A Novel Methodology to Identify and Collect Data from Relevant Blogs Leveraging Multiple Social Media Platforms and Cyber Forensics

Tuja Khaund, Kiran Kumar Bandeli, Oluwaseun Walter, Nitin Agarwal
Department of Information Science
University of Arkansas at Little Rock
Little Rock, United States
e-mail: {txkhaund, kxbandeli, oxwalter, nxagarwal}@ualr.edu

Abstract—**Blogs play a vital role in retrieving real time information, a place for users to gain insights into events and also find communities with similar interests. However, being able to identify blogs that contain honest, unbiased opinion of individuals as opposed to biased or agenda-driven coverage, is quite a challenge. Secondly, blogs are notorious for being dynamic in structure, where their owner is entitled to give them a makeover whenever they want. This changing structure of blogs can be computationally expensive for researchers and Web crawlers. In this paper, we propose a methodology to help identify relevant blogs for specific events. We provide data statistics of a few real-world events where our methodology successfully identified relevant blogs and helped us study the information discourse. We then discuss the strengths and weaknesses of this methodology and highlight the best approach to crawling blogs.**

Keywords-blog; blog identification; relevant blogs; cyber forensics; unstructured data; social media; crawling.

I. INTRODUCTION

Social media is a gold mine of valuable resources that coordinate various real-life events to a wide audience. It allows people to voice their opinions, engage in discussions and share information. However, the Internet is built to follow the power law distribution where these sources usually get buried in the Long Tail. This makes social media a valuable source for event analysis studies and to identify quality sources from the pool of information is of utmost importance. A blog site or a blog is a collection of entries, called blog posts, by individuals displayed in reverse chronological order. These posts are a combination of text, images, and Uniform Resource Locators (URL), which direct to other blogs and/or to other Web pages. Blogging has become a popular means for mass Web users to express, communicate, share, collaborate, debate, and reflect [1].

Generally, a blog has different posts written by either a single author or multiple authors on topics of interest or on events happening around the world. While blogs allow free medium to write on any events or issues, some authors use this for spreading mis/disinformation. Some of the studies on blogs look at various events such as European Union (EU) migrant crisis to analyze shift in narratives regarding migrants [2], Venezuelan Socio-Economic Crisis to gain situational awareness of the protests [3], role of blogs in disinformation campaign coordination [4], and events related to fake news in Baltic States spreading misinformation [5].

In this paper, we propose a methodology to identify relevant blogs for specific events. We use different input streams, which will obtain URLs for blog identification such as streaming Twitter based on geo-location, using Cyber Forensic analysis to detect blogs based on Google Analytics tracking codes, etc. Google Analytics tracking code monitors the activity of a website and provides insights about visitors of the website. The Analytics tracking code may be added directly to the Hypertext Markup Language (HTML) code of each page on a website, or indirectly using a tag management system such as Google Tag Manager.

The rest of the paper is organized as follows. Section II describes the related work. Section III depicts the methodology used to identify blogs from various sources. Section IV discusses the analysis and findings obtained from the methodology. We discuss the challenges to this research in Section V and conclude with intended future work in Section VI.

II. LITERATURE REVIEW

The blogosphere has generated a vast amount of content over the years making it difficult to keep track of all the resources. Identifying rich and legitimate sources of information is a challenge every researcher is trying to overcome with new methods and models. Mahata et al. [6] applied an evolutionary mutual reinforcement model to identify and rank highly 'specific' social media sources and 'close' entities related to an event. Agarwal et al [7] studied sentiments and opinions of people towards public and political events from blogs. Twitter [8][9] and YouTube [10] have been extensively used to analyze information dissemination during natural disasters and crisis. Event related contents have been found leveraging the tagging and location information associated with the photos shared on Flickr [11]. Becker et al. [12] studied how to identify events and high quality sources related to them from Twitter. In order to identify the genuine sources of information, credibility and trustworthiness of event related information were studied from Twitter [13]. New methods were investigated for filtering and assessing the variety of sources obtained from social media for journalists [14]. All these works try to explore the quality of information, in terms of relevancy, usefulness, timeliness of the content and usage patterns of authoritative users producing the content. However, only a few works involved the blogosphere. Our work will help researchers find a new direction in identifying blogs from credible sources and validate them.

 ISBN: 978-1-61208-700-9

III. METHODOLOGY

Our methodology highlights two major categories of blog data collection. First, we identify blogs and then conduct a relevance assessment to obtain the most relevant blogs. The first phase of this methodology may generate noise and also contain content from mainstream media. The main motive is to obtain rich, unbiased opinions of users to study the information discourse during crucial real-world events. The high-level design of the methodology is illustrated in Figure 1.

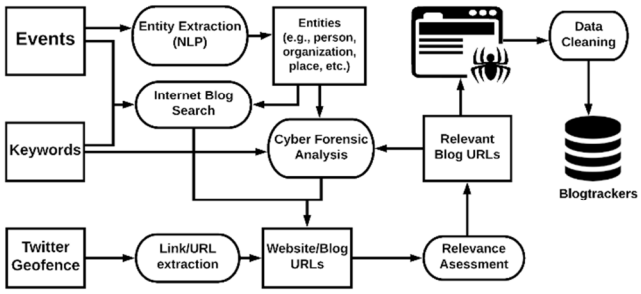

Figure 1. Blog identification and collection methodology.

Data is injected from multiple streams into the engine. The proposed methodology is scalable, meaning more such data streams can be added as they become available. The 'Events' data stream focuses mainly on local events of the country or region of interest, such as Ukraine while studying their Parliamentary Affairs, Europe while analyzing the migrant crisis, etc. The 'Keywords' data stream comprises of information provided to us by domain experts such as names of Parliament members, local political groups etc. The 'Twitter Geofence' data stream extracts tweets based on geolocation. Another laborious approach exists where data is collected from Facebook, where we analyze content dissemination based on keywords or trending topics for events of interest. Each data stream has a different shared engine that enables us to extract entities, run cyber forensic analysis and extract more blogs using a snowball approach. After we have exhausted every possible stream, we run checks for relevant content. This process is tedious but the results are promising. The relevant blogs are then crawled, cleaned and stored in our database. We provide a step-by-step procedure of the data streams used in the methodology below.

A. Blog Identification

The most important step to our methodology is to identify blogs.

1) Event Analysis:

In this process, we refer to our case study of Ukraine's Parliament Affairs where we study public discourse on social media platforms, mainly blogs. There are four different steps in this blog identification process.

a) We begin by searching for the presence of Ukrainian bloggers and blogs on the Web and different social media platforms such as Twitter, Facebook, etc., using keywords such as 'Ukraine bloggers' 'Poroschenko', 'war Donbass', 'Verkhovna Rada', 'Ukraine blog', 'Petro Poroshenko Bloc', 'People's Front', etc.

b) We then create an event dictionary using popular news website like 'Kyivpost.com' and 'Unian.info' as Ukrainian event sources. We also keep a record of keywords used by these sources for our next step.

c) We use keywords from the event dictionary to search for blog sites on social media platforms, google search engine, and these 2 websites 'searchblogspot.com' and 'search.wordpress.com'.

d) Finally, we find other relevant blogs from blogrolls of already identified blogs.

It is also important to note that converting the keywords to Ukrainian language, while searching for blogs, enabled us to discover more blogs that are specific to Ukraine. The keywords chosen are subjective but we seek guidance from domain experts for better results.

2) Twitter Daily Dumps:

There are four steps involved in this process of blog identification.

a) We set up a geofence based on coordinates of a particular country or region of interest and stream Twitter for daily tweets. The results are stored as JavaScript Object Notation (JSON) files.

b) We extract all the URLs from our Twitter daily dump and also expand all the shortened URLs.

c) We filter all the unique hyperlinks based on keywords such as 'blog', 'blogspot', 'wordpress' etc. and extract the domains (of blogs).

d) Finally, we perform relevance checks on the filtered blogs and add them to the crawling pipeline.

This methodology extracts every tweet that has been posted within the set geo-coordinates and at times, extracts noise. As a result, we run cyber forensic analysis on these blogs to discard irrelevant URLs.

3) Cyber Forensics:

There are four steps to this blog identification process.

a) Once we identify blogs from the Twitter daily dumps, we run these blogs through a cyber-forensic analysis tool, Maltego [1], to extract more blogs.

b) These blogs are identified based on common Google Analytics tracker codes.

c) External hyperlinks (out-links) are also extracted from these blogs and then, we proceed to conduct the relevance assessment.

d) Finally, relevant blogs are then added to the crawling pipeline and then stored in the database for further analysis.

This analysis can be snowballed until we have no more blogs left to identify, as shown in Figure 1.

 ISBN: 978-1-61208-700-9

B. Relevance Assessment

Identifying relevant blogs is a manual process where blogs are distributed among members of the team. This process is subjective; team members know exactly which keywords to choose in order to detect data streams delivering the most promising results. The criteria include keywords that are relevant to the events. For example, while studying the blog discourse of Ukraine, we focused on keywords such as 'Ukraine', names of parliament members, discussion of bills being passed or introduced into the legislation, etc. We were able to detect more relevant blogs from the Event Analysis and Cyber Forensics data streams. We try to eliminate content posted on mainstream media such as news sites, etc. to minimize bias. We analyze the blog's content and the links shared in it. Also, we rate a blog's severity as low, medium or high based on their content.

This methodology is open, scalable and expandable based on the number of data streams available. In order to improve the scalability of the effort, information retrieval-based relevance checks are conducted with keywords provided by domain experts. In the next section, we provide statistics of blog identification obtained for various events.

IV. ANALYSIS AND FINDINGS

A. Data Statistics

Using the methodology proposed in Section III, we have crawled 108 blog sites, at the time of writing this paper and more blogs are queued for crawling. Blogs that have been crawled are from the following datasets – The North Atlantic Treaty Organization (NATO) Trident Juncture Exercise 2018, migrant crisis in the EU, and Venezuelan Socio-Economic Crisis. Below we provide detailed statistics for each source:

1) NATO Trident Juncture Exercise 2018:

NATO's Trident Juncture exercise 2018 that happened during the period Oct. 2018 and Nov. 2018, in Norway has caused an increase in the anti NATO narratives on blogs. This sentiment was also observed during various exercises conducted by NATO (such as, Trident Juncture 2015, Brilliant Jump 2016, and Anakonda 2016). Using the methodology presented in Figure 1, we identified 46 blogs that had anti-NATO propaganda discourse. Figure 2 and Figure 3 demonstrates the statistics about this dataset.

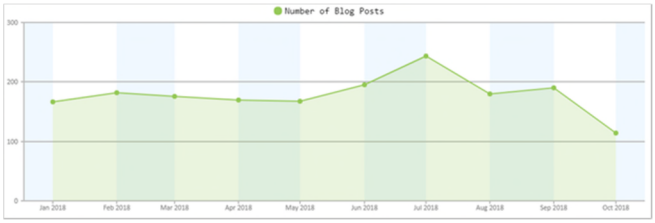

Figure 2. Blog post distribution of Anti-NATO blogs.

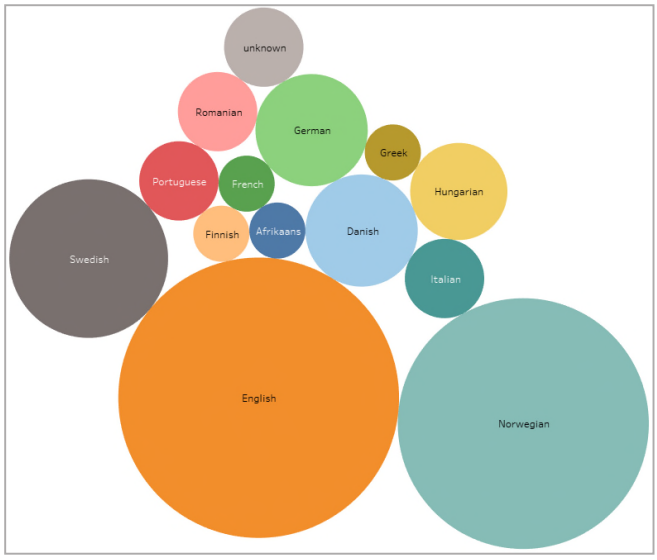

Figure 3. Language distribution of Anti-NATO blogs

2) 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 peak time of migrant crisis in Europe. Figure 4 and Figure 5 provide the details of the dataset.

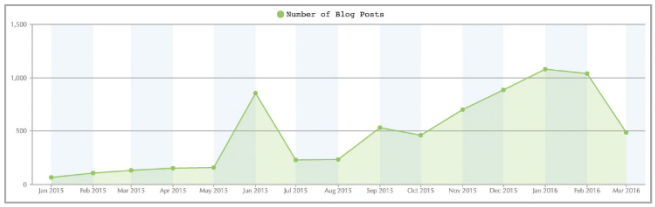

Figure 4. Blog post distribution of EU Migrant Crisis blogs.

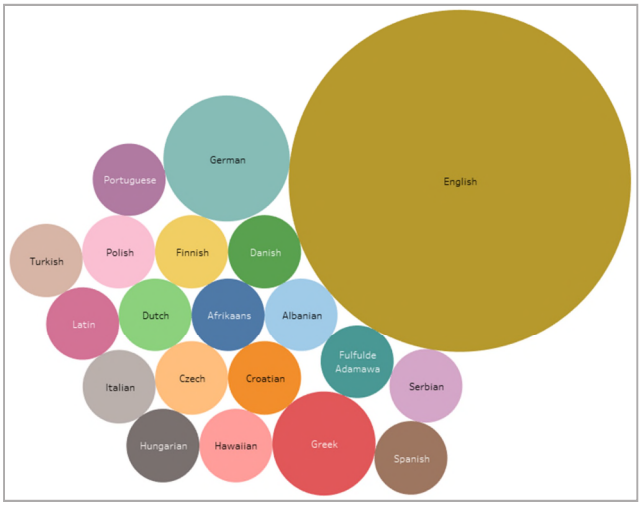

Figure 5. Language distribution of EU Migrant Crisis blogs.

3) Venezuelan Socio-Economic Crisis:

To analyze the socio-economic crisis in Venezuela from blogosphere, we collected data mainly for the period of mid-2016 and early 2017. During this period, many protests occurred covering the crisis event. Figure 6 and Figure 7 provide details about this dataset.

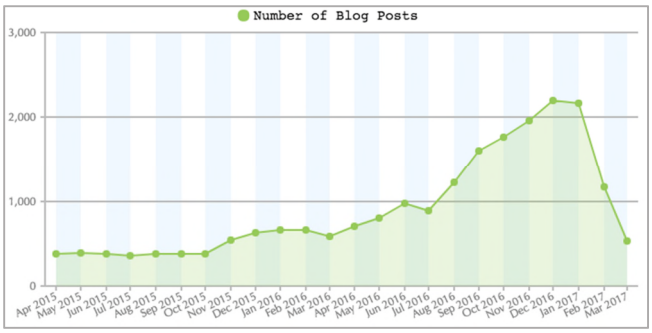

Figure 6. Blog post distribution of Venezuelan Socio-Economic Crisis blogs.

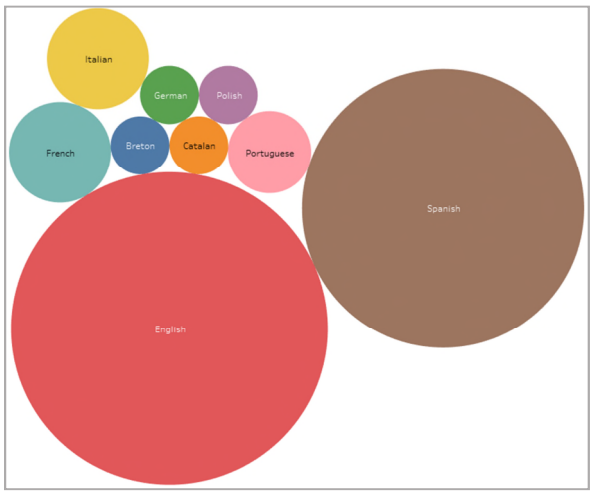

Figure 7. Language distribution of Venezuelan Socio-Economic Crisis blogs.

The analysis was conducted for three case studies that used the methodology. This is to show that we were able to get results based on the tasks listed on our methodology.

V. DISCUSSION

This paper presents a methodology, which will help researchers identify blogs. The semi-automated tasks will enable a user to obtain a much richer and prominent set of blogs, which otherwise will be extremely difficult given the dynamic nature of blog structure. However, blog identification and collection is a laborious task and has numerous challenges throughout the stages.

A. Challenges of identifying relevant blogs

The major portion of the relevance assessment is performed manually, which, in itself, is a challenge. But there exists a series of hurdles underlying the process of blog assessment.

1) Noise: Keyword-based blog searches often yields unexpected results. Blogs are not genre specific and may contain posts about world affairs, travelling, food, etc.

2) Limited Availability: A few blogs that were discovered have fewer blog posts (less than 10), while others no longer publish blog posts and a few others moved to different websites. Additionally, the content of these blog posts may or may not discuss the subject matter of interest.

3) Separating Blogs and News: During the initial stages of this methodology, differentiating blogs from mainstream websites became difficult because of the way these websites are structured.

4) Mainstream Dominance: Majority of Web links identified through search engines and social media sites were mainstream websites.

B. Challenges of blog data collection

A few challenges encountered during blog collection include the following:

1) Application Programming Interface (API) restriction: Various tools such as: BlogPulse [15], Blogdex [16], and Technorati [17], etc., were previously available to analyze blog data, but these efforts have been discontinued. As a result, there is no API available to extract blog data.

2) Dynamic blog structure: Dealing with blogs is similar to working with a moving target. Blog site owners are entitled to make changes to their blog structure any time. This confuses a trained Web crawler as it was formerly instructed to follow one structure, which has now been altered. As a result, the entire effort of blog crawling needs to be repeated for the new structure of blog site. Additionally, each blog requires its own parser to crawl the data.

3) Noise: 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.

4) No standardization: While we collect blog data, we parse important attributes for analysis. Once such attribute is date. While extracting the date field from blog posts, we noticed that it differs in format from blog site to blog site. In other words, a single standard is not followed in these blogs.

5) No automation: The process of blog crawling is not fully automated. Even the most intelligent/careful parsing may capture some noise. Manual intervention is required to identify and eliminate noise.

VI. CONCLUSION

In this paper, we proposed a methodology to help identify relevant blogs for specific events. We provided data statistics of a few real-world events where our methodology successfully identified relevant blogs and helped us study the information discourse. We then discussed the strengths and weaknesses of this methodology and highlighted the best approach to crawling blogs.

 ISBN: 978-1-61208-700-9

Conducting relevance assessment was a challenging task during this research is since it was performed manually. This is the most important task in our methodology because it helps us detect credible and important data sources. Automating this process will not only save a lot of time, but it will make blog crawling more scalable. We have tested a few blog sites that are hosted on WordPress and results are acceptable. We would like to extend this task to other blogs as a future work.

ACKNOWLEDGMENT

This research is funded in part by the U.S. National Science Foundation (IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17-1-2675), 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), the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock and Arkansas Research Alliance. 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 their support.

REFERENCES

[1] M. N. Hussain, A. Obadimu, K. K. Bandeli, M. Nooman, S. Al-khateeb, and N. Agarwal, *A Framework for Blog Data Collection: Challenges and Opportunities*. June, 2017.

[2] M. N. Hussain, K. K. Bandeli, S. Al-khateeb, and N. Agarwal, "Analyzing Shift in Narratives Regarding Migrants in Europe via Blogosphere," 2018.

[3] E. L. Mead, M. N. Hussain, M. Nooman, S. Al-khateeb, and N. Agarwal, "Assessing Situation Awareness through Blogosphere: A Case Study on Venezuelan Socio-Political Crisis and the Migrant Influx."

[4] N. Agarwal and K. K. Bandeli, "Examining Strategic Integration of Social Media Platforms In Disinformation Campaign Coordination," *Def. Strateg. Commun.*

[5] N. Agarwal and K. K. Bandeli, "Blogs, Fake News, and Information Activities," in *Digital Hydra: Security Implications of False Information Online*, NATO Strategic Communications Center of Excellence (StratCom COE), 2017, pp. 31–45.

[6] D. Mahata and N. Agarwal, "What does everybody know? identifying event-specific sources from social media," in *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, 2012, pp. 63–68.

[7] N. Agarwal, H. Liu, J. Salerno, and S. Sundarajan, "Understanding group interaction in blogosphere: a case study," in *Proc 2nd international conference on computational cultural dynamics (ICCCD), September*, 2008, pp. 15–16.

[8] T. Khaund, S. Al-Khateeb, S. Tokdemir, and N. Agarwal, "Analyzing Social Bots and Their Coordination During Natural Disasters," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 2018, pp. 207–212.

[9] F. Cheong and C. Cheong, "Social media data mining: A social network analysis of tweets during the Australian 2010-2011 floods," in *15th Pacific Asia Conference on Information Systems (PACIS)*, 2011, pp. 1–16.

[10] M. N. Hussain, S. Tokdemir, N. Agarwal, and S. Al-Khateeb, "Analyzing Disinformation and Crowd Manipulation Tactics on YouTube," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 1092–1095.

[11] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 103–110.

[12] H. Becker, M. Naaman, and L. Gravano, "Selecting Quality Twitter Content for Events." *ICWSM*, vol. 11, 2011.

[13] M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on Twitter," in *Proceedings of the 2012 SIAM International Conference on Data Mining*, 2012, pp. 153–164.

[14] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 2451–2460.

[15] N. Glance, M. Hurst, and T. Tomokiyo, "BlogPulse: Automated Trend Discovery for Weblogs," in *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*, New York, NY, USA, 2004, vol. 2004.

[16] C. Marlow, "Audience, structure and authority in the weblog community," p. 9.

[17] D. Sifry, "State of the blogosphere 2007," 2008.

 ISBN: 978-1-61208-700-9