Indonesian Journal of Science & Technology

Journal homepage: http://ejournal.upi.edu/index.php/ijost/

Determining Trending Topics in Twitter with a Data-Streaming Method in R

Melani Mediayani[1], Yudi Wibisono[1], Lala Septem Riza[1] ✉,Alejandro Rosales-Pérez[2]

[1] Department of Computer Science Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
[2] School of Engineering and Science, Tecnologico de Monterrey, Monterrey, Mexico

✉Correspondence: E-mail: lala.s.riza@upi.edu

ABSTRACT

Trending topics in Twitter is a collection of certain topics that are widely discussed by users. This study aims to design a model and strategy for finding trending topics from data streams on Twitter. The research approach was carried out in four stages, namely twitter data collection, preprocessing data, data analysis with sequential K-Means clustering and information processing. Sequential K-Means is used because it can receive input data sequentially and the cluster center can be updated. Testing of the model is carried out in three scenarios where each scenario is distinguished between the amount of data, time and parameter values. After that, evaluation of the results of clustering will be done using the Dunn Index method. Trending topics twitter application were created using the R language and produce output in the form of histograms. There are five topics being the trending topics in New York before the new year. The topic of "Times" relates to the presence of a new year's celebration night concert in Times Square. The "Hours" topic deals with the calculation of time and seconds towards 2017. "Eve" and "Party" topics relate to celebrations and the topic "Resolution" relating to hope and change for New Yorkers in in 2017.

ARTICLE INFO

Article History:
Submitted/Received 28 Aug 2018
First revised 31 Jan 2019
Accepted 06 Mar 2019
First available online 09 Mar 2019
Publication date 01 Apr 2019

Keywords:
Trending topics,
Streaming data,
Machine learning,
Large datasets,
Clustering,
Data analysis.

1. INTRODUCTION

Twitter has become one of the most popular social media today, allowing users to post short messages or tweets up to 140 characters. Twitter is presented as a means of communication to exchange information about various events in the real world in real-time (Rachmadany *et al.*, 2018).

One of the interesting features on Twitter is trending topic. It is a collection of the most popular topics discussed a lot on Twitter tweets. It has an important role in finding the hottest and most recent news or events.

Detecting trending topics is not an easy thing, it requires a special approach to analyze the data flow of tweets that come continuously from millions of Twitter users. The data used is very large so it takes large storage media, quite long processing data, and efficient algorithms for data analysis. Therefore, a streaming approach is needed as an efficient strategy for finding information from large data. Streaming data processing is the process by which data is taken within a certain period of time and then used to obtain a model after it is deleted and make room for new data. The streaming approach will generate trending topics in real-time and up-to-date (Firdaus *et al.*, 2017).

Clustering techniques can be used to analyze topics on tweets by automatically grouping tweets that have similarities. Clustering on text or documents is different from clustering on structured data. In the text clustering, a grouping algorithm is needed to handle high dimensional data. This research offers a model and strategy for finding trending topics from data streams on Twitter.

A lot of research has been done to explore and search for trending topics with a variety of techniques offered. As in the research conducted by Zubiaga *et al. (*2015) regarding the process of real-time classification of data tweets. In addition, a similar study was conducted by Becker *et al. (*2011) regarding the clustering process on twitter data streams and then carried out a classification process to distinguish cluster events and non-events.

Research on trending topics was also conducted by Aiello *et al. (*2013), namely a comparison of six topic detection methods in three Twitter datasets related to major events. Research by Sahdev and Kabra (2013), detected trending topics with a social network graph approach to determine individual behavior by connecting individual interactions with other individuals. Research by Berhandus and Kalita (2013) detected and identified trending topics from data streams. Research by Lau *et al. (*2012) presented a new modeling-based methodology for tracking events that appear on Twitter's microblog. Lu and Yang (2012) research conducted trend analysis on news topics on Twitter, which included trend prediction and analysis of the causes of trend changes. Miller *et al. (*2015) research offered a model for estimating the topic of the most popular twitter at certain time intervals or periods. Then research by Mathioudakis and Koudas (2010) presented a system that performs trend detection on Twitter streams and performs trend analysis. Research by Kim *et al. (*2013) proposed a new scheme to detect trends and keywords from the Twitter data stream.

2. METHODS

2.1. Strategy on determining trending topics on Twitter.

In this study we construct a model to detect trending topics on Twitter in four parts. The first part consists of one stage, namely the process of collecting tweet data from Twitter Streaming API automatically and online. Second, preprocessing data is processing raw data before building a model consisting of three stages, namely preprocessing text, word weighting and word selection. Third, the data analysis process uses clustering techniques, for the first iteration of the algorithm used is the k-means algorithm and for the next iteration using the k-means sequential algorithm. Fourth, information processing consists of four

| DOI: http://dx.doi.org/10.17509/ijost.v4i1.15807
| p- ISSN 2528-1410 e- ISSN 2527-8045

stages, namely cluster evaluation, topic detection, trending topics visualization and trending topics evaluation (See **Figure 1**).

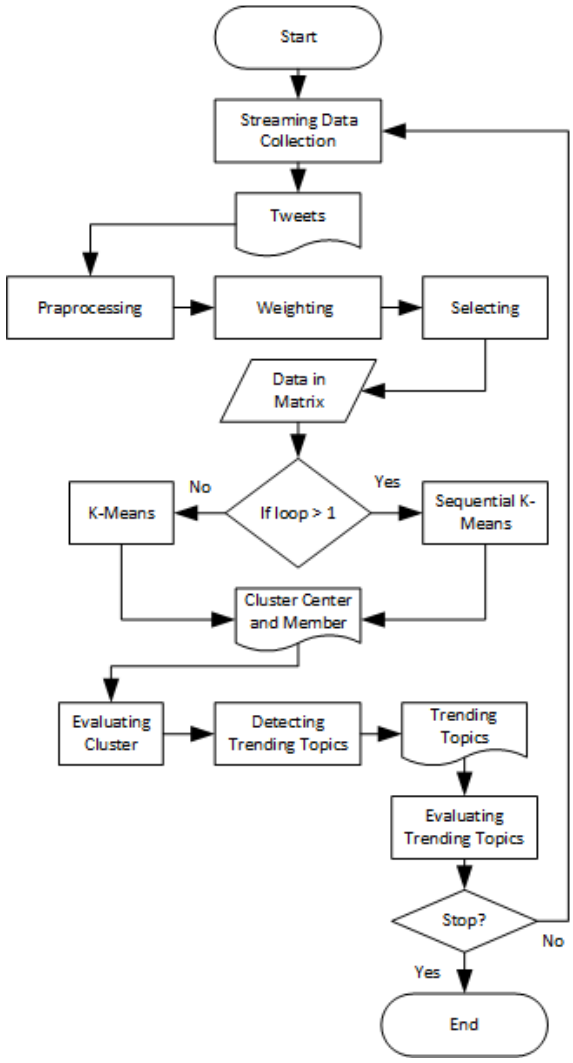

Figure 1. The Proposed Method

First, in this study ,the data used is a collection of real-time tweets obtained from Twitter Streaming API. At the stage of twitter data collection, there are two processes, namely filtering streams and parsing tweets. Filtering stream is the process of sorting the tweets that will be used. The Tweet used is an English tweet located in New York City. Based on the strategy that has been designed, data collection is carried out for 60 seconds. The tweet data obtained from Twitter Streaming API is a file with the Java Script Object Notation (JSON) format so that the tweet parsing process is carried out by converting all information from JSON files into the data frame. The tweet feature or attribute used in this study is only the "text" attribute.Twitter data collection is done by creating the *getStream()* function in the R language. The process of filtering streams and parsing tweets is done using functions contained in the "streamR" package.

Text preprocessing is the initial processing stage of tweet text data before building a model. Each data text is also called a document. In this study, preprocessing of the text carried out is as follows:

1. Choosing unique tweets, meaning tweets that have the same content will not be used in research, these tweets are usually the result of retweets.
2. Case folding or the process of uniformizing the form of Latin letters (a-z) into lowercase (lowercase).
3. Tokenizing or breaking the string based on each word that composes it.
4. Delete the parts that are considered not important from the document. The deleted parts include username and user mentions, hastags, punctuation, URLs, non alphabeth, and characters "RT".
5. Remove stopwords or non-descriptive words.
6. Delete words that are considered spam

After going through the preprocessing phase of the text, the next step is to calculate the weight for each word in the document using the Term Frequency-Inverse Document Frequency (TF-IDF) calculation (Salton and Kabra, 2013). Weighting will be done on each word that appears in the entire document. Weighting results will be used in the process of calculating the distance between documents. Also at this stage, the text document will be converted into vector form. The equation of Term Frequency can be found in (Benhardus and Kalita, 2013).

The word selection stage is the process of removing words that are not used on the

DOI:http://dx.doi.org/10.17509/ijost.v4i1.15807 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

data matrix. At this stage there are two processes, namely:

1. Add notes to the text in the form of labels on each character string or referred to as Part-of-Speech Tagging (POS Tagging) (Màrquez and Rodríguez, 1998).
2. Erase words other than labeled NN (Noun, singular or mass), NNS (Noun, plural), NNP (Proper noun, singular), and NNPS (Proper noun, plural) because the words used are words with noun categories (noun).

K-Means is one of the simplest unsupervised learning algorithms that can solve clustering problems (MacQueen, 1967). The K-Means algorithm is used to divide data sets automatically into a number of fixed clusters (assumed to be k clusters). K-means is a very well-known clustering method and is widely used in various fields because it is simple, easy to implement, has the ability to do clustering with large data and is able to handle data outliers. The stages in k-means clustering (Tan, 2018) are as follows:

1. Select k centroid point randomly. Centroid can be a vector that is considered to be the midpoint of a cluster.
2. Group each data point into the most suitable cluster based on the size of the proximity to the centroid.
3. After all data is divided into k clusters. Update the centroid using data inside the cluster.
4. Repeat steps 2 and 3 until the values from the centroid do not change.

The online clustering algorithm is used to effectively group Twitter data streams in real-time (Becker *et al.*, 2011). Therefore we need an algorithm that does not specify the number of clusters because data tweets will continue to increase with different content from time to time. Based on observations, this study proposes to use an incremental clustering algorithm, one of which is sequential k-means with threshold parameters that are arranged empirically during the experimental phase.

The sequential K-means algorithm has an algorithm similar to the classic k-means algorithm (batch mode) (MacQueen, 1967). In contrast, the classic k-means algorithm uses all training data to recalculate cluster centers. Whereas in sequential k-means, each data point contributes to the cluster center update. Sequential k-means is used to group data one by one then calculate the cluster center incrementally. The k-means sequential algorithm uses the online clustering concept of Lloyd's (Tan, 2018). This study proposes a variation of the online version of the sequential k-means algorithm and is adapted to input data, namely vector space models from text data.

Topic detection is a process used to obtain topics or keywords from documents automatically. The frequency of each word in the document is obtained by using the Term Frequency-Inverse Document Frequency (TF-IDF), illustrated in point B. The greater the TF-IDF value, the word often appears in one document and the TF-IDF value is smaller appears in many documents. Thus the word that has the greatest TF-IDF value in each document is the topic of the document.

To find out the quality of the cluster, a cluster quality testing is performed, one of which is the Dunn Index method. Dunn Index was introduced by Dunn (1974). The purpose of DI was to find good intra-cluster and inter-cluster relationships. Dunn index is defined in the below equation:

$$D = min_{1\le i\le n}\{min_{1\le j\le n, j\ne i}(\frac{d(c_i, c_j)}{max_{1\le k\le n} diam(c_k)})\}$$

| DOI: http://dx.doi.org/10.17509/ijost.v4i1.15807
| p- ISSN 2528-1410 e- ISSN 2527-8045

where $d(c_i, c_j) = min_{x \in c_i, y \in c_j}\{d(x,y)\}$ is the distance between $c_i$danc_j, $diam(c_k) = max_{x,y \in c_i}\{d(x,y)\}$is the highest distance to the cluster c_k, and *n* is the number of cluster. The value of dunn index ranges from 0 to ∞. Cluster evaluation is said to be good if the value of the obtained index is high.

In this study, trending topics are the top five topics based on the highest number of members. Topics that will become a trend are topics that Twitter users continue to talk about so that the words that represent the topic will continue to appear in the latest data stream. Therefore, the topic that has become one of the trending topics lists needs to be re-evaluated to find out whether the topic is still being discussed or not. The way to evaluate trending topics is by checking the popularity of trending topics every 3 hours by calculating the difference in the number of tweets or deltas in the last 3 hours for each cluster. If there are additions of at least 100 tweets, then the topic or cluster will be retained, otherwise the cluster will be deleted.

2.2. Experimental Study.

Tweets collected are tweets posted by users in one specific location, namely the city of New York (the location displayed by Twitter users on their profile). This location was chosen because it generates a high volume of tweets consistently. Collection of tweets is done through program code in the R language and uses the stream R package. Requests for the latest tweets continue from the Twitter Streaming API until the program stops. Components used in the data tweet are only text containing messages posted by the user. Hastag which is inserted in the tweet is not used because it is not necessarily related to the contents of the text.

Before the data of the tweet data is analyzed, preprocessing of the data is carried out first, which consists of preprocessing text (in point A), weighting the word (in point B) and word selection (in point C). After that, the data was analyzed by clustering techniques online using the k-means sequential algorithm (in point E). While the classic k-means algorithm (in point D) is only used in the first iteration for the initial initialization of the cluster. Information obtained from clustering results are cluster centers and cluster member data. After that, the information processing is done by detecting the topic (in point F) on each cluster and each topic is sorted by number of members, then a maximum of five top topics are taken to become trending topics. Trending topics that have been detected will be evaluated every time (at point H) so that there will always be changes to the trending topics list.

Experiments were carried out in three scenarios using different parameter values. This is done to find the best parameter value. The first scenario was carried out on November 26-27 2016 and collected 32,112 tweets for testing if each data request was carried out for 100 seconds and delta was 50 tweets. The second scenario was carried out on December 3-4, 2016 and collected 55,354 tweets for testing if the data request time was reduced for 60 seconds and the delta was enlarged by 100 tweets. The third scenario is similar to the second scenario, the difference in the experiment is carried out with a longer duration, namely on December 13-31 2016 and collected about 781,450 tweets. The other differences in scenario three are that each hastags inserted in each tweet are stored and accumulated per day, then a comparison of trending topics from free words with the most popular hastags from the same data tweets is compared. This is done to analyze whether hastags that often appear in harmony with words that often appear in the twitter data stream. The total dataset collected in this study is 868,916 tweets posted at the end of November to December 2016.

3. RESULTS AND DISCUSSION

Cluster evaluation (in point G) is carried out at each iteration with different input data. Evaluation values are calculated using the Dunn Index method whose values range

DOI:http://dx.doi.org/10.17509/ijost.v4i1.15807 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

from 0 - ∞. In scenario 1, out of 738 iterations produced only 64 iterations occur taking data, the rest there is no input data because of an unstable internet connection. Therefore, this program requires a good internet connection, stable and continuously connected when the program is run. The quality of the cluster is said to be good if the value of the index obtained is high. In scenario 1 only 10 iterations have more than zero evaluation values. Dunn index value is influenced by varied and complex input data. Input data in the form of text has a high complexity because the input data attributes are always different for each iteration, so cluster centers will continue to grow with increasing attributes and will be increasingly complex.

In the previous experiment, scenario 1 produced a less than optimal evaluation value. This means that many data points that are not included in the right cluster or data points that enter the cluster only have a small similarity value. This might occur because the cluster center contains words that are too broad and that the cluster center should be deleted if the addition of members is less significant. Therefore, in scenario 2, the data request time is reduced by 60 seconds and the delta is enlarged by 100 tweets and the result is 685 iterations of 1071 iterations have a dunn index value above zero. This shows that there is an increase in cluster quality from the experiments in the previous scenario. So that the parameter value in scenario 2 will be used in scenario 3 because it is considered good enough.

In scenario 3, it is determined that the cluster that has good quality is a cluster with a value of dunn index of more than 0.7. A total of 4,782 iterations of 13,479 iterations have a dunn index value above zero. This shows that around 35.48 percent of the total iteration has good cluster quality. So the experimental results in scenario 3 will be the final conclusion in this study. Trending topics generated in scenario 3 can be seen in **Table 1**.

Tabel 1.Trending topics on December 31, 2016, consisting of *Times, Hours, Eve, Party, and Resolution*.

Trending Topics	Number	*Hastags*	Number
Times	1715	#happy newyear	1047
Hours	1553	#2017	647
Eve	1213	#nye	619
Party	889	#newyear eve	520
Resolution	779	#ufc207	414

Based on **Table 1** it can be seen that the five topics are trending topics in New York before the new year. The topic of "Times" relates to the presence of a new year's celebration night concert in Times Square. The "Hours" topic deals with the calculation of time and seconds towards 2017. "Eve" and "Party" topics relate to celebrations and events/parties on New Year's Eve and the topic "Resolution" relating to hope and change for New Yorkers in in 2017. Compared to the list of hastags collected from the same tweet data, #happynewyear, #2017, #nye, #newyeareve and #ufc207, the five biggest hastags are together talking about 2017 new year.

In scenario 3, the resulting trending topics are visualized by histogram. Next is one of the histograms displayed in the last iteration of scenario 3 which can be seen in **Figure 2**.

| DOI: http://dx.doi.org/10.17509/ijost.v4i1.15807
| p- ISSN 2528-1410 e- ISSN 2527-8045

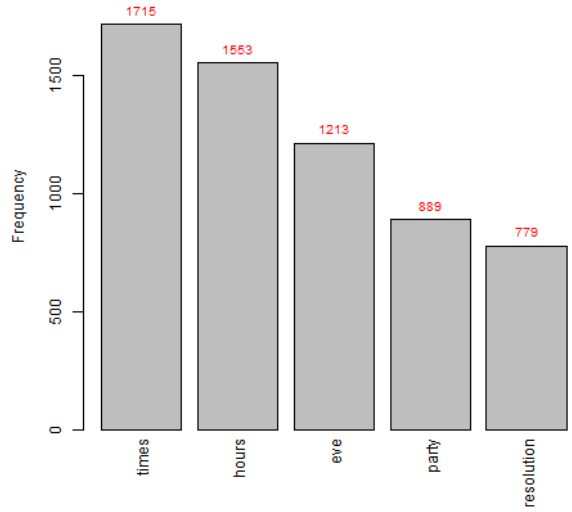

Figure 2. Histogram of trending topics on December 31, 2016

During the program, there are several conditions that occur during the trending topics detection process and can be completed by the program, as follows:

1. The formation of new clusters/topics: The data used in this study is a data stream, so it is important to do a process that can change the cluster structure including adding new clusters. This is done if the newly arrived data does not find the appropriate cluster so that it creates its own cluster. This process is determined by the value of cosine similarity, which is the similarity between the new data and the existing cluster center that occurs during the online clustering stage.
2. Delete the cluster: This process occurs during the trending topics evaluation on point F. The cluster will be deleted if there is no increase in the number of members of at least 100 tweets in a span of 3 hours.
3. Change the topic name in the cluster: Topics are words that represent clusters that are determined from the largest word weights in the cluster center. The cluster center will continue to be updated in each iteration by averaging cluster centers with new data entering the membership. So that, the cluster center will continue to grow if it experiences an increase in attributes/words and changes in value or remains at each word weight. This causes the order of word weights in the cluster center to change.
4. Changes to ranking on trending topics: Trending topics are the top 5 topics based on the number of cluster members so that the clustering or ranking process is carried out. The number of members of each cluster will continue to increase or remain at each iteration. So that the ranking process is carried out at each iteration.

Research related to determining trending topics was carried out by many researchers over the past years using different approaches. There are ten studies that will be compared with the research conducted by the authors, namely research by Zubiaga *et al. (*2015) regarding the real-time classification process of tweet data into 4 trending topics, namely news, ongoing events, memes and warnings. The classification process is carried out using the Support Vector Machine classification method. This method provides an efficient way to categorize trending topics accurately, quickly and in real-time.

In a study conducted by Becker *et al.* (2011) regarding the clustering process on twitter data streams then a classification process was carried out to distinguish cluster events and non-events. This research provides effective results in displaying real-world events on Twitter.

The research conducted by Lau *et al.* (2012) presents a new modeling-based methodology for tracking events that appear on microblogs such as Twitter. The method presented can detect events using various datasets with injections of new events, then apply to identify trending topics on Twitter. The method used is the topic modeling with the online version of the Latent Dirichlet Allocation.

In the study of Aiello *et al. (*2013) a comparison of six topic detection methods in three Twitter datasets related to major

DOI:http://dx.doi.org/10.17509/ijost.v4i1.15807 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

events. The study was conducted to observe how an event is determined naturally, the volume of activities over time, sampling procedures and data pre-processing. All of these greatly affect the quality of topic detection, which also depends on the topic detection method used. One of the proposed new topic detection methods is based on n-gram co-occurrence and ranking topics with document frequency-inverse document frequency t-time (df-idft) which consistently achieves the best performance in all conditions, making it more reliable than other techniques.

In research by Sahdev and Kabra (2013), trending topics detection was carried out with a social network graph approach to determine individual behavior by connecting individual interactions with other individuals. The social network is used as an interaction pattern that becomes a predictor of topic prediction in research. Another approach used is a non-parametric approach to solve problems in utilizing timestamps on tweets. Both approaches are combined and produce fairly good accuracy.

In Berhandus (2013) research, a methodology was used to detect and identify trending topics from data streams. Data from the Twitter Streaming API will be collected and entered into documents of the same time duration. The Term Frequency-Inverse Document Frequency and Relative Normalized Term Frequency analysis is performed on documents for identifying trending topics. Relative Normalized Term Frequency analysis identifies unigram, bigram, and trigram as trending topics while the Term Frequency-Inverse Document Frequency analysis identifies unigrams as trending topics.

In the Lu and Yang (2012)'s study, a trend analysis was carried out on news topics on Twitter, which included trend prediction and analysis of the causes of trend changes. The method used to predict trends is based on the Moving Average Convergence-Divergence (MACD) indicator. This research defines new concepts as momentum trends and uses them to predict trends in news topics. Then this research offers several causes for trend variations. The experimental results show that the process of predicting trends is simple and effective and the causes for trend variations are also verified.

Research conducted by Miller *et al. (*2015) offered an online algorithm that provides estimates of real-time frequency tags on Twitter time series data. This study offers a model for estimating the topic of the most popular twitter topics at certain time intervals or during a period. The method used is using the Naive algorithm.

Research by Mathioudakis and Koudas (2010) presented a system that performs trend detection on Twitter streams and performs trend analysis. This system is named Twitter Monitor which will identify the appearance of topics on Twitter in real time.

Kim *et al. (*2013) proposed a new scheme to detect trends and keywords from the Twitter data stream. The system prototype is applied in various experiments to show the effectiveness of the scheme created. The scheme that is made is very strong, which can handle the word abbreviation, typing errors and mistakes

4. CONCLUSION

This study presents a model and strategy for identifying trending topics from the twitter. Trending topics in Twitter is a collection of certain topics that are widely discussed by users. The research approach was carried out in four stages, namely twitter data collection, preprocessing data, data analysis with sequential *K*-Means clustering and information processing. Testing of the model is carried out in three scenarios where each scenario is distinguished between the amount of data, time and parameter values. The results show that scenario 1 produced a less than optimal evaluation value. This means that data points

| DOI: http://dx.doi.org/10.17509/ijost.v4i1.15807
| p- ISSN 2528-1410 e- ISSN 2527-8045

that enter the cluster only have a small similarity value. Scenario 2 shows that there is an increase in cluster quality from the experiments in the previous scenario. In scenario 3, it is determined that the cluster that has good quality is a cluster with a value of dunn index of more than 0.7. This shows that around 35.48 percent of the total iteration has good cluster quality. There are five topics being the trending topics in New York before the new year. The topic of "Times" relates to the presence of a new year's celebration night concert in Times Square. The "Hours" topic deals with the calculation of time and seconds towards 2017. "Eve" and "Party" topics relate to celebrations and the topic "Resolution" relating to hope and change for New Yorkers in in 2017.

5. ACKNOWLEDGEMENTS

The present review includes research evidences which were supported by Department of Computer Science Education, Universitas Pendidikan Indonesia.

6. AUTHORS' NOTE

The author(s) declare(s) that there is no conflict of interest regarding the publication of this article. Authors confirmed that the data and the paper are free of plagiarism.

7. REFERENCES

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., and Jaimes, A. (2013). Sensing trending topics in Twitter. *IEEE Transactions on Multimedia*, *15*(6), 1268-1282.

Becker, H., Naaman, M., and Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Benhardus, J., and Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities*, *9*(1), 122-139.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95-104.

Firdaus, C., Wahyudin, W., and Nugroho, E. P. (2017). Monitoring System with Two Central Facilities Protocol. *Indonesian Journal of Science and Technology*, *2*(1), 8-25.

Kim, D., Kim, D., Rho, S., and Hwang, E. (2013). Detecting trend and bursty keywords using characteristics of Twitter stream data. *International Journal of Smart Home*, *7*(1), 209-220.

Lau, J.H., Collier, N., and Baldwin, T. (2012). On-line trend analysis with topic models:\# twitter trends detection topic model online. *COLING 2012*, 10, 1519-1534.

Lu, R., and Yang, Q. (2012). Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, *2*(3), 327.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Berkeley symposium on mathematical statistics and probability*, 1(14), 281-297.

Màrquez, L., and Rodríguez, H. (1998). Part-of-speech tagging using decision trees. *European Conference on Machine Learning*, 1, 25-36.

Mathioudakis, M., and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. *ACM SIGMOD International Conference on Management of data*, 1, 1155-1158.

Miller, E., Vodrahalli, K., and Lee, A. (2015). Estimating trending topics on twitter with small subsets of the total data. *Allen Institute for Artificial Intelligence.*

Rachmadany, A., Pranoto, Y.M., Multazam, M.T., Nandiyanto, A.B.D., Abdullah, A.G., and Widiaty, I. (2018). Classification of Indonesian quote on Twitter using Naïve Bayes. *IOP Conference Series: Materials Science and Engineering*, 288(1), 012162.

Riza, L.S., Asyari, A.H., Prabawa, H.W., Kusnendar, J., and Rahman, E.F. (2018). Parallel particle swarm optimization for determining pressure on water distribution systems in R. *Advanced Science Letters*, *24*(10), 7501-7506.

Riza, L.S., Handian, D., Megasari, R., Abdullah, A.G., Nandiyanto, A.B.D., and Nazir, S. (2018). Development of R package and experimental analysis on prediction of the CO2 compressibility factor using gradient descent. *Journal of Engineering Science and Technology*, *13*(8), 2342-2351.

Riza, L.S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Śle, D., and Benítez, J.M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "roughsets". *Information Sciences*, *287*, 68-89.

Riza, L.S., Pradini, M., and Rahman, E.F. (2017). An expert system for diagnosis of sleep disorder using fuzzy rule-based classification systems. *IOP Conference Series: Materials Science and Engineering*, 185(1), 012011.

Riza, L.S., Nasrulloh, I.F., Junaeti, E., Zain, R., and Nandiyanto, A.B.D. (2016). gradDescentR: An R package implementing gradient descent and its variants for regression tasks. *International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 1, 125-129.

Riza, L. S., Zainafif, A., and Rasim, S. N. (2018). Fuzzy rule-based classification systems for the gender prediction from handwriting. *TELKOMNIKA, 16*(6), 2725-2732.

Sahdev, R. and Kabra, P. (2013). Prediction of trending topics in online social networks like Twitter. Birla Institute of Technology and Science, Hyderabad.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing and management*, *24*(5), 513-523.

Schweinberger, M. (2016). Part-Of-Speech Tagging with R.

Tan, P. N. (2018). *Introduction to data mining*. Pearson Education India.

Zubiaga, A., Spina, D., Martínez, R., and Fresno, V. (2015). Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, *66*(3), 462-473.

| DOI: http://dx.doi.org/10.17509/ijost.v4i1.xxxx
| p- ISSN 2528-1410 e- ISSN 2527-8045