A Method to Extract the Forensic about Negative Issues from Web

M K M Nasution*, M Hardi, R Sitepu and E Sinulingga

Teknologi Informasi, Program Studi Hukum, Teknik Elektro, Universitas Sumatera Utara, Padang Bualn 20155 USU Medan Indonesia

*mahyuddin@usu.ac.id

Abstract. In the social world there are many issues: positive or negative. The negative issues affect the level of social comfort. On social media such as Web, every issue positioned based on the document, which has its own attributes, such as the URL address and date of creation. Not easy to extract information from the Web, as well as to determine the origin of an issue that is flowing in the web. This paper is to derive a method for revealing the origin of an issue based on the characteristics of each webpage.

1. Introduction
In substance, the emergence of ICT has changed the flow of information from one-way to two way: From the publisher to the audience and / or from the audience to the publisher [1, 2]. In addition, the classic social ills that channeled limited in such social negative issues remain a part of social life [3]. On the Web, many documents and continues to grow unpredictably in accordance with social activities [4], and the piles of documents that can be recognized as webpages potentially changing lifestyle of the people [5]. Likewise, if the web pages containing the negative issues as the spread slander also would disrupt the fabric of society [6].

It is necessary that negative issues need to be detected its existence more early, and this requires an appropriate way to extract information related to negative issues [7, 8]. Extraction of information is a way to determine the presence information based on a predetermined reference [9]. There are several methods used for extracting social networks, but few are related to forensic of negative issues [10]. This paper aims to address the method are useful for extracting the forensic of negative issues from the Web.

2. A Review and Motivation
The Information has been recorded in the Web with various types of documents, coming from a wide variety of different sources, involves a discussion topics that wide, and the style that’s not the same [11]. Specially, Web directly has affected the management of knowledge: a variety of document formats, and news presentation also involves a lot of media [12]. However, the documents on the results of studies and research have the more trustworthy compared with documents from other sources, and is usually directly filtered by various engine indexers such as Scopus, for example [13]. The unofficial documents produced individually, generally junky, a part of them be garbage of information in the Internet [14]. Likewise, the news from newspaper publishers either the softcopy or the hardcopy. With the web, softcopy so easily spread news to the joints of the social, through online media will be so easily absorbed by the heart of social and subsequently affect the social life quickly [15]. Web allows the presentation
of information that can be read, heard, and seen in parallel. Therefore, the Web pages that can interact with the audience is increasing [16].

The webpage contains information that can be determined through the text, which consists of a set of words \( W = \{w_i\} \) [9, 17], every web page on the Internet has an address in the canonical URL. The URL address is the identity of the web page. The URL addresses replace the file names and directories (sub directories) where the file exists [18]. In addition to the file name, each file also has a document creation date and time. Web as an information document automatically contains the file name, date and time of creation. In line with the needs of information, studies on the web has given the progress that the web is becoming the smart documents, namely “documents know about itself well” [18]. The web has become the core of the internet, therefore, studies continue to be done to improve the ability of the web as a place of information, the facilities in web coupled with the implementation of the W3C (World Wide Web Consortium): an organization that assesses and standardize the attributes for structuring the information [15].

In line with the nature of natural language and other media, web contains duplication and ambiguity semantically continues to exist. So, although efforts continue to be made to achieve better, including the superficial methods is to overcome the slowness of the process by using the classification [19, 20].

3. Toward Method
Superficial method, including categories of unsupervised method, involving the search engines to access information from the Web. However, this method generally must involves the well-defined information, like use the name of social actor in query we define it as using quote: “Mahyuddin K. M. Nasution” for example [21, 22]. Therefore, to extract the forensic of negative issues from the Web, built a method involves the concept of the clustering and classification into one semi-unsupervised, namely:

- Grouping words that belonging to the negative or positive meaning based on the sentences give the mentioned discourse [23]. The approach involves a classification (supervised). Examples of words such as “korupsi” (corruption) and ”suap” (bribery).
- Selecting a keyword for each the negative meaningful word. Based on semantic concepts of the co-occurrence, keywords used to pry the web pages that are pent in repository site [20]. Example "Universitas of Sumatera Utara”, ”16 miliar”.
- Collecting all related web pages and the URL addresses into the corpus.
- Extracting information about creation date, the names of social actors, or publisher. In extracting the actor names from web pages we used the concept of Name Entity Recognition (NER) [24]. Evaluating each social actor based on creation date, the names of social actors, or publisher. In extracting the actor names from web pages we used the concept of Name Entity Recognition (NER) [24].

\[
\text{precision} = \frac{|\{\text{relevant documents}\}|n|\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (1) \\
\text{recall} = \frac{|\{\text{relevant documents}\}|n|\{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad (2) \\
\text{Fmeasure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)
\]

To get an assessment of their respective social actors we involve query that contains the name of the actor and negative issue [22].

To support the systematics of method we conduct the experiments to get the actor spreads issue [7].
4. Experiment
Suppose from a variety of Web page titles are classified and selected a set of words associated with negative issues. There is a set of words namely $W = \{"korupsi" \text{ (corruption)}, \"curang\" \text{ (cheat)}, \"suap\" \text{ (kickbacks)}, \"16 miliar\" \text{ (16 billion)}, \ldots\}$. At the time of this experiment is done, for example, using the phrase “16 billion” in which a query containing the phrase is inserted into the Google search engine acquired a collection of web pages as in Figure 1.

Table 1. Names of social actors and position

| Id | Actor Name          | Position                                                   |
|----|---------------------|------------------------------------------------------------|
| 1  | Runtuw Siptu        | Elected Rector                                             |
| 2  | Mohammad Nasir      | Minister of Research, Technology and Higher Education       |
| 3  | Todung Mulya Lubis | Chairman of the Board of Trustees (MWA)                    |
| 4  | Osmar Tanjung       | Selmas Jokowi                                              |
| 5  | Subuhhar            | Rector candidate                                           |
| 6  | Zulkifli Nasution   | Rector candidate                                           |
| 7  | T. Erry Nuradi      | North Sumatra Governor                                     |
| 8  | Randiman Tarigan    | Acting Mayor of Medan                                      |
| 9  | T. Dzulmi Eldin     | Mayor of Medan                                             |
| 10 | Luhut Binsar        | Secretary of the Rector Election Committee                 |
After that, we specify keywords that may be gouged Web pages from the repository, for example related to negative issues ”16 billion” the organization of social actors selected is ”Universitas Sumatera Utara” acquired a collection of web pages as in Figure 2, we have 9 web pages about ”16 miliar” and ”Universitas Sumatera Utara”.

Through a computer program in the system used to process the snippet as the result that returned by search engine to the submitted query. The program also we use to access the web pages one by one. The content of the webpages and their URL address gathered into the corpus. The content of each web page searched one by one to get the creation date, publisher, and the names of the social actors that appear on each web page. One snippet like Figure 3 shows the creation date is older than the others, and it has the URL address as follows,

https://plus.google.com/wm/1/107501815556102420319/posts/8p9er6DE3Tz

See Figure 2 and 3.

Based on the concept of Named Entity Recognition (NER) we use program for excursions 9 related web pages and we get the names of social actors such as Table 1.

To assess the association between negative issue ”16 miliar” and the social actors, we use precision, recall and F-measure for each actor through a query containing the name of the actor and the phrase ”16 miliar” as shown in Table 2, and we get the actor-4 (Osmar Tanjung) closest to the negative issue. Then by getting the web pages with the oldest date, acquired content-related issues in Figure 4 where translation is as follows:

”Osmar said that the current transaction outstanding issues for the election of the Rector USU is fantastic value, i.e. Rp 16 billion. "Supposition of such transactions conducted in Singapore," said Osmar’.

Thus, the social actor to be the source of the negative issue based on the method of forensic extraction is Osmar Tanjung.

5. Conclusion
To perform forensic against negative issues, we have put forward a method involving approach is supervised and unsupervised, and this method is referred to as semi-unsupervised method, and to prove the method we have performed experiments. Furthermore, this research deals with proving formally forensic extraction method of the negative issues from the Web.

References
[1] Hung T. Nguyen, Vladik Kreinovich, and Guoqing Liu, 1999, Chu spaces: Towards new foundation for fuzzy logic and fuzzy control, with applications to informatin flow on the world wide web, IEEE Proceeding of NAFIPS.

| Id of Actor | Precision | Recall | F-measure |
|-------------|-----------|--------|-----------|
| 1           | 0.06      | 0.67   | 0.12      |
| 2           | 0.05      | 0.44   | 0.08      |
| 3           | 0.02      | 0.67   | 0.04      |
| 4           | 0.86      | 0.67   | 0.75      |
| 5           | 0.20      | 0.22   | 0.21      |
| 6           | 0.03      | 0.22   | 0.05      |
| 7           | 0.03      | 0.22   | 0.05      |
| 8           | 0.02      | 0.11   | 0.03      |
| 9           | 0.25      | 0.22   | 0.24      |
| 10          | 0.33      | 0.11   | 0.17      |
[2] Zi Lu, Ruiling Han, and Jie Duan, 2010, Analyzing the effect of website information flow on realiser human flow using intelligent decision models, Knowledge-Based Systems 23 40.
[3] Michael Heberling, 2002, State Lotteries: Advocating a Social Ill for the Social Good, The Independent Review 6(4) 597.
[4] Marta Gatius, Manuel Bertran, and Horacio Rodriguez, 2004, Multilingual and Multimedia Information Retrieval from Web Documents, Proceedings of the 15th International Workshop on Database and Expert Systems Applications
[5] Alexander Garcia-Castro, Alberto Labarga, Leyla Garcia, Olga Giraldo, Cesar Montana, John A. Bateman, 2010, Semantic Web and Social Web Heading towards Living Documents in the Life Sciences, Web Semantics: Science, Service and Agents on the World Wide Web 8 155
[6] Stefan Stiegitz, Linh Dang-Xuan, Axel Bruns, and Christoph Neuberger, 2014, Social media analytics: An interdisciplinary approach and its implications for information systems, Business & Information Systems Engineering 2 89
[7] Mahyuddin K. M. Nasution, Melvani Hardi, and Runtung Sitepu, 2016, Using Social Networks to Assess Forensic of Negative Issues, IEEE International Conference on Cyber and IT Service Management.
[8] Mahyuddin K. M. Nasution, and Shahru Azman Noah, 2012, Information retrieval model: a social network extraction perspective, IEEE International Conference on Information Retrieval & Knowledge Management.
[9] Mahyuddin K. M. Nasution, Melvani Hardi, Rahmad Syah, to appear, Mining of the social network extraction.
[10] W. Mazurczyk, M. Smolarczyk, and K. Szczypiorski, 2011, Retransmission steganography and its detection, Soft. Comput. 15 505.
[11] Surya Nepal, C’ecile Paris, and Athman Bouguettaya, 2015, Trusting the social web: Issues and Challenges, World Wide Web 18 1
[12] Marco Grassi, Eric Cambria, Amir Hussain, and Francesco Piazza, 2011, Sentic Web: A New Pradigm for Managing Social Media Affective Information, Cogn. Comput. 3 480.
[13] Elizabeth S. Vieira, and Jos’e N. F. Gomes, 2009, A comparison of scopus and web of science for a typical university, Scientometrics 81(2) 587.
[14] Wei-Jen Wang, 2013, Conservative snapshot-based actor garbage collection for distributed mobile actor systems, Telecommun. Syst. 52 647.
[15] Mike Dowman, Valentin Tablen, Hamish Cunningham, and Borislav Popov, 2005, Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News, WWW 225.
[16] James N. K. Liu, Weidong Luo, and Edmond M. C. Chan, 2005, Design and Implement a Web News Retrieval System, KES 2005 LNAI 3683 149.
[17] Mahyuddin K. M. Nasution, Rahmad Syah, Marischa Elveny, to appear, Studies on Behavior of Information to Extract the Meaning behind the Behavior.
[18] Mahyuddin K. M. Nasution, 2013, Superficial method for extracting academic social network from the Web, Ph.D Dissertation, Universiti Kebangsaan Malaysia.
[19] Mahyuddin K. M. Nasution, and Syahrul Azman Mohd Noah, 2010, Superficial method for extracting social network for academic using web snippets, In J. Yu et al. (eds.) Rough Set and Knowledge Technology, LNAI 6401, Heidelberg, Springer, 483.
[20] Mahyuddin K. M. Nasution, 2014, New method for extracting keyword for the social actor, In N. T. Nguyen et al. (eds.) Intelligent Information and Database Systems, LNAI 8397, Heidelberg, Springer, 83.
[21] Mahyuddin K. M. Nasution, and Shahru Azman Noah, 2011, Extraction of academic social network from online database, In Shahru Azman Mohd Noah et al. (eds.), IEEE Proceeding of 2011 International Conference on Semantic Technology and Information Retrieval, Putrajaya, Malaysia, 64.
[22] Mahyuddin K. M. Nasution, Syahrul Azman Mohd Noah, and S. Saad, 2011, Social network extraction: Superficial method and information retrieval, Proceeding of International
Conference on Information for Development, c2-110.

[23] Hiroshi Sekiya, Takeshi Kondo, Makoto Hashimoto, and Tomohiro Takagi, 2007, Context representation using word sequence extracted from a news corpus, International Journal of Approximate Reasoning 47 424.

[24] José A. Troyano, Vicente Carrillo, Fernando Enríquez, and Francisco J. Galán, 2004, Named Entity Recognition through Corpus Transformation and System Combination, EsTAL LNAl 3230 255.