A Framework for Blog Data collection: Challenges and Opportunities

Muhammad Nihal Hussain, Adewale Obadimu, Kiran Kumar Bandeli, Mohammad Nooman, Samer Al-khateeb, Nitin Agarwal[†]

[†]Jerry L. Maulden-Entergy Chair Professor of Information Science
University of Arkansas at Little Rock, Little Rock, United States
e-mails: {mnhussain, amobadimu, kxbandeli, msnooman, sxalkhateeb, nxagarwal}@ualr.edu

***Abstract*—Blogosphere has, although slowly after the advent of Twitter, continued to rise and provides a rich medium for content framing. With no restriction on the number of characters, many users use blogs to express their opinion and use other social media channels like Twitter and Facebook to steer their audience to their blogs. Blogs provide more content than any other social media and serve as a good platform for agenda-setting. This content can be of great use to sociologists and data scientists to track opinions about events. However, the importance of blog tracking has been challenged due to the complex process of data collection and handling unstructured text data. This has caused many tracking tools to abandon blogs and move to other medium like Twitter. Nevertheless, blogs continue to be an important part of social media and cannot be ignored. In this paper, we explain the process to collect blog data, challenges we encounter, and demonstrate the importance of blog tracking through a real-world test case. The blog datasets discussed in this paper are made available publicly for researchers and practitioners through the Blogtrackers tool.**

Keywords-blog; unstructured data; web crawling; blog data collection; blog data analysis tool.

I. INTRODUCTION

The advent of participatory Web applications (or Web 2.0 [1]) have created online media that has turned the former mass information consumer to the present information producer [2]. Examples include blogs, wikis, social annotation and tagging, media sharing, and various other services. A blog site or simply blog (short for web log) is a collection of entries by individuals displayed in reverse chronological order. These entries, known as blog posts, can typically combine text, images, and URLs (Uniform Resource Locator) pointing to other blogs and/or to other Web pages. Blogging is becoming a popular means for mass Web users to express, communicate, share, collaborate, debate, and reflect. WordPress, a popular blogging platform, reports that more than 80.7 million blog posts are generated each month [3].

Blogosphere is a virtual universe that contains all blogs. Blogosphere also represents a network of blogs where each node could be either a blog or a blog post and the edges depict a hyperlink between two nodes in the Blogosphere. Bloggers, the blog post writers/authors, loosely form their special interest communities; where they share thoughts, express opinions, debate ideas, and offer suggestions interactively. Blogosphere provides a conducive platform to build virtual communities of special interests. It reshapes business models [4], assists viral marketing [5], provides trend analysis and sales prediction [6][7], aids counter-terrorism effort [8], and acts as grassroots information sources [9].

A typical blog has different posts written by one author or multiple authors on various topics of interests or activities/events occurring around the world. Blogs and other similar participatory media afford democratic spaces for people to discuss and share views that may not be endorsed by mainstream media or even traditional journalism. Additionally, the commentaries or discussions are kept for others to view and contribute further. All these features make blogs a great platform for supporting citizen journalism initiatives. Such initiatives are essential for the democratic processes of production, dissemination, and reception of news. However, one need not look farther than the current political climate to comprehend the dangers of the freedom of the Internet. Blogging and other participatory media have been strategically used to disseminate falsehoods, rumors, and gossips, to provoke hysteria, or even delegitimize governments [10][11]. Therefore, it is important to understand the blogosphere; to explore information consumption behavior of individuals, and moreover, to shed insights on how misinformation originates and spreads.

Analyzing blogs data help in understanding the pulse of a society, knowing what resonates with a community, and recognizing grievances of a group, among other reasons. Since blogs have no limit on the space available for expressing and/or discussing a topic of interest, blogs improve quality and inclusiveness of discourse and serve as a place for developing narratives. Blogs also provide a convenient platform to develop situational awareness during a socio-political crisis or humanitarian crisis in a conflict-torn region or a natural disaster struck area. While 'big' social data, especially blogs, offer promise for analysis and situational understanding [12], it also imposes significant challenges. Some of the challenges impacting this area of research are: architectural and collection issues, keeping the data up to date, processing requirements, data storage, privacy considerations, incongruities of data forms and scales, trustworthiness and reliability of the source material, and vastly varied availability of data, etc. This paper addresses key challenges pertaining to architectural and data collection issues, data cleaning, data processing, and analyzing and extracting actionable insights from blog datasets. The blog datasets discussed in this paper are made publicly available for researchers and practitioners through the Blogtrackers tool [13].

 ISBN: 978-1-61208-566-1

The rest of this paper is organized as follows. Section II describes the current state of blog data collection. Section III describes the methodology for blog data collection and curation. Section IV explains the data collected. In Section V, we demonstrate the importance of blog data analysis using Blogtrackers tool through a real-world case study. We conclude with intended future work in Section VI.

II. STATE OF THE ART

Despite the recent growth in the area of blog mining, several studies have been conducted to analyze how blog data can be effectively collected. Aschenbrenner and Miksch [14] study the development of mining techniques in a corporate environment. Their study shows a significant risk of failure due to the amount of open questions and misinformation currently available [14]. Tadanobu et al. analyzed various aspects of blog reading behavior [15]. The vast amounts of publicly available blogs have made it impossible to keep track of all of them [14]. Hence, there is a need for creating usable tools for extraction of vital information from the blogosphere.

There are some tools that were developed to analyze blogs data, but these attempts have been discontinued, such as: 1)*BlogPulse* which was developed by IntelliSeek. It was developed to provide search and analytic capabilities, automated web discovery for blogs, show the trends of information, and monitor the daily activity on weblogs. This tool was discontinued in 2012 [16][17]. 2)*Blogdex* was another service that has been discontinued; it provided a resource for understanding hot-button issues in the blogosphere. 3)*BlogScope*, was another blog tracking service developed as a research project in the department of computer science in university of Toronto, provided blog analytics and visualizations but was shut down in early 2012 [18][19]. 4)*Technorati* was originally launched as a blog index and search engine. It used a proprietary search and ranking algorithm to provide a rich directory of blogs sorted by content type and authority [20][21]. However, it did not provide blog monitoring or analytical capabilities to the end users. Furthermore, blog index and data is not available publicly to the researchers or practitioners community. The service now offers advertising platform to allow publishers to maximize their revenues without complications.

III. METHODOLOGY

To collect and store data, it is important to first identify a structure. After examining several blogs, we have identified a few common attributes such as: *title*, *author*, *date of posting*, *actual post*, *permalink*, and *number of comments* that almost all blogs have. We extract all these attributes while crawling each blog site.

For crawling blogs data, we setup crawler(s) for each blog site to extract all the required attributes. There are three main steps in crawling data from a blog site – (1) exploring the blog site, (2) crawling the blog site, and (3) cleaning and storing the data in a database for analysis and retrieval. Figure 1 shows the flow of the data crawling process.

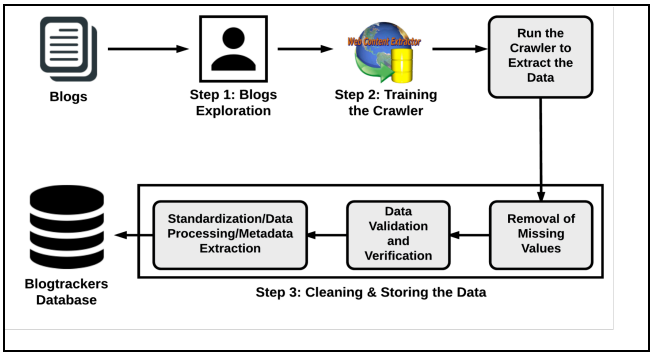

Figure 1: Data crawling process

A. *Exploring the blog site*

To train crawler(s) for a blog site, it is important to first study the structure of the blog site and understand the following:

- Type of blog site:
  - Blog on main stream media/journalist
  - Single author or multi-author blogs
  - Hobbyist blogger vs. professional blogger
  - Forum
- Site Owner(s)
- Sections of the site:
  - Archive
  - Topics or categories (e.g., news, entertainment, sports, politics, etc.)
- Language(s) of the site (inferred using AlchemyAPI [22])
- Web content structure:
  - Title of blog post
  - Author of blog post
  - Date of posting
  - Actual post/content of the post
  - Comments section
  - Tags
- Geographical location of the site (inferred using the IP address of the blog site's domain from Maltego [23])
- Description of the site
- Site navigation:
  - To identify next post or next page, if the blog is paginated
  - Search option for finding precise data.

These explorations will help us train our crawler to collect valid data and analyze them for gaining insights.

B. *Crawling the data*

Currently, we are using Web Content Extractor (WCE) tool for data collection. With this software, we train a crawler to extract data from blog sites efficiently. Figure 2 is a screenshot of the WCE.

To train the crawler, we first provide the starting or seed URLs. For example, the home page of a blog site or URL of the search page of a specific topic or section/s of a website. Then, we train the crawler to navigate to each blog post on the seed URLs as well as to the next page or older posts. Then, we take a sample post and define all the attributes we

want to collect through WCE. Here, we need to carefully select the portion of the post to avoid noises. When all the attributes are selected, WCE is ready to run for collecting the data.

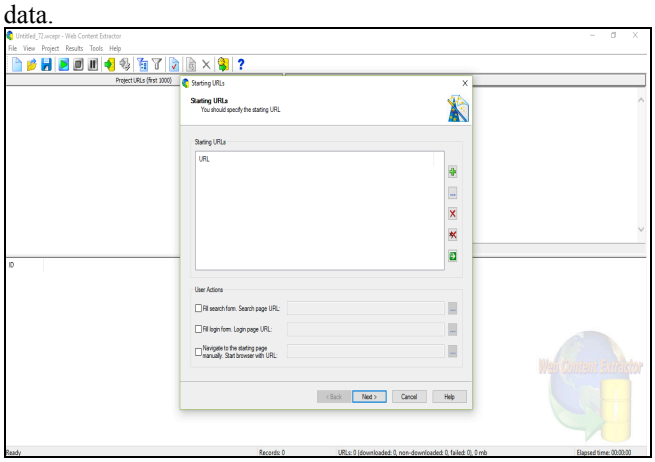

Figure 2: Web Content extractor

C. Cleaning and Storing of Data

Web crawling doesn't always give us clean data. It almost always crawls some noise. Data cleaning is required before using the data for analysis. For this, we follow a three-step cleaning process:

- *Step 1- Cleaning from WCE:* Deleting the empty fields and advertisement URLs. Later, the data is stored into a temporary database.
- *Step 2- Cleaning by SQL Queries:* Using SQL (Structured Query Language) to select validated and verified data from the temporary database. This step helps in removing the noise left unfiltered from previous step.
- *Step 3- Cleaning by Script:* In this step, a major part of cleaning along with data processing, metadata extraction, and data standardization is done. We exclude any possible noise and standardize attributes like date of posting. Followed by extracting metadata such as sentiments using LIWC (Linguistic Inquiry and Word Count) [24][25], outbound URLs, entities and their types, language of the post, proper author name etc. This is all performed before pushing the clean data into Blogtrackers database for analysis.

D. Challenges

Some of the challenges that we face during blog crawling process are:

- Changing blog structure – Blog site owners can change their blog structure any time and the crawler trained for one structure does not work for the other. This causes us to repeat the effort of training the crawler for the new structure of blog site.
- Noises – Irrespective of how well a crawler is trained, noise is always crawled. Social media plugins (such as Facebook share plugins, Twitter share plugins, etc.) and advertisements from the blog site could be crawled as JavaScript.
- Limitations of WCE – WCE sometimes fails to crawl dynamic pages that are loaded using JavaScript.

IV. DATA STATISTICS

Following the methodology proposed in Section III, we have crawled 194 blog sites, at the time of writing this paper and more blogs are being crawled. Blogs have been crawled for Ukraine-Russia Conflict, anti-NATO (the North Atlantic Treaty Organization) narratives, migrant crisis in the European Union, and the fake news blogs in the Baltic States. Below, we provide details for each dataset:

A. Ukraine-Russia Conflict

This blog dataset was collected from mid 2014 to mid 2016, during the political and military tension between Ukraine and Russia. A total of 57 blogs discussing the conflict were crawled. Tables 1 and 2 give location and language statistics for this dataset. Some blogs may have posts in more than one language.

TABLE 1. LOCATION DISTRIBUTION

Location	Blogs	Blog Posts
USA	43	15145
RUS	6	627
UKR	6	157
GBR	1	25
FRA	1	20

TABLE 2. LANGUAGE DISTRIBUTION

Language	Blogs	Blog Posts
English	42	15357
Russian	6	233
Danish	3	4
Spanish	2	9
Italian	2	2
Ukrainian	2	2
Swedish	1	45
Croatian	1	19
Norwegian	1	1
Polish	1	1
Portuguese	1	1
French	1	1

TABLE 3. LOCATION DISTRIBUTION

Location	Blogs	Blog Posts
USA	51	51436
RUS	4	3187
DEU	4	1609
NLD	3	5579
FRA	2	174
SVK	1	285
SRB	1	36
IRL	1	26
POL	1	16

 ISBN: 978-1-61208-566-1

UKR	1	6
ZWE	1	1

B. Anti-NATO

NATO's support of Ukraine during the Ukraine-Russia conflict caused an increase in the anti NATO narratives in the blogs and this sentiment was also observed during various exercises conducted by NATO (such as, Trident Juncture 2015, Brilliant Jump 2016, and Anakonda 2016). We crawled 70 blogs that had an anti-NATO propaganda from mid 2015 to late 2016. Tables 3 and 4 give statistics for this dataset.

TABLE 4. LANGUAGE DISTRIBUTION

Language	Blogs	Blog Posts
English	62	51467
French	16	223
Spanish	14	1542
German	13	1389
Russian	11	605
Polish	10	2228
Italian	9	449
Romanian	5	287
Danish	5	8
Catalan	4	104
Arabic	4	101
Czech	3	11
Finnish	3	6
Portuguese	3	4
Ukrainian	2	42
Afrikaans	2	9
Swahili	2	3
Dutch	2	3
Serbian	2	3
Welsh	2	3
Turkish	2	2
Croatian	1	1704
Greek	1	25
Basque	1	6
Hungarian	1	2
Albanian	2	2
Central mam	1	2
Faroese	1	2
Maltese	1	1
Indonesian	1	1
Tagalog	1	1
Slovak	1	1
Latvian	1	1

C. EU migrant crisis

Due to the conflict in Eastern Europe and Middle East during late 2015 and 2016, many people were migrating from war torn regions to stable regions in Europe. This dataset was collected in early 2016 during the height of migrant crisis in Europe. Tables 5 and 6 give statistics for this dataset.

TABLE 5. LOCATION DISTRIBUTION

Location	Blogs	Blog Posts
USA	21	9002
DEU	1	181

TABLE 6. LANGUAGE DISTRIBUTION

Language	Blogs	Blog Posts
English	22	7996
German	3	3
Greek	2	1039
Italian	1	4
Albanian	1	3
Croatian	1	3
Danish	1	2
Fulfulde Adamawa	1	2
Portuguese	1	2
Serbian	1	2
Czech	1	1
Finnish	1	1
Hawaiian	1	1
Polish	1	1
Turkish	1	1
Afrikaans	1	1
Hungarian	1	1
Dutch	1	1
Latin	1	1
Spanish	1	1

D. Fake News Blogs in Baltic States

There is a rising concern for fake news. Subject matter experts had identified 26 fake news blogs from the Baltic States, especially of Latvian, Estonian or Lithuanian origin suspected for disseminating fake news. We crawled 16667 blog posts from 21 blogs. This dataset was collected in early 2017. Tables 7 and 8 give basic statistics for this dataset.

TABLE 7. LOCATION DISTRIBUTION

Location	Blogs	Blog Posts
EST	7	2592
DEU	5	2156
LVA	3	1976
USA	3	3156
LTU	2	6595
NLD	1	192

TABLE 8. LANGUAGE DISTRIBUTION

Language	Blogs	Blog Posts
Latvian	10	3793
English	8	738
Russian	7	4599

 ISBN: 978-1-61208-566-1

Estonian	6	809
Lithuanian	2	6590
Bulgarian	1	1

V. DATA ANALYSIS

Blogs provide immense amount of content that can be analyzed to extract insights and sometimes gain situational awareness during conflicts. In this section, we explain the importance of analyzing blog data by extracting insights from our in-house developed tool, Blogtrackers (available for public use [13]).

We used Blogtrackers to understand the anti-NATO narratives disseminated in blogs during the NATO's Trident Juncture exercise conducted in October 2015. We started exploring our anti-NATO blogs dataset by studying the posting trends for the year 2015. Figure 3 is the posting frequency chart generated by Blogtrackers for the said period. We observed an increase in activity during August 2015 – December 2015, roughly 2 months before and after the exercise.

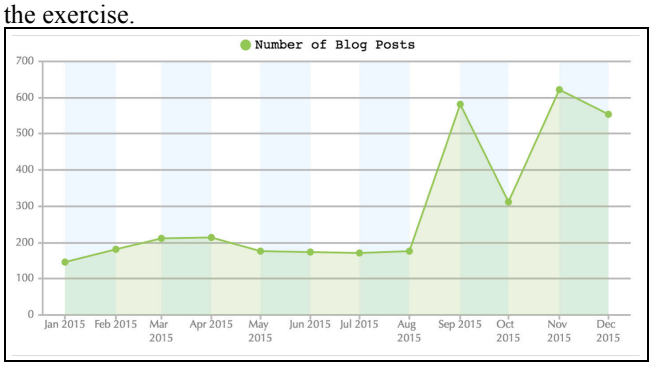

Figure 3: Posting Frequency of Anti-NATO blogs from January 2015 to December 2015.

We generated sentiment trends to understand the overall tonality of the bloggers' postings during this period. Figure 4 is the sentiment trend generated by Blogtrackers for the said period. We found that the sentiment was majorly positive up until the exercise, i.e., October 2015, and negative after the exercise, demonstrating that bloggers did not see the exercise in a good light. There was a strong anti-NATO sentiment stemming from anti-NATO propaganda.

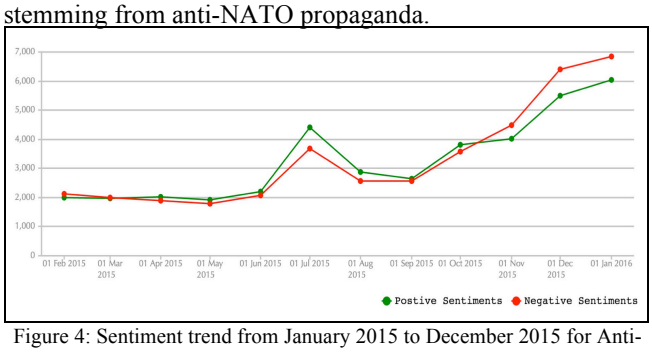

Figure 4: Sentiment trend from January 2015 to December 2015 for Anti-NATO blogs

We also generated the influence trend to understand the variation of blogger's influence over the community. This is helpful to know how the narratives in blogosphere resonated with the readers. Influence score for each blog post is based on the chatter it generates in the blogosphere. It is computed using the amount of discussion it generated (comments) and outbound URLs [26][27]. Influence of a blogger is assessed by how influential his/her posts are. Figure 5 is the influence trend of the top 5 (influential) bloggers in the said period.

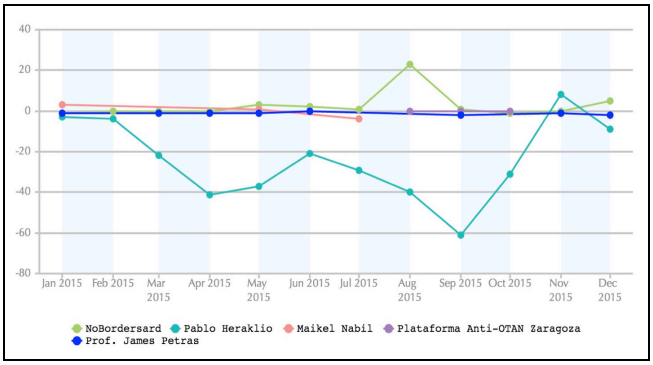

Figure 5: Influence trend for top 5 bloggers

Figure 6: Blog browser showing the most influential blog post.

From the trend chart in Figure 5, we identified NoBordersard was the most influential. As depicted in Figure 6, we used blog browser feature of Blogtrackers to know what has influenced the community. We found a blog post written by NoBordersard in Italian was calling for civil disobedience march against the NATO exercise. This blog post generated a considerable amount of chatter and had the highest influence.

VI. CONCLUSION AND FUTURE WORK

This article presents a novel approach on blog data collection. Currently, the approach followed is to - manually collect, clean and save blog data to relational databases for further analysis. This is helpful in many ways to benefit the user with the process of carefully analyzing the blog site structure and its changing nature, noises, and myriad other challenges. Obtaining a cleaner blog data sample is an extremely time consuming process and involves significant human intervention. Understandably, this is not a scalable approach, given the speed with which the blogosphere is expanding. Therefore, we are developing an automated crawling mechanism to overcome the challenges presented by blog data collection thereby significantly increasing the efficiency of the overall process from data to decisions.

The article also presented a case study on how Blogtrackers, tool for analyzing blogs, had sift through more than 60,000 blog posts from 70 anti-NATO blogs to identify

 ISBN: 978-1-61208-566-1

a blog post calling for civil disobedience; explaining the significance of studying blogs in analyzing information dissemination through social media to identify blogs and bloggers calling for deviant activities. Going further, we would like to add content analysis features to Blogtrackers, such as: topic modeling, network analysis, and cyber forensics features, to not only study blogs individually, but also to understand their coordination structure and information dissemination structure.

ACKNOWLEDGMENT

This research is funded in part by the U.S. Nation Science Foundation (IIS-1110868 and ACI-1429160), U.S. Office of Naval Research (N000141010091, N000141410489, N0001415P1187, N000141612016, and N000141612412), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059) and the Jerry L. Maulden/Entergy Fund at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

REFERENCES

[1] T. O'Reilly, "What is Web 2.0 - design patterns and business models for the next generation of software," 30-Sep-2005. [Online]. Available: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html. [Accessed: 18-May-2017].

[2] D. Gillmor, "*We the media: Grassroots journalism by the people, for the people*", O'Reilly Media, Inc., 2006.

[3] "Stats — WordPress.com." [Online]. Available: https://wordpress.com/activity/. [Accessed: 04-Apr-2017].

[4] R. Scoble and S. Israel, "*Naked conversations: how blogs are changing the way businesses talk with customers*", John Wiley & Sons, 2006.

[5] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2002, pp. 61–70.

[6] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, 2005, pp. 78–87.

[7] G. Mishne and M. de Rijke, "Deriving wishlists from blogs show us your blog, and we'll tell you what books to buy," in *Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, 2006, pp. 925–926.

[8] T. R. Coffman and S. E. Marcus, "Dynamic classification of groups through social network analysis and hmms," in *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, 2004, vol. 5, pp. 3197–3205.

[9] M. Thelwall, "Bloggers during the London attacks: Top information sources and topics," in *Proceedings of the 3rd International Workshop on the Weblogging Ecosystem (WWE 2006)*, 2006.

[10] S. Al-khateeb, M. Hussain, and N. Agarwal, "Analyzing Deviant Socio-technical Behaviors using Social Network Analysis and Cyber Forensics-based Methodologies," in *Big Data Analytics in Cybersecurity and IT Management*, CRC Press, Taylor & Francis, in press.

[11] S. Al-khateeb and N. Agarwal, "Understanding Strategic Information Maneuvers in Network Media to Advance Cyber Operations: A Case Study Analyzing pro-Russian separatisits' Cyber Information Operations in Crimean Water Crisis," *J. Balt. Secur.*, vol. 2, no. 1, pp. 6–27, 2016.

[12] J. Kopecky, N. Bos, and A. Greenberg, "Social identity modeling: past work and relevant issues for socio-cultural modeling," in *Proceedings of the 19th Conference on Behavior Representation in Modeling and Simulation, Charleston, SC*, 2010, pp. 203–210.

[13] "Blogtrackers." [Online]. Available: http://blogtrackers.host.ualr.edu/. [Accessed: 18-May-2017].

[14] A. Aschenbrenner and S. Miksch, "Blog mining in a corporate environment," *Vienna Univ. Technol. Inst. Softw. Technol. Interact. Syst. Res. Studio Austria Smart Agent Technol.*, 2005.

[15] T. Furukawa, M. Ishizuka, Y. Matsuo, I. Ohmukai, K. Uchiyama, and others, "Analyzing reading behavior by blog mining," in *Proceedings of the National Conference on Artificial Intelligence*, 2007, vol. 22, p. 1353.

[16] M. Hurst, "Farewell To BlogPulse | SmartData Collective," *SmartData Collective*, 14-Jan-2012. [Online]. Available: http://www.smartdatacollective.com/matthewhurst/44748/farewell-blogpulse. [Accessed: 18-May-2017].

[17] "BlogPulse," *Wikipedia*. 08-Mar-2017.

[18] N. Bansal and N. Koudas, "Blogscope: a system for online analysis of high volume text streams," in *Proceedings of the 33rd international conference on Very large data bases*, 2007, pp. 1410–1413.

[19] "BlogScope," *Wikipedia*. 29-Nov-2015.

[20] "Technorati—the World's Largest Blog Directory—is Gone," *Business 2 Community*. [Online]. Available: http://www.business2community.com/social-media/technorati-worlds-largest-blog-directory-gone-0915716. [Accessed: 04-Apr-2017].

[21] "About Us | Technorati." [Online]. Available: http://technorati.com/company/about-us/. [Accessed: 18-May-2017].

[22] "AlchemyAPI," *Wikipedia*. 02-May-2017.

[23] "Paterva Home." [Online]. Available: https://www.paterva.com/web7/. [Accessed: 18-May-2017].

[24] I. LIWC, "Linguistic Inquiry and Word Count (LIWC)." [Online]. Available: www.liwc.wpengine.com. [Accessed: 12-Apr-2016].

[25] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway Lawrence Erlbaum Assoc.*, vol. 71, no. 2001, p. 2001, 2001.

[26] N. Agarwal, H. Liu, L. Tang, and S. Y. Philip, "Modeling blogger influence in a community," *Soc. Netw. Anal. Min.*, vol. 2, no. 2, pp. 139–162, 2012.

[27] N. Agarwal, D. Mahata, and H. Liu, "Time- and Event-Driven Modeling of Blogger Influence," in *Encyclopedia of Social Network Analysis and Mining*, R. Alhajj and J. Rokne, Eds. New York, NY: Springer New York, 2014, pp. 2154–2165.

 ISBN: 978-1-61208-566-1