clear/vague), Certainty (confident/doubtful) and Intensifiers (more/less)

[5] presents a supervised learning based technique for blog classification without the data in the same domain. In this work authors make use of several features of three principal types for our classification task: textual features (exclamation points and question marks), part-of-speech features, and lexical semantic features. No part-of-speech features showed an impact on the classifications; hence for reasons of space we do not discuss them further. Each post is then represented as a feature vector in an SVM classification.

[6] presents an overall system of learning based classification and explains the accuracy measure of such a system explained by figure 1.

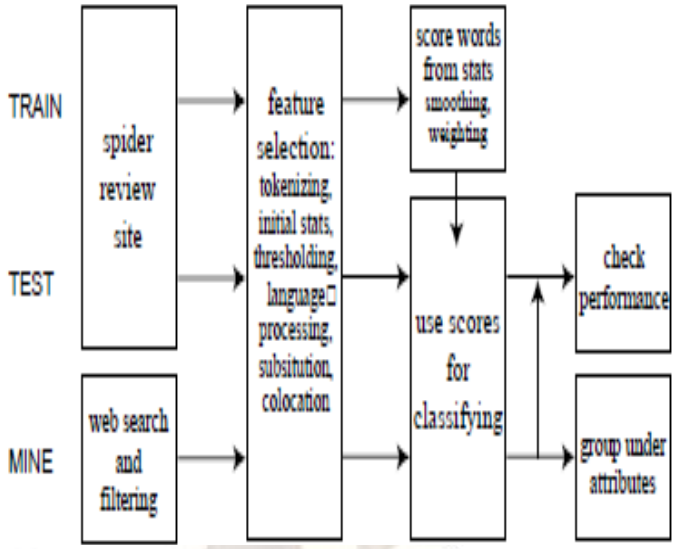

Figure 1: Classifier based sentiment detection technique as proposed by [6]

The work of [6] is motivated by the fuzzy based sentiment classification as proposed by [7] which uses a manually constructed rule set for the same.

[8] presents a query specific summarization without the use of any learning rules. The system is based on scoring of the opinion pattern against the query specified by the user. The system is explained in figure 2.

[9] explains the feature extraction process for the opinion polarity mining. It explains that for any opinion mining first there must be an association with the genere or the heading or the subject of the base. This forms the preprocessing step in the proposed system.

[10] presents another mechanism of subjective analysis of any document or post. They present a part of speech based algorithm for identification of the subjective with respect to the document.

Product ReputationMiner [12] extracts positive or negative opinions based on a dictionary. Then it extracts characteristic

words, co-occurrence words, and typical sentences for individual target categories. For each characteristic word or phrase they compute frequently co-occurring terms. However, their collocation-based association of characteristic terms and co-occurring terms is known to be highly noisy [11].

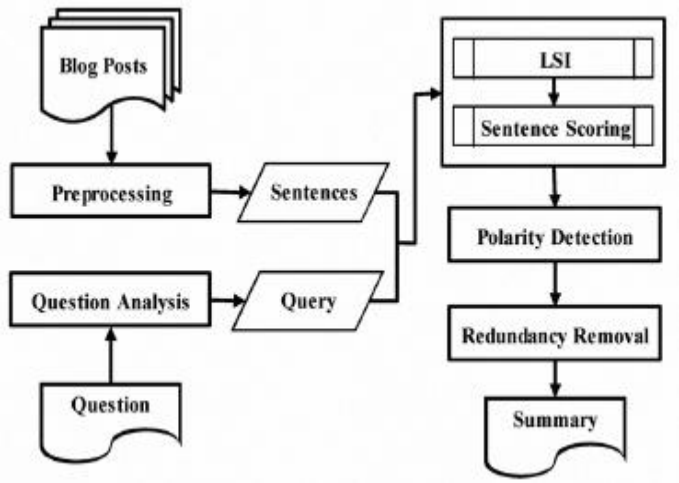

Figure 2: Query based opinion polarity detection as suggested by [8]

[13] explains the challenges in opinion mining and sentence extraction in blogs. They emphasized that opinion polarity mining in blogs are difficult due to several aspects like mis-spelled words, synonyms, emotions, short terms and so on. This are the challenges that the proposed work is designed to overcome to.

III. PROPOSED WORK

Various polarity detection techniques are being proposed in the text as summarized in the related work section. The main problem with most of the techniques is that they depend upon the distance analysis and clustering result based on the occurrence of the words. The polarity detection is purely a syntactic outcome of a sentence interpretation and many a document may not have a clear polarity. The techniques have not proposed a clear mechanism of extracting a polarity of a given subject. In short polarity detection is presented as an aggregation result of distance in terms of sentences and not as a natural language processing technique. No past work has defined finite automata for polarity detection, though numerous tree based approaches are proposed. The present system of polarity detection technique is broadly categorized into two categories: 1) technique based on machine learning and 2) Technique based on clustering. In 1) A machine learning system like support vector machine is trained with known blogs with and without opinion. Large databases are used as training sample in such techniques. The given blogs are classified into various groups of opinionistic sentences based on various distance measure by the classifier. The type 2 category of methods depends upon building a decision tree based on the clustering and occurrence of interrelated words and the words that presents the various opinion representation. Moreover the technique are tested against standard databases like Trec blog database.

The system is modeled in two test sets. Firstly we extract the live blogs from the news feeds like various yahoo sites. Here the subject matter is considered as the news item itself. The live blogs are extracted and stored offline for analysis. Secondly we consider standard blogs for analysis of the strength of the algorithm and to verify the correctness of the proposed system.

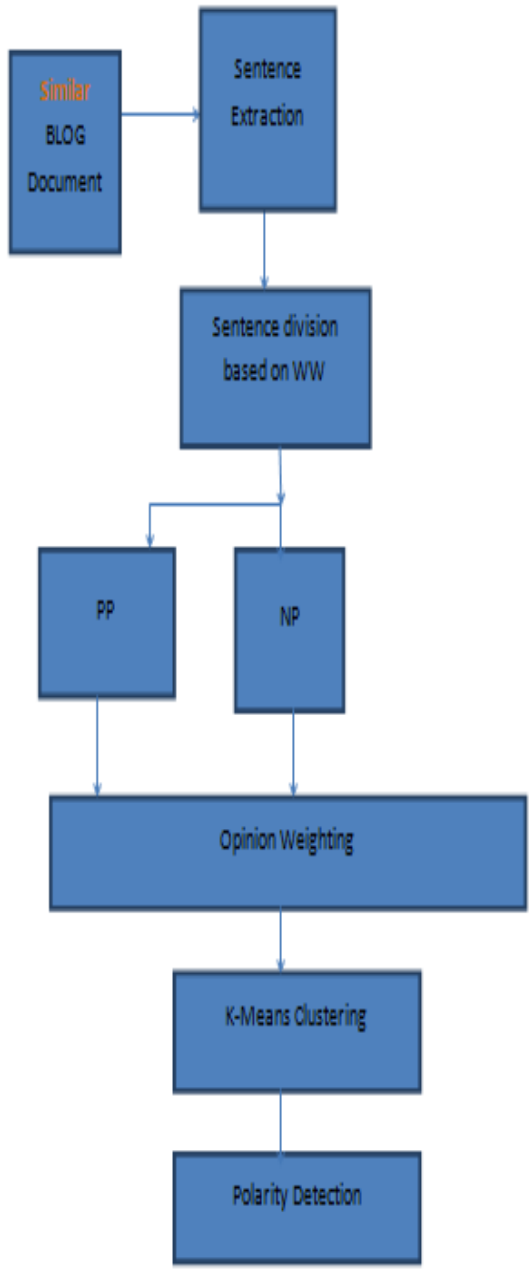

Figure 3: block diagram of the system

The main stages and functioning of the system is elaborated as bellow.

1) First segment the blogs into sentences and sentences into words. The words are tagged based on wordNet tool for sentence segmentation and tagging.

2) Once the words are tagged, find the similarity of the blogs with respect to a specific subject matter based on the tags of the blogs.

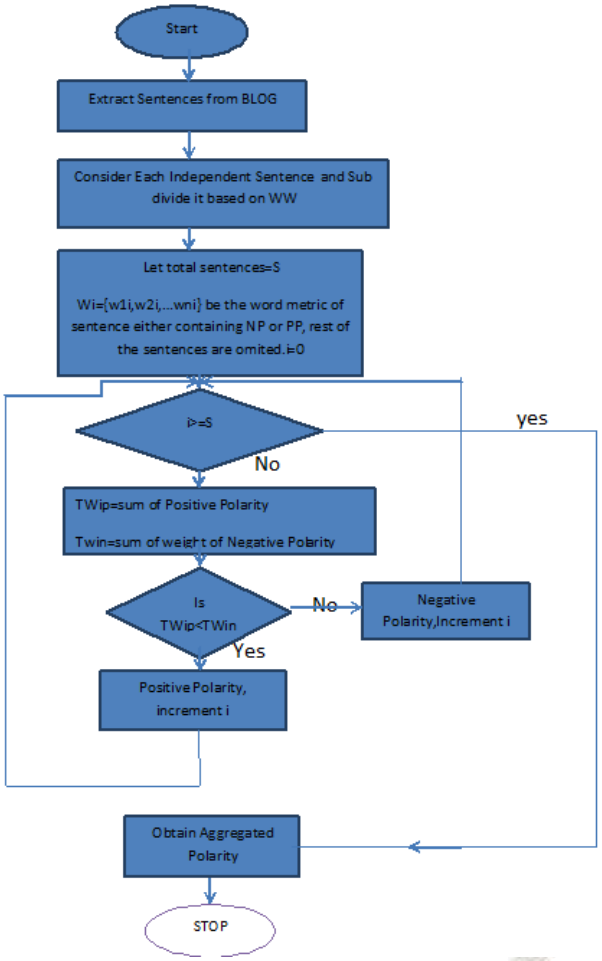

Figure 4: Flow Chart of the System

3) The similar items to the headings are ranked higher and are sorted at the top in comparision with the other blogs.

4) The high ranked blogs are forward scanned for the deterministic words like " I believe", "I think" and so on. The closeness measure with such words are performed on the high ranked blogs and they are further categorized into blogs with opinion and blogs without opinion.

5) logs related to a certain heading and that posses a opinionHis now scanned for type of opinion.

6) Based on 4, the sentences are weighted from the start to end based on segment fragments as elaborated in 5.

7) O Based on the positive, negative or zero scores the blogs are classified as positive, negative or neutral opinion blogs.

The block diagram and the flow chart are presented in figure 3 and 4.

IV. METHODOLOGY

The systems developed without any Learning based classifiers are generally classified in two major categories: 1) Occurrence based Thresholding and weighted thresholding as proposed by the system. Table 1 presents both the system in short. The general sentiment database[16] is used for constructing the set of positive and negative words.

Occurrence Based Detection	Regular Expression Driven Score based Detection
1. Scan The Blog 2. Generate Tokens from the Blog 3. Remove the common words like "is","and","the" 4. Normalize Document 5. Find the relevancy of the Heading with Posts using N-Gram Algorithm 6. If blogs are irrelevant, exit 7. Count Number of Negative and Positive words 8. Generate the Opinion Summery based on the difference between the scores. Result: Efficiency :68%	1. Scan The Blog 2. Generate Token 3. Remove the common words like "is","and","the" 4. Normalize Document 5. Find the relevancy of the Heading with Posts using N-Gram Algorithm 6. If Relevant than load the database of words. 7. Load the Slang words 8. Load the expression icons 9. Find the occurrence of each word in the database using regular expression. So "gud" and "good" are both matched. 10. Extract the weight of each words and add to the polarity count. 11. If any strong words like "very"," too much" appears before the polarity word, add the score of the weighted word with the polarity word. 12. If the words are preceded by negative words like "not", the reverse the polarity of the word. Result : Efficiency 79%

Table 1: Comparative steps of Existing technique and our approach.

V. RESULTS

Experiments are conducted for various categories of blogs which are firstly categorized into blogs with different post word count. The feeds are extracted from blogs with different Google page rank. It was observed that for short blogs comments were also short which lead to many misdetection. Another observed fact was that for Low PR sites comments are less moderated which resulted in many general comments and out of the context comments that did not help analyzing the opinion about the subject. Experimental results are presented in figure 4.

T

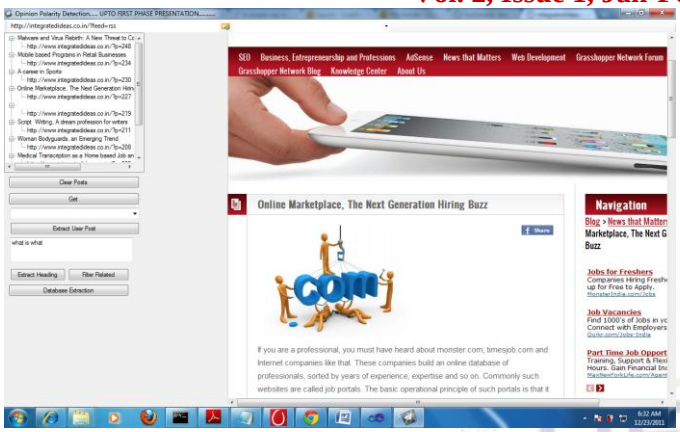

Figure 3: Interface of the work showing the extraction of the Blogs.

Word count	blogs with PR3	blogs with PR4	blogs with PR5	PR6 and above
less than 200	51	53	63	66
200-400	67	71	75	80
400-600	88	90	91	90
600-800	92	92	94	95
above 800	94	94	95	95
Overall Accuray	78.4	80	80.75	82.75

Figure 4: Results depicting the comparision of results in present and proposed system

VI. CONCLUSION

Opinion polarity is an important aspect of web mining and the web data analysis because major issues and news are posted as blogs these days. Number of blog readers and blog commenters are also increasing by each passing day. Therefore it becomes important to develop tools which can not only extract correlated blogs but also gets an overview of independent and in turn generalized overview of the blogs. Many algorithms are proposed in this direction. Most of these works are organized to detect the opinion in the blogs only and do not present a comprehensive overview of the entire technique of fetching the RSS blog data and analyze them on the fly. In this work we developed an entire lifecycle of fetching and analyzing the blogs and the comments for opinion. The technique is based on similarity of the blog with its subject matter and the presence of opinion in such correlated blogs. The result shows a significant similarity with human perception. The only limitation of these method was that the xml parsing had to be separately designed for different blog machines like drupal , wordpress, blogspot and blogger. A comprehensive framework can be built that can extract the comments and posts irrespective of the type Blog Engine. Further a machine learning technique can be used alongside the proposed technique for better results.

REFERENCES

[1] A.Tiwari, K.Pathak & N.S. choudhary "Real Time Opinion Polarity Detection in Blogs by Weighted Ranking TF-IDF Algorithm" *International Conference on Communication and Networks from 4th-6th December 2011 at Udaipur*

[2]Esuli A, Sebastiani F. SentiWordNet: "A Publicly Available Lexical Resource for Opinion Mining". *LREC 2006*

[3] Khurshid Ahmad*,"Multi-lingual Sentiment Analysis of Financial News Streams, Grid Technology for Financial Modeling and Simulation" *February 3/4, 2006 – Palermo, (Italy)*

[4] Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens, "Automatic Sentiment Analysis in On-line Text", *Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria – June 2007*

[5] Paula Chesley, Bruce Vincent, Li Xu, and Rohini K. Srihari, "Using Verbs and Adjectives to Automatically Classify Blog Sentiment" *In Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*

[6] Kushal Dave, Steve Lawrence, David M. Pennock, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, WWW2003, May 20–24, 2003, Budapest, Hungary.

[7] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. IEEE-FS, 9:483–496, Aug. 2001.

[8] Feng Jin, Minlie Huang, Xiaoyan Zhu, A Query-specific Opinion Summarization System, Prlc.lib *IEEE (ICCI'09)*

[9] Jeonghee Yi, Wayne Niblack, Sentiment Mining in WebFountain, *IEEE 2005*

[10] Amitava Das, Sivaji Bandyopadhyay, Theme Detection an Exploration of Opinion Subjectivity, *IEEE 2009*

[11] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of the Int. WWW Conference, 2003.*

[12] S. Morinaga, K. Yamanishi, K. Teteishi, and T. Fukushima. Mining product reputations on the web. In *Proceedings of the ACM SIGKDD Conference, 2002.*

[13] Malik Muhammad Saad Missen, Mohand Boughanem, Guillaume Cabanac, Challenges for Sentence Level Opinion Detection in Blogs, 2009 *Eigth IEEE/ACIS International Conference on Computer and Information Science*

[14]Sentence-Level Opinion-Topic Association for Opinion Detection in Blogs, 2009 *International Conference on Advanced Information Networking and Applications Workshops*

[15]Farhad Oroumchian, Abolfazl Aleahmad , Parsia Hakimian$_a$, Farzad Mahdikhani , "N-Gram And Local Context Analysis For Persian Text Retrieval" *Signal Processing and Its Applications, 2007. ISSPA 2007.*

[16] *Kerstin Denecke*, How to Assess Customer Opinions Beyond Language Barriers?, *IEEE 2008*

[17] http://sentistrength.wlv.ac.uk/: Sentiment words database.