Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 4 (2013) 212 – 222

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Intelligent Systems and Control

Computational Exploration of Theme-based Blog Data using Topic Modeling, NERC and Sentiment Classifier Combine

V. K. Singh[a]*, P. Waila[b], R. Piryani[a], A. Uddin[a]

[a]*Department of Computer Science, South Asian University, New Delhi 110021, India*
[b]*DST Centre for Interdisciplinary Mathematical Sciences, Banaras Hindu University, Varanasi 221005, India.*

Abstract

This paper presents findings of our exploratory research work on a novel combine of Topic Modeling, Named Entity Recognition and Sentiment Classification for sociological analysis of blog data. We have collected more than 500 blog posts on the broader theme of 'Discrimination, Abuse and Crime against Women'. We employed topic discovery to identify top keywords and key themes and implemented the 7-entity model Named Entity Recognition process to identify the key persons, organizations and locations discussed in the blog posts. Thereafter we performed sentiment classification of the entire blog data into positive and negative categories using SentiWordNet. The results obtained are very interesting and validate the usefulness of our approach for computational analysis of social media data. The key contribution of the paper is to propose a novel Text Analytics combine and demonstrate its applicability for computational exploration of the social media data for sociological analysis purposes.

Keywords: Social Media; Text Analytics; Topic Modeling; Named Entity Recognition; Sentiment Classification

* Corresponding author. Tel.: +91-11-24195148; fax: +91-11-24122511.
E-mail address: vivek@cs.sau.ac.in.

doi:10.1016/j.aasri.2013.10.033

1. Introduction

In the journey of Time magazine's naming "You" as person of year in 2006 to "The Protester" as the person of year in 2011, social media is the common entity and driving force. While 2006 saw social media platforms started becoming popular and rapidly attaining huge user base; in 2011 social media platforms were used to organize and coordinate protests in various countries ranging from Libya to Tunisia at unprecedented scale. The blogging phenomenon is a major part of this social revolution. A blog (or more technically a weblog) is a website that allows one or more individuals to write about things they want to share with others. Blogs can be individual blog sites, like personal diaries; or community blog sites, which are like discussion forums & collaborative platforms. A typical blog post can have text, images, and links to other media. The universe of all blog sites is typically referred to as the blogosphere.

The blogging phenomenon has grown at an unprecedented rate. The blog tracking company *Technorati* tracked about 4 million blogs in September 2004, which has grown to about 164 million blogs in July 2011(Technorati Statistics, 2013), an increase of 41 times in just a period of seven years. A recent statistics by Wordpress (Wordpress Statistics, 2013) reports that it has a user base of more than 383 million people and more than 3.5 billion blog page views every month. The users of Wordpress alone produce 33.9 million new blog posts and about 40.9 million new comments every month. Blogosphere is now a huge collection of discussions, commentaries and opinions on virtually every topic of interest, and its size and magnitude is increasing every day. As of now approximately 66% blogs are in English. As the penetration of Internet will increase more in the other parts of the world, the number of blog posts (including in other languages) will further increase. This is further strengthened by the fact that now 4 out of every 5 internet users use some kind of social network and media (including blogs).

The large number of blog posts and comments created on blogging sites become more interesting when we look at the user profile statistics. According to an estimate by *Blogging.org,* approximately 60% bloggers are hobbyists (Infographic, 2012). They are not guided by business or professional motives, but voluntarily write about various things ranging from politics, religion to society. While a major reason behind the popularity of blogging is the ease of creating blog posts, low barrier to publication, open standards of content generation and the free-form writing style; but it is the urge of expression and availability of a convenient platform, which has made blogosphere to attain the current size and scale.

One aspect of the blogosphere that remained relatively unexplored till the recent times is that it is a rich and unique treasure house for cross-cultural psychological & sociological analysis as well. This paper presents our computational exploration of the blog text data with this broader aim. The section 2 describes the motivation for this work. The computational formulation combine used by us is described in section 3. The section 4 explains the dataset collection and its properties and section 5 presents the detailed experimental setup and results. The paper concludes with a summary of observations illustrated in section 6.

2. Motivation

The unprecedented rate of growth of blogging and the huge amounts of data (mostly textual) that is now contained in the blogosphere is a unique treasure not only for commercial exploitation but also for sociological and political analysis. Two key observations that make this statement more relevant are that (a) the Internet has reduced the distance between people across the world and allowed them to express themselves and interact with others, irrespective of the geographical, demographic, religious and cultural boundaries; and (b) the free form, unedited, first hand and relatively more emotionally laden expressions of various people on various social, political, cultural issues. Blog sites are now a very rich source for cross-cultural and diverse socio-political account of bloggers on varied issues and events.

During the last few years researchers from different domains have started exploring the blogosphere for non-commercial aspects. This analytical work has two broad flavours. A more computer science oriented flavour includes tasks like finding influential bloggers (Agarwal et al., 2008) and blog sites about an event (Mahata and Agarwal, 2012), community discovery, filtering spam blogs etc. (Liu et al., 2010), (Agarwal and Liu, 2008). The other flavour is oriented more towards socio-political analysis of blog posts (Singh et al., 2012), (Singh et al., 2010), (Singh, 2010). This includes tasks like mapping the blogosphere around a particular socio-political event (Mehrav et al., 2012), analysis of blog posts relevant to an important event/ personality/ organization or process (Moe, 2011), (Suhara et al., 2007), (Adamic and Glanse, 2005), (Lin and Halavais, 2004). Since we wanted to study the blog posts related to the topic of discrimination of women at workplaces and crime against them, our approach is a socio-political analysis one. Our aim was primarily to explore the major entities (persons, organizations etc.) discussed in the blog posts; identify key issues about the theme and to understand how bloggers perceive the theme in general. Blog text was chosen because it's the best source to obtain un-inhibited, first hand, un-edited expression, thoughts and viewpoint of people across the world.

3. Computational Formulations

We have used a novel combine of Topic Modeling, Named Entity Recognition and Sentiment Classifier for the blog text analysis. Here we briefly describe these three computational formulations as popularly perceived and used by us.

3.1. Topic Modeling

Topic Modeling (also known as topic discovery) identifies the themes inherent in a collection of documents or in other words it tries to annotate a large collection of documents with thematic information. It employs a set of statistical methods that analyze the words of the text documents in a collection, use the information of word usage patterns and connect documents that exhibit similar patterns. It uses a probabilistic model based on hierarchical Bayesian analysis of the text documents (Blei, 2012). Topic modeling can be used not only for discovering themes, but also to figure out how these themes are connected to each other and possibly how they change over time. The simplest kind of topic model is Latent Dirichlet Allocation.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a statistical method that uses a generative process (an imaginary process by which the model assumes that documents are generated by the topics). A topic is formally considered as a distribution over a fixed vocabulary. The main idea behind LDA is thus to model documents as arising from multiple topics, more specifically we assume that some k topics are associated with the documents collection and that each document exhibits these topics with different proportions. All the documents in the collection thus share the same set of topics, but each document exhibits those topics with different proportions. With a good topic modeling algorithm, the inferred hidden topic structure resembles the thematic structure of the document collection. The Bayesian Non-parametric Topic Model, Dynamic Topic Model, and the Correlated Topic Model are few other variants of Topic Models (Blei and Lafferty, 2009).

3.2. Named Entity Recognition

Named Entity Recognition (NER) is a task in IR that tries to identify and classify words in a text into some predefined categories, such as names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The term 'Named Entity' is in use since the sixth Message Understanding Conference (MUC) in 1996. It is also known as Named Entity Recognition and Classification (NERC), for it

performs dual task of identifying proper names in text and classifying them into predefined categories. NER generally adopts one of the following three approaches: rule based, machine learning based and hybrids (Nadeau and Sekine, 2007). Modern NER systems perform the entity identification and classification task either by using a hand crafted linguistic grammar based technique for the concerned language, or by using statistical models of classification. The first one involves a good amount of work by computational linguists to hand code entity location and identification rules for a particular language. While the second one uses a statistical model, such as Naïve Bayes classifier, Conditional Random Fields, Hidden Markov Models or Maximum Entropy method, however they require sufficient amount of human annotated training data.

In this work we have used the 7-entity model Stanford Named Entity Recognizer (NER, 2012). It's a Conditional random Field based Machine Learning classification approach that has following seven classes: Person, Organization, Location, Time, Money, Percent, and Date. For example, consider the text: "South Asian University was established by the SAARC nations with a financial support of $1M, 50% of the finance is supported by India, located in New Delhi". The NER algorithm interprets this text in the following indicative form: <*ORGANIZATION*> South Asian University <*/ORGANIZATION*> was established by the <*ORGANIZATION*> *SAARC* <*/ORGANIZATION*> nations with a financial support of <*MONEY*>*$1M*<*/MONEY*>, <*PERCENT*> *50%* <*/PERCENT*> of the finance is supported by <*LOCATION*> *India* <*/LOCATION*>, located in <LOCATION> New Delhi </LOCATION>.

3.3. Sentiment Classification

Sentiment classification aims to assign a 'positive' or 'negative' label to every opinionated text. There are primarily three types of approaches for sentiment classification of opinionated texts: (a) using a machine learning based text classifier -such as Naïve Bayes, SVM or kNN; (b) using Semantic Orientation scheme of extracting relevant n-grams of the text; and (c) using the SentiWordNet based publicly available library that provides positive, negative and neutral scores for words. In this work we have used the SentiWordNet approach for classifying the sentiments expressed in blogs. The main reason behind this choice was that it does not require any training data, can be readily implemented and obtains reasonable accuracy levels. The SentiWordNet approach involves use of the publicly available library of SentiWordNet (SentiWordNet, 2012). To make use of SentiWordNet we need to first extract relevant opinionated terms and then lookup for their scores in the SentiWordNet. Past works in the area have shown that adjectives, adverb+adjective and adjective+verb combinations are few reasonable choices for the terms to be extracted.

We have used a simple version of the SentiWordNet approach, which extracts adverb+adjective combines and compute their SentiWordNet scores. These scores are computed in a way that gives some weight to adverb scores as well, in addition to adjective scores (Singh et al., 2013). In fact presence of adverbs prior to adjectives, modify their SentiWordNet scores. We also considered for occurrence of the term 'not' and negated the sign of SentiWordNet score value of a term if a 'not' precedes it. A text is classified as 'positive' or 'negative' based on the aggregate sum values for extracted 'Adv+Adj' combine being positive or negative.

4. Dataset Collection

We have collected blog text on the broader theme of "Discrimination, Harassment and Crime against Women". We collected a total of 512 blog posts from very popular blog sites like Wordpress, Blogspot, Thefword, Feministblog and Blogher, during June 2012. We have used a completely automated process for data collection. We wrote a search client program which uses Google search api for finding relevant blog posts from the blog sites mentioned above and result the url link of blog texts that satisfy our given query. We made a persistent repository of url links. As second phase we executed a JAVA program to fetch the text of

the blog posts for the url's in our repository. The entire data is collected in an xml format. Essentially, for every blog post we store the following xml tags:

```
<blog>
    <url>  </url>
    <language>  </language>
    <author>  </author>
    <title>  </title>
    <text>  </text>
</blog>
```

In order to find the relevant blog posts, we supplied following seven hand coded search queries to the crawler: "discriminating weaker sex at workplaces", "sexual harassment of women at workplaces", "unfavorable conditions for women at workplaces", "sexual abuse and discrimination of women", "preventing discrimination and harassment of women", "women in third world countries", and "crime against women". As can be observed from the queries, we tried to collect highly relevant and at the same time relatively factual, argumentative or opinionated blog posts. Some of the blog post url links returned by the crawler contained very small amount of texts or were in language other than English, therefore we filtered them out and our resultant dataset finally comprised of about 485 blog posts out of 512 posts collected on this theme. A summary of the blog posts collected (before filtering) from the different sites is given in Table 1 below.

Table 1. Details of Dataset Collected

Blog Site	No. of Blog Post	Word Count	Unique Word Count	Average Word Count
Blogspot	213	4749098	22348	22296.234
Wordpress	194	5308684	26801	27364.351
Blogher	10	11104	1212	1110.4
Thefword	22	54693	2894	2486.045
Feministblog	60	373641	4185	6227.35
Miscellaneous	13	42518	3455	3270.615

5. Experimental Work and Results

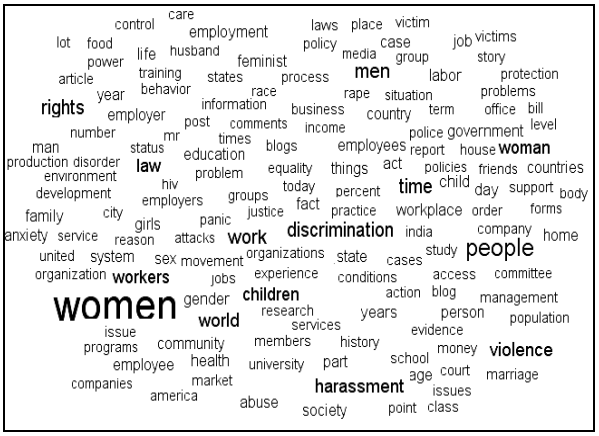
This section describes the step by step process of running the different parts of computational formulation, its aim and the result obtained. We developed an integrated java program capable of doing topic modeling, vector space model conversion from text, parts of speech tagging of texts, named entity recognition from text document and sentiment analysis by using the SentiWordNet library. The first task that we carried out across the entire collection of blog posts was topic modeling, which we did using Stanford Topic Modeling Toolbox (Stanford Topic Modeling, 2012). The main goal of topic modeling for us was to find the key themes running across the entire blog data collection. This would not only help in identifying the major themes expressed by bloggers but also capture the top representative keywords that one may expect to find in literature and texts on this theme. We extracted top 50 keywords from the topic modeling result and filtered this list to arrive at top 20 most relevant keywords. After obtaining the reduced list of 20 keywords, we computed the frequency of their occurrence across the entire dataset. These top 20 keywords along with their frequency counts are displayed in table 2. While some of the top occurring words like 'work', 'sexual', 'violence', 'harassment', 'discrimination' and 'women' were also part of the hand coded search queries, we obtained fairly different top keywords as well. The overlapping words with our queries are a good measure of the relevance of blog data collected with respect to the theme of analysis. The other top keywords obtained included words like 'men',

'social', 'law', 'years', 'life' and 'state'. This is a good indicator of the key entities/ points involved in the writings on this theme. While many writings emphasize that it's the 'men' who are primarily responsible for discrimination and harassment; others express that its problem which needs support of 'social institutions', 'state' and 'law' to resolve the problems. Some other writings, as captured through the computational approach adopted here, also express that this discrimination and harassment is not a problem of only modern days but it has been there since many 'years' and women suffer it in various forms throughout their 'life'. We plotted a tag cloud of the top keywords obtained to have a better visualization of the main themes/ actors/ institutions etc., expressed through keywords. Figure 1(a) presents this plot drawn from the top keyword list transformed into a graph data file (gdf) for use by Gephi (Gephi, 2012). The size of the word in the tag cloud plot represents its frequency of occurrence, i.e. words with higher frequency of occurrence appear bigger than those with lower occurrence frequency.

Table 2. Top 20 Keywords with Frequency

Word 1-10	Count	Word 11-20	Count
Work	1365	make	688
Sexual	1205	law	687
Men	1128	woman	681
Time	1015	workers	620
discrimination	946	state	611
Rights	910	gender	607
Violence	869	life	582
harassment	792	don	568
World	758	labor	568
Social	709	years	567

(a)



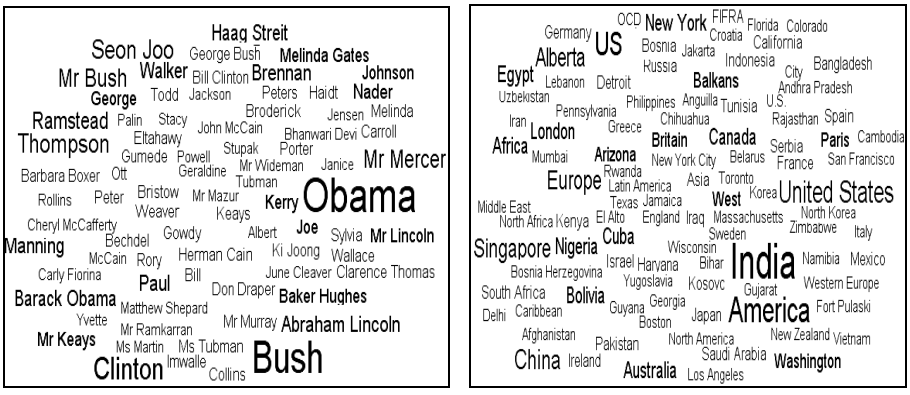
(b)

Fig. 1. (a) Tag Cloud visualization of the top 50 keywords; (b) Tag Cloud Plot of Noun Occurrences in the data

After doing the topic modeling, we thought it may be a good idea to see the relative strength of frequency of occurrence of nouns in the entire dataset. For this we performed POS tagging of all the blog posts and extracted those with the tag <noun>. After extracting nouns, we counted the occurrence frequencies of these nouns across the entire dataset and used the frequency counts to plot the nouns in a tag cloud plot, shown in figure 1(b) above. The idea behind extracting nouns and plotting their relative occurrence frequency strengths is to identify the key important entities in the dataset on the theme. As is clearly visualized in the plot, few notable noun occurrences are 'women', 'people', 'rights', 'children', 'people', 'society', 'community', 'men', 'employer', 'school', 'court', 'state', 'management' etc. This plot thus clearly brings in perspective, all major actors, institutions and entities related to the theme of discussion.

The next task that we performed to further illustrate the key entities in the blog data was Named Entity Recognition (NER). We implemented the Stanford 7-entity NER model and obtained the entire list for entities of the class person, location and organization etc. Thereafter, we extracted all the entities of the three important classes – person, location and organization. NER recognized a total 2012 person entities, 907 location entities and 2722 organization entities in the entire dataset. The figure 2(a) plots the tag cloud for person class, figure 2(b) for the location class and the figure 2(c) for the organization class; showing their relative strength of occurrence in the entire dataset. In order to draw these plots, we have taken only entities with minimal occurrence threshold (4 for person class, 8 for location class and 7 for organization class).

The tag cloud for person class shows prominent entities like 'Obama', 'Bush', 'Clinton', 'Melinda Gates', 'Stacy', 'Mr Murray'. These are the most referred names by the bloggers on the theme. A possible reason is that these are the people who have been advocating policies on this theme and/or people look with hope towards these people to get rid of the menace of discrimination of women. Similarly, the location cloud has prominent entities ranging from 'America', 'Canada', 'Britain' to 'Singapore', 'Australia' and those like 'India', 'China', 'Nigeria'. It can therefore be clearly inferred that it's a worldwide problem which is seen in almost all societies and countries at different stages of development. It may also be a depiction that almost all countries in the World are trying to address the issue with legislation, community awareness and other kinds of affirmative action. The tag cloud plot of entities of organization class mention few important organization names ranging from States, Governments, Corporations and Universities to individuals.



(a) (b)

(c)

Fig. 2. Tag cloud plots of (a) Person-class entities; (b) Location-class entities; (c) Organization-class entities

The next computational task that we performed was the sentiment analysis of the blog data. We first extracted all adverb+adjective combines in a blog posts and then used SentiWordNet based formulation to compute the sentiment score of a blog posts by aggregating the sentiment values of individual adverb+adjective combines. We have plotted the sentiment polarity strengths of all the blogs in the dataset in figure 3. The sentiment value plotted is sum of 'Positive' and 'Negative' scores of that blog post with the sign being used for computation.

6. Conclusion

The computational framework that we have designed for the exploratory analysis of the blog data on a particular theme has been able to obtained very interesting and relevant results. Through the topic modeling implementation we are able to find out the major thematic keywords from the entire blog data collection. These thematic keywords depict the major issues, role players and entities associated with the writings on the theme of 'discrimination and harassment of women'. The POS tagging helps in identifying the nouns

occurring in the dataset, which further elaborates the key issues/ institutions and other entities connected with this theme. The NER implementation helps in identification of entities a level further by allowing extracting persons, locations and organizations mentioned in the dataset. We are able to identify the major persons, locations and organizations that are frequently talked about or are found connected in a strong way to the issue in all writings on this theme. Similarly the results of sentiment analysis show that the entire dataset is relatively equally distributed on both 'positive' and 'negative' sentiment scale, thereby showing that the writings on this theme are not only negatively oriented, rather a good number of them express hope and optimism of improvement in the situation. An entity-based sentiment analysis can further focus it by helping in obtaining sentiment orientation on all major entities talked about in the dataset.

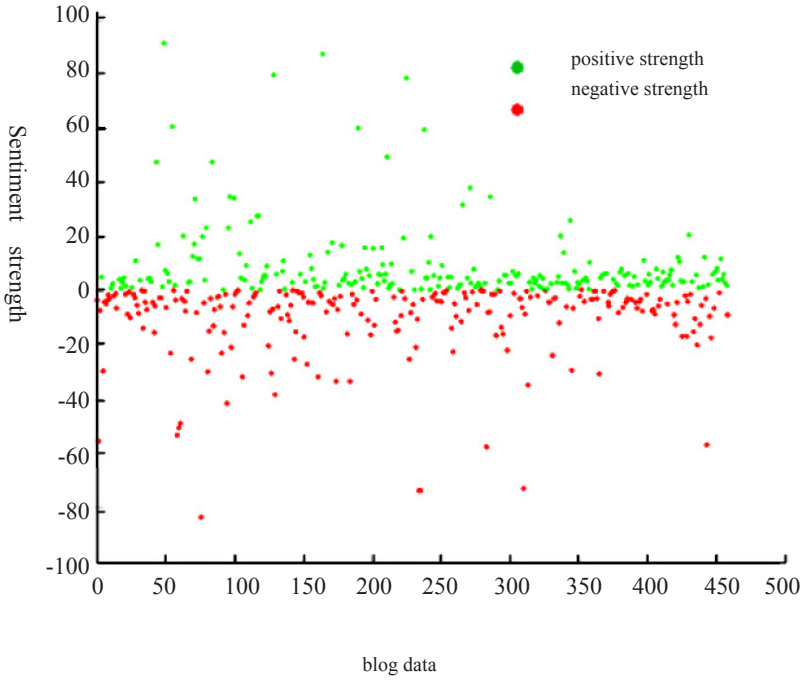


Fig. 3. Sentiment Polarity Strength Constellation for the blog data

The analytical task that we undertook is on a very relevant theme in today's world and we proposed to approach this analytical task through a computational formulation that uses a combine of Topic Modeling, NER and Sentiment Analysis. This approach has many advantages over a traditional subjective analysis, though it does not aim to be a substitute for the traditional subjective analysis. First of all, our computational formulation automatically collects relevant texts written by people across the world, thereby allowing an inherent cross-cultural and demographic perspective on the issue. Secondly, we are not limited in the amount of data that we can analyze quickly. Analyzing this scale of data in a traditional manual way requires a much higher amount of effort and time. Thirdly, this formulation can identify the major themes running across the entire text collection and also measure their relative strengths. This computational formulation thus presents a unique framework for automatic analysis of text documents, in much less effort and time as compared to traditional subjective methods, and inherently provides for cross-cultural sociological and socio-political

perspective of analysis along any important theme/ issue of interest. The findings can also provide an initial starting point (or food for thought) for a detailed subjective analysis around the theme.

References

[1] Technorati & Blogpulse Blogging Statistics, Retrieved from http://www.socialmediaexaminer.com/tag/blogging-statistics/ on Jan 15, 2013.
[2] Wordpress Blogging Statistics, Retrieved from en.wordpress.com.stats/ on Jan 15, 2013.
[3] Blogging Stats 2012 (Infographic), Retrieved from http://blogging.org/blog/blogging-stats-2012-infographic/ on Jan 17, 2013.
[4] Agarwal N, Liu H, Tang L, and Yu PS. Identifying the Influential Bloggers in a Community. In Proceedings of International Conference on Web Search and Web Data Mining; ACM Press, Palo Alto, USA 2008, pp. 207-218.
[5] Mahata D and Agarwal N. What Does Everybody Know? Identifying Event-specific Sources from Social Media. In Proceedings of the fourth International Conference on Computational Aspects of Social Networks (CASoN 2012); November 21-23, 2012; Sao Carlos, Brazil.
[6] Liu H, Yu PS, Agarwal N and Suel T. Social Computing in the Blogosphere. IEEE Internet Computing; April 2010; pp. 12-14.
[7] Agarwal N and Liu H. Blogosphere: research Issues, Tools and Applications. SIGKDD Explorations; Vol. 10, No.1; pp. 18-31; 2008.
[8] Singh VK, Mukherjee M, Mehta GK, Tiwari N and Garg S. Opinion Mining from Weblogs and its Relevance for Socio-political Research. In M Natarajan, C Nabendu and N Dhinaharan (Eds.) Advances in Computer Science and Information Technology. Computer Science and Engineering; Part II, Jan. 2012, LNICST 85, Springer, pp. 134-145.
[9] Singh VK, Mahata D and Adhikari R. Mining the Blogosphere from a Socio-political Perspective. In Proceedings of International Conference on Computer Information Systems and Industrial Management Applications (CISIM), 2010, pp. 365-370.
[10] Singh VK. Mining the Blogosphere for Sociological Inferences. In S Ranka et al. (Eds.): Contemporary Computing; CCIS Vol. 94, Springer-Verlag, Heidelberg; 2010, pp. 547-558.
[11] Mehrav Y, Mesquita F, Barbosa D, Yee WG and Fireder O. Extracting Information Networks from the Blogosphere. ACM Transactions on the Web; Vol. 6; No. 3; September 2012.
[12] Moe H. Mapping the Norwegian Blogosphere: Mthodological Challenges in Internationalizing Internet Research. Social Science Computer Review 29(3) 313-326, 2011.
[13] Suhara Y, Toda H and Sakurai A. Event Mining from the Blogosphere using Topic Words. Proceedings of ICWSM; 2007.
[14] Adamic L and Glanse N. The Political Blogosphere and the 2004 US Election: Divided they Blog. Proceedings of 3rd International Workshop on Link Discovery; ACM; 2005.
[15] Lin J and Halavais A. Mapping the Blogosphere in America. In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics; 2004.
[16] Blei D. Probabilistic topic models. Communications of the ACM; 55(4); pp.77–84, 2012.
[17] Blei D, Ng A and Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3; pp.993–1022; January 2003.
[18] Blei D and Lafferty J. Topic Models. In A Srivastava and M Sahami (eds.) Text Mining: Classification, Clustering, and Applications, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series; 2009.
[19] Nadeau D and Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes 30.1; pp. 3-26, 2007.
[20] Stanford Named Entity Recognizer, Retrieved from http://nlp.stanford.edu/software/CRF-NER.shtml on

15 Dec. 2012.
[21] SentiWordNet, Retrieved from http://sentiwordnet.isti.cnr.it/ on 15 Dec., 2012.
[22] Singh VK, Piryani R, Uddin A and Waila P. Sentiment Analysis of Movie Reviews and Blog Posts: Evaluating SentiWordNet with different linguistic features and scoring schemes. In Proceedings of 3rd IEEE International Advanced Computing Conference; Ghaziabad; India; Feb. 2013.
[23] Stanford Topic Modeling Toolbox, Retrieved from nlp.stanford.edu/software/tmt/tmt-0.4/ on 15 June 2012.
[24] Gephi: The Open Graph Viz Platform, Retrieved from https://gephi.org/ on 1 Dec. 2012.