Journal of Physics: Conference Series

PAPER • OPEN ACCESS

DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru

To cite this article: Mustakim *et al* 2019 *J. Phys.: Conf. Ser.* **1363** 012001

View the article online for updates and enhancements.

You may also like

- Twitter Sentimentality Examination Using Convolutional Neural Setups and Compare with DCNN Based on Accuracy
S. Subbiah and R. Dheeraj
- Customer Critique Analysis System for PT. KCI's Twitter
Ahmad Husen, Sari Widya Sihwi and Esti Suryani
- Improve topic modeling algorithms based on Twitter hashtags
Hayder M Alash and Ghaidaa A Al-Sultany

Breath Biopsy® OMNI®
The most advanced, complete solution for global breath biomarker analysis

TRANSFORM YOUR RESEARCH WORKFLOW


Expert Study Design & Management


Robust Breath Collection

Reliable Sample Processing & Analysis


In-depth Data Analysis


Specialist Data Interpretation

This content was downloaded from IP address 103.168.127.2 on 01/02/2023 at 02:37

Journal of Physics: Conference Series **1363** (2019) 012001 doi:10.1088/1742-6596/1363/1/012001

DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru

Mustakim[1,2*], Nurul Gayatri Indah Reza[1,2], Rice Novita[1,2], Oktaf Brillian Kharisma[1], Rian Vebrianto[3], Suwanto Sanjaya[1], Hasbullah[4], Tuti Andriani[3], Wardani Purnama Sari[3], Yulia Novita[3], Robbi Rahim[5]

[1]Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia.
[2]Puzzle Research Data Technology, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia.
[3]Faculty of Education and Teaching, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia.
[4]Faculty of Ushuluddin, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia.
[5]Informatic Management, Sekolah Tinggi Ilmu Manajemen SUKMA, Medan, Indonesia.

*mustakim@uin-suska.ac.id

Abstract. Social media is one of the most common sources used to communicate, such as Twitter. Every tweet on Twitter contains data such as text which when collected can be processed into information. Data processed from Twitter tweet will create a trend which can be used for information such as in education, economics, politics, etc. This then created the concept of text mining. Text mining techniques are needed to find an interesting pattern in search of trends based on Twitter text with topics related to Pilkada Pekanbaru 2017. This research is intended to cluster Twitter text data using Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm. This research was conducted with several experiments using different Eps and MinPts parameters for 2,184 text data which has been through several stages, such as cleaning, duplication removal, pre-processing like stemming and stopwords. Based on the highest average of Silhouette Index, Eps 0.1 and MinPts 10 with SI = 0.413 were chosen as paramaters, thus forming 31 clusters. According to the frequency of word occurrences in the cluster, the highest are "kpu", followed by "firdaus", "kota", "pasang", and "ayat". As can be seen that the candidate pairs most often appear on cluster results are Firdaus-Ayat, and based on the results of Pilkada 2017, Firdaus-Ayat was chosen as Mayor and Vice Mayor of Pekanbaru.

1. Introduction

Social media is one of the most common sources which is used to communicate, documents sharing, photos, and videos with large communities. One example of social media is Twitter which is a microblogging service with monthly active users reaching 317 million worldwide [1] and Indonesian user by 50 million users in 2015 alone [2]. Every tweet on Twitter contains data such as text that when collected can be processed into information. This data may contains personal user's data, opinions on

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

current issues, opinions about a product or service which they use or their views on political and religious issues. Data processing from Twitter will create a trend that can be used for information such as in education, economics, politics, etc. This then created the concept of text mining.

The large amount of data generated from Twitter users is a source of knowledge and will be worth to consider its use in analyzing online social communities and user activities. The data can be used to analyze user opinion related to an event or aspect from certain characteristics according to the user's perception. Sentences on tweets which may contain non-standard word and symbols. Which causes the user difficulty in manually interpreting the thousands of text tweets. Text mining is a technique to analyze large text data. Text mining is able to find trend interpretations from user opinions, or analyze sentiments based on Twitter tweet.

Text mining is one of the special area of data mining. According to Feldman and Sanger (2007) text mining is the process of digging information where the user interacts with a set of documents from time to time using a set of tools analysis [3]. Text mining merupakan teknik yang digunakan untuk menangani masalah klasifikasi, clustering, information extraction dan information retrieval [4]. Text mining techniques are needed to find an interesting pattern in searching for trends based on Twitter text with topics related to Pilkada Pekanbaru 2017. Pilkada Pekanbaru 2017 is a topic which recently happened in Pekanbaru. This topic was chosen because the research is expected to provide insight or knowledge to the public about politics of Pekanbaru in cyberspace.

One of the techniques known in text mining is Text Clustering. Text clustering is the process of dividing text content into different groups according to their similarities, such as cosine similarity or dice similarity so that text clustering will find which documents have the most common words [5]. The Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm is used because it can identify clusters arbitrarily, able to cope with noise and outliers, and does not require the specified number of clusters expected in the data [6].

In the DBSCAN algorithm, clusters are identified as dense areas of data objects surrounded by low density areas [6]. The density is evaluated based on Eps and MinPts parameters determined by user. If the amount of data within a radius of Eps is greater than or equal to MinPts, the data falls into the desired density category, the amount of data within that radius including the data itself [7].

Based on previous research, text mining has been used by Himanshu Suyal et al, about Text Clustering Algorithms: A Review with the result of research is the comparison and complexity calculation of some clustering algorithms such as cluster hierarchy, K-Means, K-Medoid, DBSCAN, DENCLUE, and Self -Organizing Map (SOM) based on parameters so it can be used to construct new cluster algorithm which can cluster data efficiently [5]. Another study was A Case Study in Text Mining: Interpreting Twitter Data From World Cups by Daniel Godfrey et al which compared Cluster K-Means and Non-Matrix Factorization (NMF) techniques for analysis, proving that NMF is faster and easier in interpreting the results. In this research DBSCAN algorithm combined with consensus matrix was also applied to reduce noise from data [8]. According to Godfrey, et al, DBSCAN was able to eliminate noise in the data.

Based on the description above and supported by some previous research, then in this research text mining in searching trend Twitter using DBSCAN algorithm will be experimented. It is expected that the text data generated from Twitter can be processed to find trends based on the topic and can provide convenience in reading the data about the pattern of communication of Pekanbaru citizen related to Pilkada Pekanbaru 2017. The purpose of this research is to impelement DBSCAN algorithm as one of the clustering method to find Twitter trend from text as well as knowledge of trend information on interesting pattern based on "Pilkada Pekanbaru 2017" which describes the political situation in cyberspace especially Twitter. The dataset was generated from Twitter since September 1st, 2016 to February 14th 2017 with a total of 8,158 data.

2. Material and Method

2.1. Twitter

Twitter was founded in March 2006 by Jack Dorsey, and its social networking site was launched in July. Twitter is a website owned and operated by Twitter Inc. Twitter offers a social network of microblogs allowing users to post and read tweets [9].

2.2. Election of Regional Head (Pilkada) Pekanbaru

On February 15th, 2017, there was a simultaneous elections for several regions in Indonesia organized by KPU. Pekanbaru is one of that regions. Pilkada Pekanbaru 2017 was held to elect the Mayor and Vice Mayor period 2017-2022. There are 5 pairs of candidates who participated in Pilwako Pekanbaru 2017, namely Syahril and Said Zohrin number 1 from the independent, Herman Nazar and Defi Warman number 2 from the independent, Firdaus and Ayat Cahyadi number 3, supported by the Democratic Party, Gerindra, and PKS, M. Ramli and Irvan Herman number 4 supported by PKB, PAN, Hanura, Golkar and NasDem parties, and the last Dastrayani Bibra and Said Usman Abdullah number 5 from the PDIP party and PPP.

Based on the results of Pilkada, 273,342 (48.1%) of Pekanbaru residents chose not to exercise their voting rights [10]. Although this number was lower than the elections in 2011 with a total of 277,838 voters or about 51.82% [11], the number of voters who did not exercise their voting rights remained close to half of all voting residents in Pekanbaru. The results of Pilkada 2017 can be seen in the following table.

Table 1. Result of Pilkada Pekanbaru 2017

No	Candidate Names	Voters	%
1	DR. H. Syahril, S.Pd, MM and H. Said Zohrin, SH, MH	22,149	7.78
2	H. Herman Nazar, S.H ., M.Si and Defi Warman, S.Pd., M.Pd	46,472	16.33
3	Dr. H. Firdaus, ST,. MT and H. Ayat Cahyadi, S.Si	94,118	33.07
4	DR. H. M. Ramli, S.E., M.Si. and Dr. Irvan Herman	59,613	20.95
5	Drs. H. Dastrayani Bibra, M.Si and H. Said Usman Abdullah	62,249	21.87
	Total	284,601	100%

2.3. Text Mining

Text mining is one of the special areas of data mining. Text mining is one of the steps from text analysis which is done automatically by a computer to dig up quality information from a series of texts summarized in a document [12]. According to Feldman and Sanger (2007) text mining is the process of digging information where the user interacts with a set of documents from time to time using a set of tools analysis [3]. Text mining is a technique used to handle the problem of classification, clustering, information extraction and information retrieval [4].

The main procedure in this method is related in finding words that can represent the contents of the document for subsequent analysis of connectedness between documents using certain statistical methods such as cluster analysis, classification and association. The initial stages in text mining are called preprocessing text.

2.4. Text Preprocessing

The preprocessing text includes all routines, and a process for preparing the data to be used on the operation of the discovery knowledge of the text mining [3]. Steps of preprocessing text in general are tokenizing, filtering, stemming, tagging, and analyzing [13]. Tokenizing is the process of deciphering the original description of sentences into words and eliminate delimiters such as periods (.), commas (,), spaces and numbers characters that exist in the word [14]. Filtering is the process of word selection resulting from tokenizing process, it can be done with the stopword algorithm (stop list) or word list. Stopword is a vocabulary that is not a feature (unique word) of a document [15]. Stemming is the process of mapping and decomposing various forms (variants) from a word to its basic form (stem) [16].

2.5. Text Reprecentation

There are several ways to model a text document. For example, by presenting the document in bag-of-words, in which words are assumed to appear independently and the sequence is unimportant [17]. Text data is represented in Vector Space Model by using TF-IDF weighting.

Term Frequency-Inverse Document Frequency (TF-IDF) is often used as a weighting factor in information retrieval and text mining. The value of TF-IDF increases proportionally based on how many words appear in the document (term frequency), but is neutralized by the word frequency in the corpus (inverse document frequency) [18]. In this method, the calculation of term t weight in a document is done by multiplying the value of Term Frequency with Inverse Document Frequency. The term frequency value is generated from the number of terms in a document. Meanwhile, to generate Inverse Document Frequency value is by Equation 1 [18]:

$$idf_j = \log\left(\frac{D}{df_j}\right) \quad (1)$$

Vector Space Model (VSM) is a representation of a document as a vector in vector space. VSM is a basic technique to obtain information used to assess the relevance of documents against query(keyword), document classification, and document clustering [19]. In the Vector Space Model, the document collection is represented as a document-term matrix or also known as a bag-of-words model to reduce the dimensions of data and rank the documents by relevance by ignoring the order of words in document [20]. Here is an example of the document-term matrix [20]:

$$\begin{Bmatrix} & T_1 & T_2 & \dots & T_t \\ D_1 & W_{11} & W_{21} & \dots & W_{t1} \\ D_2 & W_{12} & W_{22} & \dots & W_{t2} \\ \dots & \dots & \dots & \dots & \dots \\ D_n & W_{1n} & W_{2n} & \dots & W_{tn} \end{Bmatrix}$$

Figure 1. Example of term-document matrix for database with n documents and t term

Through the vector space model and TF-IDF weighting, resulting in representation of the numerical value of documents thus the proximity between documents can be calculated. The closer the two vectors in a VSM, the more similar the two documents represented by the two vectors. The similarities between documents can be calculated using a function of similarity measure.

One of the most popular measures of text similarity is cosine similarity. The equation of document vector with the query vector is the cosine angle between the two. This measure calculates the cosine value of the angle between two vectors. If there are two document vectors dj and query q, and t term is extracted from the document collection the cosine value between dj and q is defined by the following equation 2 [20][21]:

$$similarity\left(\vec{d_j},\vec{q}\right) = \frac{\vec{d_j}.\vec{q}}{\left|\vec{d_j}\right|.\left|\vec{q}\right|} = \frac{\sum_{i=1}^{t}(W_{ij}.W_{iq})}{\sqrt{\sum_{i=1}^{t} W_{ij}^2 \cdot \sum_{i=1}^{t} W_{iq}^2}} \quad (2)$$

2.6. Density-Based Spatial Clustering of Application with Noise (DBSCAN)

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a clustering method that builds areas based on its density connected. DBSCAN is a type of partition clustering where high density areas are considered clusters whereas low-density or incompatible clusters are regarded as noise [22]. DBSCAN requires two input of parameters, epsilon (Eps) and minimum point (MinPts). Eps-point around is defined by equation 3:

$$N_{Eps}(x) = \{y \in D \mid dist(x, y) \leq Eps\} \quad (3)$$

Where N Eps (x) is the dot of x in the radius of Eps, D is the data cluster, dist (x, y) is the euclidean distance of the objects x and y, and Eps is the radius or threshold, in this study, the distance measure used is cosine distance. Cosine distance is a standard calculation to find the distance between text data [6]. Cosine distance is obtained by subtracting 1 and cosine similarity [4]:

$$cosine\ distance = 1 - cosine\ similarity \quad (4)$$

2.7. Silhouette Coefficient

The purpose of cluster validation technique is to evaluate the cluster results, the results of this evaluation can be used to determine the number of clusters in the dataset [23]. One of the cluster validation techniques is Silhouette Coefficient or Silhouette Index (SI) which is an interpretation method to cluster validation on objects. This technique provides a brief graphical representation of how well each object lies within its cluster. Silhouette Coefficient was first developed by Rousseeuw in 1986. Silhouette Coefficient is a ratio type index that relies on cohesion and cluster segregation [23]. To get the value of Silhouette Coefficient of i-th data using the following equation:

$$b_i = \frac{b_i - a_i}{\text{Max}\{a_i, b_i\}} \quad (5)$$

3. Research Methodology

Method applied in this research can be seen in Figure 2 bellow:

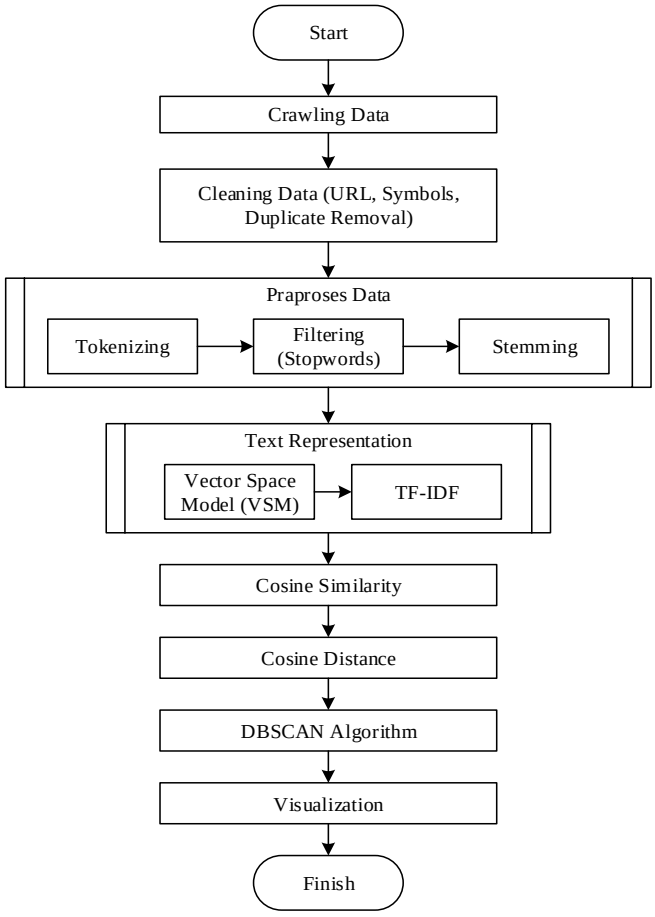

Figure 2. Research Methodology

4. Results and Discussion

According to research methodology in the previous discussion, several important things which will be done in this research consist of data collection, cleaning, pre-processing, DBSCAN algorithm analysis and Silhouette Index analysis.

4.1. Data Collection

In data collection, the data used in the form of secondary data. The data obtained from Twitter crawling with query search "Pilkada Pekanbaru", "Pilwako Pekanbaru" and "Paslon Pekanbaru" thus produce the total of 8,469 data from September 2016 until February 2017.

Before data can be processed, tweets will go through the cleaning process because a lot of data in Indonesian tweets use non-standar sentences and words, using symbols and numbers, and URLs. From this stage, the initial data tweet of 8,158 data was reduced to 2,523 data, which consist of 1,449 data from the query "Pilkada Pekanbaru", 667 data from "Pilwako Pekanbaru" and 407 data from "Paslon Pekanbaru" due to duplication removal.

Table 2. Data After Cleaning from Query "Pilkada Pekanbaru"

No	Text
1	*laporan dugaan keterlibatan asn di pilkada pekanbaru panwaslu tak cukup bukti*
2	*walau paska pilkada asn pekanbaru dituntut tetap bekerja profesional*
3	*pasca pilkada sekdako bilang tingkat disiplin asn pemko pekanbaru menurun*
4	*kalah di pilkada bekasi ahmad dhani melayani permintaan jadi kepala daerah*
5	*tak capai target pj bupati puas dengan partisipasi pemilih pilkada kampar*
…	…
1,449	*artis david chalik maju di pilkada pekanbaru beritasatu*

4.2. Data Preprocess

Text data will go through several pre-process stages such as tokenizing, filtering, stopwords, and stemming. The stopwords list wast taken from Tala (2003) of 740 words and added with some common abbreviations so that the total stop words are 765 words. Term Frequency-Inverse Document Frequency (TF-IDF)

Each data from a different query is weighted by TF-IDF weighting with the condition of eliminating terms with a maximum proportion of frequency documents exceeding 90% and a minimum less than 3% resulting in 23, 38 and 32 terms for "Pilkada, Pilwako, and Paslon Pekanbaru" respectively.

Table 3. Calculation Result of TF-IDF from Keyword "Pilkada Pekanbaru"

No. Doc	Term						
	aman	***ayat***	***calon***	***dukung***	***firdaus***	***…***	***walikota***
1	*2*	*3*	*5*	*6*	*7*	*…*	*23*
1	0	0	0	0	0	…	0
2	0	0	0	0	0	…	0
3	0	0	0	0	0	…	0
4	0	0	0	0	0	…	0
5	0	0	0	0	0	…	0
…	…	…	…	…	…	…	
1.449	0	0	0	0	0	…	0

4.3. Cosine Similarity

Similarity measure with cosine similarity aims to eliminate data that has no similarity to the query. Data with a cosine value of 0 is omitted because it is considered to have no similarity. Resulting in 1,182 data for query "Pilkada Pekanbaru", 653 data for "Pilwako Pekanbaru" and 343 data for "Paslon Pekanbaru". The total data is 2,184 which will be used for cluster analysis using DBSCAN algorithm.

4.4. DBSCAN Analysis

DBSCAN is a clustering method that builds areas based on density connected. Every object of a radius area (cluster) must contain at least a minimum amount of data. All objects not included in the cluster is considered as noise. DBSCAN generally uses Euclidean distance calculation to find the distance of each data. However, in this study, the distance calculation used is the cosine distance obtained by subtracting the value of 1 with cosine similarity of each documents. So the range of cosine distance is 0-1. From DBSCAN algorithm by using several different Eps and MinPts, the following cluster results:

Table 4. Cluster Result

MinPts	Eps		
	0.1	**0.2**	**0.3**
5	57	19	1
10	31	19	1
15	25	15	1
20	19	16	1
30	12	12	3
40	10	8	6
50	7	7	6

To obtain optimal cluster results, the Silhouette Index (SI) analysis is performed with the following results:

Table 5. SI Analysis

MinPts	Eps			Average
	0.1	**0.2**	**0.3**	
	0.5210	0.1670	0.1140	0.2673
10	0.4130	0.3120	0.1140	0.2797
15	0.3640	0.3120	0.1140	0.2633
20	0.3240	0.2870	0.1140	0.2417
30	0.3030	0.2740	0.1380	0.2383
40	0.3210	0.2560	0.1750	0.2507
50	0.2950	0.2480	0.1710	0.2380
60	0.2410	0.3310	0.2310	0.2677
Average	0.3478	0.2734	0.1464	0.2558

Based on SI analysis, Eps and MinPts were chosen from the highest average SI value, i.e. Eps = 0.1 and MinPts = 10 resulting in 31 cluster trends. The trend from cluster result can be seen in the following table where cluster -1 is labeled as noise.

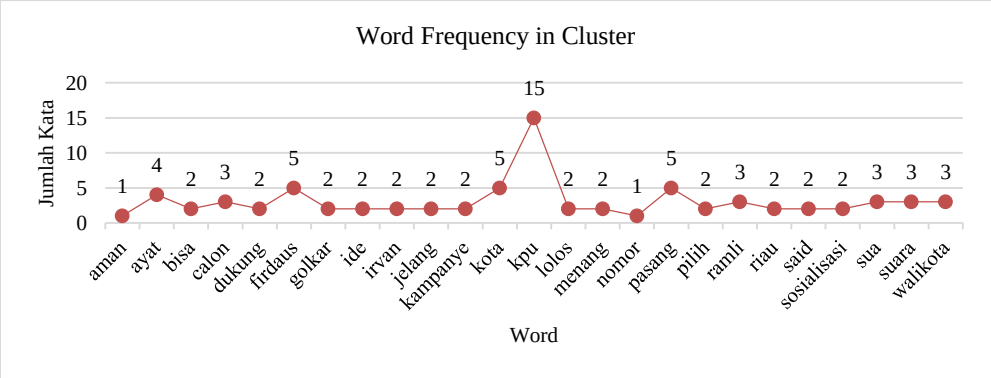

Figure 3. Word Frequency in Cluster

Based on the graph above, the highest frequency of words in the cluster is "kpu", followed by "firdaus", "kota", "pasang", and "ayat". The disadvantage of this research lies in the process of cleaning and preprocessing data, where a lot of text data from Indonesian Twitter do not use standard language. And also the sentiment from trends in clusters can not be known whether they are positive or negative.

5. Conclusion

From clustering experiments using a combination of different Eps and MinPts from DBSCAN algorithm, Eps = 0.1 and MinPts = 10 were chosen based on the average validation of Silhouette Index thus produce 31 cluster, with cluster -1 labeled as noise and the value of SI is 0.413. The highest

frequency of word occurrences in cluster are "kpu", followed by "firdaus", "kota", "pasang", and "ayat". As can be seen that the candidate pairs most often appear on cluster results are Firdaus-Ayat, and based on the results of Pilkada 2017, Firdaus-Ayat was chosen as Mayor and Vice Mayor of Pekanbaru with 33.07% votes. The trend of social media in Pilkada Pekanbaru represent the real situation based on the result of Pekanbaru Mayor Election period 2017-2022, so it can be concluded that the pattern on predicting social media Twitter is similiar with the pattern of society in Pekanbaru. Furthermore, this research can be improved by optimizing the cleaning stage, pre-processing stages in text mining, and determination of Eps and MinPts parameters in order to obtain optimal results. And combine other algorithms such as association rules to know the correlation between words in each cluster and the classification algorithm to know the sentiments of word analysis.

Acknowledgment
A biggest thanks to Puzzle Reseach Data Technology (Predatech) Team for their feedbacks and their assistance in implementing these activities so that research can be done well.

References

[1] Twitter, "Company Facts," 2017. [Online]. Available: https://about.twitter.com/company. [Accessed: 12-Mar-2017].
[2] CNN Indonesia, "Jumlah Pengguna Twitter Di Indonesia Akhirnya Terungkap," 2017. [Online]. Available: http://www.cnnindonesia.com/teknologi/20150326141025-185-42076/jumlah-pengguna-twitter-di-indonesia-akhirnya-terungkap/. [Accessed: 12-Mar-2017].
[3] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.
[4] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "KNN based machine learning approach for text and document mining," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 61–70, 2014.
[5] H. Suyal, A. Panwar, and A. S. Negi, "Text Clustering Algorithms: A Review," *Int. J. Comput. Appl.*, vol. 96, no. 24, 2014.
[6] E. Baralis, T. Cerquitelli, S. Chiusano, L. Grimaudo, and X. Xiao, "Analysis of twitter data using a multiple-level clustering strategy," in *International Conference on Model and Data Engineering*, 2013, pp. 13–24.
[7] E. Prasetyo, "Data Mining Konsep dan aplikasi menggunakan matlab," *Yogyakarta Andi*, 2012.
[8] D. Godfrey, C. Johns, C. Meyer, S. Race, and C. Sadek, "A case study in text mining: Interpreting twitter data from world cup tweets," *arXiv Prepr. arXiv1408.5427*, 2014.
[9] Twitter, "Twitter," 2017. [Online]. Available: https://support.twitter.com/. [Accessed: 12-Mar-2017].
[10] KPU, "Pilkada," 2017. [Online]. Available: https://pilkada2017.kpu.go.id/hasil/t2/riau/kota_pekanbaru. [Accessed: 20-Jun-2017].
[11] Antara Riau, "Golput Pilkada Pekanbaru Capai 50 Persen Lebih," 2011. [Online]. Available: http://www.antarariau.com/berita/14521/melayu. [Accessed: 12-Mar-2017].
[12] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
[13] M. W. Berry and M. Castellanos, "Survey of text mining," *Comput. Rev.*, vol. 45, no. 9, p. 548, 2004.
[14] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media, 2010.
[15] E. Dragut, F. Fang, P. Sistla, C. Yu, and W. Meng, "Stop word and related problems in web interface integration," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 349–360, 2009.
[16] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," *Inst. Logic, Lang. Comput. Univ. van Amsterdam, Netherlands*, 2003.
[17] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49–56.
[18] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[19] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*, vol. 39. Cambridge University Press, 2008.
[20] R. M. Ravindran and A. S. Thanamani, "K-Means Document Clustering using Vector Space Model," *Bonfring Int. J. Data Min.*, vol. 5, no. 2, p. 10, 2015.
[21] S. Shamshirband, A. Amini, N. B. Anuar, M. L. M. Kiah, Y. W. Teh, and S. Furnell, "D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks," *Measurement*, vol. 55, pp. 212–226, 2014.
[22] P. B. Nagpal and P. A. Mann, "Comparative study of density based clustering algorithms," *Int. J. Comput. Appl.*, vol. 27, no. 11, pp. 421–435, 2011.
[23] S. Saitta, B. Raphael, and I. F. C. Smith, "A bounded index for cluster validity," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 2007, pp. 174–187.