**ISBN** 978-981-11-0008-6
**Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)**
*Tokyo, Japan, 17-19 June, 2016, pp. 575-579*
doi: 10.18178/wcse.2016.06.098

# Automatic Summarization from Indonesian Hashtag on Twitter Using TF-IDF and Phrase Reinforcement Algorithm

Willyh Hariardi, Novita Latief, David Febryanto and Derwin Suhartono[+]

Bina Nusantara University, School of Computer Science, Jakarta, Indonesia

**Abstract.** The objective of this research is to produce a summary about what is currently happening from Indonesian hashtag on Twitter. Combination of TF-IDF (term frequency-inverse document frequency) and Phrase Reinforcement Algorithm are used as the methodology to do the automatic summarization. We use 2 sentences as the final summary result. It contains all essential information given by Twitter data. At the end of this paper, we describe the evaluation result by analyzing result using Precision and ROUGE. Based on the result, we conclude that TF-IDF and Phrase Reinforcement Algorithm can successfully generate summary and it works well enough on hashtags that do not have such lot variants of the word. Generally, summary results quality is quite low because the data still contains too much noise. The precision is 0.327 and ROUGE-1 is 0.3087.

**Keywords:** summary, hashtag, twitter, phrase reinforcement algorithm, automatic summarization, tf-idf.

## 1. Introduction

Twitter is a social media that has the highest number of users in Indonesia after Facebook. It is included as a part of social service category named microblogging. Microblogging is a blog that enables users to write a short text less than 200 words length. Twitter is quite different to other social media, especially in the message content. It is interactive, with open-source API, user can write, read, and also send a short message that is called as a tweet. A tweet has a maximum length of 140 characters and it can be seen by all the Twitter users. Therefore, Twitter is called as a microblogging social media.

One interesting point for Twitter user is that user can access hashtag (#) locally or globally (worldwide). Hashtag is a symbol used in marking or grouping topic on Twitter. It eases user to search tweets by topics. Hashtag is closely related to trending topic, yet not all hashtags is able to be trending topics. Trending topic is a topic in a short phrase which is most discussed by Twitter users. However, the presence of many hashtags that contains unnecessary information becomes problem. Sometimes it does not correspond with the hashtag tweets thus making it ineffective. As a result, Twitter users meet difficulties to find information inside the hashtag. Information adopted from hashtag can also have overloaded condition due to the number of tweets which always grows over time (real-time). Character length limitation also makes hashtag contains many words or phrases that are not standardized, concise and normal. Thus, it is so difficult to understand what information is going to be conveyed.

The problem described above can be solved by using summarization. Summarization (Lloret, 2009) is a technique for generating document that contains important information from some documents. Automatic summarization can be used to give comprehensive explanation and help users to understand the meaning of hashtag on Twitter.

## 2. Related Work

[+] Corresponding author. Tel.: +62215345830; fax: +62215300244.
*E-mail address*: dsuhartono@binus.edu.

Focus of summarization is to take out important information from data source or information in large quantities and display them on a consistent, short and representative summary (Kageback et al., 2014). Commonly, a proper summarization has to be done by human. However, the issue of huge amount of information or data sources makes human is no longer able to produce manual summary. It needs much time and energy. Summarization has two main techniques; they are abstractive summarization and extractive summarization. The most widely used approach is extractive summarization. Extractive summarization is a technique for making a summary by getting words or sentences from the original document (Cheung, 2008). This technique has a result which looks similar with the original sentence structure.

Research in summarization has been started since 1950s and it continues until now. Summarization refers to a text or document whether it summarizes single or multiple documents, in order to produce information that represents the original document content. Automatic text summarization in Indonesian language has been done by Silvia et al. (2014). Aside from that, automatic summarization on microblog is a quite new research area. Sharifi et al. (2010) used TF-IDF and Phrase Reinforcement Algorithm in text summarization which are applied for tweet summarization. In other related research, Winatmoko and Khodra (2013) used TF-IDF to summarize Indonesian trending topic. Automatic summarization technique that we proposed combines two algorithms: TF-IDF and Phrase Reinforcement Algorithm. TF-IDF counts weight in every sentence inside the document. Cosine similarity is used after TF-IDF to look for relation between sentences before they are processed in the summarization. The other technique, Phrase Reinforcement Algorithm (PRA) is an algorithm to generate summary by searching for the most common or most often arise words by building a word graph. Summary is generated by the topics that have clear information and does not have a lot of variety in the written data. It makes PRA can work well and produce a good summary.

## 3. Methodology

We separate all steps in this research into two main phases; they are pre-processing phase and summarization phase. Pre-processing phase is a process that covers streaming, tokenization, normalization, TF-IDF, and cosine similarity; while summarization phase is a process that involves getting sentence, initializing token, making graph, calculating token weight, searching graph using Depth-First Search (DFS), and generating summarization result. Figure 1 explains us about pre-processing phase.

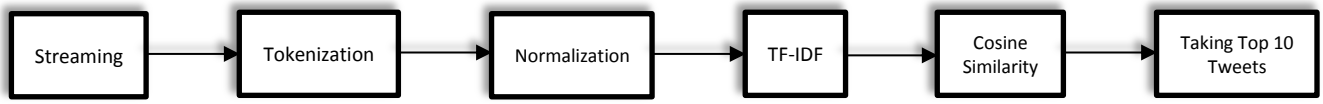


Fig. 1: Pre-processing phase.

Streaming is the initial process which gets all Twitter data. Data was taken using Twitter API (Application Programming Interface) that was provided by Twitter. We took tweets which used Indonesian language. It also a real-time process. Tokenization is a process of dividing the sentence or document into several tokens so that they can be used for the next process (Attia, 2007). Normalization is the process of transforming abbreviated word into a normal word so it can form a structured sentence. In this experiment, we use the normalization scheme as proposed by Naradipha and Purwarianti (2012). TF-IDF is a process of calculating weight for every sentence based on term frequency inside one document and whole documents. Furthermore, cosine similarity is a process of calculating relation between sentences to get the best result. The range of result value in cosine similarity is 0 to 1. If the result is 1, then the relation between the sentences is perfectly connected and vice versa. Lastly, 10 highest results from previous process (cosine similarity) were taken for summarization phase.

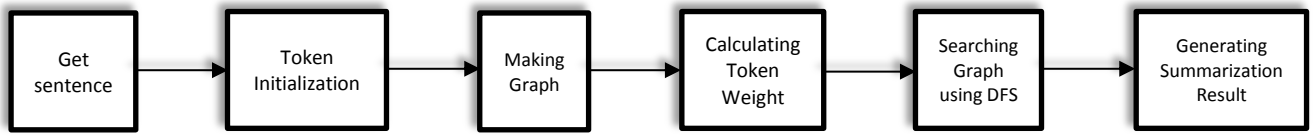


Fig. 2: Summarization phase.

Result from pre-processing phase is used in summarization phase as attached in Fig. 2. Every token in each sentence is initialized before entering the array. This process consists of assigning start node, end node, and root node. Making graph is a process of building graph which is the main process of PRA. PRA builds the graph for showing the word position before and after the topic. The graph has left and right node. The left node contains token that appears before the topic, while the right node contains the token that appears after the topic. Subsequently, weight is calculated for each token. Later, it continues with process of searching graph using DFS. It is a process of finding the best path using Depth First Search (DFS). It looks for the highest weight at first but if only the weights of all tokens are the same, it will look for the distance of each node. Finally, the summarization result is generated. It chooses the best path result from previous process and it is selected as the final result or output.

## 4. Results and Discussion

We collected tweets whose date range was from 12th to 26th January 2016. They consist of 30 hashtags which was only Indonesian language tweets. To make the steps can be understood easier, we give an example of processing one (1) hashtag from the beginning step till the end.

The following sample was taken from Indonesian Trending Topic #1, which it was '*Koruptor'* (literally means corruptor). The data was taken at the end of October 2015 using Twitter API. Fig. 3 shows the shortened data after be processed through pre-processing phase.

| No | Data |
|---|---|
| 1 | Selain lemahkan KPK, DPR juga usulkan draf RUU untuk ampuni koruptor.<br>*(literally means: Besides weakening Corruption Eradication Commision (KPK), House of* |
| 2 | Kalau koruptor sudah diampuni, usulkan juga pembuatan buku dan seminar.<br>*(literally means: If the corruptor was forgiven, then propose book and seminar making as well)* |
| 3 | Kontroversi DPR: Pasal Kretek, Revisi UU KPK sampai Pengampunan koruptor.<br>*(literally means: House of Representative (DPR) Controversy: Clove Subsection, revision of* |
| 4 | Ketua PDIP: setelah diampuni, koruptor Akan Jadi Orang Baik.<br>*(literally means: Chairman of PDIP: After be forgiven, corruptor will be a good people)* |

Fig. 3: Example of shortened data after through the pre-processing phase.

After getting the result above, we go to summarization process using PRA to construct graph according to the previous phase that is explained before. The word graph is shown in Fig. 4.

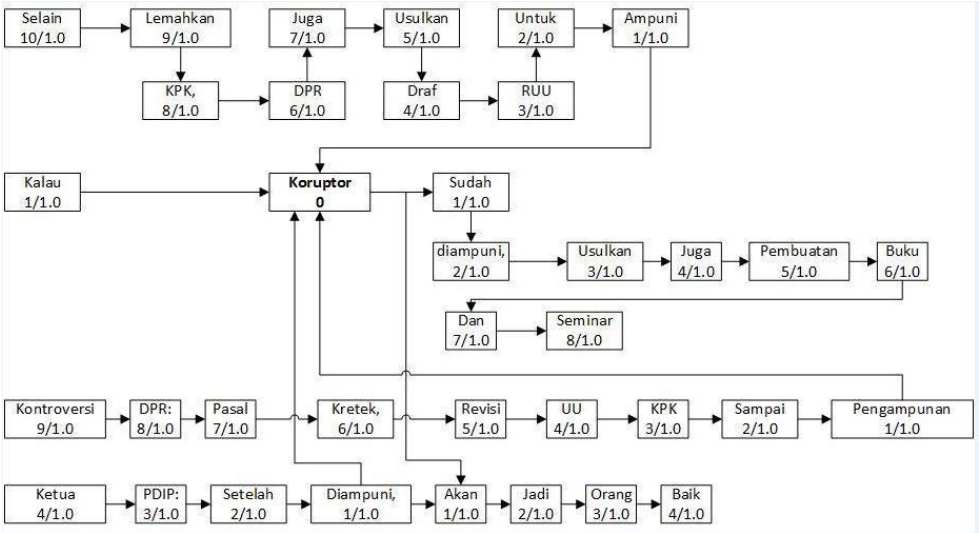


Fig. 4: The word graph.

The next step is to specify the best start node. It is necessary to check weight on the first word from each sentence on graph. If there is a word that has the same weight, then we need to check the distance between the word and root. After that, specify the best path as the summarization result. When specifying the best path, we use DFS and weighting process so there will be no repetition in the same node. It can be seen in Fig. 5.

We evaluated 30 hashtags by calculating their precision and ROUGE-1. In calculating precision and ROUGE-N for each hashtag, we use manual summary made by 26 person involving one (1) language expert in it. The manual summary was compared with the automatic summary which was made by our system. Some examples of the result are shown in Fig. 6.

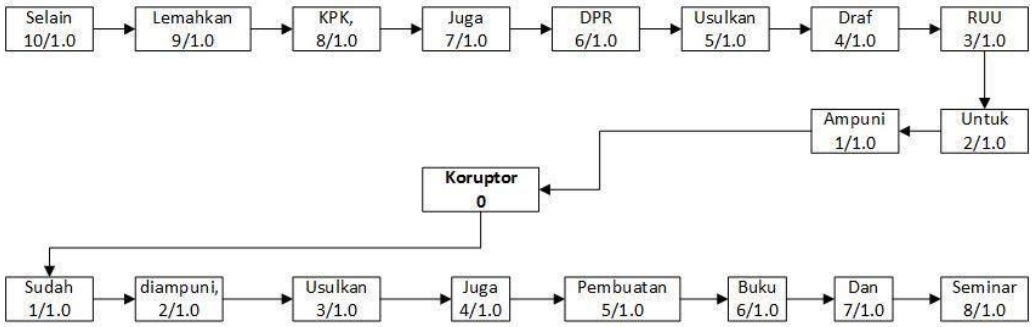


Fig. 5: Result shown as the best path from search results using DFS.

The average values of precision and ROUGE-1 from 30 hashtags (only 10 samples are provided in figure 6) are relatively low, resulted on 0.327 and 0.3087. We conclude that the higher the precision value, the less relation between the automatic summary and manual summary. Meanwhile, the higher the value of ROUGE-1, the more important information in the manual summary that contained in the automatic summary. Ideally, good summary is one with the high value of ROUGE-1 and precision. From the evaluation above, it shows that the increment of precision value on each hashtag will be followed by the decrement of ROUGE-1 on each hashtag.

| No | Hashtag | Precision | ROUGE-1 |
|---|---|---|---|
| 1 | *#revisiuuterorisme* | 1.0 | 0.511 |
| 2 | *#gafatar* | 0.923 | 0.333 |
| 3 | *mempawah* | 0.92 | 0.6 |
| 4 | *setyo novanto* | 0.69 | 0.61 |
| 5 | *partai perindo* | 0.5 | 0.75 |
| 6 | *bripka taufik* | 0.57 | 0.62 |
| 7 | *#korupsi* | 0.57 | 0.44 |
| 8 | *#BinusEvent* | 0.556 | 0.625 |
| 9 | *Zheng si wei* | 0.545 | 0.5 |
| 10 | *Soekarno Hatta* | 0.5 | 0.129 |

Fig. 6: Sample of precision and ROUGE-1 values from the experiment.

## 5. Conclusion

Based on the results and discussion, we conclude that:

- The algorithm can produce summary that represents the important information from hashtag. The summary result will be better if tweets are relevant with the hashtag. In certain circumstances, hashtags do not represent the tweet content.
- Phrase Reinforcement Algorithm can work well on data that do not have lots of word variety.
- The summarization algorithm was tested using precision and ROUGE-1 on 30 hashtags from Twitter with Indonesian language. The average values of precision and ROUGE-1 were 0.327 and 0.3087. This shows us that the quality of summary result is quite low. This is because most of the data still contains a lot of noise. Algorithm which can resolve the noise will be advantageous.

## 6. Acknowledgement

1. Mr. Marcus Bambang Walgito as Indonesian language lecturer in Bina Nusantara University.

2. All respondents that participate in the evaluation phase.

## 7. References

[1] E. Lloret, and M. Palomar. Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. *International Journal of Informatica*. 2009, 34 (1).

[2] M. Kageback, O. Mogren, N. Tahmasebi, and D. Dubhashi. Extractive Summarization using Continuous Vector Space Models. *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL*. 2014, 31–39.

[3] J.C.K. Cheung. Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection. *Thesis of University of British Columbia, Minors in Linguistics and German*. 2008.

[4] Silvia, P. Rukmana, V.R. Aprilia, D. Suhartono, R. Wongso, and Meiliana. Summarizing Text for Indonesian Language by Using Latent Dirichlet Allocation and Genetic Algorithm. *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2014), Yogyakarta, Indonesia*. 2014, pp. 148-153.

[5] B. Sharifi, M.A. Hutton, and J.K. Kalita. Experiments in microblog summarization. *In Proc. of IEEE Second International Conference on Social Computing*. 2010.

[6] Y.A. Winatmoko, and M.L. Khodra. Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation. *The 4th International Conference on Electrical Engineering and Informatics*. 2013, pp. 1027–1033.

[7] M. Attia. Arabic tokenization system. Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, (CIR' 07), ACM, USA, 2007, pp. 65-72.

[8] A. R. Naradhipa, and A. Purwarianti. Sentiment classification for Indonesian message in social media. *International Conference on Cloud Computing and Social Networking* (ICCCSN), Bandung, Indonesia, 2012. pp. 1-5.