A SENTIMENT-BASED FILTERATION AND DATA ANALYSIS FRAMEWORK FOR SOCIAL MEDIA

Norjihan Abd Ghani[1] and Siti Syahidah Mohamad Kamal[2]

[1]University of Malaya, Malaysia, *norjihan@um.edu.my*
[2]University of Malaya, Malaysia, *syaeika@gmail.com*

ABSTRACT. This paper describes a framework that explains the processes involved in the filteration and analysis of data for user generated content in social media. Previous researches have put their focus in leveraging high quality data from social media data stream, but there are many opportunities that need to be explored. This paper proposes a sentiment-based filteration and data analysis framework in identifying relevant information from data generated by users in social media. Based on the textual contents generated and spread through social media, it is assumed that each of the set of text streams/corpora might carry a sentiment associated with it regardless of its polarity bias. Due to this, the proposed framework introduces the idea of data filtering that exploits information and sentiment captured in text while at the same time adapts text analysis methods overcoming the noisy and unstructured nature of social media textual content.

Keywords: filteration, data analysis, sentiment analysis, social media

INTRODUCTION

Social media (SM) has now become an important medium of communication and interaction tools such as social networks, web communities, wikis, blogs and other online collaborative media. All the statuses, tweets, comments, posts and reviews are the user generated content that are directly created by users of the online system or service that are made available via online sites such as social media, forums, blogs and more. The amounts of these types of data that are produced continue to grow at a staggering rate. To a certain extent, this type of content might actually contain useful information; within huge amount of data, to be benefited from. Furthermore, texts in user generated content are mostly non-standard in nature as it is directly produced and created by humans. These non standards token may be deviated from standard vocabulary in term of its syntax, or even the semantic aspect of the texts. Thus, it's quite a challenge in order to extract useful information from SM content.

SOCIAL MEDIA DATA ANALYSIS

In dealing with noisy and unstructured nature of user generated content in SM data stream, text analysis is an important component that is to be adapted in filtering high quality data from this type of content. The main goal is to turn the texts into data for analysis. Text classification is one of text analysis process. It is a process that attempts to classify or categorizing element in text into categories from a predefined set. There are three approaches of categorizing text which are supervised, unsupervised and semi supervised method (Pawar & Gawande, 2012). Example of previous text classification studies are Nagaraj *et al*. (2014) who proposed

a novel approach on semantically classify text, short text classification in Twitter (Sriram, 2010) and also on hashtag classification based on topical classification on twitter (Asbagh *et al*., 2014).

Another important aspect in analyzing SM data is sentiment analysis. Sentiment analysis (SA) fundamental in identifying opinion or detecting emotion in user generated content. SA is usually used in product reviews by business holders to identify customers' feedbacks and reviews regarding their product public opinions, financial prediction or even monitoring real-world events. Accordingly, there are three levels of classification in SA, which are document-ed-level SA, sentence-level SA and aspect-level SA. Previously, there are many researches about techniques and algorithms used to analyze sentiment. Based on a survey by Medhat *et al*. (2014), these techniques and algorithms are categorized into two main approaches which are machine learning and lexicon-based approaches. Machine learning approach indicates the building of classifier from labeled instance of texts to automated the classification process (Brynielsson *et al*., 2013; Medhat *et al*., 2014). While on the other hand, lexicon-based approach is an unsupervised learning approach that exploits the use of annotation in analyzing the sentiment that rely on sentiment lexicon which is a precompiled sentiment terms collection (Medhat *et al*., 2014) or the use of opinion words (Ding *et al*., 2008) such as 'good', 'cheap', 'amazing' or even 'rich' to perform tasks.

A SENTIMENT-BASED DATA FILTRATION AND ANALYSIS FRAMEWORK

In this paper, we adapt a concept of data conditioning by Kalampokis et al., (2013). The concept refers to the process of transforming noisy and raw SM content into a high quality data. In accordance to this, this paper discusses an approach that depicts data conditioning concept through the explaination of series of processes in filtering and analysis of data in extracting useful and high-quality information from SM data stream in order for researcher to perceive real word events. Thus, to precisely projecting the intention to be represented in retrieving information from unstructured and noisy data generated in SM, we propose a framework that combines data filteration together with sentiment analysis that is not just based on the literal meaning of the filtered search term but also the sentiment polarity bias of the intended information.

Figure 1 shows the proposed framework which refine and manipulate the effectiveness of data filteration through the integration of sentiment lexicon into search term while at the same time analysis on the textual content and its sentiment will be conducted and is expected to increase the precision in filtering relevant and valuable information from SM content. The framework portrays six important phases that capture the components and tasks involve in filtering and analyzing of information from SM data. 1.) Data Input, 2.) Pre-filtering 3.) Text Analysis, 4.) Content Filtering, 5.) Sentiment Analysis and 6.) Data Output.

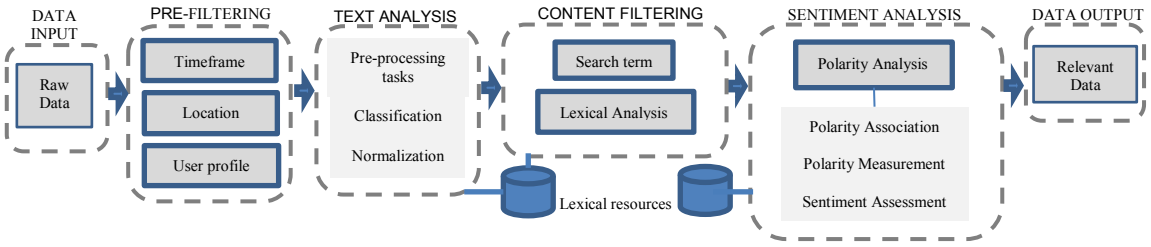

Figure 1. A Sentiment-based Filteration and Data Analysis Framework

Phase 1 : Data Input

In the first phase, the data input for the proposed framework will be extracted directly from SM medium. It is a stream of textual contents generated by users in SM which is referred as *raw data*; data yet to be processed.

Phase 2 : Pre-filtering

From the given input of *raw data* harvested from social media, a sequence of steps will be pursued for the filtration process to take place in second phase; *pre- filtering*. Accordingly, there are three aspects to be exercised in order to perform this process as described below:

i. ***Determination of timeframe***: Timeframe is the time window that is determined for the duration of collection activity period. It is refering to the specified time for the *raw data* to be filtered based on the timestamps of the generated textual content.
ii. ***Identification of location***: Another important aspect in filtering data is the *location* of the user whom the data being generated into SM data stream. In order to make the filtered data to be more specific, *location* aspects will be specified. Data generated out of the specified *location* will be filtered and discarded.
iii. ***Identification of user profile***: Characteristic of the *user profile* whom posted the textual context will be analzed and filtered.

Phase 3 : Text Analysis

This phase includes the processes involved in refining textual information that are already filtered during the previous phase. Pre-filtered data will be thoroughly processed and analyzed in order to prepare it for *content filtering* phase before further analyze its sentiment in *sentiment analysis* phase. This phase involve the process of classification and normalization of the noisy-and-raw-form text as shown in Figure 2.

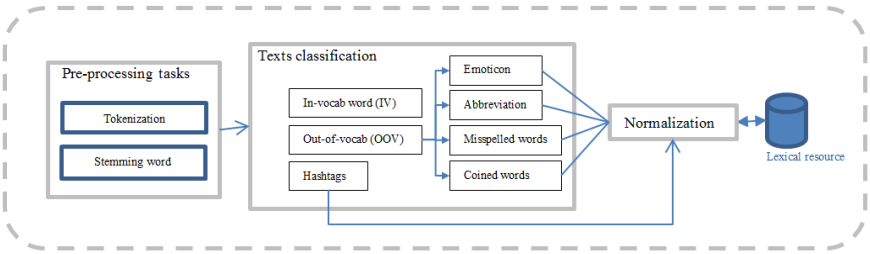

Figure 2. Text Analysis

i. ***Pre-processing task***. This task is conducted in order to prepare the text for the classification process. Pre-processing task includes *tokenization*; remove punctuations, remove-stop-word, and also *stemming word*; process word into their root word, for example 'honesty' to 'honest'. The 'cleaning' process of text from 'unnecessary' word such as tokenization, stop words and white space. Irfan *et al.* (2004) argue that this task should be conducted in order for text analysis to be successfully implemented.

ii. ***Text classification***. The purpose of this classifation is to identify categories in the texts that are to go through 'text normalization'. In this framework, the text will be classified into categories which are, abbreviation, emoticon, hashtag, in-vocabulary word (IV) , out of vocabulary word (OOV) or even symbols.

iii. ***Text normalization***. *Text normalization* process is required in order to normalize any noisy words as in text generated by users in SM. Example of text condition that needs normalization is out-of-vocabulary word(s) that might be due to missed spelling; *bokk* instead of *book*, or coined; *sleepzzz* instead of *sleep*, abbreviation or even emoticon symbols. Therefore, in order to be passed to sentiment analysis phase, the text should be normalized in order to transform it into standard words.

Phase 4 : Content Filtering

Content filtering phase is a phase where the acquired data; which have been passed through the *pre-filtering* and *text analysis* phase, being filtered based on the specified *search term*.

i. ***Initiation of search term***. The most important aspect is the *search term* that represent the focused theme of the required content. Based on our proposed framework, the *search term* comprises of two types of lexicon; the *subject* that is to be retrieved, and secondly, *sentiment lexicon* (sentiment bearing words) that represent the type of intended polarity bias of sentiment that are to be filtered. For example, in detecting the *outbreak of dengue*, one of the possible *search term* is "dengue fever". Referring to this, "dengue" is the *subject* while "fever" is the *sentiment lexicon*. Both types of lexicon should be initially specified to form the *search term* in the framework.

ii. ***Lexical Analysis*.** The *search term* will be analyzed and the subject and its related contexts will be extracted. There are two aspects that will be processed and extracted in *lexical analysis*; *domain context(DC)* of the *subject* and *supporting context(SC)* of the *sentiment lexicon*. Example of *DC* of *subject* 'dengue' are *disease, Aedes, fever, mosquito*, while example of *SC* maybe *fever, infection, diagnose* and many more. *SC* can be either noun, verb, adjective or adverb as part-of-speech (PoS) tagging. A stream of texts will be filtered based on the *DC* formulated from the *subject*. Both collection of stream of texts is filtered based on *DC* and *SC* of the *sentiment lexicon* will be prepared to be assessed based on lexical resource, such as WordNet (Miller, 1995) of its sentiment polarity in next phase. It will eventually determine its bias and be used as the basis of the user intention to match and measure the sentiment polarity bias of the obtained corpora. For instance, referred to the example of *search term* above 'dengue fever'. In this analysis, the *subject* of the *search term* is 'dengue' while the *sentiment lexicon* is 'fever'. Base on the *search term*, the *DC subject* is analyzed in order to retrieve related corpora.

For example, Table 1 below shows the sample of corpora/tweets filtered through previous process and retrieved based on *DC* of the *subject* specified in *search term*. The filtered corpora might not necessarily consist the literal word 'dengue' but it might contain the related context of the domain 'dengue' such as 'Aedes' which can be referred to *Tweet 2*.

Table 1. Example of corpora related to *subject* 'dengue'

ID	Tweets
Tweet1	"Scientists here discover important role of a type of white blood cell which can destroy dengue cells.
Tweet2	"It's totally sad that I'm so so so busy and having a fever and I missed a lot of things ☹☹☹☹. Its all Aedes fault!"
Tweet3	"it is a relief! his friend doesn't get dengue fever!"
Tweet4	"Earghhh..no way! Here comes the real deal! Down with fever ☹ #dengue"

Phase 5 : Sentiment Analysis

In this phase, the analysis of sentiment polarity will be exercised. There are two types of input that is to be processed in this phase; the collection of text corpora that is filtered based on *DC* of the *subject* in *content filtering* phase, and the *SC* of the *sentiment lexicon* which extracted in previous phase. During *polarity analysis*', the valuation of sentiment will be done to the specified *sentiment lexicon* and also the stream of texts that have gone through previous analyses. There are three processes involve as following:

i. ***Polarity association***. *SC* of the *sentiment lexicon* provided in the specified *search term* in *content filtering* phase, will be referred and based on a lexical resources such as SentiWordNet 3.0 (Baccianella *et al*., 2010) in order to determine the bias of its polarity association which can either positive, negative or neutral. The polarity bias is be used as the indication of the intended sentiment projection to be matched with the polarity bias of the acquired corpora later. It is referred to as *SC polarity bias*. Example of *SC* of the sentiment lexicon provided in *search term* "dengue fever" is 'fever' when it is referred to a selected lexical resource (Baccianella *et al*., 2010), the lexicon is being associated to *negative* polarity bias.
ii. *Polarity measurement*. Polarity score will be assigned to the acquired and filtered stream of text resulted *content filtering* phase. The calculation is conducted by calculating the overal sentiment score for each corpus in order to determine its polarity bias. The sentiment polarity bias should be regarded as a whole for each corpus, thus, the sentiment will be analyzed at the document level. The assignment and calculation of overall polarity bias will be based on a lexical resource (Baccianella *et al*., 2010).
iii. *Sentiment assessment*. Sentiment polarity bias of each accessed corpus will be measured and *matched* against the *SC polarity bias*. As an example, after going through filteration and analysis process, the polarity bias for all of the corpora (as being sampled in Table 1) is identified and assigned to each corpus (Table 2).

Table 2. Eample of polarity bias assigned for each corpus

ID	Polarity Bias
Tweet1	*neutral*
Tweet2	*negative*
Tweet3	*positive*
Tweet4	*negative*

Therefore in order to finally derive with relevant result, the polarity bias of all the corpora being compared and matched with *SC polarity bias* which is *negative* as determined before. Hence, *Tweet2* and *Tweet4* are the matched as they measured as *negative* polarity bias. On the other hand, *Tweet1* and *Tweet3* which resulted as *neutral* and *positive* polarity bias do not really represent the outbreak of dengue.

Phase 6 : Data Output

In the final phase, the matched corpora from previous phase will be presented as the output resulted from the series of tasks and processes in the propose framework. This will be the most relevant result formed based on the filteration and analysis phases conducted to process raw data from data input phase.

CONCLUSION

This paper propose a framework for filtering and analysis of data in order to obtain useful and high quality information from SM text stream through the implementation of sentiment-

based approach. The proposed framework will improve the relevance and precision in filtering and analysis of useful information from SM.

Acknowledgement

This research was supported by the UMRG Programme-AET (Innovative Technology (ITRC)) at the University of Malaya (RP029A-14AET).

REFERENCES

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, *0*, 2200–2204. http://doi.org/citeulike-article-id:9238846

Brynielsson, J., Johansson, F., & Westling, A. (2013). Learning to classify emotional content in crisis-related tweets. *2013 IEEE International Conference on Intelligence and Security Informatics*, 33–38. http://doi.org/10.1109/ISI.2013.6578782

Ding, X., Ding, X., Liu, B., Liu, B., Yu, P. S., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the International Conference on Web Search and Web Data Mining - WSDM '08*, 231. http://doi.org/10.1145/1341531.1341561

JafariAsbagh, M., Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2014). Clustering memes in social media streams. *Computers and Society; Learning; Physics and Society*, 25. Retrieved from http://arxiv.org/abs/1411.0652

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559. http://doi.org/10.1108/IntR-06-2012-0114

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. http://doi.org/10.1016/j.asej.2014.04.011

Miller, G. a. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. http://doi.org/10.1145/219717.219748

Nagaraj, R., Thiagarasu, V., & Vijayakumar, P. (2014). A novel semantic level text classification by combining NLP and Thesaurus concepts, *16*(4), 14–26.

Pawar, P. Y., & Gawande, S. H. (2012). A Comparative Study on Different Types of Approaches to Text Categorization. *International Journal of Machine Learning and Computing*, *2*(4), 423–426. http://doi.org/10.7763/IJMLC.2012.V2.158

Roy, S., Dhar, S., Bhattacharjee, S., & Das, A. (2013). A Lexicon Based Algorithm for Noisy Text Normalization as Pre-Processing for Sentiment Analysis, 2319–2322.

Sriram, B. (2010). Short Text Classification in Twitter to Improve Information Filtering.

Zablith, F., Antoniou, G., Aquin, M., Flouris, G. O. S., Kondylakis, H., Motta, E., & Sabou, M. (2013). *Ontology Evolution : A Process Centric Survey*. *The Knowledge Engineering Review* (Vol. 00). http://doi.org/10.1017/S000000000000000