International Journal of Advanced Research and Review

www.ijarr.org

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

Chevula Rekha, Mr. Merugu Anand Kumar, Lankala Mounika,

Dr. Godagala Madhava Rao

Assistant Professor[1,2,3], Professor[4]

rekhavenkat16@gmail.com[1], meruguanand502@gmail.com[2],
lankala.mounikareddy@gmail.com[3],madhavaraog175@gmail.com[4]

Department of CSE, A.M. Reddy Memorial College of Engineering and Technology, Petlurivaripalem, Narasaraopet, Andhra Pradesh

Abstract

Our preliminary research results on a unique combination of Topic Modelling, Named Entity Recognition, and Sentiment Classification for social analysis of blog data are shown here. More than five hundred blog entries on the topic of "discrimination, abuse, and crime against women" have been compiled. Here, we used topic discovery to using a combination of keyword analysis and a Named Entity Recognition method based on the 7-entity model, we were able to zero in on the most important topics and people covered in the blog postings. We then used SentiWordNet to classify all of the blog data as positive or negative depending on the prevailing tone. The findings produced are fascinating and provide strong evidence for the efficacy of our method for computational analysis of social media data. This paper's main contribution is the suggestion of a new Text Analytics combination and the subsequent demonstration of its usefulness for computational investigation of the data gleaned from social media for sociological research.

1. Introduction

The unifying thread and impetus behind Time magazine's progression from naming "You" as person of the year in 2006 to "The Protester" as person of the year in 2011 is social networking. In 2006, social media platforms began to gain popularity and quickly amass a large user base; by 2011, these platforms had utilized for the unprecedented coordination of demonstrations in nations ranging from Libya to Tunisia. An important aspect of this cultural shift is the rise of the blogosphere. A blog, sometimes known as a weblog, is a website where one or more people may make public postings about their own experiences and opinions.

Personal blogs function similarly to online diaries, while community blogs serve as online discussion boards and teamwork hubs. In addition to text, photos, and links to other media are common in blog posts. The term "blogosphere" is often used to refer to the whole network of blogs. The popularity of blogs has increased at an unprecedented pace. Over the course of only seven years, the number of blogs monitored by Technocratic increased from 4 million in September 2004 to 164 million in July 2011 (Technocratic Statistics, 2013). Word press has over 383 million users and over 3.5 billion monthly blog page views, according to the most current figures provided by Word press (Word press Statistics, 2013). It is estimated that there are 33.9 million new blog posts and 40.9 million new comments made each month alone by Word press users. The blogosphere has

grown into a vast archive of articles covering a wide range of subjects, and its scope and scale continue to expand daily. About 66% of blogs are written in English

Right now. Increases in Internet access in developing regions will lead to a corresponding rise in the quantity and diversity of blog entries, both in English and locally spoken languages. Four out of five people who have access to the internet nowadays utilize some kind of social media, lending credence to this trend (including blogs). Looking at the user profile data adds further intrigue to the massive amount of blog posts and comments made on blogging sites. Blogging.org claims that just around 40% of bloggers are professional writers (Info graphic, 2012). They are not compelled to write by financial or career concerns, but do it of their own own in relation to a wide range of topics, from politics and religion to social issues. While the low barrier to entry, open standards of content generation, and free-form writing style are all contributing factors to logging's rising popularity, it is the need to get one's thoughts out into the world and the availability of a suitable platform that have propelled the blogosphere to its current size and scope. The blogosphere is a rich and unique treasure trove for cross-cultural psychological and sociological study, which has been largely untapped until recently. For this purpose, we provide here the results of our computational investigation into the blog text data. As stated in Section 2, this work's inspiration is discussed. The section 3 describes the computational formulation mix that we used. Section 4 describes the datasets and their characteristics, whereas Section 5 gives the experimental design and data. Section 6 of the study provides an overview of the findings.

2. Motivation

The blogosphere is a treasure trove for more than just economic gain; its explosive development and massive volumes of data (mainly textual) make it ideal for studying society and politics. There are two major considerations that strengthen the validity of this statement: (a) free-form, unfiltered, first-hand, and relatively more emotionally laden expressions of various people on various social, political, cultural issues; and (b) the fact that the Internet has reduced the distance between people all over the world and allowed them to express themselves and interact with others regardless of geographical, demographic, religious, or cultural boundaries. Web logs today provide a wealth of information from a wide range of people from different cultures and political perspectives on a wide range of subjects and events. In recent years, academics from a variety of fields have begun investigating the blogosphere for purely academic purposes. There are two main themes in this body of analysis. Finding notable bloggers and blog sites about an event (Agarwal et al., 2008; Mahatma and Agarwal, 2012); community discovery; screening spam blogs; etc. (Liu et al., 2010; Agarwal and Liu, 2008) are examples of the more computer science-oriented activities that fall under this category. The second variety focuses more on the political and social context of blog postings (Singh et al., 2012), (Singh et al., 2010). (Singh, 2010). Tasks like these include mapping the blogosphere around a specific political or social event (Mehrav et al., 2012) and analyzing blog posts that are pertinent to a significant event, person, organization, or process (Moe, 2011), (Suhara et al., 2007), (Adamic and Glanse, 2005), (Lin and Halavais, 2004). Our method is a socio-political analysis of relevant blog articles on gender-based job discrimination and violent crime against women. Our primary goal was to investigate the many entities (individuals, groups, and institutions) that were key to the topic at hand, as well as to isolate major concerns related to the subject and get insight into the bloggers' overall perspective on the matter. Blog posts were selected because they provide the most authentic, first-hand account of people's feelings, ideas, and opinions from all around the globe.

3. Computational Formulations

For this blog text analysis, we employed a new combination of Topic Modelling, Named Entity Recognition, and Sentiment Classifier. In this article, we provide a concise summary of how these three computational formulations are often understood and used in our lab.

Topic Modelling

When applied to a large set of texts, Topic Modelling (also known as topic discovery) attempts to extract the underlying themes by way of semantic annotation. The system uses a suite of statistical techniques to examine the text documents in a corpus, including the language usage data to link related publications. It analyzes the text documents using a probabilistic model based on hierarchical Bayesian analysis (Blei, 2012). In addition to locating topics of interest, topic modelling may be used to investigate the relationships between them and, perhaps, how they have evolved over time. Latent Dirichlet Allocation is the most elementary topic model.

LDA, or latent Dirichlet allocation, is a generative statistical approach (Blei et al., 2003). (An imaginary process by which the model assumes that documents are generated by the topics). In formal definitions, a subject is modelled as a distribution over a constant set of words. The core tenet of LDA is, thus, to represent documents as emerging from various topics, and more precisely, to suppose that certain k themes are linked with the documents collection, and that each document shows these topics in varying amounts. This means that there is a consistent set of subjects throughout the whole collection, even if the relative emphasis of those themes varies from page to document. The thematic organization of the documents is reflected in the hidden topic structure determined by an effective topic modelling technique. Other forms of Topic Models include the Bayesian Non-parametric Topic Model, the Dynamic Topic Model, and the Correlated Topic Model (Blei and Lafferty, 2009).

Sentiment Classification

The goal of sentiment analysis is to categorize all expressions of opinion as either "positive" or "negative." When it comes to determining how an author feels about a certain topic, three main methods have emerged: (a) a text classifier based on machine learning techniques, such as Naive Bayes, Support Vector Machines, or k-Nearest Neighbours; (b) a Semantic Orientation scheme. Of via (a) identifying important n-grams in the text, (b) analyzing those n-grams, and (c) using the SentiWordNet-based public library that offers positive, negative, and neutral ratings for words. In this study, we use the SentiWordNet method to categorize the emotional tone of blog posts. The key reasons for selecting this method were its simplicity, ease of use, and high levels of accuracy achieved with little to no training data. The SentiWordNet method utilizes the SentiWordNet library, which is open to the public (SentiWordNet, 2012). We must first extract pertinent opinionated phrases and then query for their scores in SentiWordNet before we can utilize it. Adjectives, adverb adjective, and adjective verb combinations have all been proven to be good candidates for the phrases to be retrieved in previous research. We've implemented a basic version of the SentiWordNet method, which uses SentiWordNet scores to determine the quality of adverb+adjective combinations. There is considerable consideration given to adverb scores in the computation of these ratings with adjective scores (Singh et al., 2013). The SentiWordNet ratings of adjectives are changed when adverbs come before them. We also looked for the word "not," and if we found it, we reversed the sign of the SentiWordNet score value of the preceding phrase. The overall value of the retrieved 'Adv+Adj' combination determines whether a sentence is positive or negative.

4. Dataset Collection

We have compiled blog posts on the topic of "Discrimination, Harassment, and Crime against Women." In the month of June 2012, we collected a total of 512 blog articles from widely read sites including Word press, BlogSpot, The word, Feminist blog, and Blogher. We have relied only on an automated system for Information gathering. We developed search client software that makes use of the Google Search API to get the URL link of blog texts that match our supplied query from the aforementioned blog sites. To keep track of all the places you may go online, we built a database that will remember them. In the second step, we ran a JAVA application to get the full contents of the blog entries linked to in our data warehouse. All of the information is saved in xml. For each blog post, we save the following xml tags:

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

We fed the crawler the seven hand-coded search queries listed below to locate applicable weblog posts: "discriminating weaker sex at workplaces," "sexual harassment of women at workplaces," "unfavourable conditions for women at workplaces," "sexual abuse and discrimination of women," and "preventing sexual abuse and discrimination of women." issues of sexism, misogyny, and violence against women; mostly affecting women in the developing nations. As you can see from the questions we asked, we aimed to find blog entries that were extremely relevant and somewhat factual, controversial, or opinionated. After filtering out articles with insufficient data or posts written in languages other than English, we were left with a dataset consisting of 485 blog entries from a total of 512 posts gathered on this subject. Table 1 below provides an overview of the blog entries that were gathered (before to filtering) from the various sites.

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 details the methodology behind a computational formulation's execution, including its purpose and the results it produces. We created unified java software that can model topics, convert text to a vector space model, tag texts for parts of speech, and identify named entities in text. Sentiment analysis of documents with the help of the SentiWordNet database. Topic modelling was the first thing we performed with the complete set of blog entries, and we used Stanford's Topic Modelling Toolbox to do it (Stanford Topic Modelling, 2012). The primary reason we used topic modelling was to identify common threads within the blog data. This would be useful for two reasons: first, it would capture the most typical terms one may anticipate to find in books and texts on this topic, and second, it would assist identify the important themes stated by bloggers. After sifting through the top 50 keywords obtained from the topic modelling result, we settled on the top 20 most relevant keywords. After narrowing the sample down to 20 keywords, we calculated how often those terms appeared overall. Table 2 shows the most often used 20 keywords and the number of times each term appears in the text. Although several of the most frequent terms were also included in the manually coded search queries (such as "work," "sexual," "violence," "harassment," "discrimination," and "women"), we also received quite distinct top keywords. When evaluating the usefulness of the gathered blog data in light of the subject of study, the degree to which its contents coincide with our queries is a good indicator of its usefulness. Words like "men," "society," "law," "years," "life," and "state" were also among the most sought-after phrases.

There is a clear pattern here, which serves as an indication of the important ideas and characters discussed in the literature on this topic. While many texts place blame on "guys," others claim that "social structures" should help address the issue of discrimination and harassment. "State" and "law" to find answers to the issues. Through the computational method used here, we can see that other authors have also pointed out that discrimination against women and harassment in all its manifestations has been an issue for many years, and that it manifests differently for different women at different times in their lives. In order to better visualize the most important topics, people, institutions, etc., represented via keywords, we plotted a tag cloud of the top keywords gathered. This graph is shown in Figure 1(a) and was created using the most popular keyword list and a graph data file (gdf) for usage in Gephi (Gephi, 2012). The tag cloud plot shows the relative size of terms based on how often they appear; more often used words are larger in the plot.

Refer to Table 2 for a listing of the 20 most often used terms.

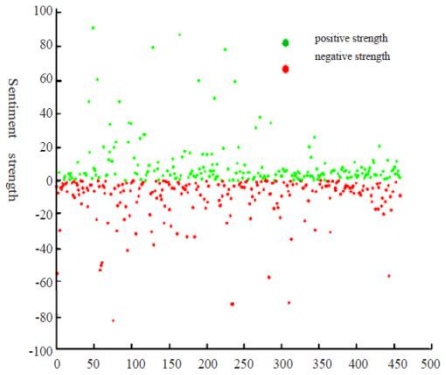
Word 1-10	Count	Word 11-20	Count
Work	1365	make	688
Sexual	1205	law	687
Men	1128	woman	681
Time	1015	workers	620
discrimination	946	state	611
Rights	910	gender	607
Violence	869	life	582
harassment	792	don	568
World	758	labor	568
Social	709	years	567

Fig. 1. Tag cloud plot of noun occurrences (b) Tag cloud visualization of the top 50 keywords Having completed the topic modelling, we figured it would be instructive to examine the frequency with which certain nouns occurred throughout the full dataset. With this in mind, we POS-tagged every single blog article and the nouns that were identified by the tag were extracted. Following noun extraction, we measured how often each noun occurred throughout the whole dataset, and then used those counts to generate the tag cloud plot in figure 1(b). Using a noun extraction and frequency plot, we can see which entities in the dataset are most essential to understanding the topic at hand. Women, people, rights, children, people, society, community, men, employer, school, court, state, management, etc. are some of the most prominent nouns seen throughout the story. Every key person, organization, and entity relevant to the topic at hand is therefore placed in context by this plot. Next, we used Named Entity Recognition to better depict the major players in the blog data (NER). When we used the Stanford 7-entity NER model, we were able to compile a comprehensive set of names for nouns belonging to the classes "person," "place," and "organization," among others. Then, we pulled out every person, place, and thing that fit those three criteria.

In conclusion, NER identified 2012 unique people, 907 unique places, and 2722 unique organizations throughout the whole dataset. Plots of tag clouds for the person class (figure 2(a)), the place class (figure 2(b)), and the organization class (figure 2(c)) are provided to illustrate the distribution of terms associated with these entities. Relative frequency in the whole data collection. We have used only entities with a low occurrence threshold to create these charts (4 for person class, 8 for location class and 7 for organization class). Names like Obama, Bush, Clinton, Melinda Gates, Stacy, and Mr. Murray stand out in the tag cloud for the person class. Bloggers on the topic most often cite these authors. Possible explanations include the fact that they are the individuals who have been pushing for policy changes on this issue, and/or that people are putting their faith in them to end the pervasive problem of discrimination against women. In a similar vein, the location cloud has several outsized nouns, such as "America," "Canada," and "Great Britain," as well as "Singapore," "Australia," and even "India," "China," and "Nigeria." As a result, it's easy to deduce that this is a universal issue, present in almost every society and country, although at varying degrees of severity. Another possible interpretation is that the vast majority of nations are making an effort to deal with the problem by passing laws, raising public awareness, and taking other forms of positive action. Several notable nations, governments, corporations, academic institutions, and private citizens are represented in the tag cloud plot of entities belonging to the class "organization."

6. Conclusion

We have created a computational framework for exploratory analysis of blog data on a certain topic, and it has yielded some really intriguing and useful findings. The introduction of topic modelling allows us to identify the most important thematic keywords across all of the available blog material. These topical terms represent significant ideas, individuals, and organizations discussed in works grouped under the rubric of "discrimination and harassment of women." In order to better comment on the primary problems/institutions/entities associated with this subject, the dataset has been POS-tagged to assist identify the nouns that appear within it. The NER implementation provides a higher degree of entity identification by enabling the extraction of named people, places, and organizations. Important people, places, and organizations mentioned repeatedly or determined to be strongly tied to the problem may be isolated from the corpus of literature on the topic. Also, the sentiment analysis findings demonstrate that the whole dataset is rather well distributed on both the "positive" and "negative" sentiment scales, indicating that the articles written about this topic are not entirely pessimistic and pessimistic. When combined with an entity-based sentiment analysis, the results become even more laser-focused, allowing researchers to get a sense of how people feel about each of the most prominent things mentioned in the dataset.



Blog data

Figure 3: A sentiment polarity strength constellation derived from blog data

We presented a computational framework for tackling this analytical challenge, one that makes use of a trisect of techniques drawn from Topic Modelling, Network Analysis, and Knowledge Graphs, all of which are very applicable to the analysis we did. Natural Language Processing and Emotional Insights. Although this method does not intend to replace the standard subjective analysis, it does provide certain benefits over that method. To begin, our computational formulation inherently incorporates a cross-cultural and demographic viewpoint on the topic by collecting pertinent writings written by individuals around the globe. Second, there is no limit on the volume of data we can process in a reasonable length of time. It takes a lot more time and effort to analyze this volume of data manually. Finally, this formulation may be used to determine the relative strength of various themes over an entire corpus of texts. Thus, this computational formulation offers a novel framework for the automatic analysis of text documents, with much less time and effort required compared to conventional subjective methods, and which inherently provides for a sociological and socio-political analysis along any theme or issue of interest, regardless of cultural background. The insights might also serve as inspiration for further introspection on the topic.

References

[1] Technocratic & 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- info graphic/ 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] Mahatma 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.): Contemporaryn 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.