A Clustering and Opinion Mining Approach to Socio-political Analysis of the Blogosphere

Vivek Kumar Singh, Rakesh Adhikari, Debanjan Mahata[1]

[1]Department of Computer Science, Banaras Hindu University,
Varanasi, India
vivek@bhu.ac.in, adhikari.rakesh@gmail.com, debanjanbhucs@gmail.com,

Abstract - **Blogosphere is the name associated to universe of all the blog sites. A blog is a website that allows people to write about topics they want to share with others. The ease & simplicity of creating blog posts and their free form and unedited nature have made blogging a happening thing. The blogosphere today contains a large number of posts on virtually every topic of interest. This rich and unique source of data has attracted people and companies across disciplines to exploit it for varied purposes. However, it not only provides a rich data source for commercial exploitation but also for psychological & socio-political analysis purposes. This paper tries to identify the methodology and technical framework required for an analytical experiment of this kind, and demonstrates the plausibility of this idea through our clustering & opinion mining experiment on analysis of blog posts about a recent socio-political issue.**

Keywords- **Blogosphere, Blog Mining, Text Mining, Clustering, Nepal Republic.**

I. INTRODUCTION

A blog is website that allows individuals to write about different topics. Users can both post their views about any topic on the site and comment on the posts written by other users. A blog post may comprise only of text, or it may contain images and links to other media. A blog site typically consists of blog posts arranged in reverse chronological order. Blog sites can be individually owned or they may be a community blog site. Over the past few years, blogs have become a very popular medium for expressing opinions, communicating with others, providing suggestions, sharing thoughts on different issues and to debate over them. The universe of all blogs written by people all over the world in various languages is referred to as Blogosphere. 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].

The large amount of valuable data contained in the blogosphere is attracting interest of people from industry, academicians, professional artists, mass media etc. The studies on blogosphere are now helping in reshaping business models, assist viral marketing, providing trend analysis & sales prediction, aiding counter terrorism efforts etc. The blogosphere has now become an ideal platform from which we can extract information, opinions, moods & emotions of people on various topics. The topics may include anything from political issues, social issues, product reviews to market surveys. However, most of the previous analytical work in blogosphere has focused only on marketing applications. Efforts for socio-political exploitation of the blogosphere are almost negligent.

The fact that the contents of blogs are original free form writings and that they capture uninhibited, unedited expressions of a wide variety of people on very contemporary topics; makes a genuine case for serious efforts for socio-political analysis of the blogosphere. In this paper, we have described our experimental work on socio-political analysis of the blog posts about constitutional developments in the new democratic republic of Nepal. Section II of the paper describes the methodology and technical framework that is used for the analysis. Section III explains the experimental setup and section IV presents the results. The paper concludes (section V) with a short discussion of the relevance of results obtained and the approach in general.

II. METHODOLOGY

The socio-political analysis of the blogosphere requires an appropriate technical framework comprising of suitable search techniques, text processing algorithms and sentiment and mood analysis schemes. Search helps identifying & collecting relevant blog posts, whereas text processing and sentiment analysis techniques help in extracting useful inferences from the collected blog data. During last few years the 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. They try to identify suitable blogs where they can place their advertisements so as to achieve maximum and relevant viewership. The huge size and scale of blogging phenomenon, however, makes these tasks difficult and requires development and use of automated techniques. Fortunately, there has been a noticeable progress towards this end recently. Some of the search and mining techniques developed for marketing & advertising companies can be extended and used for social-political analysis as well.

The analytical task involves a two step process of collecting data (searching) and extracting inferences from the collected data (mining). Searching for relevant blog posts on

978-1-4244-5967-4/10/$26.00 ©2010 IEEE

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 relevant data. The experimenter 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 translates the query into a format acceptable to the blog tracking provider and sends it to blog tracking system. The blog tracking provider then processes the query received from the blog search program and sends back a response, usually in XML. The XML response received is parsed by the blog search program and the retrieved results are displayed. More complex blog search programs can send queries simultaneously to multiple blog tracking providers and aggregate the retrieved responses.

Once the relevant data is available, next step is to process it using suitable text processing algorithms and sentiment & mood analysis techniques to obtain useful inferences. Since the collected blog is primarily textual in nature, it is transformed into an appropriate data structure suitable for clustering and opinion mining. The text contained in different fields of each blog post, including its title, body, comments and user tags, is converted into a term vector structure. The term vector consists of distinct tokens (terms) in the data along with their collection frequency. Tokenization process captures the terms and converts them into a normalized form; where it manages stop words, hyphens, phrases and synonyms etc. Stop words are terms such as and, or to, the, are, their etc. Text processing also requires stemming and preserving multi-word phrases.

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. Once the term vector structure is prepared, second part of the analysis is performed. This is often context specific and may involve clustering the posts into different groups, identifying main topics discussed in posts, aggregating and summarizing the posts, extracting opinions expressed in blog posts and identifying bloggers mood and sentiments,. There are few computational tools and APIs available which can be customized to help tasks like mood & sentiment analysis etc.

III. EXPERIMENTAL WORK

We have chosen to analyze the blog posts about the constitutional developments in democratic republic of Nepal, owing to its current socio-political importance. There are numerous political and sociological issues and people's concerns around the new constitution of Nepal. Our approach involved a content based analysis. We have collected a large number of blog posts on the topic and used the collected blog data for deriving interesting inferences. Two sets of parallel tasks have been performed. First task involved collecting the blog data, storing it in database, extracting term vector from the data and then performing text based analysis. The text analysis included clustering the posts into distinct groups, identifying opinionated words through parts of speech tagging and collection frequency computation and obtaining positive, negative and objective scores of terms along each cluster. The second task involved performing mood and sentiment analysis on the entire text of 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 new democratic republic of Nepal. We did two simultaneous tasks: collecting full blogs through manual search and creating Java program for retrieving feeds of recent blogs from Google Blog Search. This was to help the process of internal validation. 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. 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 search of blogs involved identifying 70-75 high rank blog posts.

B. Creating Term Vectors

We have used the Vector Space Model [2] for representing transformed blog posts. Every blog post was 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 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 a high *tf* value, representing the fact that the theme of the post may be somewhat closely related to this word. The words that appear more often may represent the key idea conveyed in the post. However, to arrive at a definite conclusion, we need to find those terms as well that have the less common occurence. This is the idea behind computing *inverse document frequency (idf)*. If the total number of documents of interest be N, and df_t be the number of documents that contain the term t, the *idf* of the term t is computed as $idf_t = log\ (N/df_t)$. Thus *idf* of a rarely occurring term is higher than a frequently occurring term. 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 $\text{tf-idf}_{t,d} = \text{tf}_{t,d} \text{ X } \text{idf}_t$. 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 [3] were used for tokenizing the document, normalizing them, eliminating the stop words and for stemming. The resulting lists of terms were processed with the help of RiTa WordNet Java library [4]. Parts of speech tagging and synonym injection were also performed using WordNet based infrastructure. The nouns, adjectives, verbs and adverbs identified from blog contents helped us to find out the keywords used by the bloggers in different blog posts, which in turn helped in drawing inferences about their opinions and concerns.

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 [5], [6] 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. Mood and Sentiment Analysis

Like most of the research efforts to extract the opinions from a text being based on finding opinionated words, we have also relied on identifying key words and then computing their positivity, negativity and objectivity scores. Thus determine the "PN-polarity" of subjective terms involves identifying whether the term has a positive or a negative connotation. SentiWordNet [7], [8], [9] is the lexical framework where each wordNet synset s is associated to three numerical scores Obj(s), Pos(s) and Neg(s). These scores describe how objective, positive and negative the terms contained in the synset are. We have also performed mood and sentiment analysis using online APIs and tools such as uClassify and Sentiment Analysis Test Beta version to validate our results.

IV. RESULTS

The computed term vectors, *tf* and *idf* measures for terms, cosine similarity between documents produced interesting inferences. Clustering the term vectors into various groups; identifying nouns, adjectives, verbs and adverbs; computing positive, negative and objective scores of important words have produced important results. The most frequent words (determined from *tf* measure) in the blog data included *Nepal, Constitution, Rights, Discrimination, Fundamental, Human, Territorial, Guarantee, Equal, Ensure, Protect* etc. The collected blog data was satisfactorily grouped into clusters using K-means clustering. Table I presents the key themes of the clustered blog data, where the value of K is 5.

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. We have computed the average positive and negative scores of various clusters to obtain a deeper insight into the reactions of the bloggers. As discussed earlier, 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 III presents the average positive, negative and objective scores of nouns, adjectives, verbs and adverbs for cluster with id 0.

The identified clusters represent viewpoints of bloggers on different aspects of constitutional development in the new democratic republic of Nepal. The terms in cluster with id 0 have variations in scores of positivity, negativity and objectivity. However, most of the terms in cluster with id 1 (having posts about 28th May deadline for framing the constitution) yield higher scores on negativity thereby

emphasizing the fact that a majority of bloggers are not much hopeful about the deadline of 28th May. Similarly most of the terms in cluster with id 2 have higher scores for objectivity. Parts of speech tagging of words helped identifying the reactions and expressions of bloggers. Positive, negative and objective score computations of the terms in term vector accurately identify the type and magnitude of bloggers' reactions on various issues. Mood and sentiment analysis also confirmed the implications obtained through the term vector based analysis.

TABLE I. THEMES IDENTIFIED IN THE COLLECTED BLOG DATA WHEN 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
fundamental	0.375	0	0.625
democratic	0.5	0.25	0.25

TABLE III. AVERAGE SCORES FOR CLUSTER HAVING 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

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 for socio-political inference goals. The blog data classified into clusters provided 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 popular political literature and other media including newspapers and magazines. The extracted tags and their classification into different parts of speech help identifying the reactions of bloggers. 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 preliminary results can be used for a detailed analysis by political scientists and sociologists.

Blogosphere no doubt has immense potential for socio-political studies and research, as is clearly evident from the analytical work (also seen in an earlier work on event based blogosphere analysis [10]). The first hand, un-edited, free-form writings of bloggers; and their geographical, demographic and cultural diversity; provide a valuable and incomparable input data. This data is rich in the variety of perspectives of the topic of concern. The current limitations of syntactic mechanisms, due to the textual nature of the blog data and lack of suitable mining and language processing tools, make the task challenging. With new structuring and semantic representations becoming popular on the World Wide Web, we will see more analytical experiments of this kind which will produce high quality inferences.

REFERENCES

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

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

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

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

[5] 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.

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

[7] 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.

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

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

[10] V. K. Singh, "Mining the Blogosphere for Sociological Inferences", In S. Ranka et al. (Eds.): IC3 2010, Part I, CCIS Vol. 94, Springer-Verlag, Heidelberg, pp. 547-558, 2010.