Blog Text Analysis Using Topic Modeling, Named Entity Recognition and Sentiment Classifier Combine

Pranav Waila*, V.K. Singh** and M. K. Singh*
* DST-CIMS, Banaras Hindu University, Varanasi, India
** Department of Computer Science, South Asian University, New Delhi, India
pranav.waila@gmail.com, vivek@cs.sau.ac.in, mks_kjist@yahoo.co.in

***Abstract*- This paper describes our experimental work on computational analysis of socio-political blog data through a novel combine of sophisticated language processing and visualization techniques. We have designed an integrated framework by utilizing Topic Modeling, Entity Extraction and Sentiment Analysis; to draw sociologically relevant inferences from unstructured free form blogosphere data. The dataset comprised of more than 9290 blog posts on social-political events related to the Arab spring. We have tried to extract important inferences from the dataset; such as key themes, persons, places, organizations and overall sentiment orientation of the content around different entities in the texts. We have tried to validate the inferences obtained through manual and Google search trends statistics. The results obtained are quite relevant and demonstrate the usefulness of our approach for computational analysis of social media data.**

Keywords— Blogosphere; Information Extraction; Sentiment Analysis; Social Media Analytics; Social Computing.

I. INTRODUCTION

Few years ago blogosphere and social media were the hot topics for discussions. With the technological extravaganza in social media and common man's involvement, now blogosphere is a fruitful data repository for understanding customer preferences and attitudes. These weblogs and user interaction are being analyzed by top commercial companies, to understand customer's choices, their economic behaviour and identify market trends. Recent disclosure of controversial prism project shows, how governments are utilizing web content for electronic surveillance or spying. Blogosphere is constantly changing its shape and structure; especially at the time of big political or social movements a huge amount of growth in blog data is witnessed.

Blogs are one of the most powerful and trusted text media channels available in recent time, which provides platform to common man and allow to independently express his/her opinion without any discrimination. Blogging platforms like blogger, blogspot, worldpress with easy maintainable content management systems have provided freedom to users to express themselves. Such free and easy publishable blogging services have resulted in large amount of content being created and posted in the blogosphere. The blog tracking company *Technorati* tracked about 4 million blogs in September 2004, which has grown to about 164 million blogs in July 2011[1], an increase of 41 times in just a period of seven years. A recent statistics by Wordpress [2] (the top ranking blogging site) reports that it has a user base of more than 383 million people, with 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. People write about variety of topics on blogs ranging from technical discussions to socio-political events and institutions. In 2011 and 2012 Social media was used to coordinate and manage sociopolitical protests against governments in many countries. People were using blog platforms to control and coordinate movements which were never seen before in the history. It is this motivation, which prompted us to take up this task of computational analysis of blog data about a set of socio-political events. Now blogs are considered as most trusted and original data source which is full of emotions and uninhibited expression of people's views. It is evident from the fact that more than 60% of the bloggers are hobbyists, with no commercial motive behind their writings [3]. Further, the speed of content creation in blogosphere, makes it most up-to-date and a true repository of immediate reaction of people about an event.

Thus Blogosphere analysis is now an important and interesting interdisciplinary research area. For blog analysis language processing, machine learning and pattern recognition techniques are utilized to discover hidden information patterns, identify entities and other important aspects. This paper presents our experimental work on computational analysis of blog data on socio-political events. We have designed an integrated framework for the purpose of this analysis. The rest of the paper is organized as follows. Section II describes the background and the motivation. Section III describes the computational formulation. Section IV explains the dataset and the section V presents the experimental setup and results. The paper concludes with a summary of observations illustrated in section VI. The key contribution of this paper is to propose a novel integrated text analytics framework for content-based analysis of socio-political blog data. The framework has been tested with sufficiently large amount of blog data and thus demonstrated the applicability and usefulness of our work.

II. MOTIVATION

Most of the early work on blogosphere analysis has been aimed for a commercial exploitation. Whether it be the much talked about strategy- to get early feedback about the newly released operating system Vista by Microsoft- by contacting influential bloggers in early 2007 and persuading them to share their experiences in return of Laptops as free gifts, or tacit marketing strategies by a number of companies; blogosphere is now a widely acknowledged platform for commercial exploitation. However, 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. The unprecedented rate of growth of blogging and the huge

978-1-4673-6217-7/13/$31.00 ©2013 IEEE

amount of data contained in the blogosphere is a unique treasure not only for commercial exploitation but also for socio-political analysis. 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 [4] and blog sites about an event [5], community discovery, filtering spam blogs etc. [6], [7]. The other flavour is oriented more towards socio-political analysis of blog posts [8], [9], [10]. This includes tasks like mapping the blogosphere around a particular socio-political event [11], analysis of blog posts relevant to an important event/ personality/ organization or process [12], [13], [14], [15]. Our analytical approach is a socio-political inference oriented approach. We have focused on exploring 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.

III. COMPUTATIONAL FORMULATION

We have evaluated the socio-political blog data with an integrated computational formulation, which combines topic modeling, named entity recognition and sentiment classifier. Each of these tasks is very briefly described in the following paragraphs.

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 some specific topic. Most basic topic modelers apply clustering algorithms for topic detection. Our topic modeler employs a set of statistical methods which analyzes the words of the text documents in a collection, and uses the information of word usage patterns and connects all documents with similar patterns. It uses a probabilistic model based on hierarchical Bayesian analysis of the text documents [16]. Now topic modeling are not only used for discovering themes, but also to find out how these themes are connected to each other and how they change over time. The Latent Dirichlet Allocation (LDA) based topic modeler [17] assumes documents as mixture of multiple topics. More specifically, it assumes 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. The Bayesian Non-parametric Topic Model, Dynamic Topic Model, and the Correlated Topic Model are few other variants of Topic Models [18].

Named Entity Recognition (NER) is a popular technique of information retrieval which automatically annotates the entities of textual document. This can be seen as a classification task or well known tagging problem. Generally Named Entity Recognizers are supervised classifiers. There are various well established approaches available for named entity recognition such as training corpus based statistical technique which identifies words based on its neighboring keywords, or rule based approach which works strictly on defined rules [19], [20]. Some standard well known NERC libraries are available with good accuracy to identify entities such as person, organization, date-time, location etc. We have utilized alchemy web-service [21] for recognizing important entities in each post.

Sentiment classification is a task of identifying sentiment scores for text. Sentiment classification can be done at various levels in case of text, namely document-level, aspect-level and entity-level. There are various approaches for the sentiment classification ranging from machine learning classifiers to lexicon based methods. SentiWordNet based approach is one of the well-known semi-supervised approaches based on SentiWordNet lexicon [22]. This approach targets selected feature occurrences and associates the sentiment polarities based on their score in the dictionary. We have designed a SentiWordNet based algorithmic formulation, described in detail in [23] and [24], for computing sentiment polarity of the important entities occurring in the in the blog data. For this task, we have first calculated frequency and the polarity of the entities associated in the blog posts and then aggregated the polarity of these entities by adding the sentiment scores.

In order to depict the results obtained in an easy to understand format, we have utilized visualization tools. Information visualization is a task of representing some raw data in an informative visual form. In this work we used various techniques such as tag cloud plots using Gephi [25] and plotting graphs to represent topic proportion information. Entities are represented as the tags, whereas frequency of an entity determines its size in the tag cloud plot drawn. The sentiment around an entity is represented by depicting the entities in two different colors to mark them as entities having 'positive' or 'negative' orientation. We have used the full range of color spectrum of 'Red' and 'Green' colors to show the strength of sentiment polarities around various entities described in the blog data.

IV. DATASET

We have applied our algorithmic formulation on a sufficiently large dataset collected originally for work reported in [5]. The dataset consists of total 9290 blog posts on the broader theme of 'Arab Spring' having blogs written on various revolutions held in and around 2011 and 2012 in Arabic region to mark fight for democracy. Most of this protest has been managed and coordinated though social media. The dataset contains of a total of 7343883 words and average word count of a blog is 790.5148547. Dataset contains 3 categories of blog posts with the following statistics. The main motivation behind using this dataset is to identify and measure how representative the social media data is about the actual socio-political events occurring in our real world.

TABLE I. DATASET

	No. of blog posts	*Word count*	*Average word count*
Egyptian Revolution	5799	4228360	729.1533
Libyan Revolution	2271	2032937	895.1726
Tunisian Revolution	1220	1082586	887.3656

V. EXPERIMENTAL WORK AND RESULTS

We extracted topics on whole dataset to identify hidden underlying theme in the entire collection of the blog data. For doing this operation we utilized Stanford Topic Modeling Toolkit based on Conditional Random Field approach classifier. We did the LDA based topic modeling with 1000 iteration. Initially we executed topic modeler with different number of topics to find the topic perplexity. Number of topics with high perplexity is better candidate for topic selection. In table II, the perplexity values are shown.

TABLE II. TOPIC PERPLEXITY

Number of Topic	*Perplexity*
5 topic	4877.675154561823
10 topic	4582.594127526477
15 topic	4354.035194905415
20 topic	4175.618282787014

Based on inputs obtained from the topic perplexity value, we executed topic model with 5 topics and recorded the top 50 words, as the representative term profile, for each of the 05 topics. A sample list of keywords for the 05 topic themes observed is given in table III below.

TABLE III. TOP TOPIC REPRESENTING KEYWORDS

Serial	*Topic keywords*
Topic 0.	World, Political, Protest, Democracy, Movement
Topic 1.	Gaddafi, Serian, Libyan, War, Government
Topic 2.	Militry, Cairo, Protesters, Power, Citizens.
Topic 3.	Time, Her, She, Life, Revolution, behind, war.
Topic 4.	Die, Death, War, 2011, 2012, Revolution, Man.

After pursuing the top 50 keywords recorded for each of the 05 topic themes, we assigned thematic names to each of the 05 topics. The table IV below shows the assigned topic theme labels. Since the blog dataset is on socio-political movements in Egypt, Libya and Tunisia, identified topics are closely congruent to the major issues, role players and the events associated with the three revolutions, identified as 'Arab Spring'. The topic themes like protest for democracy, military against citizens, war against government; all of them are representative of socio-political scenario of the said issue.

TABLE IV. TOPICS IN THE COMBINED DATASET

serial	*Topic*
Topic 0.	Protest for democracy.
Topic 1.	War against government.
Topic 2.	Military against citizens.
Topic 3.	Women and war.
Topic 4.	Deaths in social protest (2011 and 2012).

In the figure 1, we have shown the topic distribution over documents, i.e. how closely related the themes are in the blog data corresponding to the three events. We can clearly see that the topics 'military against citizens' and 'women and war' are common across the three sets of blog data. For the data on 'Egyptian revolution', these are the most prevalent themes, which is what has been very widely reported in other media as well. For the dataset on 'Libyan revolution', 'military against citizens' is a prevalent theme, a correct depiction as known from the news stories from other media sources about the actual events that happened in Libya.

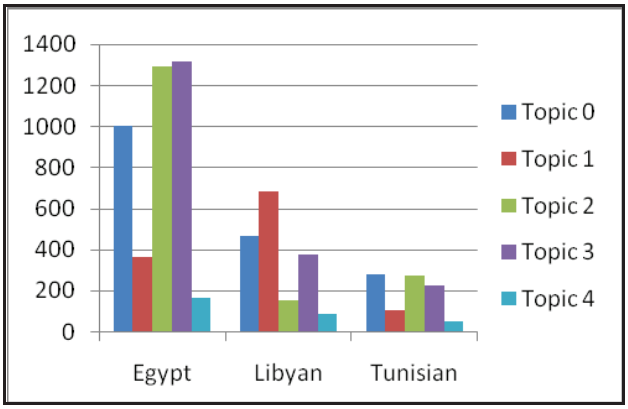

Fig. 1 Topic distribution over documents

In order to have more confidence in the results obtained by the topic modeling exercise and also to validate the results, we explored Google trend with the given topics to validate our results. Google Trend visualizes people's search interest over time. The figure 2 below plots the graph for search trend as recorded by Google Trend. The graph shows high search trend during the concerned period of these events. It starts taking higher value from late 2010 and continues till the recent time. The topic 'war against government' seems to have the highest search trend value in the graph followed by the topic 'military against citizens'. The topics 'protest for democracy' and 'women and war' also record many surges during this period. The search trend figures thus strengthen our confidence in the topic modeling results obtained.

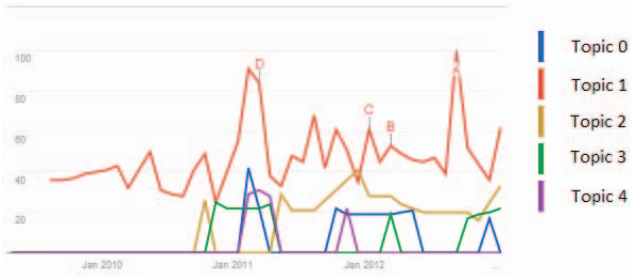

Fig. 2 Topic search trend

The next analytical task that we performed was identification of named entities referred to in the blog data. The main goal for this task has been to obtain a deeper understanding of the role players, places and organizations associated with the three socio-political events, for which we

have the blog dataset. The key entities observed in the full blog dataset are extracted and identified as belonging to different classes, such as person, place and organization. We have also obtained the sentiment polarity of the text written about a particular entity in the blog dataset. The sentiment polarities for a particular named entity from different blog posts are aggregated to have an aggregate sentiment polarity value around the entity. The tag cloud plots have some words represented in larger size as against others. The size of the words in the tag cloud plots is a measure of their relative frequency of occurrence in the blog dataset. The identified entities are shown in shades of either red or green, depending upon the fact that whether the sentiment polarity around an entity is 'negative' or 'positive', respectively. The figure 3 below plots the entities of the person class as observed in the dataset.

Fig. 3 Tag Cloud of person class entities in the full dataset

As is seen in the figure 3 above, the person class entities 'Muammar Gaddafi', 'Hosni Mubarak', and 'president bashar al-assad' are shown prominently and with negative sentiment. This is a well understood result as these are the main persons related to the three revolutions and they played a negative oppressive role. We also see that 'ban ki-moon' and 'president obama' are marked as red. This may be due the reason that bloggers see them as persons who could have helped the oppressed people of these countries but who, in their view, may not have acted as per their expectation. People like 'wael ghonim', an Internet activist and computer engineer and 'mohamed albaradei' are shown in green, expressing 'positive' sentiment around them shown by the bloggers. The size and color assignment for person class entities are quite accurate.

The figure 4 shows the tag cloud plot of important organizations mentioned by the bloggers in their posts. These organizations are also color coded to depict the sentiment polarity associated with them as expressed by the bloggers. Many social networking and media sites also figure as organizations mentioned by the bloggers in the tag cloud plot. We can clearly see that most of the social networking websites like 'facebook', 'youtube', 'flicker' and 'google' are shown to have 'positive' sentiment. This is primarily due to the reason that social networking sites have been the only platform that people used to raise their voice, mark their protest and express their anger against the oppressive regimes in this region. The other forms of media have been largely censored by these governments. Most other entities are shown to be close to 'neutral' and some tv channels shown as of 'negative' class.

Fig. 4 Tag Cloud of organization class entities in the full dataset

The figures 5 and 6 show the place class of entities, with figure 5 depicting the places for the full dataset and the figure 6 for places in the dataset on 'Egyptian revolution'. The capital cities of the region are shown to have negative sentiment, a measure of peoples association of these cities to the oppressive governments and expressed negativity around these cities. Most of the other cities are shown to have sentiment polarity close to the 'neutral' class. Some cities shown in 'grey' shades are the ones where people have been largely vocal and supportive of the people's right in the region. In figure 6, 'Cairo' and 'Alexandria' figure prominently and in red color, a true depiction as these are the two cities most affected and having seen violent activities. We have also recorded the occurrence of the continent references in the dataset, shown in table V. Interestingly 'Europe' and 'North America' are seen as having negative polarity as opposed to others, a measure of expectations of the affected people from these regions.

Fig. 5 Tag Cloud of place class entities in the full dataset

Fig. 6 Tag Cloud of place class entities in the Egyptian revolution dataset

TABLE V. CONTINENTS WITH FREQUENCY AND SENTIMENT SCORE

Continent	*Frequency*	*Sentiment score*
Asia	101	-.99845
Antartica	2	-0.02651
Europe	660	-3.29104
Latin America	73	.392666
North america	78	-2.68954
South America	46	.325686
Sub-Sahara Africa	2	.118754

VI. OBSERVATIONS

The computational framework that we have designed for the exploratory analysis of the blog data on a particular theme has been able to obtain very interesting and relevant results. The analytical task that we undertook is on a very important set of socio-political events of the contemporary world and we demonstrated an algorithmic setup to approach this analytical task through a computational formulation that uses a combine of Topic Modeling, NER and Sentiment classification. Through the topic modeling implementation we are able to identify the major thematic keywords from the entire blog data collection. These topic oriented keywords depict the major issues, role players, places and entities associated with the blog data on themes such as "Protest for democracy", "War against government", "Military against citizens", "Women and war" and "Deaths in social protest". 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. The results of sentiment analysis present a further level of insight into the blog data. It shows the sentiment of writings of bloggers about different entities (persons, places and organizations) across the full and individual blog data sets. We can thus observe and infer which entities are associated with a 'positive' sentiment and which with a 'negative' sentiment by the bloggers. The entity-based sentiment analysis obtains the sentiment orientation on all major entities talked about in the dataset. The sentiment polarity results are shown using an easy to understand color coded scheme, depicting the strength of sentiment polarities.

The computational analysis approach proposed by us 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. Further, we are able to capture the overall mood of the society (represented by bloggers) towards various issues and aspects around socio-political events. 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.

We seek to extend this work further by transforming it into a generalized computational analyzer of social media text. We are working towards a full integrated setup, which is easy to use by social scientists, performs all the analytical tasks without requiring user intervention, and which allows a social scientist to tune in various analysis parameters. Our aim is to design a system which can work as an exploratory tool for a social scientist interested in analysis of any socio-political event or phenomena. The social scientist will decide on what socio-political event or phenomenon s/he has to collect the blog data and then what kind of analytical tasks s/he would be interested in. We hope to introduce a temporal tracking capability in the system as well, where an event or phenomenon analytics could be mapped over a desired period of time. Such a tool will no doubt be extremely useful for performing a quick analysis of the social media data in a very short span of time and without much effort required. The exploratory results can then be subjected to and/or validated through a detailed analysis.

REFERENCES

[1] Technorati & Blogpulse Blogging Statistics, Retrieved from http://www.socialmediaexaminer.com/tag/blogging-statistics/ on 15 Jan., 2013.

[2] Wordpress Blogging Statistics, Retrieved from en.wordpress.com/stats/ on 15 Jan., 2013.

[3] Blogging Stats 2012 (Infographic), Retrieved from http://blogging.org/blog/blogging-stats-2012-infographic/ on 17 Jan., 2013.

[4] N. Agarwal, H. Liu, L. Tang, and P.S. Yu, "Identifying the Influential Bloggers in a Community", In proceedings of International Conference on Web Search and Web Data Mining, pp. 207--218. ACM Press, Palo Alto, U.S.A.. 2008

[5] D. Mahata and N. Agarwal, "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] H. Liu, P.S. Yu, N.Agarwal and T. Suel, "Social Computing in the Blogosphere", IEEE Internet Computing, April 2010, pp. 12-14.

[7] N. Agarwal and H. Liu, "Blogosphere: Research Issues, Tools and Applications", SIGKDD Explorations, Vol. 10, No.1, pp. 18-31, 2008.

[8] V. K. Singh, M. Mukherjee, G. K. Mehta, N. Tiwari and S. Garg, "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] V.K. Singh, D. Mahata and R. Adhikari, "Mining the Blogosphere from a Socio-political Perspective", In Proceedings of the 6th International Conference on Computer Information Systems and Industrial Management, IEEE press, Nov. 2010, pp. 365-370.

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

[11] Y. Mehrav, F. Mesquita, D. Barbosa, W.G. Yee and O. Fireder, "Extracting Information Networks from the Blogosphere", ACM Transactions on the Web, Vol. 6, No. 3, September 2012.

[12] H. Moe, “Mapping the Norwegian Blogosphere: Mthodological Challenges in Internationalizing Internet Research”, Social Science Computer Review 29(3) 313-326, 2011.

[13] Y. Suhara, H. Toda and A. Sakurai, “event Mining from the Blogosphere using Topic Words”, Proceedings of ICWSM, 2007

[14] L. Adamic and N. Glanse, “The Political Bolosphere and the 2004 US Election: Divided they Blog”, Proceedings of 3rd International Workshop on Link Discovery, ACM, 2005.

[15] J. Lin and A. Halavais, “Mapping the Blogosphere in America”, In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004

[16] D. Blei, “Probabilistic Topic Models”, Communications of the ACM, 55(4), pp.77–84, 2012.

[17] D. Blei, A. Ng and M. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 3, pp.993–1022, January 2003

[18] D. Blei and J. Lafferty, “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] D. Nadeau and S. Sekine, "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 Jan., 2013.

[21] Alchemy API, retrieved from http://www.alchemyapi.com/ on 15 Jan., 2013.

[22] SentiWordNet, Retrieved from http://sentiwordnet.isti.cnr.it/ on 15 Jan., 2013.

[23] V.K. Singh, R. Piryani, A. Uddin and P. Waila, “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, India, Feb. 2013.

[24] V.K. Singh, R. Piryani, A. Uddin and P. Waila, “Sentiment Analysis of Movie Reviews: A new feature-based Heuristic for Aspect-level Sentiment Classification”, In Proceedings of International Muli-Conference on Automation, Communication, Computing, Control and Compressed Sensing, Kerala-India, March 2013

[25] Gephi: The Open Graph Viz Platform, Retrieved from https://gephi.org/ on 15 Jan. 2013.