Mining the Blogosphere from a Socio-political Perspective

Vivek Kumar Singh, Debanjan Mahata, Rakesh Adhikari
Department of Computer Science,
Banaras Hindu University,
Varanasi, India
vivek@bhu.ac.in, debanjanbhucs@gmail.com, adhikari.rakesh@gmail.com

Abstract- Blogs are websites that allow one or more individuals to write about things they want to share with others. The universe of all blog sites is referred to as Blogosphere. The ease & simplicity of creating blog posts and their free form and unedited nature have made the blogosphere a rich and unique source of data, which has attracted people and companies across disciplines to exploit it for varied purposes. The valuable data contained in posts from a large number of users across geographic, demographic and cultural boundaries provide a rich data source not only for commercial exploitation but also for psychological & socio-political research. This paper tries to demonstrate the plausibility of the idea through our clustering and opinion mining experiment on analysis of blog posts on recent socio-political developments in the new democratic republic of Nepal; and to elaborate the broader technical framework & tools required for this kind of analysis.

Keywords- Blogosphere; Blog Mining; Blog Clustering; Opinion & Sentiment Analysis; Nepal Constitution.

I. INTRODUCTION

Blogosphere is the virtual universe of all blogs written by people all over the world, in various languages. Blogosphere is now a huge collection of discussions, commentaries and opinions on virtually every topic of interest. The phenomenal growth in the blogosphere is evident from the fact that more than two blog posts on an average are created every second [1]. A blog is a website that allows individuals to write about different topics. A blog site typically consists of blog posts arranged in reverse chronological order, along with the comments by their readers. These sites contain a series of posts typically characterized by brief texts, which have minimal editing. A blog post may comprise only of text, or it may contain images and links to other media. Unlike websites, the contents of blogs are random, unstructured and chaotic. A blog site may be owned individually, or it may be a community blog site.

Blogs have become a very popular medium for expressing opinions, communicating with others, providing suggestions, sharing thoughts on different issues and also to debate over them. The large amount of valuable data contained in the blogosphere is making it an important field of research, not only for academicians but also for people from industry and other disciplines. The study of blogosphere has helped in reshaping business models, assist viral marketing, providing trend analysis & sales prediction, and aiding counter terrorism efforts. The blogosphere is an ideal platform from which we can extract information, opinion, moods and emotions on various topics. It may include topics like political issues, social issues, product reviews or market surveys. The current research in blogosphere includes areas like blog classification & clustering, community discovery, analysis of relationship among bloggers, topic discovery & tagging, blog mining, trend discovery, bloggers' sentiment & interest analysis, filtering spam blogs and modeling the blogosphere.

Most of the contemporary analytical research in blog mining, however, focuses more on marketing applications. Efforts on socio-political exploitation of the blogosphere are almost negligent. The fact that the content of the blogs are original free-form writings, and that they are very contemporary & emotion laden; makes the blogosphere an ideal platform for socio-political analysis. The personality, expertise, views and moods of bloggers are well reflected in their posts. The postings are on real time basis and information in the posts is current and relevant. It is, therefore, beyond doubt that the blogosphere now depicts the accurate views and sentiments of the common people, in an electronic media.

In this paper, we have tried to analyze the blogosphere from a socio-political perspective by collecting and analyzing the blog posts about political and constitutional developments in the new democratic republic of Nepal. Section II of the paper presents the general technical framework that may be used for this kind of analysis. Section III describes the experimental formulation and setup and section IV presents the preliminary results. The paper concludes (section V) with a short discussion of the approach and its relevance.

II. ANALYTICAL FRAMEWORK

A successful socio-political analysis of the blogosphere requires appropriate technical framework with suitable searching and mining techniques. Searching helps in identifying and collecting relevant blog posts whereas mining techniques extract meaningful inferences from the collected data. Over the past few years, blogosphere has been the target of data mining efforts by marketing companies and advertisers. Marketing companies view the blogosphere as a potential source for tracking consumer beliefs and opinions, understanding language of customers, knowing consumer preferences, knowing quick initial reactions to product launches and to track performance of different products. Advertisers view blogosphere as a platform to disseminate information about variety of products and services and try to identify suitable blogs where they can place their advertisements for maximum and relevant viewership. The

978-1-4244-7818-7/10/$26.00 ©2010 IEEE

huge size and scale of blogging phenomenon makes it a difficult task, calling for use of automated techniques. The techniques & tools developed by marketing & advertising companies for searching and mining can be extended and used for social-political analysis as well.

The analysis task can be viewed as a two step process of: collecting data (searching) and obtaining inferences from the collected data (mining). The quality of analysis and mining of blogs depends largely on how the blogs are extracted from the blogosphere. Searching for relevant blog posts on a topic has been made simple by availability of several blog tracking companies. Most of these blog tracking companies provide free tracking service which can be used by a blog search program to find high authority score data. The experimenter thus needs to devise a blog search program which can accept user queries and send it to a blog tracking provider through HTTP Get/Post. The blog search program needs to translate the query into a format acceptable to the blog tracking provider before sending. The blog tracking provider processes the query and sends back a response, usually in XML. The XML response is then parsed by the blog search program and the retrieved results are displayed. Figure 1 illustrates the steps and message exchanges involved in the process. Sophisticated blog search programs can send queries simultaneously to multiple blog tracking providers and aggregate the retrieved responses.

After the relevant data is obtained by the search program, it is subjected to the mining program for extracting useful inferences. However, since the collected blog is primarily textual in nature it needs to be organized into appropriate data structures to perform the mining task. This is often termed as preprocessing. The text contained in different fields of each blog post, including its title, body, comments and user tags; is transformed into a term vector structure with frequency of occurrence of different terms. Tokenization is the first step towards this end and involves identifying valid terms contained in the document. Tokenization includes managing hyphens, converting the tokens into lowercase and dropping stop words (such as and, or, to, the, are, their) etc.

Stemming and preserving multi-word phrase may also be necessary. Stemming removes occurrences of the same word in multiple forms, for example 'person' and 'persons'. Multi-word phrases are those which are important and convey exact and relevant meaning, such as 'collective intelligence'. The token set is sometimes added with the synonyms of valid tokens, a process called synonym injection. The identified tokens are then either accepted as valid tokens or discarded based on their frequency of occurrence in the weighted sets of title, body, comments and user tags. The computed measures of Term Frequency (*tf*) and Inverse Document Frequency (*idf*) are used to compute *tf-idf* weighting. The *tf-idf* weighting is then used to represent the data as vectors in common vector space (known as the vector space model). Depending on the goal of analysis the preprocessing task may also employ techniques like summarization, blog statistics collection (such as time interval between posts) and other techniques like stochastic graph based methods.

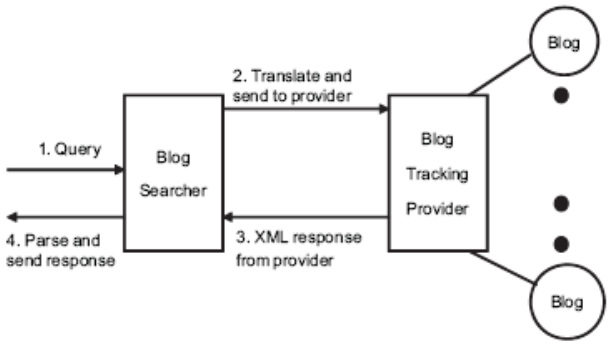

Figure 1. Steps in searching the blogosphere (Courtesy [9])

Once the preprocessing part is over, the mining task begins. The methods, techniques and tools used in this phase are varied and task specific. The analysis & mining task may involve identifying bloggers' mood & sentiments, classifying the blog posts, clustering the posts into different groups, identifying main topics discussed in posts, summarizing the posts and to extract opinions expressed in blog posts. Clustering & Classification techniques (such as K-means & SVM based clustering methods), Natural Language Processing techniques (such as LSI & LDA) and statistical & probabilistic methods are employed during this phase. Most of these methods can be readily implemented in popular high level languages like JAVA and are also available as APIs and code libraries. There are many other tools and packages available, which can be customized and integrated together to help in the analysis tasks.

III. EXPERIMENTAL WORK

The aim of our experiment was to demonstrate that blog posts can be a valuable source for socio-political analysis. We have chosen to analyze the blog posts related to the drafting of the new constitution of the democratic republic of Nepal. We have chosen this topic for its current socio-political importance. There are many political and social issues, as well as varied reactions of people, around this theme. We devised a content analysis approach where we collected a good number of blog posts on the topic and used the collected blog data for deriving interesting inferences. We followed the two step process as discussed in the previous paragraphs. *First step* involved collecting the blog data, storing it in database, extracting term vector & other structures from the data, generating vector space model for the collected data and then clustering the collected blog data into different categories. The categories of clusters represent different aspects of the issue and also different viewpoints presented by the bloggers. The *second step* involved content analysis using a variety of techniques. The blog posts clustered into different groups have been subjected to parts of speech tagging, identifying opinionated words, mood & sentiment analysis of the posts in each cluster, and computing positive-objective-negative scores of the key words of the posts in each cluster.

A. Collecting Relevant Blog Data

We have collected a good number of recent blog posts related to the socio-political issues around the constitutional development in the democratic republic of Nepal. We did two tasks in parallel: collecting full blogs manually and retrieving feeds of recent blogs from Google Blog Search [2] through a Java program. The automated process of retrieving the blog feeds (through Java program) has been kept versatile, making it capable of retrieving feeds from any blog tracking company. The program also had the ability to integrate the various APIs made available by different blogging sites. The blog search was a four step process as described in section II and used various search engines, tools and APIs [3], [4], [5], [6], [7], [8]. The collected blog data was stored in a MySQL database. The stored blog data comprised of name of the blog site, permalink of the blog post, author's name, title of the blog post, its body and comments, and user tags. The manual collection of blogs involved identifying 70-75 high rank blog posts.

B. Preprocessing

We have used the Vector Space Model [9], [10] to represent each document. Every blog post is represented in the form of a term vector. A term vector consists of the terms appearing in a blog post and their relative weights. The weight associated with each term (*tf-idf* measure) is the product of two computations: Term Frequency and Inverse Document Frequency. Term Frequency (*tf*) is a count of how often a term appears within the document. Words that appear often have high *tf* value, representing that the document has its main theme related to this term. Given a particular domain, some words appear more often than others. However, to capture the entire picture, we have to be also more discriminating to find terms that have less common occurrence. This is the motivation behind inverse document frequency (*idf*). It aims to boost terms that are present in only few documents. If the total number of documents of interest be N, and df_t be the number of documents that contain the term t, then the *idf* for the term t is computed as $idf_t = \log (N/df_t)$. Thus the *idf* of a rare term is high, whereas the *idf* of a frequent term is likely to be low. If a term appears in all documents, then its idf is log (1) which is 0.

The *tf* and *idf* values are used to produce a composite weight for each term in each document (say a term t in document d), defined as $tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t$. The vector V(d) derived from the document *d* thus contain one component for each term. Once the entire vector space model is obtained, the documents can be clustered in groups based on their degree of relatedness. However, since a large document will have different vector representation than a similar document with smaller length, the similarity between documents cannot be computed simply from vector space representation of the two documents. Therefore, the *cosine similarity* of the two vector representations is computed as:- $sim(d_1,d_2)=V(d_1).V(d_2) / |V(d_1)||V(d_2)|$. The numerator represents the dot product of the vectors $V(d_1)$ and $V(d_2)$ and the denominator is product of their *Euclidean lengths*. The denominator length-normalizes the vectors $V(d_1)$ and $V(d_2)$ to unit vectors $v(d_1)=V(d_1)/|V(d_1)|$ and $v(d_2)=V(d_2)/|V(d_2)|$. The $sim(d_1,d_2)$ can thus be written as $v(d_1).v(d_2)$. This value is then used to compute similarity between documents and hence cluster them into related groups. Lucene Tokenizers, Analyzers and Filters [11], [12] were used for tokenizing the document; that is to extract the terms, normalizing them, eliminating the stop words and for stemming. The resulting lists of terms were processed with the help of RiTa WordNet Java library [13]. Parts of speech tagging and synonym injection were also performed.

C. Clustering the blog posts

The next step in the analytical effort was to cluster the collected posts into different groups based on their similarity and dissimilarity. The different clusters so identified were expected to elaborate various dimensions of the topic including different issues and concerns. We used a k-means clustering scheme [14], [15], [16] to classify the data into distinct groups, where value of k was obtained through iterative trials supported by human inputs. The scheme is a hard flat clustering method with a goal defined as follows: Given (i) a set of documents $D=\{d_1,d_2 \ldots d_N\}$, (ii) a desired number of clusters K, and (iii) an objective function that evaluates the quality of clustering, we want to compute an assignment γ: $D \rightarrow \{1 \ldots K\}$ that minimizes the objective function. The objective function γ is often defined in terms of similarity or distance between documents. The objective in K-means clustering is to minimize the average squared Euclidean distance of documents from their cluster centers (centroids or mean).

A measure of how well the centroids represent the members of their clusters is the *residual sum of squares* or RSS, defined as the squared distance of each vector from its centroid summed over all vectors. The first step in K-means is to select as initial cluster centers K randomly selected documents, the *seeds*. The algorithm then moves the cluster centers around in space to minimize RSS. This is done iteratively by repeating two steps until a stopping criterion is met: reassigning the documents to the cluster with the closest centroid and recomputing each centroid based on current members of its cluster. The algorithm produced clusters of blog posts, each of them identified by an Id. Different groups largely represented variation in issues discussed and reactions expressed by people.

D. Opinion Extraction and Sentiment Analysis

After clustering the blog posts, we perform opinion mining and mood & sentiment analysis. Most of the contemporary research works for extracting the opinions from a text are based on finding opinionated words. Once the opinionated words are obtained a suitable technique for analysis could be to determine the "PN-polarity" of

subjective terms. This involves identifying whether the term has a positive or a negative connotation. SentiWordNet [17], [18], [19] is a lexical framework in which each WordNet synset is associated to three numerical scores Obj(s), Pos(s) and Neg(s), as shown in the Figure 2. These scores describe how objective, positive and negative the terms contained in the synset are. We obtained positive, objective and negative scores for identified words. We have also performed mood and sentiment analysis on the entire text of posts using uClassify [20] and Sentiment Analysis Test Beta version [21] respectively. Mood analysis, using uClassify, associated with each blog post happy and upset scores corresponding to the positivity and negativity in bloggers' moods. The sentiment analysis labeled each blog post as positive, negative or neutral depending on the contents of the blog post. The mood and sentiment analysis targeted the entire post unlike the term based analysis. Besides the analytical results, it also provided a way to validate the findings of the term structure based analysis.

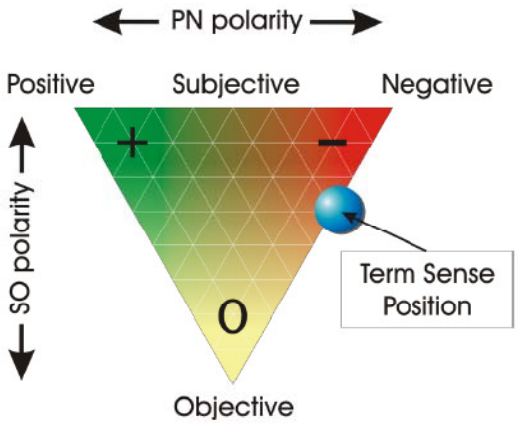

Figure 2. Notion of PN polarity of words.

IV. RESULTS

We have obtained the vector space representation of blog data along with *tf, idf, tf-idf weights* and *cosine similarity* between documents. The blog data was grouped into different clusters based on their similarity and dissimilarity. The most frequent words (determined from *tf* measure) in the blog data included *Nepal, Constitution, Rights, Discrimination, Fundamental, Human, Territorial, Guarantee, Equal, Ensure, Protect.* The blog data was clustered with the appropriate value of K being 5. Table I presents the key themes of the clustered blog data. The terms contained in term vectors have also been subjected to parts of speech tagging where words were grouped into nouns, adjectives, verbs and adverbs. The nouns, adjectives, verbs and adverbs identified from blog posts helped us to find out the key words used by the bloggers in different blog posts, which in turn helped in drawing inferences about their concerns.

The positive, negative and objective scores of different words have been obtained using SentiWordNet. The adjectives and adverbs were expected to have less objective and more subjective scores. Similarly, the sentiment analysis and positive-negative score computation of terms provided the social and emotional contents of reactions of bloggers. We have computed the average positive and negative scores of various clusters to obtain a deeper insight into the reactions of the bloggers around various themes. It was done through a three step process involving parts of speech tagging; computing positive, negative and objective scores of different words; and then averaging the computed scores for different clusters. Table II gives an example of positive, negative and objective score computations for few frequently occurring words. The adjectives mostly have significant values along positivity and negativity whereas nouns mostly have significant values along the dimension of objectivity. Table III lists the key adjectives and their positive, negative and objective scores for cluster with id 0. It represents reactions of people expressed through words. Table IV presents the normalized average positive, negative and objective scores of nouns, adjectives, verbs and adverbs for cluster with id 0.

The identified clusters represent viewpoints of bloggers about different aspects of constitutional development in the democratic republic of Nepal. The terms in cluster 0 have variations in scores of positivity, negativity and objectivity thereby representing the variation in opinion and sentiments of bloggers. However, most of the terms in cluster 1 (having posts about 28th May deadline for framing the constitution) yield higher scores on negativity. This demonstrates that most of the bloggers do not believe that it's a realistic deadline or that it can be achieved. Similarly most of the terms in cluster 2 have higher scores for objectivity as they are primarily factual in nature and raise the various issues of concern. The mood and sentiment analysis task done on the full text of the blog posts also yielded interesting results. Table V displays a part of the result of mood and sentiment analysis. The blog posts are labeled as positive, negative or neutral based on sentiment analysis and similarly they are associated with happy and upset scores based on mood analysis. For example, the blog post titled 'Unlocking Nepal's future through entrepreneurship' was labeled as positive and yielded 95.5% on happy and 4.5% on upset scores. The post titled 'Nepal's rising religious intolerance and communal divide' was labeled as negative and was associated with 5.2% on happy and 94.8% on upset scores. A closer look at the text of the posts makes it clear that the posts have been appropriately labeled and the scores of happy and upset accurately represent the bloggers' mood. The labels and scores associated with the blog posts were also congruent to the parts of speech tagging and resulting PN polarity computations along the clustered blog data.

TABLE I. THEMES IDENTIFIED IN THE COLLECTED BLOG DATA GROUPED INTO 5 CLUSTERS

Cluster Id	Theme
0	Rights of different ethnic groups. Rights of Dalits, NRN, press freedom, gay rights etc.
1	Issues around whether the constitution would be framed by 28th May 2010 or not.
2	Discussion about prioritizing issues and concentrating on them with a deadline of 28th May 2010.
3	Reactions and views of various diplomats on the framing of the constitution
4	Constitution making process, the delays in it and major challenges.

TABLE II. POSITIVITY, NEGATIVITY AND OBJECTIVITY OF WORDS

Words	Positive	Negative	Objective
Important	0.125	0	0.875
Great	0.25	0	0.75
Gay	0.375	0.125	0.5
Himalayan	0	0	1
Extreme	0.25	0.125	0.625
Violate	0	0.375	0.625
Militates	0	0.625	0.375

TABLE III. SCORES FOR ADJECTIVES OF CLUSTER ID 0

Words	Positive	Negative	Objective
temporary	0.75	0	0.25
gay	0.375	0.125	0.5
great	0.25	0	0.75
official	0	0	1
progressive	0	0	1
full	0.375	0	0.625
silent	0	0	1
timely	0	0	1
himalayan	0	0	1
territorial	0	0	1
expert	0.75	0	0.25
general	0	0	1
fundamental	0.375	0	0.625
human	0	0	1
extreme	0.25	0.125	0.625
accessible	0.625	0	0.375
reproductive	0.25	0.125	0.625
convinced	0	0	1
spousal	0	0	1
international	0	0.625	0.375
regional	0	0	1
good	0.625	0	0.375
democratic	0.5	0.25	0.25
crucial	0.375	0.25	0.375
recent	0	0	1

TABLE IV. AVERAGE SCORES FOR CLUSTER ID 0.

	Nouns	Adjectives	Verbs	Adverbs
Avg. Positive Score	0.026	0.174	0.002	0.083
Avg. Negative Score	0.018	0.037	0.045	0.083

TABLE V. MOOD AND SENTIMENT ANALYSIS SCORES & LABELS

Blog Title	Mood Classification Happy (%)	Upset (%)	Sentiment Classification Positive/ Negative/ Neutral
Nepal Constitution timetable amended.	81.9	18.1	Neutral
Monarchy Coming Back? In Your Wildest Dreams Maybe.	36.2	63.8	Positive
Nepal Constitution will Guarantee Equal Rights to LGBTs	79.3	20.7	Positive
New Constitution unlikely to be on time.	15.4	84.6	Neutral
Nepal: Failure to Comply with the Recommendations Issued by UN!	23	77	Negative
Nepal Constitution as the stepping stone for Revolution?	32.9	67.1	Negative
Unlocking Nepal's Future Through Entrepreneurship.	95.5	4.5	Positive
Nepal's rising religious intolerance and communal divide!	5.2	94.8	Negative
Nepal's Constitution must enshrine Press freedom.	38	62	Positive
To have peace and development in Nepal, the rule of law must prevail.	46.4	53.6	Negative

V. CONCLUSION

The experimental work and the analytical results obtained, support our hypothesis that blogosphere analysis is an important and interesting task not only from commercial perspective but also from socio-political angle. The blog data classified into clusters provide a reasonably

good representation of the viewpoints and issues & concerns about the constitutional developments in the new democratic republic of Nepal. The important topics identified are similar to those found in commonly available political literature and socio-political writings in newspapers and magazines. The extracted terms and their classification into different parts of speech help knowing the reactions of people. This includes the emotive elements as well. The computed positivity, negativity and objectivity values of different words provide an insight into the expectations and feelings of people about the topic of analysis. The mood and sentiment analysis on the entire blog post data support the findings of clustering and term structure based analysis. The preliminary data obtained through this work can be used for detailed analysis by political scientists and sociologists. The plausibility of this method of analysis has also been supported by another experiment performed earlier for mining sociological inferences from event based blog posts [22].

Blogosphere has immense potential for socio-political studies and research, which needs to be exploited in a useful manner. The first hand, un-edited, free-form writings of bloggers provide an incomparable input data. The geographical, demographic and cultural diversity of the bloggers make the data still more valuable, and rich in the various perspectives of the topic of concern. Most of the contemporary work on blogosphere analysis (primarily for commercial exploitation), however, is limited to syntactic mechanisms. The textual nature of the blog data and the lack of sophisticated natural language processing tools (having semantic orientation), make the task very challenging. Availability of good text mining and language processing tools will definitely make the analysis results much more valuable. Moreover, with new structuring and semantic representations becoming popular on the World Wide Web, much of the limitations may be overcome and analytical experiment of this kind may produce more accurate results and hence more relevant inferences.

REFERENCES

[1] Technorati Blogosphere Statistics, 2008, http://technorati.com/blogging/state-of-the-blogosphere/

[2] Google Blog Search, http://google.com/help/about_blogsearch.html retrieved May 2010.

[3] Blog Search Engine, http://blogsearchengine.com/ retrieved May 2010.

[4] Bloglines API, http://www.bloglines.com/services/api/ retrieved May 2010.

[5] Java API for XML Processing (JAXP), Sun Microsystems, http://jaxp-sources.dev.java.net/ retrieved April 2010.

[6] Technorati API, http://www.technorati.com/developers/api/ retrieved April 2009.

[7] C.D. Manning, P. Raghvan and H. Schutze, "Introduction to information retrieval", Cambridge University Press, 2008.

[8] S. Chakrabarti, "Mining the Web: Discovering knowledge from hypertext data", Morgan Kaufmann, CA, 2005.

[9] S. Alag, "Collective intelligence in action", Manning, New York, pp. 30-106, 2009.

[10] T. Segaran, "Programming collective intelligence", O'Reilly, CA, pp. 30-84, 2007.

[11] Lucene Full Text Search Engine, http://lucene.apache.org retrieved May 2010.

[12] O. Gospodnetic and E. Hatcher, "Lucene in action", Manning, New York, 2004.

[13] RiTa WordNet Java Library, Part of RiTa Toolkit for Generative Literature, http://rednoise.org/rita/wordnet/ retrieved May 2010.

[14] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297, 1967.

[15] J. A. Hartigan, "Clustering algorithms", Wiley, 1975.

[16] WEKA-Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka/ retrieved May 2010.

[17] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly availabe lexical resource for opinion mining", In Proceedings of fifth Conference on Language Resources and Evaluation, Geneva, 2006.

[18] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Journal of Foundation and Trends in Information Retrieval 2, 2008

[19] G. A. Miller, "Wordnet: A Lexical database for english", Communications of the ACM 38 (11), pp. 39-41, 1995.

[20] Uclassify Mood Analysis tool, http://www.uclassify.com/browse/prfekt/Mood, retrieved Apr. 2009.

[21] Scott Piao, Sentiment Analysis Test Site (Beta Version), http://texto.mib.man.ac.uk:8080/opminpackage/opinion_analysis, 2008, retrieved Jun 2010.

[22] V. K. Singh, "Mining the blogosphere for sociological inferences", In S. Ranka et al. (Eds.): Contemporary Computing, Part I, Communications in Computer and Information Science, Vol. 94, Springer-Verlag, Heidelberg, pp. 547-558, 2010.