Manuscript Details

Manuscript number	JKSUCIS_2017_231_R1
Title	Trending Topics Detection of Indonesian Tweets Using BN-grams and Doc-p
Short title	Trending Topics Detection of Indonesian Tweets Using BN-grams and Doc-p
Article type	Full Length Article

Abstract

Researches on trending topics detection, especially on Twitter, have increased and various methods for detecting trending topics have been developed. Most of these researches have been focused on tweets written in English. Previous researches on trending topics detection on Indonesian tweets are still relatively few. In this paper, we compare two methods, namely document pivot and BN-grams, for detecting trending topics on Indonesian tweets. In our experiments, we examine the effects of varying the number of topics, n-grams, stemming, and aggregation on the quality of the resulting trending topics. We measure the accuracy of trending topics detection by comparing both algorithms with trending topics found in local news and Twitter trending topics. The results of our experiments show that using ten topics produces the highest topic recall; that using trigrams in BN-grams results in the highest value topic recall; and that using aggregation reduces the quality of trending topics produced. Overall, BN-grams has a higher value of topic recall than that of document pivot.

Keywords	trending topics detection; Twitter; BN-grams; document pivot
Taxonomy	Data Mining, Software Engineering

Corresponding Author	Indra Indra
Corresponding Author's Institution	Universitas Budi Luhur
Order of Authors	Indra Indra, Edi Winarko, Reza Pulungan

Submission Files Included in this PDF

File Name [File Type]

Covering-Lettter (Dear JKSUCI editors).pdf [Cover Letter]

Jawaban-Review-21-12-2017-v1.docx [Response to Reviewers (without Author Details)]

TitlePage.pdf [Title Page (with Author Details)]

Jurnal-Komparasi-23-Des2017-v3-rev.docx [Manuscript (without Author Details)]

author agreement-5-9-17.pdf [Author Agreement]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Dear JKSUCI editors,

Herewith, we submit a manuscript titled “Trending Topics Detection of Indonesian Tweets Using BN-grams and Doc-p” for consideration for publication by the Journal of King Saud University - Computer and Information Sciences, Elsevier.

We confirm that this work is original and has not been submitted/published earlier in any journal and is not being considered for publication elsewhere. All authors have seen and approved the manuscript and have contributed significantly to the paper.

In this paper, we report on comparison BN-Grams and Doc-P to trending topic detection that has two main contributions. First, a comparison of BN-Grams and document pivot with LSH clustering methods applied on tweets in Indonesian language. Second, An analysis on the effect of the usage of n-grams variations in BN-Grams method on the quality of the result of trending topics in tweets in Indonesian language. The n-grams types used in this study are unigrams up to six-grams. In our experiments, we examine the effects of varying the number of topics, n-grams, stemming, and aggregation on the quality of the resulting trending topics. We measure the accuracy of trending topics detection by comparing both algorithms with trending topics found in local news. The results of our experiments show that using ten topics produces the highest topic recall, that using trigrams in BN-grams results in the highest value topic recall, and that using aggregation reduces the quality of trending topics produced. Overall, BN-grams has a higher value of topic recall than that of document pivot. This paper should be of interest to readers in the areas of data mining, text mining especially for topic detection and tracking (TDT) on social media.

Thank you for your consideration of this manuscript. Please address all correspondence concerning this manuscript to me at indra@mail.ugm.ac.id or indra@budiluhur.ac.id

Sincerely,

Indra

OUR RESPONSES TO THE COMMENTS OF EDITORS AND REVIEWERS

-Reviewer #1:

1. The authors propose a method for detecting trending topics on Indonesian tweets. In their experiments, they examine the effects of varying the number of topics, n-grams, stemming, and aggregation on the quality of the resulting trending topics.
2. Modeling of the problem with some proves are good presented in this paper. The author shows clarity of expression and the technical level of the paper is well presented. The paper also has an authoritative list of references.
3. There is still lacking on survey and analysis of related work especially on the benchmark approach and discuss their strength and weaknesses. There is no verification done between the proposed method and benchmark method.

Response:

In Section 2 (Related Work), we have revised the related work about document pivot and explained the verification done between the proposed method and benchmark method. In our paper, we only applied BN-Grams and Doc-P for Indonesian tweets. In the work of Aiello et al. (2013), BN-Grams and Doc-p have been compared but for English tweets. Detection of the trending topics in Indonesian tweets requires different stemming and stop words during the preprocessing stage. While Aiello et al. (2013) uses Porter stemming and English stop words, in this study, we use Adriani et al.'s (2007) stemming and Tala's Indonesian stop words (Tala, 2003). The impact of the use of different stemming and stop words will be investigated in this paper. (We have revised Introduction, accordingly).

4. Twitter has its own trending detection mechanism, which is one of the several natural and essential features of Twitter. However, the paper doesn't mention Twitter's own trending feature at all, which means it also doesn't mention what differentiate this paper with Twitter's trending detection, or how is the performance of this paper's approach/algorithm, compared to Twitter's.

Response:

There are two differences between the trending topics of our paper and Twitter's. First, the trending topics of our method is generated for general topics, but in this paper, we want to produce trending topics especially in political events. Second, trending topics of our method are shown with a set of keywords, while Twitter's trending topics are formatted with hashtags and a set of keywords.

We have compared the trending topics of our method with Twitter's trending topics in Section 4.4.5.

-**Reviewer#2:**

1. I think it is interesting approach for analyzing Twitter messages for obtaining the emotional reaction of persons and their sentiments. The graphics have a good explanation There are an excellent comparison between Doc-p. BN-grams and Doc-p in all three datasets.

Response:
Thank you very much for the suggestion. The suggestion for analyzing Twitter messages for obtaining the emotional reaction of persons and their sentiments will be studied in our next paper. In this paper, our scope is for comparing BN-Grams with Doc-p in Indonesian tweets for trending topics detection.

-Reviewer#3:

Overall this is a well-written paper on an interesting and important topic. However, the paper would benefit greatly from improvement in above mentioned aspects.

The manuscript would benefit from the following considerations:
1) The benchmark and comparisons
a. The benchmark with local news needs justification: The main thrust of the evaluation rests on comparing the results of the trending algorithms investigated by the authors with trending topics found in local news, as a bench mark. This is a reasonable approach but it is only one side of the story. There may well be trending topics that originate in social media and then get reported in mainstream news media. e.g. particularly in political news and elections. [See: ……] There is a two relationship: items in mainstream news can be reported in social media and vice versa, in other words.

Response:
We have compared trending topics of our method with Twitter's trending topics in Section 4.4.5.

The suggestion has been added in Section 5 (The experiments also indicate that trending topics generated by our method and trending topics in local news complement each other. The trending topics of our method form a material for the trending topics in local news, beside the tweets collected based on the trending topics in local news. There are two relationships: the trending topics in local news can be reported in trending topics of our method and vice versa.).

We have revised Section 4.1 (The ground truth consists of ten topics, built based on the trending topics in local news. The trending topics in local news is the most read news by news readers as called the most popular news. Ground truth contains a set of keywords based on the most popular news taken the next day after the trending topic is detected.).

There does not seem to be an acknowledgement of the above, in the paper. There should be some discussion and justification as to why the comparison with local news has been chosen.

Although it may be less likely, if there is any comparison of topics focused on breaking news originating from social media then that would be very good. Otherwise, a strong rationale of the approach, along with some acknowledgement of improvements possible / weaknesses would be good.

Response:
Breaking news is news that has either just happened or is currently happening (Phuvipadawat & Murata, 2010). While, trending topic, which is also called emerging trend or emerging topic (Becker, 2011), is a topic area that is growing in interest and utility over time (Kontostathis, Galitsky, Pottenger, & Phelps, 2004). Based on the definitions, trending topics and breaking news are two different meanings. Therefore, our paper does not contain a comparison of trending topics with breaking news.

b. The use of one expert needs justification and background: The paper states that one expert was used. Why only one and not 2 or 3 as is sometimes the case, and use the commonly agreed ones? This may be an issue of resource but awareness of this is important. What is the background of the expert. Is it an Indonesian news agency worker? An editor? Or is it someone expert in tweet headlines? Or is it a computer science student (in which case further discussion is needed)? Or is it a journalism student? Etc.

Response:
We have clarified in Section 4.3 (In this study, we employ two experts, namely a lecturer with a Ph.D. in political sciences, and an Indonesian news agency worker who has contributed to the three most popular news websites in Indonesia (Kompas.com, Tempo, and Detik.com).).

2) The language and cultural context: There is some good relevant referencing to other work on Indonesian tweets. However, more could be said on what is interesting to investigate and why with Indonesian tweets. Is there something about the language and grammar that we could learn and extrapolate to other languages? Is there something about the topics investigated that we should pay attention to when making comparisons in other languages, including English, for example?

- However, more could be said on what is interesting to investigate and why with Indonesian tweets

Response:
We have added in Section 1, Introduction (Detection of the trending topics in Indonesian tweets requires different stemming and stop words during the preprocessing stage. While Aiello et al. (2013) uses Porter stemming and English

stop words, in this study, we use Adriani et al.'s (2007) stemming and Tala's Indonesian stop words (Tala, 2003). The impact of the use of different stemming and stop words will be investigated in this paper.).

- Is there something about the language and grammar that we could learn and extrapolate to other languages?

Response:
We have added in Conclusion (The pattern in Indonesian writing is similar to the language pattern in the Indonesian subgroups: Melayu (Malaysia), Malagasy (Madagascar), Formosa and Philippines (Darmini, 2012). Therefore, Indonesian trending topics research has an excellent opportunity to be applied to trending topics in Indonesian language subgroup. Also, experimental results show that topics generated by BN-grams and Doc-P from Indonesian tweets do not have subject, predicate, object and adverb (SPOK) pattern, as Indonesian sentences should be; this will become a challenge for future research.).

Is there something about the topics investigated that we should pay attention to when making comparisons in other languages, including English, for example?

Response:
The focus of this study is limited only to comparing trending topics of our method with trending topics of mainstream news media and social media based on Indonesian tweets. Comparisons with other languages other than Indonesian will serve as input for future research.

3) Tweet examples: The paper would benefit greatly from more sample tweets, topics, and headlines – both in Indonesian and where possible English translation and phonetics. Table 3 seems to be the only example provided and that is only in Indonesian so it is difficult to conclude topically relevant assessments for the readers with limited or no knowledge of Indonesian.

Response:
We have revised in Table 3.

The language and style of the paper is clear.

Structurally the paper would benefit from a more focused discussion and analysis section prior to conclusions (which should be a highlight/summary of what has already been said). In other words, there could be a clear account of the results presented with the figures and tables as well as comparison with others' works and the implications. For example, this is done with the comparison and of results relating to stemming in section 4. More of this type of cross-referencing would benefit the paper. This can be done either within relevant parts in section 4 and/or by adding another section before conclusion as appropriate.

Response:
We have added discussion and analysis in Section 5.

Literature: There may be a few more papers the authors find relevant in last few years' ECIR, SIGIR, and JASIST papers on social media and news where the topic has featured more lately too.

Response:
During literature study, we did not find papers that are related to document pivot and feature pivot in the journals.

More information on the datasets is needed. Over what period were the tweets gathered for each collection? Were there any omitted and if so how/why? The authors state that keywords from 'political figures, executive agencies, …. ' were used. Were the same keywords applied for all collections or were some keywords used for one or two of the collections and not the other(s)? If so, why?

Response:
We have clarified in Section 4.1 (The datasets are constructed based on keywords from political figures, executive agencies, legislative assembly, judicial bodies, political event hashtags, names of the governor or vice governor candidates, and names of political parties. Moreover, in the absence of keywords relevant to political events emerging during the period of dataset collection, new keywords were added.).

The possible effect of the nature of the collection on any results discussed in section 4 should then be discussed accordingly too.

Response:
1. We have revised in Section 4.4.1 (Topic recall values increase as the number of topics increases. P1 produces more certain trending topics than P2 and P3. This is because P1 has fewer tweets and also has relatively shorter drawing period than P2 and P3. Thus, P1 has less difficulty and higher accuracy than P2 and P3.).
2. We have explained in Section 4.4.2 (Fig. 4 depicts the accuracy of trending topics detection by varying the n-grams used. The n-grams used are unigram to sixgrams. Trigrams produces higher accuracy than other n-grams. In P1 and P2 trigrams has the highest topic recall compared to unigram and bigrams; in P3 however trigrams produces lower topic recall than bigrams. This is because the number of tweets is smaller and the period of tweet collection is also shorter for P1 and P2 compared with those of P3.).

The authors refer to campaign accounts 'always generating same tweet for a specific purpose'. Work on the Scottish Referendum showed that different campaigns used their accounts and topics differently (see Pedersen et al further below).

Response:
We have clarified in Section 1, Introduction.

Actually, campaign accounts do not always generate the same tweet. However, campaign accounts in Indonesia generate different tweets, but still have the same meaning for a particular campaign objective. While in Pedersen et al.'s research (Pedersen et al., 2015a), Twitter was used to see the public response of a political debate aired on television.

Also, as in point 1) mentioned early above, what are the limitations of a collection that is based on news title form mainstream media. (i.e. this disadvantages consideration of news items discussed in society because of the good exposure in social media – even if it is not reported in mainstream media. So what is the implication of this?)

Response:
We have added in Section 5 (The critical finding in our experiments is the contrary between trending topics of our method and the trending topics in local news, which is evident in the local election of governor and vice governor of Jakarta in 2017. The trending topics generated by our method is a direct opinion of the society without any manipulation. Therefore, the trending topics produces by our method can be an early warning system for political events in Indonesia.).

Further clarity on the topics. The paper refers to 'The objective of the first category is to generate topics from an event, political movements, and the rate of urbanization..' (p2). Is this a reference to a category of topics that the authors have created/used in their work described or is this a category of topics mentioned in the literature that authors simply want to refer to or both? The authors refer to 'disruptive events, popular topics, and daily routines' (earlier on in p1). How do these categories relate to those? What exactly are the categories for each collection?

Response:
We have clarified in Section 1, Introduction (In this research, we want to apply BN-grams and document pivot for trending topics detection for general use, even though our case studies are limited only to detection of trending topics in the political field.).

Hence, as mentioned above, overall this is a well-written paper on an interesting and important topic. However, the paper would benefit greatly from improvement in above mentioned aspects.

References from other domains (media and journalism) that the authors may find relevant:

Pedersen, S., Baxter, G., Burnett, S., Göker, A., Corney, D., and Martin, C. (2015). Backchannel Chat: Peaks and Troughs in a Twitter Response to Three Televised Debates during the 2014 Scottish Independence Referendum Campaign.

Response:
We have added (Pedersen et al., 2015) in Section 1, Introduction.

Schifferes, S., Newman, N., Thurman, N., Corney, D.P.A., Göker, A. and Martin C. (2013) "Identifying and verifying news through social media: Developing a user-centred tool for professional journalists," The Future of Journalism Conference, 12-13 September 2013, Cardiff, UK.

Response:
Based on the literature study, this paper is too far from our topic.

Trending Topics Detection of Indonesian Tweets Using BN-grams and Doc-p

Indra*, Edi Winarko, Reza Pulungan

Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia

ABSTRACT

Researches on trending topics detection, especially on Twitter, have increased and various methods for detecting trending topics have been developed. Most of these researches have been focused on tweets written in English. Previous researches on trending topics detection on Indonesian tweets are still relatively few. In this paper, we compare two methods, namely document pivot and BN-grams, for detecting trending topics on Indonesian tweets. In our experiments, we examine the effects of varying the number of topics, n-grams, stemming, and aggregation on the quality of the resulting trending topics. We measure the accuracy of trending topics detection by comparing both algorithms with trending topics found in local news and Twitter trending topics. The results of our experiments show that using ten topics produces the highest topic recall; that using trigrams in BN-grams results in the highest value topic recall; and that using aggregation reduces the quality of trending topics produced. Overall, BN-grams has a higher value of topic recall than that of document pivot.

Keywords: trending topics detection, Twitter, BN-grams, document pivot.

1. INTRODUCTION

Trending topic, which is also called emerging trend or emerging topic (Becker, 2011), is a research area that is growing in interest and utility over time (Kontostathis et al., 2004). Trending topics can be categorized into three types (Cvijikj and Michahelles, 2011): disruptive events, popular topics, and daily routines. Disruptive events are events or phenomena that draw global attention, such as earthquakes and tsunami. Popular topics might be related to some past events, celebrities, products, or brands that remain popular over a long period of time, such as Coca Cola and Michael Jackson. Daily routines are trending topics related to some common phrases, such as "good night" or birthday wishes. In this paper, we want to generate trending topics based on disruptive political events in Indonesia.

Based on textual content of the news, there are two main approaches for detecting trending topics, namely trending topics detections based on document pivot, feature pivot, and probabilistic topic model (Aiello et al., 2013; Petkos et al.,

*Corresponding author: Indra, Faculty of Information Technology, Universitas Budi Luhur, 12260 Jakarta, Indonesia. E-mail: indra@budiluhur.ac.id, indra@mail.ugm.ac.id

Trending Topics Detection of Indonesian Tweets Using BN-grams and Doc-p

ABSTRACT

Researches on trending topics detection, especially on Twitter, have increased and various methods for detecting trending topics have been developed. Most of these researches have been focused on tweets written in English. Previous researches on trending topics detection on Indonesian tweets are still relatively few. In this paper, we compare two methods, namely document pivot and BN-grams, for detecting trending topics on Indonesian tweets. In our experiments, we examine the effects of varying the number of topics, n-grams, stemming, and aggregation on the quality of the resulting trending topics. We measure the accuracy of trending topics detection by comparing both algorithms with trending topics found in local news and Twitter trending topics. The results of our experiments show that using ten topics produces the highest topic recall; that using trigrams in BN-grams results in the highest value topic recall; and that using aggregation reduces the quality of trending topics produced. Overall, BN-grams has a higher value of topic recall than that of document pivot.

Keywords: trending topics detection, Twitter, BN-grams, document pivot.

1. INTRODUCTION

Trending topic, which is also called emerging trend or emerging topic (Becker, 2011), is a research area that is growing in interest and utility over time (Kontostathis et al., 2004). Trending topics can be categorized into three types (Cvijikj and Michahelles, 2011): disruptive events, popular topics, and daily routines. Disruptive events are events or phenomena that draw global attention, such as earthquakes and tsunami. Popular topics might be related to some past events, celebrities, products, or brands that remain popular over a long period of time, such as Coca Cola and Michael Jackson. Daily routines are trending topics related to some common phrases, such as "good night" or birthday wishes. In this paper, we want to generate trending topics based on disruptive political events in Indonesia.

Based on textual content of the news, there are three main approaches for detecting trending topics, namely trending topics detections based on document pivot, feature pivot, and probabilistic topic model (Aiello et al., 2013; Petkos et al., 2014a, 2014b). Trending topics detection based on document pivot is performed by clustering documents based on similarities among the documents (Aiello et al., 2013; Andoni et al., 2014; Charikar, 2002; Indyk and Motwani, 1998; Petrović et al., 2010; Ravichandran et al., 2005). Feature pivot is based on documents clustering using some features from the documents, such as terms and n-grams (Aiello et al., 2013; Benhardus and Kalita, 2013; Martin and Göker, 2014; Petkos et al., 2014a). Probabilistic topic model, on the other hand, is based on the probability of some features, such as terms or n-grams, in the documents (AlSumait et al., 2008; Blei et al., 2003; Ge et al., 2013; Wang et al., 2012).

According to its objective, trending topics detection in Indonesia can be divided into two categories. The first objective is to generate topics from events, political movements, urbanization, etc. (Mazumder et al., 2013; Oktafiani et al., 2012; Purwitasari et al., 2015; Sitorus et al., 2017). The method discussed in Oktafiani et al. (2012) used a combination of NLP, graph concept, and network analysis methods to generate topics for a flooding event and the gubernatorial election event in Jakarta. Radical political movements in several provinces of Indonesia were detected based on radical sentiments expressed in tweets with data validation from the Wahid Institute (Mazumder et al., 2013). Trending topics detection for urban monitoring in several areas, such as Jakarta, Bogor, Tangerang, and Bekasi, was carried out in (Sitorus et al., 2017). Furthermore, Purwitasari et al. (2015) aimed to make a summary of various issues showing up on Twitter by classifying them into clusters using K-Medoids method. The results of the clustering were then used as abstracts of news articles published in Kompas.

The second objective is to describe a trending topic in detail (Hariardi et al., 2016; Winatmoko and Khodra, 2013). The trending topics show up on Twitter with hashtags and keywords; and this only makes them more difficult to understand. Hence, researches were conducted to make a summary of a group of hashtags to generate more detailed information on Twitter using the combination of TF-IDF and phrase reinforcement (Hariardi et al., 2016). Subsequently, Winatmoko and Khodra (2013) aimed to provide a more comprehensive description of a trending topic generated based on three main stages, namely topic categorization (using cosine similarity), sentence extraction (using sum-basic and hybrid TF-IDF), and sentence clustering (using TF-IDF and distance-based method). In this research, we want to apply BN-grams and document pivot for trending topics detection for general use, even though our case studies are limited only to detection of trending topics in the political field.

According to their sources, trending topic accounts are classified into real accounts and campaign accounts (Mafrur et al., 2014a, 2014b). Real accounts are used for communicating or tweeting but not for spamming, promoting or campaigning. Campaign accounts, on the other hand, are used for political campaign purpose. Real or campaign accounts can be identified based on some features, such as creation date, tweet contents, periods of tweeting, followers, and friends. For example, tweets from campaign accounts usually contain the same meaning for a specific purpose. Campaign accounts in Indonesia usually generate tweets with different sentences but with the same meaning and are not based on political shows in television. This is in contrast to Pedersen et al. (2015), who showed that Twitter has been widely used to see public responses to political debates shown on television.

In this paper, we compare two methods to detect trending topics of tweets in Indonesian language. The compared methods are document pivot method and BN-grams, which is a feature pivot method. In previous studies (Aiello et al., 2013; Kaleel and Abhari, 2015; Petrović et al., 2010), both methods have been compared to detect trending topics in English tweets. Document pivot with Locality-Sensitive Hashing (LSH) clustering was used to detect trending topics in (Kaleel and Abhari, 2015; Petrović et al., 2010), while BN-grams was used to detect trending topics in (Aiello et al., 2013; Martin and Göker, 2014; Tembhurnikar and Patil, 2015). As reported by Aiello et al. (2013), BN-grams achieves higher accuracy in detecting trending topics than document pivot.

Detection of the trending topics in Indonesian tweets requires different stemming and stop words during the preprocessing stage. While Aiello et al. (2013) uses Porter stemming and English stop words, in this study, we use Adriani et al.'s (2007) stemming and Tala's Indonesian stop words (Tala, 2003). The impact of the use of different stemming and stop words will be investigated in this paper.

The contributions of this paper are:

1. A comparison of BN-grams and document pivot with LSH clustering applied on tweets in Indonesian.
2. An analysis on the effect of the use of n-grams variations in BN-grams on the quality of trending topics of tweets in Indonesian. The n-gram types used in this study are unigram up to six-grams.

The rest of the paper is organized as follows: Section 2 presents the related work and Section 3 provides description of basic concepts of trending topics detection. In Section 4, we present the experimental result, followed by discussion and analysis in Section 5. Section 6 concludes the paper.

2. Related Work

Trending topics detection based on textual content is a derivative of topics detection based on text for a set of data in a corpus (Petkos et al., 2014a, 2014b). Text-based topics detection uses three approaches: based on document pivot, feature pivot, and probabilistic topic model (Aiello et al., 2013; Panagiotou et al., 2016b).

Document pivot based approach is a trending topics detection technique that uses document clustering based on similarities among the documents (Aiello et al., 2013; Panagiotou et al., 2016b). This technique was developed based on the research into First Story Detection (FSD) using Locality Sensitive Hashing (LSH) method (Allan et al., 1998). LSH is used to differentiate between events and non-events, and generates clusters of high precision. However, the recall value this technique produces is low (*i.e.*, the size of the resulting cluster is too small). This low recall value is improved by Petrovic et al. (2010) by modifying the former LSH method into new LSH. The new LSH accelerates the process of clusterization among documents using Nearest Neighbour. Aiello et al. (2013) developed *Doc-p*, which includes new LSH and incorporates the stages of clusters ranking to detect trending topics in English tweets.

Feature pivot-based approach performs clustering on documents based on feature selection (Aiello et al., 2013). Feature selection on documents uses two approaches, based on determination of threshold values and probabilistic model.

One of the features based on threshold values approach is TF-IDF (Benhardus and Kalita, 2013; Cvijikj and Michahelles, 2011; Phuvipadawat and Murata, 2010). Meanwhile, one of the features based on probabilistic topic model is document burst (Aiello et al., 2013; Fung et al., 2005; Kleinberg, 2002; Mathioudakis and Koudas, 2010).

Burst is a document with higher frequency of appearance than other documents and the frequency exceeds a particular threshold (Panagiotou et al., 2016a). Bursts in a group of documents that appear consecutively can be modeled by infinite state automata (Kleinberg, 2002). Features in the form of detected bursts are clustered with other burst features to detect the same event (Fung et al., 2005). On the formed clusters, trend analysis detection is performed to identify every event that becomes a trend of each cluster (Mathioudakis and Koudas, 2010).

The study of Mathioudakis and Koudas (2010) is further developed by Aiello et al. (2013), where the formed event clusters are developed into trending topics on Twitter, where features are clustered into n-grams; hence *BN-grams*. In (Martin et al., 2013), BN-grams is developed by adding a new formula into topic ranking: while in (Aiello et al., 2013), topic ranking is based on document frequency-inverse document frequency (DF-IDF), in (Martin et al., 2013), it is based on the computation of the total terms in every topic and the total tweets related to the topic (weighted based on the topic length). Experimental result with the new formula indicates that the produced topic recall is higher than that of ranking by DF-IDF. Unlike in (Martin et al., 2013), BN-grams in (Martin and Göker, 2014) is developed by adding topic labeling and diversity measurement to remove tweets that are not related to a particular topic in the cluster; this has not been considered in (Aiello et al., 2013).

A study on trending topics detection methods were performed by Aiello et al. (2013), in which BN-grams was compared with LDA, Doc-p, graph-based feature pivot (GFeat-p), frequent pattern mining (FPM), and soft frequent pattern mining (SFPM). Their result shows that BN-grams achieves the highest accuracy in terms of topic recall, keyword precision, and keyword recall.

BN-grams and Doc-P in Aiello et al. (2013) were used to detect trending topics in English tweets. This research proposes to use BN-grams and Doc-P to detect trending topics in Indonesian tweets. For English tweets, BN-grams produces higher topic recall than does Doc-P. To determine whether or not BN-grams applied to identify trending topics in Indonesian tweets remains higher than Doc-P constitutes the challenge of this research.

3. Trending Topics Detection

This section describes the basic concepts of document pivot and BN-grams.

3.1 Document Pivot

Document pivot method consists of four steps (Aiello et al., 2013; Kaleel and Abhari, 2015; Petrović et al., 2010): clustering of tweets using LSH, elimination of clusters whose members are under a threshold, calculation of each cluster's score, and topic ranking. Before these four steps begin, tweets that have been grouped into several time intervals using time aggregation are preprocessed using tokenization and stemming.

Step 1. Clustering of tweets using LSH

Clustering tweets using LSH has five steps (Kaleel and Abhari, 2015), as shown in Fig. 1. First, a dictionary, which consists of a unique glossary of collected tweets, is created. Every entry in the dictionary has an index term, which is a single word in a sentence (El-Fishawy et al., 2013). Second, based on the index term in the dictionary, every collected tweet is converted into a bit array signature and is included into a collection of hash tables S (Martin et al., 2015). The LSH method uses k bits and L hash tables and two documents are considered *collided* if and only if those two documents have the same bit array signature (Kaleel and Abhari, 2015). A document is several tweets posted in a certain constant length of time (Benhardus and Kalita, 2013). In this research, the bit array signature is 17 bits long. Third, collided tweets, namely those having the same bit array signature as other tweets, are included into the same bucket in the hash tables collection S. Fourth, a cosine similarity is calculated on the tweets in S. In the fifth step, if the cosine similarity score exceeds a certain threshold, the tweets will be included in the same cluster; if the cosine similarity is below the threshold, a new cluster will be formed.

Step 2. Elimination of clusters whose members are under the threshold

The threshold used in this research is 2; hence the resulting clusters whose members are less than 2 will be eliminated.

Step 3. Calculation of each cluster's score

The score of a cluster is defined by Equation (1):

$$Score_c = \sum_{i=1}^{|Docs_c|} \sum_{j=1}^{|words_i|} \exp\left(-p(w_{ij})\right) \qquad (1)$$

where $p(w_{ij})$ is the probability of the frequency of occurrences of term j in document i in the cluster given the used corpus (see Equation (2)) and it is given by (Aiello et al., 2013; O'Connor et al., 2010) :

$$p\left(w \,\middle|\, corpus\right) = \frac{N_w + \delta}{\left(\sum_{u}^{n} N_u\right) + \delta n} \qquad (2)$$

where N_w is the total occurrences of term w in the corpus, N_u is the total occurrences of term u, and δ is the constant smoothing. In this research, δ is set to 0.5 (Aiello et al., 2013). A corpus is simultaneously a collection of words and a collection of document (Rzeszutek et al., 2010).

Step 4. Topic ranking

A trending topic is represented in the form of a set of keywords in each cluster. The clusters are ordered based on the score of each cluster and the cluster with the highest score will become a trending topic.

3.2 BN-grams

BN-grams method consists of three steps, as shown in Fig. 2, namely calculation of DF-IDF$_t$, n-grams clustering, and topic ranking. Before tweets are processed in the first step, tweets collected in the current and previous time slots based on aggregation proximity of time undergo tokenization preprocessing, stemming and aggregation. In this research, two kinds of aggregation are used, time and topic aggregations. Time aggregation is performed to collect tweets based on the proximity of time in each time slot. After collecting tweets inside of the time slot, topic aggregation will be carried out in each time slot to combine tweets based on their similarity using LSH method (Petrović et al., 2010).

Step 1. Calculation of DF-IDF$_t$

For each extracted n-gram from the collection of tweets, its DF-IDF$_t$ is computed. An n-gram is generalized words consisting of *n* consecutive grams (symbols, letters or even words), as they are used in a text (Egghe, 2005). DF-IDF$_t$ is based on the frequency of n-grams occurrences in some tweets at a certain time slot compared to the frequency of n-grams occurrences in some previous time slots. DF-IDF$_t$ is defined by Equation (3):

$$\mathrm{DF\text{ - }IDF_t} = \frac{\mathrm{df_i} + 1}{\log\left(\frac{\sum_{j=i}^{t} \mathrm{df_{i-j}}}{t} + 1\right) + 1}.\mathrm{boost} \qquad (3)$$

where df_i is the frequency of n-grams occurrences in some tweets at time slot i, df_{i-j} is the frequency of n-grams occurrences in some tweets in the previous *i-j* time slots, and t is the number of all time slots. The *boost* score is the score of certain terms that can be classified as a person, location or organization in each sentence in the tweet. If the term is in the categories of person, location or organization, it has boost score 1.5, otherwise 1 (Aiello et al., 2013).

Step 2. N-grams clustering

Merging some n-grams to become clusters gives more factual, complete and reliable information about the trending topic. The merging of the n-grams is carried out using hierarchical clustering of group average. N-grams are classified into clusters based on their distance, which is defined by Equation (4):

$$\mathrm{d(g_1,g_2)} = \quad 1 - \frac{\mathrm{A}}{\min\{B,C\}} \qquad (4)$$

where $d(g_1,g_2)$ is the distance between n-grams g_1 and g_2, A is the number of tweets that contains n-grams g_1 or g_2, and B and C are the number of tweets that contain n-grams g_1 and n-grams g_2, respectively.

Step 3. Topic ranking

Every cluster represents a topic or an event that happens in social media. An event is something that happens at specific time and place along with all necessary conditions and unavoidable consequences (Kaleel and Abhari, 2015). A topic is a seminal event or activity, along with all directly related events and activities (Kaleel and Abhari, 2015). The clusters are ordered based on their scores of DF-IDF_t. The cluster that contains n-grams with the highest score of DF-IDF_t represents the topic that is most widely discussed. This cluster is the representation of the trending topic.

4. EXPERIMENTAL EVALUATION

4.1. Datasets

This study uses six datasets, namely P1, P2, P3, P4, P5 and P6, each consisting of 6,630, 21,306, 74,790, 5,327, 807, and 2,527 tweets, respectively. P1, P2, and P3 are collected in June 23, November 14, and November 28 until December 1, 2016, respectively. P4, P5, and P6 are crawled in December 13, 14, and 16, 2017, respectively. Of the tweets collected, some were omitted as they did not contain any text (null) and were not written in Indonesian. The datasets are constructed based on keywords from political figures, executive agencies, legislative assembly, judicial bodies, political event hashtags, names of the governor or vice governor candidates, and names of political parties. Moreover, in the absence of keywords relevant to political events emerging during the period of dataset collection, new keywords were added. The detection of trending topics in the present research did not refer to a particular event. Conversely, it is expected that the detection of trending-topics in this research can generate events that have not gained coverage in the leading news media. The ground truth consists of ten topics, built based on the trending topics in local news. The trending topics in local news is the most read news by news readers as called the most popular news. Ground truth contains a set of keywords based on the most popular news taken the next day after the trending topic is detected.

4.2. Preliminary dataset analysis

Preliminary dataset analysis is carried out to examine the suitability of a dataset for trending topics detection. Three tests are performed, *i.e.*, determining the percentage of relevant tweets, determining proportions of media and non-media tweets, and calculating entropy distribution in each dataset.

The percentage of relevant tweets is determined by manual labeling or training. Labeling is performed by randomly selecting a sample of 250 tweets from each dataset. The selected tweets are identified for their relevance or irrelevance with political events in Indonesia in the period of the dataset and were performed by experts. Table 1 shows the percentage of the relevance of tweets collected randomly from the three datasets. The percentage of relevant tweets in P1, P2, and P3 is 83.6%, 80%, and 88%, respectively.

The proportion of tweets from media and individual accounts is determined by manual labeling. Labeling is performed by randomly selecting 200 tweet accounts in each dataset. The selection was performed by experts. The experts compared the accounts with media's emails and names listed in the national press data in 2016. The percentage of tweets from media and individual accounts is shown in Table 2, which shows that the percentage of tweets from media account in P1, P2, and P3 is 11.5%, 7.5%, and 6%, respectively.

Entropy distribution is used to measure the diversity of terms in a dataset. Entropy with a high value means uncertainty and terms very widely in the corpus. This expands the possibility of forming topics and influences the difficulty of detecting trending topics. Entropy is defined by Equation (5):

$$Entropy = -\sum_{i} \frac{n_i}{N} \log\left(\frac{n_i}{N}\right) \qquad (5)$$

where n_i is the number of appearances of term i in a dataset and N is the total number of terms in the dataset. In this study, the entropy value of P1, P2, and P3 is 38.89, 53.87 and 104.29, respectively, which means that trending topics detection for P1 will be easier than for P2 and P3.

4.3. Evaluation method

The performance of BN-grams and Doc-p methods was evaluated by comparing the number of topics produced by the method with the ground truth created by experts. In this study, we employ two experts, namely a lecturer with a Ph.D. in political sciences, and an Indonesian news agency worker who has contributed to the three most popular news websites in Indonesia (Kompas.com, Tempo, and Detik.com). The keywords used as ground truth are keywords describing the essence of news in the media and selected by using three criteria: related to a trending topic, around the period of the emergence of the trending topic, coming from official media and becomes popular news afterward. Several examples of the keywords in the ground truth shown in Table 3.

All evaluations in this paper use three metrics: topic recall (TR), keyword precision (KP), and keyword recall (KR). Topic recall (TR) is the ratio of trending topics to the topics in the ground truth (Equation (6)). Keyword precision (KP) is the ratio of trending topics keywords that are consistent with ground truth keywords to all keywords in the trending topics (Equation (7)). Keyword recall (KR) is the ratio of trending topics keywords that are consistent with ground truth keywords to all keywords in the ground truth (Equation (8)). Formally, they are defined as:

$$\mathrm{TR} = \frac{|\mathrm{GT} \cap \mathrm{BT}|}{|\mathrm{GT}|} \quad (6),\ \mathrm{KP} = \frac{|\mathrm{KGT} \cap \mathrm{KBT}|}{|\mathrm{KBT}|} \quad (7),\ \text{and}\ \mathrm{KR} = \frac{|\mathrm{KGT} \cap \mathrm{KBT}|}{|\mathrm{KGT}|} \quad (8),$$

where GT is the set of topics in the ground truth, BT is the set of trending topics, KGT is the set of keywords in the ground truth, and KBT is the set of trending topics keywords.

4.4. Evaluation result

4.4.1 The effect of the number of topics

The first performance of BN-grams and Doc-p methods is evaluated with the ground truth on the same time slots. The number of topics that we measure in our experiment is up to 10 topics. The ground truth consists of 10 topics in each time slot. The performance of the methods is measured by comparing the accuracy score in each number of topics. In this experiment, we want to analyze whether increasing the number of topics produced by the methods also increases the overall accuracy.

Fig. 3 depicts topic recall values produced by BN-grams and Doc-p for different number of topics. BN-grams method produces trending topics with higher accuracy than Doc-p, indicated by the topic recall (TR) value of BN-grams is higher than that of Doc-p for the three datasets. Keyword precision and recall stay constant when the number of topics varies (not shown for conciseness).

Topic recall values increase as the number of topics increases. P1 produces more certain trending topics than P2 and P3. This is because P1 has fewer tweets and also has relatively shorter drawing period than P2 and P3. Thus, P1 has less difficulty and higher accuracy than P2 and P3. This is in line with the result of entropy reported in Section 4.2, where P1 had the smallest entropy values compared to P2 and P3. Experiments also indicate that BN-grams produces more topics that are consistent with real life news than Doc-p. This is because the principle of clustering based on frequency in BN-grams increases the accuracy of trending topics detection compared to Doc-p, which is based on threshold and similarity.

4.4.2 The effect of n-grams variations in BN-grams

Fig. 4 depicts the accuracy of trending topics detection by varying the n-grams used. The n-grams used are unigram to sixgrams. Trigrams produces higher accuracy than other n-grams. In P1 and P2 trigrams has the highest topic recall compared to unigram and bigrams; in P3 however trigrams produces lower topic recall than bigrams. This is because the number of tweets is smaller and the period of tweet collection is also shorter for P1 and P2 compared with those of P3.

Topics containing more factual keywords describing events in real life come from bigrams. Bigrams has higher keyword precision and keyword recall values than other n-grams. The use of bigrams in P1 and P3 also produces higher keyword precision and keyword values than other n-grams. Therefore, the use of trigrams results in trending topics which better describe events in real life. However, to obtain trending topics with more factual keywords and containing topics consistent with local news, bigrams and trigrams are recommended.

Fig. 4 shows the detail accuracy of trending topics detection using unigram up to sixgrams. In the three graphs, the use of trigrams up to sixgrams produces the same accuracy for P1 (with the values of topic recall, keyword precision, and keyword recall 0.556, 0.921, and 0.824, respectively) and P3 (with the values of topic recall, keyword precision, and keyword recall 0.5, 0.692, and 0.353, respectively), and relatively close in P2 (with the values of topic recall, keyword precision, and keyword recall around 0.3, 0.6, and 0.9, respectively). We can conclude that the use of trigrams produces nearly the same accuracy level as fourgrams, fivegrams, and sixgrams. This is because trigrams contain three terms, which in Indonesian grammar represents subject, predicate, and object pattern. The use of trigrams therefore produces sentence structure that is easier to understand in Indonesian compared to those produced by unigram or bigrams. The topics generated by trigrams are also highly similar with news in local media.

4.4.3 The effect of stemming

In this third experiment, we determine the effect of stemming on the accuracy of trending topics detection. Overall, the use of stemming in P1 worsens the accuracy of trending topics detection. The number of topics used in this experiment is 5. In BN-grams and Doc-p, topic recalls produced with stemming and non-stemming are 20%, 40% and 0%, 63.6%, respectively. Experiments indicate that the quality of trending topics from P1 using stemming in BN-grams is worsened by 20% compared to those produced without stemming. In Doc-p, the use of stemming also worsens the quality of the trending topics produced. This is evident in the topic recall of Doc-p with stemming, which is 0%; meaning the produced trending topics are not relevant at all to topics in local media. This is because several prefix or suffix of Indonesian words, which should not be removed, is removed during stemming. Another reason is that Adriani et al.'s (2007) stemming is still unable to detect new vocabularies today and therefore produces inaccurate stemming;for instance the term "jokowi" becomes "jokow".

4.4.4 The effect of aggregation variation on preprocessing

Detection of topics generated from Twitter has the problem of poor quality information because tweets usually contain short sentences, slang words and abbreviations. To solve it, we aggregate tweets into datasets to create documents that contain a lot more information and thus produce better topic result. Tweet aggregation produces four datasets. First, by combining every 2,000 tweets contiguously in time (Time Aggregation 2000). Second, by collecting every 4,000 tweets contiguously in time (Time Aggregation 4000). Third, by combining tweets based on their similarity using LSH method (Topic Aggregation) in each time slots. Fourth, by concatenating tweets at specific time slots regardless of the similarity and proximity of time (No Aggregation). On each dataset trending topics detection with BN-grams and Doc-p methods is applied.

Overall, the use of aggregation in P2 reduces the accuracy of trending topics detection. The aggregation types that are compared are topic aggregation, time aggregation with 2,000 tweets, time aggregation with 4,000 tweets and no aggregation. The result is shown in Fig. 5. The use of no aggregation with 10 topics produces the highest accuracy of trending topics, namely with topic recall value of 38.1 % in BN-grams. The use of topic aggregation and no aggregation (as opposed to time aggregation) increases the accuracy of BN-grams compared to Doc-p. This is because topics produced with topic aggregation and no aggregation contain a set of tweets with higher similarity than those produced with time aggregation, so the topics produced are more specific, focused and not mixed up.

The use of time aggregation 2,000 and 4,000 only increases the accuracy of Doc-p. In P2, topic recall of Doc-p using time aggregation 2,000 and 4,000 has a similar accuracy of 33.3 %. Conversely, time aggregation reduces the accuracy of BN-grams. This is because time aggregation contains more multiple tweets with more complex term distribution. Therefore, the topics produced by time aggregation contain a mixture of several topics, and fewer produced topics are consistent with local news.

4.4.5 Comparison of the proposed trending topics and Twitter's trending topics

To compare our proposed method for trending topics detection with Twitter's trending, we perform the following three steps. First, in a particular day we generate trending topics using BN-grams or Doc-p methods. Second, in the following days we create a ground truth that contains a set of keywords based on

Twitter's trending topics. Third, we measure the accuracy based on topic recall from the results of both trending topics. In P4, evaluation with 10 trending topics only produces two similar trending topics for our method and Twitter's trending topics, so topic recall is 0.2; while in P5 and P6 the topic recall is 0 and 0.1 respectively.

Fig. 6 depicts the topic recall values for various numbers of topics produced by BN-grams and Doc-P with Twitter's trending topics. Doc-p produces trending topics with higher accuracy than BN-grams, indicated by the topic recall value of Doc-p is higher than that of BN-grams for the three datasets.

Topic recall values stay constant when the number of topics increases. P4 produces more certain trending topics than P5 and P6. This is because P4 is larger than P5 and P6. Experiments also indicate that Doc-p produces more topics that are consistent with Twitter's trending topics than BN-grams. This is because the principle of clustering based on similarity and threshold in Doc-p increases the accuracy of trending topics detection compared to BN-grams, which is based on the number of frequency.

5. Discussion And Analysis

Evaluation of each experimental result generates several findings. An increase in the topic recall value is consistent with the increase in the number of topics tested. This happens as the higher the number of topics generated is, the higher the probability of similarity between the trending topics in this paper and those of the popular news media. In Aiello et al. (2013), an increase in the number of trending topics, which is directly proportional to the increase in the topic recall value, was found only in Doc-P.

In general, the use of trigrams in BN-grams generates the highest topic recall in two of the three datasets. This is because trending topics using trigrams accommodates the pattern of subject, predicate, and object (SPO). These three components form the basic formation of an Indonesian sentence. Therefore, the trending topics generated by trigrams have a higher level of similarity to the popular news in the local media. The research by Aiello et al. (2013) had not tested trending topics using variation in the n-grams, which differentiates the present testing from the research of Aiello et al. (2013).

The stemming in BN-grams and Doc-P negatively affects the resulting trending-topics. This strengthens the research by Aiello et al. (2013). The reason is that the use of stemming results in the omission of prefixes and suffixes from any Indonesian terms, making Indonesian trending topics have a shallow level of similarity to local news.

Variation in the type of aggregation has a different effect on each method under study. The implementation of no aggregation in BN-grams generates the highest topic recall value among all types of aggregations. This corroborates results of aggregation testing in Aiello et al. (2013). Furthermore, the application of time aggregation in Doc-P generates the highest topic recall value among all types of aggregations, while topic aggregation in Aiello et al. (2013) produced the highest topic recall in Doc-P. This difference exists as time aggregation contains a set of similar tweets posted relatively close to one another.

The result of the comparison of our proposed trending topics and Twitter's is contradictory to the comparison of our proposed trending topics with local news trending topics. The comparison with Twitter's Doc-p produces a higher accuracy. But, with local news trending topics, BN-grams has a higher accuracy than Doc-p. This is because clustering based on similarity and threshold is more applicable in Twitters, while clustering based the frequency is more suitable in local news.

The experiments also indicate that trending topics generated by our method and trending topics in local news complement each other. The trending topics of our method form a material for the trending topics in local news, beside the tweets collected based on the trending topics in local news. There are two relationships: the trending topics in local news can be reported in trending topics of our method and vice versa.

The critical finding in our experiments is the contrary between trending topics of our method and the trending topics in local news, which is evident in the local election of governor and vice governor of Jakarta in 2017. The trending topics generated by our method is a direct opinion of the society without any manipulation.

Therefore, the trending topics produces by our method can be an early warning system for political events in Indonesia.

6. Conclusion

Generally, trending topics detection in Indonesian tweets is influenced by preprocessing and the total number of collected tweets. Experiments show that trending topics detection in Indonesian tweets is more accurate when using BN-grams than Doc-p. BN-grams produces higher accuracy in detecting trending topics than Doc-p in all three datasets. However, for keyword precision, Doc-p is better than BN-grams.

The use of preprocessing, especially stemming and aggregation, also influences the quality of the produced trending topics. The use of stemming in preprocessing worsens the accuracy, while aggregation also reduces the quality of produced trending topics.

The use of n-grams variations influences the quality of trending topics produced by BN-grams. Experiments using unigram result in the worst quality of the produced trending topics, while the use of trigram results in the highest quality. It is concluded that trending topics detection in Indonesian tweets especially by BN-grams should use trigrams to produce trending topics with high accuracy and nearly the same accuracy as four-grams, five-grams, and six-grams.

The pattern in Indonesian writing is similar to the language pattern in the Indonesian subgroups: Melayu (Malaysia), Malagasy (Madagascar), Formosa and Philippines (Darmini, 2012). Therefore, Indonesian trending topics research has an excellent opportunity to be applied to trending topics in Indonesian language subgroup. Also, experimental results show that topics generated by BN-grams and Doc-P from Indonesian tweets do not have subject, predicate, object and adverb (SPOK) pattern, as Indonesian sentences should be; this will become a challenge for future research.

Acknowledgement

This research is supported by the Domestic Postgraduate Education Scholarship (BPPDN) and Doctoral Dissertation Grant (grant number

0426/K3/KM/2017) of Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

References

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M., Williams, H.E., 2007. Stemming Indonesian: A confix-stripping approach. ACM Transactions on Asian Language Information Processing (TALIP) 6, 1–33. doi:10.1145/1316457.1316459

Aiello, L.M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., Member, S., IEEE, Jaimes, A., 2013. Sensing Trending Topics in Twitter. IEEE Transactions on Multimedia (2013) 15, 1268–1282.

Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y., 1998. Topic detection and tracking pilot study final report.

AlSumait, L., Barbará, D., Domeniconi, C., 2008. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. 2008 Eighth IEEE International Conference on Data Mining 3–12. doi:10.1109/ICDM.2008.140

Andoni, A., Indyk, P., Nguyen, H.L., Razenshteyn, I., 2014. Beyond locality-sensitive hashing, in: Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, pp. 1018–1028.

Becker, H., 2011. Identification and Characterization of Events in Social Media. Columbia University.

Benhardus, J., Kalita, J., 2013. Streaming trend detection in twitter. International Journal of Web Based Communities 9, 122–139. doi:10.1504/IJWBC.2013.051298

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022.

Charikar, M.S., 2002. Similarity estimation techniques from rounding algorithms. Proceedings of the thiry-fourth annual ACM symposium on Theory of computing - STOC '02 380–388. doi:10.1145/509907.509965

Cvijikj, I.P., Michahelles, F., 2011. Monitoring Trends on Facebook. 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing 895–902. doi:10.1109/DASC.2011.150

Darmini, W., 2012. Perbedaan Kata Bahasa Indonesia dengan Bahasa Melayu (Malaysia) dalam Sistem Ejaan. Widyatama 21, 103–108.

Egghe, L., 2005. The exact rank-frequency function and size-frequency function of N-grams and N-word phrases with applications. Mathematical and Computer Modelling 41, 807–823. doi:10.1016/j.mcm.2003.12.016

El-Fishawy, N., Hamouda, A., Attiya, G.M., Atef, M., 2013. Arabic summarization in Twitter social network. Ain Shams Engineering Journal. doi:10.1016/j.asej.2013.11.002

Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H., 2005. Parameter free bursty events detection in text streams. VLDB '05 Proceedings of the 31st international

conference on Very large data bases 1, 181–192. doi:10.1.1.60.2671
Ge, G., Chen, L., Du, J., 2013. The research on topic detection of microblog based on TC-LDA. 2013 15th IEEE International Conference on Communication Technology 722–727. doi:10.1109/ICCT.2013.6820469
Hariardi, W., Latief, N., Febryanto, D., Suhartono, D., 2016. Automatic Summarization from Indonesian Hashtag on Twitter Using TF-IDF and Phrase Reinforcement Algorithm, in: International Workshop on Computer Science and Engineering (WCSE 2016). pp. 17–19.
Indyk, P., Motwani, R., 1998. Approximate nearest neighbors: towards removing the curse of dimensionality, in: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing. pp. 604–613.
Kaleel, S.B., Abhari, A., 2015. Cluster-discovery of Twitter messages for event detection and trending. Journal of Computational Science 6, 47–57. doi:10.1016/j.jocs.2014.11.004
Kleinberg, J., 2002. Bursty and Hierarchical Structure in Streams, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’02. ACM, New York, NY, USA, pp. 91–101. doi:10.1145/775047.775061
Kontostathis, A., Galitsky, L.M., Pottenger, W.M., Phelps, D.J., 2004. A Survey of Emerging Trend Detection in Textual Data Mining. Springer Science and Business Media New York, New York, New York, USA. doi:http://dx.doi.org/10.1007/978-1-4757-4305-0_9
Mafrur, R., Fiqri Muthohar, M., Bang, G.H., Lee, D.K., Kim, K., Choi, D., 2014a. Twitter mining: The case of 2014 indonesian legislative elections. International Journal of Software Engineering and its Applications 8, 191–202. doi:10.14257/ijseia.2014.8.10.17
Mafrur, R., Muthohar, M.F., Bang, G.H., Lee, D.K., Choi, D., 2014b. Who are Tweeting in the 2014 Indonesia ’ s Legislative Elections ?
Martin, C., Corney, D., Goker, A., 2015. Mining newsworthy topics from social media, in: Advances in Social Media Analysis. Springer International Publishing, pp. 21–43. doi:10.1007/978-3-319-18458-6_2
Martin, C., Corney, D., Göker, A., MacFarlane, A., 2013. Mining Newsworthy Topics from Social Media, in: British Computer Sociey (BCS) The Specialist Group on Artificial Intelligent (SGAI) Workshop on Social Media Analysis. pp. 35–46.
Martin, C., Göker, A., 2014. Real-time topic detection with bursty n-grams: RGU’s submission to the 2014 SNOW Challenge, in: Proceedings of the Social News on the Web (SNOW) 2014. pp. 9–16.
Mathioudakis, M., Koudas, N., 2010. Twittermonitor: trend detection over the twitter stream, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. New York, NY, USA, pp. 1155–1158. doi:10.1145/1807167.1807306
Mazumder, A., Das, A., Kim, N., Gokalp, S., Sen, A., Davulcu, H., 2013. Spatio-temporal signal recovery from political tweets in Indonesia, in: Proceedings - SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013. pp. 280–287. doi:10.1109/SocialCom.2013.46

O'Connor, B., Krieger, M., Ahn, D., 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter., in: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 384–385.
Oktafiani, P.., Jariyah, A., Fitri, S.., Hashimoto, T., 2012. Social media analysis for Indonesian language: case study flood in Jakarta. International Conference on Advanced Computer Science and Information Systems (ICACSIS) 161–166.
Panagiotou, N., Ioannis, K., Gunopulos, D., 2016a. Detecting Events in Online Social Networks: Definitions, Trends and Challenges 9580, 42–84. doi:10.1007/978-3-319-41706-6
Panagiotou, N., Katakis, I., Gunopulos, D., 2016b. Detecting Events in Online Social Networks: Definitions, Trends and Challenges, in: Michaelis, S., Piatkowski, N., Stolpe, M. (Eds.), Solving Large Scale Learning Tasks. Challenges and Algorithms. Springer International Publishing, Cham, pp. 42–84. doi:10.1007/978-3-319-41706-6_2
Pedersen, S., Baxter, G., Burnett, S., Göker, A., Corney, D., Martin, C., 2015. Backchannel Chat: Peaks and Troughs in a Twitter Response to Three Televised Debates during the 2014 Scottish Independence Referendum Campaign, in: Proceedings of the 5th Conference for E-Democracy and Open Government (CeDEM 2015). pp. 105–117.
Petkos, G., Papadopoulos, S., Aiello, L., Skraba, R., Kompatsiaris, Y., 2014a. A soft frequent pattern mining approach for textual topic detection, in: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14). p. 25.
Petkos, G., Papadopoulos, S., Kompatsiaris, Y., 2014b. Two-level Message Clustering for Topic Detection in Twitter., in: Social News on The Web (SNOW) 2014 Data Challenge. http://ceur-ws.org, Seoul, Korea, pp. 49–56.
Petrović, S., Osborne, M., Lavrenko, V., 2010. Streaming first story detection with application to twitter, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 181–189.
Petrović, S., Osborne, M., Lavrenko, V., 2010. Streaming first story detection with application to twitter, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 181–189.
Phuvipadawat, S., Murata, T., 2010. Breaking News Detection and Tracking in Twitter. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 120–123. doi:10.1109/WI-IAT.2010.205
Purwitasari, D., Fatichah, C., Arieshanti, I., Hayatin, N., 2015. K-Medoids Algorithm on Indonesian Twitter Feeds For CLustering Trending Issue as Important Terms in News Summarization, in: 2015 International Conference on Information, Communication Technology Systems and System (ICTS). pp. 95–98.
Ravichandran, D., Ravichandran, D., Pantel, P., Pantel, P., Hovy, E., Hovy, E., 2005. Randomized algorithms and NLP: using locality sensitive hash function for high speed noun clustering. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics 622–629.

doi:10.3115/1219840.1219917
Rzeszutek, R., Member, S., Androutsos, D., Member, S., 2010. Self-Organizing Maps for Topic Trend Discovery. IEEE Signal Processing Letters 17, 607–610.
Sitorus, A.P., Murfi, H., Nurrohmah, S., Akbar, A., 2017. Sensing Trending Topics in Twitter for Greater Jakarta Area. International Journal of Electrical and Computer Engineering (IJECE) 7, 330–336. doi:10.11591/ijece.v7i1.pp330-336
Tala, F.Z., 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. M.Sc. Thesis, Appendix D. University van Amsterdam The Netherlands.
Tembhurnikar, S.D., Patil, N.N., 2015. Topic detection using BNgram method and sentiment analysis on twitter dataset, in: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions). pp. 1–6.
Wang, Y., Agichtein, E., Benzi, M., 2012. TM-LDA : Efficient Online Modeling of the Latent Topic Transitions in Social Media, in: KDD ’12 Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, pp. 123–131. doi:10.1145/2339530.2339552
Winatmoko, Y.A., Khodra, M.L., 2013. Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation. Procedia Technology 11, 1027–1033. doi:http://dx.doi.org/10.1016/j.protcy.2013.12.290

Tables

Table 1. Percentage of relevance of tweets in the three datasets

Dataset	Relevant	Not Relevant	% Relevant
P1	209	41	83.6%
P2	200	50	80.0%
P3	220	30	88.0%

Table 2. Proportion of total tweets from media and individual accounts

Dataset	Media	Individual	% Media
P1	23	77	11.5%
P2	15	185	7.5%
P3	12	188	6.0%

Table 3. Examples of the ground truth

Dataset	Time Period	Title (Headline) News	Keyword
P1	Jun. 23, 2016 (09:25-10:25)	Kata richard eks teman ahok soal fotonya dengan seragam pdip dan ormas prospera (*Richard (former friend Ahok) said about his picture with pdip uniform and prospera organizations*)	richard; sukarno; soal; fotonya; seragam; pdip; ormas; prospera. (*Richard; Sukarno; picture; uniform; pdip; organizations; prospera*)
P2	Nov. 14, 2016 (10:30-13:30)	Setya Novanto Layangkan Teguran Tertulis untuk Aburizal Bakrie (*Setya Novanto give letter reprimand for Aburizal Bakrie*)	evaluasi; pendukung; ahok; goyah; golkar; fadel; muhammad. (*evaluation; supporter; ahok; faltering; Golkar; fadel; muhammad*)
P3	Nov. 28 – Dec. 1, 2016	Terbukti Korupsi 12 Juta Dollar AS, Brigjen Teddy Divonis Seumur Hidup (*Proven Corruption 12 Million US Dollar, Brigadier General Teddy sentenced for life*)	brigjen; teddy; korupsi jutaan; dollar; vonis; seumur; hidup. (*brigjen; teddy; corrupt; million; dollar; sentenced for life*)

Figure Captions

Figure

Fig. 1. Clustering tweets with LSH (Kaleel and Abhari, 2015)

Fig. 2. Clustering tweets in BN-grams (Aiello et al., 2013)

Fig. 3. The effect of the number of topics on topic recall

Fig. 4. The effect of n-grams variation on accuracy

Fig. 5. The effect of aggregation variation on total recall

Fig. 6. Comparison of the proposed trending topics and Twitter's trending topics

Figures

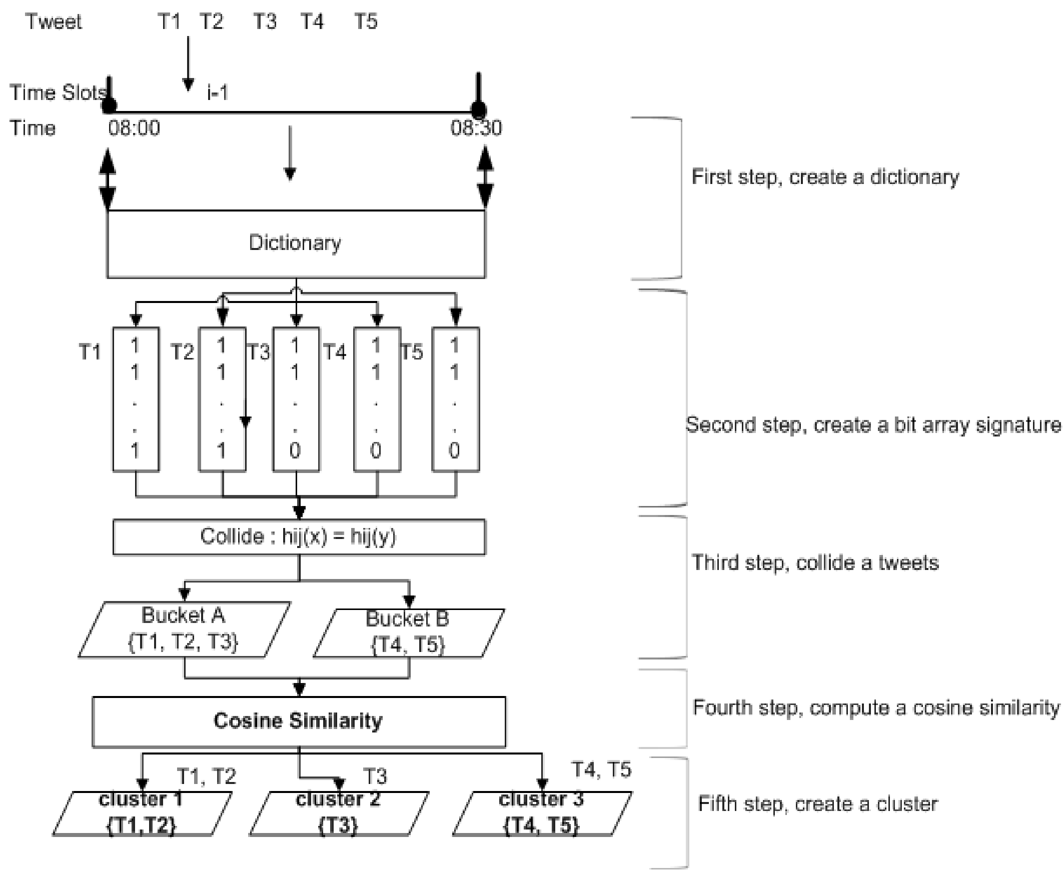
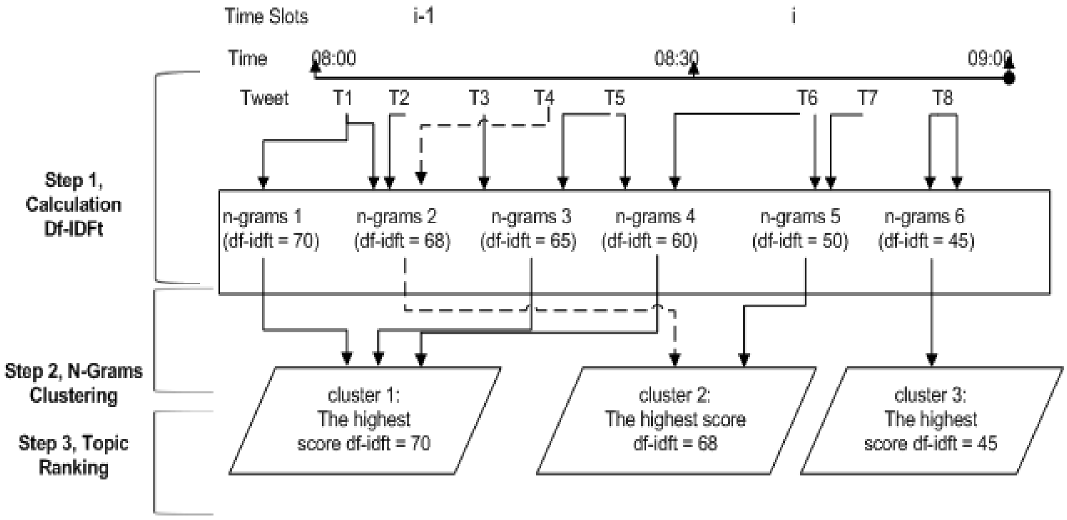
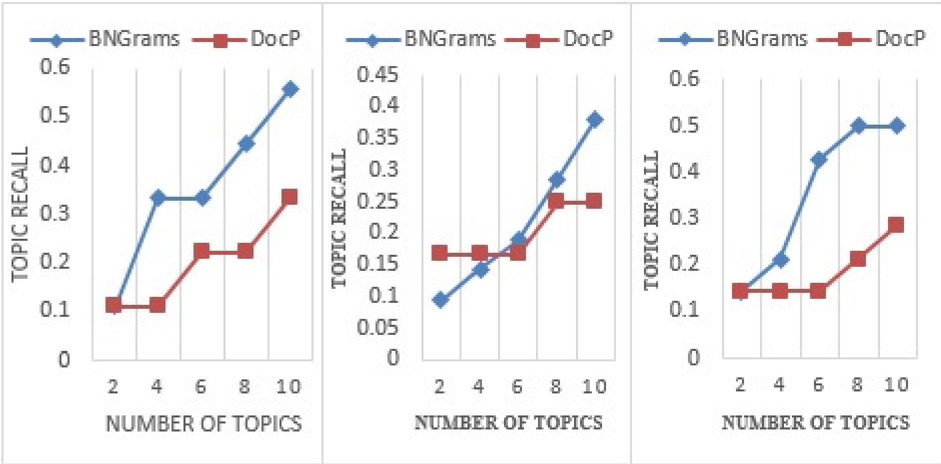


Fig. 1.



Time Slots
i-1
i
Time
08:00
08:30
09:00
Tweet
T1
T2
T3
T4
T5
T6
T7
T8
Step 1, Calculation Df-IDFt
n-grams 1 (df-idft = 70)
n-grams 2 (df-idft = 68)
n-grams 3 (df-idft = 65)
n-grams 4 (df-idft = 60)
n-grams 5 (df-idft = 50)
n-grams 6 (df-idft = 45)
Step 2, N-Grams Clustering
Step 3, Topic Ranking
cluster 1: The highest score df-idft = 70
cluster 2: The highest score df-idft = 68
cluster 3: The highest score df-idft = 45

Fig. 2.



a) P1 b) P2 c) P3

Fig. 3.

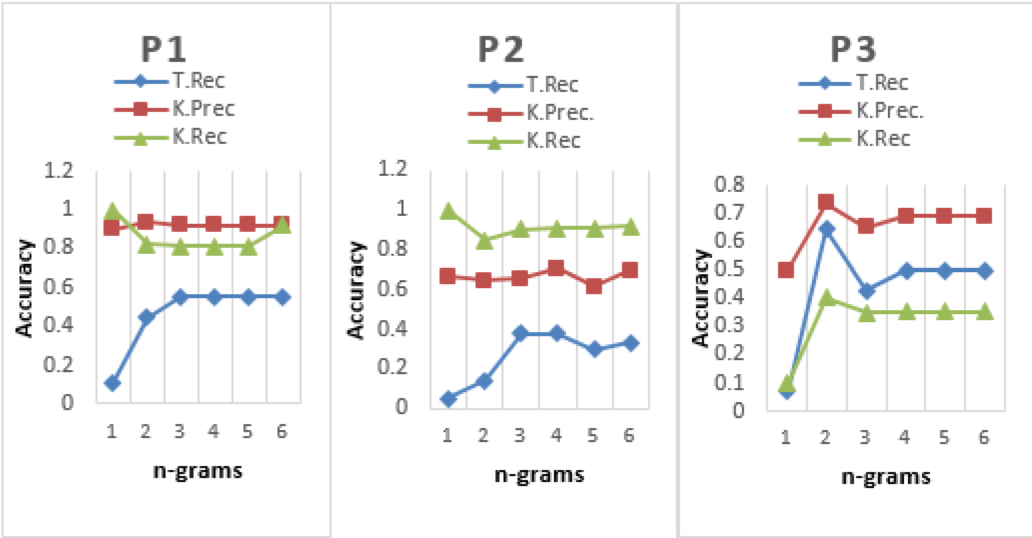

P1
T.Rec
K.Prec
K.Rec
Accuracy
n-grams
P2
T.Rec
K.Prec.
K.Rec
Accuracy
n-grams
P3
T.Rec
K.Prec.
K.Rec
Accuracy
n-grams

Fig. 4.

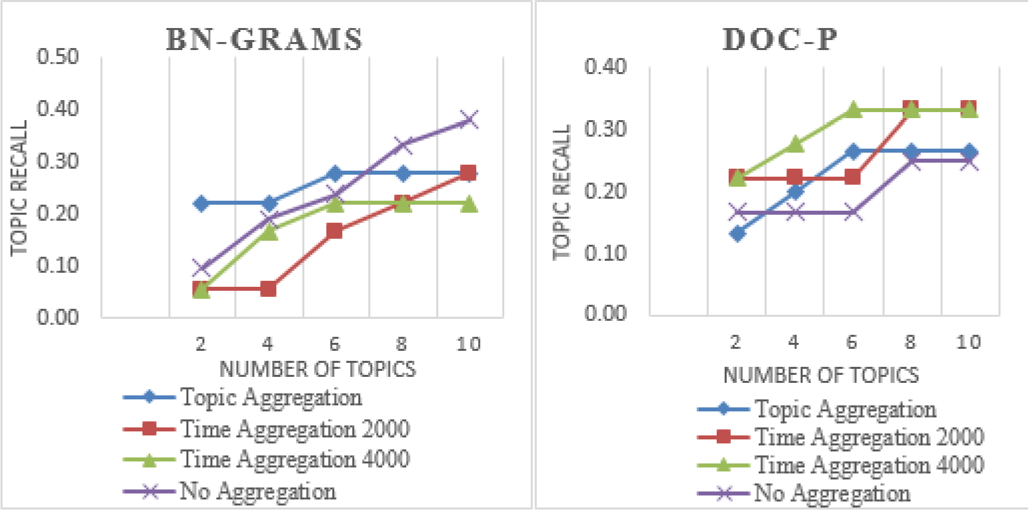
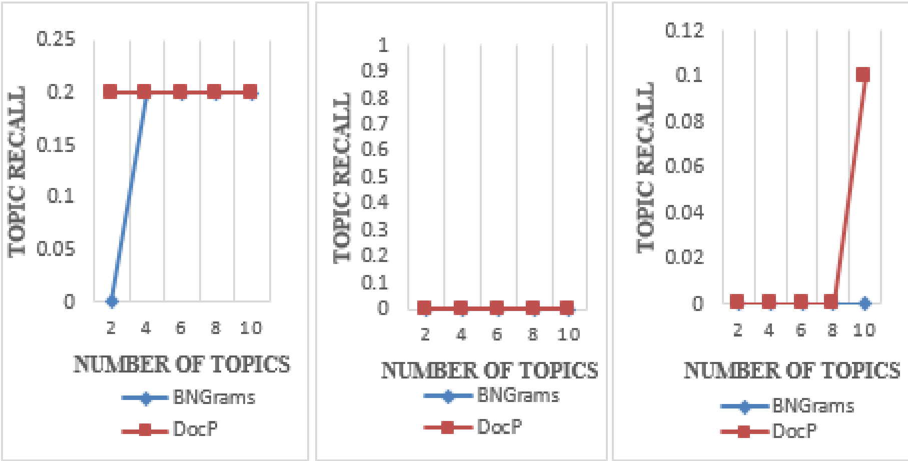

BN-GRAMS
TOPIC RECALL
0.50
0.40
0.30
0.20
0.10
0.00
2
4
6
8
10
NUMBER OF TOPICS
Topic Aggregation
Time Aggregation 2000
Time Aggregation 4000
No Aggregation
DOC-P
TOPIC RECALL
0.40
0.30
0.20
0.10
0.00
2
4
6
8
10
NUMBER OF TOPICS
Topic Aggregation
Time Aggregation 2000
Time Aggregation 4000
No Aggregation

Fig. 5.



a) P4 b) P5 c) P6

Fig 6.

Author Agreement

Submission of work requires that the piece to be reviewed has not been previously published. Upon acceptance, the Author assigns to the Journal of King Saud University – Computer and Information Sciences (JKSUCI) the right to publish and distribute the manuscript in part or in its entirety. The Author's name will always be included with the publication of the manuscript.

The Author has the following nonexclusive rights: (1) to use the manuscript in the Author's teaching activities; (2) to publish the manuscript, or permit its publication, as part of any book the Author may write; (3) to include the manuscript in the Author's own personal or departmental (but not institutional) database or on-line site; and (4) to license reprints of the manuscript to third persons for educational photocopying. The Author also agrees to properly credit the Journal of King Saud University – Computer and Information Sciences (JKSUCI) as the original place of publication.

The Author hereby grants the Journal of King Saud University – Computer and Information Sciences (JKSUCI) full and exclusive rights to the manuscript, all revisions, and the full copyright. The Journal of King Saud University – Computer and Information Sciences (JKSUCI) rights include but are not limited to the following: (1) to reproduce, publish, sell, and distribute copies of the manuscript, selections of the manuscript, and translations and other derivative works based upon the manuscript, in print, audio-visual, electronic, or by any and all media now or hereafter known or devised; (2) to license reprints of the manuscript to third persons for educational photocopying; (3) to license others to create abstracts of the manuscript and to index the manuscript; (4) to license secondary publishers to reproduce the manuscript in print, microform, or any computer-readable form, including electronic on-line databases; and (5) to license the manuscript for document delivery. These exclusive rights run the full term of the copyright, and all renewals and extensions thereof.

I hereby accept the terms of the above Author Agreement.

Author :-	Date :-
Indra, S. Kom, M.T.I	5 September 2017
Editor in Chief:- Nasser-Eddine Rikli	Date:-