Opinion Mining from Weblogs and Its Relevance for Socio-political Research

Vivek Kumar Singh[1], Mousumi Mukherjee[2], Ghanshyam Kumar Mehta[2], Shekhar Garg[2], and Nisha Tiwari[2]

[1] Department of Computer Science, South Asian University, New Delhi-110067, India
[2] Department of Computer Science, Banaras Hindu University, Varanasi-221005, India
{vivek,mou.sonai,ghanshyam4u2000,shekharbhumca08, nisha.bhumca08}@gmail.com

Abstract. This paper presents our experimental work on mining of opinions from a large number of blog posts and its relevance for socio-political research. The experimental work involves collecting blog data on three interesting topics, transforming the collected blog data into vector space representation, and then performing opinion mining using both a machine learning text classifier and an unsupervised semantic orientation approach. We implemented Naïve Bayes and SO-PMI-IR algorithms for opinion mining. We obtained interesting results, which have been evaluated for correctness and also cross-validated with the outcomes of multiple techniques employed. The paper concludes with a short discussion of the results and relevance of the experimental work.

Keywords: Blogosphere, Computational Sociology, Opinion Mining, Sentiment Analysis, Social Computing.

1 Introduction

The increased penetration of the Internet and the new participative Web 2.0 has facilitated a large number of users to use & interact with web applications. The users are now interacting with web applications and contributing in a variety of forms, such as rating, tagging, writing blogs, social networking and sharing items with friends. This is creating huge amount of user generated information. Weblogs (also termed Blogs) provide an important platform for users to express themselves. The ease of creating blog posts, low barrier to publication, open standards of content generation and the free-form writing style allow large number of people to create their own blogs or post their contributions on community blog sites. People express their opinions, ideas, experiences, thoughts, and wishes in blog posts. The Blogosphere (universe of all blog sites) is now a huge source of user posted data [1]. The tremendous amount of valuable information contained in blogs has attracted the attention of people in academics as well in industry. Over the years, social media platforms such as Blogosphere have become a widely acknowledged platform for commercial exploitation. Companies use it to know the reactions of users on its products; and

N. Meghanathan et al. (Eds.): CCSIT 2012, Part II, LNICST 85, pp. 134–145, 2012.

advertisers use it to place their advertisement. The high user activity, wide visibility and large amount of user posted data, is attracting people and companies to exploit it in numerous other ways.

The immense potential of the data in Blogosphere has opened new areas of research in and around the blogosphere, with key areas of research including efforts to model the Blogosphere, blog clustering, mining the blog posts, community discovery, searching influential bloggers, filtering spam blogs etc [2]. However, one area that is relatively still unexplored is to use the data on the Blogosphere for socio-political analysis. Blogosphere is a rich and unique treasure house of cross-cultural data that can be used for psychological & socio-political analysis. There are two important motivating factors for this purpose. First, 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. Therefore, we now have opinions on a topic from people across the continents. Secondly, the free form, unedited, first-hand and relatively more emotionally laden expressions of various people on blog sites provide a rich source of original data for cross cultural & sociological analysis. Moreover, one may no longer be required to travel distances to get the cross-cultural perspective on different issues. Blog posts on an issue may be used for an analytical experiment of this kind, which can result in at least a preliminary picture, if not a rigorously validated one. Hence, we get a more original data at a much lower cost.

The unstructured nature of data in blogs, however, presents computational challenges that require sophisticated search and mining techniques. Opinion Mining which has traditionally been an area of exploration by Linguists, is now becoming a mature technique for use with data on the Web. During the last few years there have been interesting works on opinion mining from the user posted data on the Web. Experiments to identify opinions have been performed with product reviews, movie reviews, restaurant reviews, user's posts on social media and about prospects of candidates in elections etc. This paper presents our experimental work on opinion mining on three different collections of Weblogs. We chose the topics of posts taking into account the highly opinionated & emotion laden nature of the prospective posts. The primary aim of the experimental work is to evaluate and demonstrate that the Blogosphere is a platform worth socio-political and cross-cultural psychological research. The rest of the paper is organized as follows. Section 2 describes opinion mining techniques used and the section 3 presents the experimental setup and results. The paper concludes with a short discussion (Section 4) of the relevance of experimental work for socio-political and cross-cultural psychological research.

2 Opinion Mining

The opinion mining problem in its most popular sense can be formally defined as follows: Given a set of documents D, the opinion mining algorithm classifies each document $d \in D$ into one of the two classes, positive or negative. Positive means that d expresses a positive opinion and negative means that d expresses a negative opinion. Most of the experiments performed so far employed one of the following two approaches: (a) using a text classifier (such as Naïve Bayes, SVM or kNN) that takes a machine learning approach to categorize the documents in positive and negative

groups; and (b) computing semantic orientation of documents based on aggregated semantic orientation values of selected opinionated POS tags extracted from the document. Some of the past works on opinion mining and sentiment analysis have also attempted to determine the strengths of positive and negative orientations. Few prominent researches works around these themes can be found in [3], [4], [5], [6], [7], [8], [9] & [10].

2.1 Machine Learning Approach

A simple scheme for opinion mining is to use supervised machine learning based text classification approach to classify the documents in two classes (positive and negative). Naïve Bayes and SVM are two widely used machine learning approaches. Naïve Bayes [11] is a probabilistic learning method which computes the probability of a document *d* being in class *c* as in eq. 1 below.

$$\mathrm{P}\,(c|d)\ \alpha\ \mathrm{P}(c)\prod_{1\le k\le nd} P(t_k|c) \tag{1}$$

where, P (t_k/c) is the conditional probability of a term t_k occurring in a document of class c. P(t_k/c) is thus a measure of how much evidence t_k contributes that c is correct class. P(c) is the prior probability of a document occurring in class c. The goal in text classification is to find the best class for a document. The key idea in this classification is thus to categorize documents based on statistical pattern of occurrence of terms. The selected terms are often called features. Therefore, in order to do topic categorization we may use features like terms having frequency greater than a value. For categorizing documents into two categories of 'positive' and 'negative', a good choice is to use selected terms with specific POS tags such as adjectives (or adjective and adverb combination). The class membership of a document can be computed as in eq. 2

$$c_{map} = arg\max_{c\in C}\widehat{\mathrm{P}}(c|d)\ = arg\max_{c\in C}\widehat{\mathrm{P}}(c)\prod_{1\le k\le n_d}\widehat{\mathrm{P}}\,(t_k|c) \tag{2}$$

where, P` is an estimated value obtained from the training set. In order to reduce the computational complexity resulting from multiplication of large number of probability terms, the eq. 2 can be transformed to eq. 3.

$$C_{map} = arg\max_{c\in C}[\log\hat{P}(c) + \sum_{1\le k\le n_d}\log\hat{P}(t_k|c)] \tag{3}$$

Each conditional parameter P(t_k/c) is a weight that indicates how good an indicator the term t_k is for class c and the prior log P(c) indicates relative frequency of class c.

2.2 Semantic Orientation Approach

In semantic orientation approach we first extract phrases that conform to a specific pattern [12]. Thereafter, the semantic orientation of extracted phrases is computed using the Pointwise Mutual Information (PMI) measure given in Eq. 4 below,

$$PMI(term_1, term_2) = \log_2\{\mathrm{Pr}\,(term_1\Delta term_2)/\mathrm{Pr}(term_1)\,.\,\mathrm{Pr}\,(term_2)\} \tag{4}$$

where, Pr($term_1$ ▲ $term_2$) is the co-occurrence probability of $term_1$ and $term_2$ and Pr($term_1$).Pr($term_2$) gives the probability that two terms co-occur if they are statistically independent. The ratio between Pr($term_1$ ▲ $term_2$) and Pr($term_1$).Pr($term_2$) measures the degree of statistical independence between them. The semantic orientation (SO) of a phrase can thus be computed by using the Eq. 5,

$$SO(phrase) = PMI(pharse, \text{"excellent"}) - PMI(phrase, "poor") \quad (5)$$

where, PMI (phrase, "excellent") measures the association of the phrase with positive reference word "excellent" and PMI (phrase, "poor") measures the association of phrase with negative reference word "poor". These probabilities are calculated by issuing search query of the form "phrase * excellent" and "phrase * poor" to a search engine. The number of hits obtained is used as a measure of PMI value. The SO value for all the extracted phrases is computed using this scheme. To determine the semantic orientation of the entire document, the SO values of the opinionated phrases in it is aggregated. Every term having SO value greater than a threshold value is assigned a score of '+1' and '-1' otherwise. The SO values of all the extracted subjective terms are then added and if the sum is greater than a threshold value, the document is labeled as 'positive' and 'negative' otherwise. This algorithm is often referred to as SO-PMI-IR.

A variation of this scheme is SO-PMI-LSA [13], which uses Latent Semantic Analysis. In this scheme, the term-document matrix is first reduced using Singular Value Decomposition (SVD) and then LSA($word_1$, $word_2$) is computed by measuring the cosine similarity (described in Eq. 7) of the compressed row vectors corresponding to $word_1$ and $word_2$. Then, the Semantic orientation of a word is computed by Eq. 6,

$$SO(word) = LSA(word, \{positive\ terms\}) - LSA(word, \{negative\ terms\}) \quad (6)$$

where, positive terms refer to words like 'good', 'superior', 'excellent' etc. and negative terms refer to words like 'bad', 'poor', 'inferior' etc. The LSA of a word is computed with term vectors of positive and negative words occurring in the document set. Experimental results have shown that SO-PMI-IR and SO-LSA have approximately the same accuracy on large datasets.

2.3 Mood Analysis

The task of classifying a document by mood involves predicting the most likely state of mind with which the document was written i.e., whether the author was depressed, cheerful, bored, upset etc. The task is similar to opinion mining and has been used for productive purposes viz. filtering results of a search engine by mood, identifying communities and possibly even to assist behavioral scientists in behavioral research and training. Most of the past research on mood analysis used style-related as well as topic-related features in the text for identifying the mood of the author. Usually 'bag of words' or POS tags are extracted along with their frequency counts for subsequent use in mood classification. There has been several interesting works on mood analysis [14], [15], [16]. In an important experiment with mood classification in blog posts Mishne used a Mood PMI-IR based approach. This approach is conceptually similar

to SO-PMI-IR scheme. However, instead of using "excellent" and "poor" as reference terms, he used terms corresponding to various moods (such great, annoyed, cheerful, sleepy etc.) for calculating PMI measure.

3 Experimental Setup and Results

We have performed opinion mining, mood analysis and gender analysis on three different blog datasets. We collected blog data on three topics. The collected data was transformed into vector space model representation and bag of words based feature selection was used for opinion mining. We implemented both Naïve Bayes text classifier and SO-PMI-IR algorithms for classifying blogs. To implement SO-PMI-IR, we extracted selected POS tags and computed their SO values. The SO values of terms in a blog were then aggregated using two different schemes to classify the blog as positive and negative. To implement Naïve Bayes, we used a three-fold scheme. One part of the data was used for training and the other two are then classified accordingly. We used both manual opinion labels and opinion labels assigned by SO-PMI-IR for training. The results of both these methods were cross validated.

3.1 Collecting the Blog Data

We obtained a large number of blogs on following three topics: 'Women's Reservation in India', 'Regionalism' and 'Three International Terror Events'. We collected full blogs from Google Blog Search through a Java program. Everyïblog was stored as separate text files. Every blog entry comprised of name of the blog site, permalink of the blog post, author's name, title of the blog post, its text and user comments. In the third dataset we also stored the country (or region) information of the author. In the first dataset we collected blog posts on 'Women's Reservation in India'. We obtained a large number of posts written by male as well as female authors. The second dataset was a collection of blog posts on 'Regionalism'. We did not restrict the data collection to the posts originating from India only; still we obtained a good number of posts in Indian perspective.

The third dataset was a collection of blog posts on three terror events, namely '26/11/08 Mumbai hotel attack in India', 'the March 2009 twin terrorist attacks in Pakistan on Lahore hotel & on Srilankan cricket team', and the '9/11/01 attack on World Trade Centre in USA'. This data was originally collected in a previous experimental work [17]. These events were chosen due to their socio-political relevance, highly discussed nature, and the demographic & social variations of the three different societies in which they occurred. Our goal of analysis for the third dataset was a cross-cultural opinion mining task. It involved observing the variations of opinions and sentiments of bloggers from different demographic & social groups on the events of high social & political concern. The blog data in the third dataset was grouped in three categories, each corresponding to the three demographic areas in which these events took place. These three groups were termed as IND, WAC and USE corresponding to bloggers belonging to India, West Asian Countries and United States & Europe (West) respectively. Table 1 summarizes the scheme of grouping the collected data.

Table 1. Clustering the blog data of the third dataset into different groups

	Event 1	**Event 2**	**Event 3**
Blog Data	IND	IND	IND
Blog Data	WAC	WAC	WAC
Blog Data	USE	USE	USE

3.2 Preprocessing the Data

We have used the Vector Space Model to represent each blog post. Every blog post is represented in the form of a term vector. A term vector consists of the distinct terms appearing in a blog post and their relative weights. There are a number of ways to represent the term vectors. Commonly used ones are *tf*, *tf.idf* and *Boolean presence*. We used tf.idf measure, defined as $\text{tf-idf}_{t,d} = \text{tf}_{t,d}$ X idf_t, where *tf* is the term frequency and *idf* is the inverse document frequency. The vector V(d) derived from the document *d* thus contain one component for each distinct term. Once we have the entire set of posts represented as document vectors, their degree of similarity can also be computed using *cosine similarity* measure as in equation 7 below.

$$\text{Cosine Similarity } (d_1, d_2) = \{V(d_1).V(d_2)\} / \{|V(d_1)||V(d_2)|\} \quad (7)$$

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 thus length-normalizes the vectors $V(d_1)$ and $V(d_2)$ to unit vectors $v(d_1) = V(d_1) / |V(d_1)|$ and $v(d_2) = V(d_2) / |V(d_2)|$ respectively. The Cosine Similarity measure is used for clustering and classification tasks of text documents.

In addition to representing the blog posts as term vectors of tf.idf values, we preserved the original value and position of terms in the blog post for extracting suitable POS tags for opinion mining. We also performed stop word removal from the collected blog but no stemming was done. POS tagging refers to assigning a linguistic category (often termed as POS tag) to every term in the document based on its syntactic and morphological behavior. Common POS categories in English language are: noun, verb, adjective, adverb, pronoun, preposition, conjunction and interjection. We have used Penn Treebank POS Tags [12]. We extracted adjectives from the blog data for opinion mining. The selected adjectives were then used to compute a SO score using PMI value computations with reference words “excellent” and “poor” as in eq. 4 & 5. The opinion label assigned to a post was an aggregate of the SO values of the terms in the post.

3.3 Opinion Mining

We have used both a Naïve Bayes machine learning text classifier and the unsupervised semantic orientation approach for opinion classification. First we implemented the unsupervised semantic orientation approach to assign opinion labels

to blog posts in all the three datasets collected. As stated earlier, the SO label of a blog post is computed using an aggregation of SO values of the selected terms in the blog post. We tried two schemes of aggregation. In one scheme we associated '+1' for every term having SO value above a threshold (0.8 in most of the cases) and '-1' for every term having SO value below it. The SO value of a blog post is then computed by obtaining the sum of SO values of all extracted terms of that blog post. If the sum is positive (or greater than a '+ve' reference value say 2 - the one used in our work), the blog post is labeled as positive and negative otherwise. In the other scheme we simply added the SO values of all the extracted terms in a blog post and then divided the sum by the total number of extracted terms of that blog post. For example, if the sum of SO values of n extracted terms is x, then the aggregate SO value of that blog

Table 2. A snapshot of results for a subset of the first data set. The thresholds for two SO aggregation schemes are 0 and 0.70 respectively.

Title of the Blog	Mood	Gender	Aggregate SO Value	Semantic Orientation
Do women need reservation?	happy (70.1 %)	female (74.8 %)	+3	Positive
	upset (29.9 %)	male (25.2 %)	0.76653093	Positive
Reservation for Women: The	happy (54.7 %)	female (52.8 %)	-7	Negative
icing on the cake	upset (45.3 %)	male (47.2 %)	0.73098755	Positive
Reservation to power for Indian	happy (76.8 %)	female (64.9 %)	+4	Positive
women	upset (23.2 %)	male (35.1 %)	0.91197723	Positive
Reservation by custom and	happy (50.1 %)	female (65.0 %)	+12	Positive
tradition is acceptable	upset (49.9 %)	male (35.0 %)	0.9886512	Positive
Women's bill should lead on to	upset (81.4 %)	male (80.3 %)	-6	Negative
real political reform	happy (18.6 %)	female (19.7 %)	0.6081681	Negative

Table 3. A snapshot of results for a subset of the second data set. The thresholds for two SO aggregation schemes are 0 and 0.70 respectively.

Title of the Blog	Mood	Aggregate SO Value	Semantic Orientation
Is Your Region More Important	upset (66.8 %)	1	Positive
than your Nation?	happy (33.2 %)	0.76783	Positive
Is it legitimate to give J&K the	upset (81.2 %)	-2	Negative
status of a special state?	happy (18.8 %)	0.68771	Negative
The whole of India does not	upset (75.9 %)	-3	Negative
belong to all Indians	happy (24.1 %)	0.62420	Negative
Do we need to fear Regionalism?	upset (93.9 %)	+2	Positive
	happy (6.1 %)	0.68932	Negative
A Sense of where you are!	happy (74.1 %)	+10	Positive
	upset (25.9 %)	1.09861	Positive

post is x/n. This aggregate value is then compared with a threshold and the blog post is labeled as positive or negative accordingly. We also performed mood and gender analysis using an online tool uClassify [18]. A snapshot of the opinion label assignments of few blog posts of the first and second datasets is presented in tables 2 and 3. The tables show aggregate SO values using both aggregation schemes and resulting opinion label assignments (columns 4 and 5).

The second implementation for opinion mining used Naïve Bayes text classifier. We used two methods for training. In first method we used the manually assigned opinion labels and in the second scheme we used opinion label assignments generated by the SO-PMI-IR approach as training labels. We employed a three-fold classification scheme. The blog data is divided into three subsets, with one of the three subsets used for training and the remaining two subsets classified using Naïve Bayes run. This is done by taking different parts as the training set. Our implementation of Naïve Bayes used term frequencies and hence is a multinomial Naïve Bayes implementation. A snapshot of the Naïve Bayes run on the first and second datasets is presented in table 4. Table 5 presents a summary of accuracy of the opinion mining results of both the methods (SO-PMI-IR and Naïve Bayes). Table 6 presents the cross-validation of results of opinion label assignments for a sample of 50 blog posts on the first and second datasets using Naïve Bayes and SO-PMI-IR approaches.

Table 4. Classification results of a sample of 50 of blog posts of the first and second datasets. The Naïve Bayes results are on a training data size of 25 and run on 50 blog posts respectively. The SO approach result is on the same 50 blog post data.

50 Blog data subset on	**Number of blogs classified as positive**		**Number of blogs classified as negative**		**Precision, Recall and F-measure values (NB)**
	SO Approach	**NB Classifier**	**SO Approach**	**NB Classifier**	
Women's Reservation in India	36	42	14	08	Precision: 0.5147 Recall: 0.9138 F-measure: 0.8873
Regionalism	38	44	12	06	Precision: 0.5392 Recall: 0.9642 F-measure: 0.8981

Table 5. Classification accuracy of a sample subset of 50 blog posts of the first and the second blog datasets

50 Blog Data subset on	**Classification accuracy using first SO aggregation scheme**	**Classification accuracy using second SO aggregation scheme**	**Classification accuracy using Naïve Bayes Machine Learning approach**
Women's Reservation in India	84%	81%	69%
Regionalism	77%	76%	67%

Table 6. Cross-validation of opinion label assignments of a sample set of 50 blog posts of the first and the second datasets, by both the methods

50 Blog Data subset on	Number of matching opinion label assignments	Number of mismatches in opinion label assignments
Women's Reservation in India	40 (80%)	10 (20%)
Regionalism	42 (84%)	08 (16%)

The third dataset involved data based on three different events and hence was subjected to analysis along two different dimensions: *vertical* and *horizontal*. Vertical analysis involved comparison of the blog posts of IND, WAC and USE groups on one event (say 26/11 Mumbai attack event). This gives an analytical perspective of the reactions of the different demographic groups on a particular event. Horizontal analysis involved comparing the posts of a particular group along three different events. For example comparison of the blog posts of WAC group for all the three events. This would give the variation in opinion, sentiment and mood of a particular group along the three events that happened in three different regions. While vertical analysis could have important inferences about the difference in opinions of IND, WAC and USE groups on a particular event (say 26/11 Mumbai attack); horizontal analysis was expected to have implications about the variations of opinion of the same group about the three different terror events that occurred at different places (for example observing IND group's reaction on 26/11 Mumbai attack, Lahore bombing and 9/11 WTC attack). Table 7 shows the result of mood analysis based on the blog data of the three groups along 9/11 WTC attack event (vertical analysis). Similarly, the horizontal analysis showed varied opinions of the three groups on different events. While IND and USE group's opinions matched to some extent on the 26/11 Mumbai attack and 9/11 WTC attack, there was a slight degree of variation in case of Lahore twin terror attack event. WAC group on the other hand showed quite different opinion on the three different events considered.

Table 7. Mood analysis of the IND, WAC and USE groups on 9/11 WTC event in U.S.A.

Mood	Upset	Happy
IND	85%	15%
WAC	36.6%	63.4%
USE	97.6%	2.4%

3.4 Inferences

The opinion, mood and gender analysis results obtained on the three datasets present an interesting picture. For the first dataset we performed opinion, mood and gender analysis. The blog posts of the first dataset being centered around 'Women's Reservation in India', there were a good number of posts that expressed positive opinion. Moreover, the negative opinion blog posts were largely by male bloggers. The results of mood analysis were however not in perfect congruence with the opinion mining results, with some blog posts classified as positive by both Naïve Bayes and SO-PMI-IR approaches but attributed higher score of 'upset' on mood analysis. This variation in results may be due to the use of different reference words for opinion mining and mood analysis. Also the gender analysis results were not very much accurate for the first dataset, possibly because of the topic of discussion itself.

The second dataset on 'Regionalism' has been subjected to opinion mining and mood analysis. The results of opinion mining by both the techniques produced similar results. Though the accuracy of classification for this dataset was slightly less than that on the first dataset; it has more congruent results on opinion mining and mood analysis. We assumed gender neutrality in the second dataset. Most of the blog posts of this dataset were labeled positive by both the approaches. A cursory look at the contents of the blog show that most of the bloggers express their concern about regionalism but did not use strong negative words (which bloggers did in the posts on Women's reservation).

The vertical and horizontal analysis on the third dataset resulted in one of the most interesting findings. Our focus in this dataset was purely cross-cultural. We wanted to observe how opinions of bloggers vary when they write about three gruesome terror events that took place in three different regions (one being their own country/region). Analysis along vertical dimension show that by and large the IND and USE groups tend to agree on the same set of opinions for a particular event and this was true for all the three events. WAC group's reaction was varied and differed a lot across the three events, namely the 26/11 Mumbai attack event, 9/11 WTC attack event and the twin terror events in Lahore. The findings were also supported by the mood analysis along vertical dimension. While vertical analysis helped in understanding the variation of perceptions and opinions of bloggers from different region on a highly emotion-laden event; horizontal analysis showed opinion variations of the same blogger group across events in different societies. This analysis is more valuable since it uses more original data at almost negligible physical cost.

4 Discussion

The experimental work carried out by us on opinion mining from Weblogs produced interesting results. The results of both the techniques employed were accurate to a good degree. However, it would be worth stating that unsupervised semantic orientation scheme is a better choice for opinion mining (at least for Weblogs). There appears a clear reason for this. The semantic orientation tries to capture the subjectivity in the blog posts by identifying the positive and negative terms, and

labeling the blog post accordingly. The Naïve Bayes classifier, however, only uses term occurrence statistics to classify the blog posts into two groups. Since the terms present in the actual data may be quite different from the terms in the training data, it may result in more mis-classifications. This is a more severe problem in case of blog posts as different bloggers tend to use entirely different terms (possibly varied vocabulary). Hence, unsupervised semantic orientation approach may produce better results on opinion labeling of data from social media. However, it is more computationally challenging approach. The experimental work also demonstrates the applicability of opinion mining techniques on the blog data.

It is beyond doubt that the large amount of data in blogosphere is an extremely valuable source for commercial exploitation. What remains however relatively unexplored is to see how good is the huge amount of the data in the Blogosphere for cross-cultural psychological & sociological analysis. Blogosphere being a platform where people across the continent express themselves; we may use it for socio-political and cross-cultural psychological and sociological analysis. In this paper we have demonstrated one such effort using opinion and mood analysis on three different datasets. The preliminary results are interesting and worth further analysis (possibly by a social scientist). This experimental work demonstrates the use of opinion mining techniques for exploiting the blog data for sociologically relevant analysis. Advances in Natural Language Processing, Information Retrieval techniques for mining unstructured data will make this task more relevant and valuable.

References

1. Technorati Blogosphere Statistics (2008), http://technorati.com/blogging/state-of-the-blogosphere/
2. Agarwal, N., Liu, H.: Data Mining and Knowledge Discovery in Blogs. Morgan & Claypool Publishers (2010)
3. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002, Philadelphia, US, pp. 417–424 (2002)
4. Esuli, A., Sebastiani, F.: Determing the Semantic Orientation of Terms Through Gloss Analysis. In: 14th ACM International Conference on Information and Knowledge Management, CIKM 2005, Bremen, DE, pp. 617–624 (2005)
5. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1), 1–47 (2002)
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Conference on Empirical Methods in Natural Language Processing, Philadelphia, US, pp. 79–86 (2002)
7. Kim., S.M., Hovy, E.: Determining Sentiment of Opinions. In: COLING Conference, Geneva (2004)
8. Durant, K.T., Smith, M.D.: Mining Sentiment Classification from Political Web Logs. In: WEBKDD 2006. ACM Press, New York (2006)
9. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, Heidelberg (2008)

10. Dave, K., Lawerence, S., Pennock, D.: Mining the Peanut Gallery-Opinion Extraction and Semantic Classification of Product Reviews. In: 12th International World Wide Web Conference, pp. 519–528. ACM Press, New York (2003)
11. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
12. Penn Treebank Project, http://www.cis.upenn.edu/~treebank/home.html
13. Turney, P., Littman, M.L.: Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word corpus. NRC Publications Archive (2002)
14. Mishne, G., Rijke, M.D.: MoodViews: Tools for Blog Mood Analysis. In: AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI-CAAW 2006 (March 2006)
15. Balog, K., Rijke, M.D.: Decomposing Bloggers' Moods. In: 3rd Annual Workshop on the Web Blogging Ecosystem, At WWW 2006 (2006)
16. Mishne, G.: Experiments with Mood Classification in Blog Posts. In: 2005 Stylistic Analysis of Text for Information Access Conference (2005)
17. Singh, V.K.: Mining the Blogosphere for Sociological Inferences. In: Ranka, S., Banerjee, A., Biswas, K.K., Dua, S., Mishra, P., Moona, R., Poon, S.-H., Wang, C.-L., et al. (eds.) IC3 2010. Communications in Computer and Information Science, vol. 94, pp. 547–558. Springer, Heidelberg (2010)
18. Uclassify Mood Analysis Tool, http://www.uclassify.com/browse/prfekt/Mood (retrieved, April 2009)