Combination of Singular Value Decomposition and K-means Clustering Methods for Topic Detection on Twitter

Khumaisa Nur'aini[1], Ibtisami Najahaty[2], Lina Hidayati[3], Hendri Murfi[4], Siti Nurrohmah[5]
Department of Mathematics
Universitas Indonesia
Indonesia
[1]khumaisa.nur@sci.ui.ac.id, [2]ibtisami.najahaty@sci.ui.ac.id, [3]lina.hidayati@sci.ui.ac.id,
[4]hendri@ui.ac.id, [5]nurohmah@ui.ac.id

***Abstract*—Online social media are growing very rapidly in recent years, such as Twitter. Even the interaction and communication in the social media can reflect on the events of the real world. This causes the value of the information increasing significantly. However, the huge amount of the information requires a method of automatically detecting topics, one of which is the K-means Clustering. Moreover, the large dimensions of data become obstacles. So, we used singular value decomposition (SVD) to reduce the dimension of the data prior to the learning process using the K-means Clustering. The accuracy of the combination of SVD and K-means Clustering methods showed comparative results, while the computation time required is likely to be faster than the method of K-means Clustering without any reduction in advance.**

Keywords*—*topic detection; K-means clustering; dimension reduction; singular value decomposition; Twitter

I. Introduction

The rapid development of social media on the Internet has increased the amount of information available. One example of the popular online social media in recent years is Twitter. Twitter is a social media that allows users to send and read information in textual documents. The ease of obtaining and sharing information makes Twitter user number increases every year. In 2012, 140 million people are registered as users of Twitter, this number increased by 80 million within a period of two years. In Indonesia, at the end of 2014 the number of Twitter users reached 15.3 million users, surpassing UK for the first time for a number of users[1].

Twitter users get more active in producing tweets about real-world events almost in real-time, therefore, Twitter becomes accurate sensors of real-world events. Determining which are the topics being discussed on Twitter is very useful e.g. for journalists who use the Twitter to find news for their media [1]. However, finding topic on a very large number of tweets is hard to do manually so that automatic methods are needed. Automatic method to detect topics in textual documents is known as Topic Detection and Tracking (TDT) [2]. TDT is the study of detecting topics and tracking the evolution of those topics in a constantly updated collection. TDT can be applied to several domains such as news, research papers, or digital library. There are some methods for topic detection, i.e. latent Dirichlet allocation (LDA) [3][4], and nonnegative matrix factorization [5][6]. Another approach used in topic detection is clustering [7][8]. Clustering is a technique of grouping data so that members of the same group more homogeneous or similar to each other than with members of different groups. One of popular clustering methods is K-means clustering [9]. Given tweets of Twitter, this method will split the tweets into k clusters in which each tweet belongs to the nearest cluster center. The cluster centers are means of their members. Then, the cluster centers or centroids are interpreted as topics of the tweets. Some papers consider the singular value decomposition (SVD) as another topic detection method. However, SVD results candidate topics with both positive and negative values. These negative values make interpretation are difficult or even impossible.

Clustering in high dimensional data, such as tweets, requires quite a long time. For some applications, this situation is not so applicable. They need a reasonable response time in order to serve more usable. Therefore, a method to reduce the computational time is needed for this case. Dimensionality reduction is one of the approaches that usually used to improve time complexity by reducing the dimension into a smaller size. SVD is a popular method for the dimensionality reduction. For textual data, this method is also known as latent semantic analysis (LSA) [10][11]. In this paper, we combine SVD and K-means clustering as a topic detection method on Twitter. Firstly, the dimensions of tweets are reduced using SVD. Having these reduced forms of tweets, we use the K-means clustering method to partition the tweets. The centroids of the clusters form topics of the given tweets. Our experiments show that the combination method gives comparative accuracy with the method without any dimensionality reduction, while its computation time is likely to be faster.

The rest of the paper is organized as follows: Section II reviews K-mean clustering and SVD. Section III describes the

[1] Emerging Markets Drive Growth Twitter Users Worldwide http://www.emarketer.com/Article/Emerging-Markets-Drive-Twitter-User-Growth-Worldwide/1010874#sthash.Amee9JvH.dpuf

This research was supported in part by PUPT 2015

978-1-5090-0363-1/15/$31.00 ©2015 IEEE

combination method for topic detection. In Section IV, we show our experimental results. We conclude and give a summary in Section V.

II. Basic Theory

A. K-means Clustering

The purpose of the K-means clustering is to group a set of data into K clusters with K values that have been given. Suppose there are M data $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_M\}$ where $\boldsymbol{a}_i$, for $i = 1, 2, \dots, M$, have D dimension. If $\boldsymbol{\mu}_k$, a vector that has D dimension, denotes the center of the cluster (centroid) $\boldsymbol{\mu}_k$, $k = 1, 2, \dots, K$, then the objective function of K-means clustering, called the distortion measure, is defined as,

$$J=\sum_{m=1}^{M}\sum_{k=1}^{K} r_{m,k}\left\|\boldsymbol{a}_m\text{-}\boldsymbol{\mu}_k\right\|^2,$$

where $r_{m,k} \in \{0,1\}$ is a binary indicator [9]. $r_{m,k}$ will be equal to 1 if $\boldsymbol{a}_m$ have a smaller squared distance to the centroid $\boldsymbol{\mu}_k$ than centroid $\boldsymbol{\mu}_j$ for $j \neq k$. That is, if $\boldsymbol{a}_m$ are in cluster k then $r_{m,k}= 1$, and $r_{m,j}= 0$ for $j \neq k$.

J is minimized by finding the values of $r_{m,k}$ and $\boldsymbol{\mu}_k$ in an iterative process. K-means algorithm begins by selecting an initial value for $\boldsymbol{\mu}_k$. The first step, find the value $r_{m,k}$ that minimizes J by keeping the value of $\boldsymbol{\mu}_k$ fixed. The second step, find the value of $\boldsymbol{\mu}_k$that minimizes J by keeping the value of $r_{m,k}$ fixed. These steps are repeated until they reach the convergence [9]. This means, these two steps are repeated until there is no more change of data in each cluster or reaches a maximum number of iterations. Every step will minimize the value of the objective function J, then the convergence of the algorithm can be guaranteed. However, the K-means algorithm will converge to a local minimum of J rather than to the global minimum of J [9].

As in [9], formal form to get the value $r_{m,k}$ is

$$r_{m,k}=\begin{cases}1 & \text{if } k=\arg\min_j \left\|\boldsymbol{a}_m\text{-}\boldsymbol{\mu}_k\right\|^2 \\ 0 & \text{others}\end{cases} \tag{1}$$

and $\boldsymbol{\mu}_k$ can be found by

$$\boldsymbol{\mu}_k=\frac{\sum_m r_{m,k}\boldsymbol{a}_m}{\sum_m r_{m,k}}. \tag{2}$$

In general, the K-means algorithm can be described as follows:

a. Determine the number of clusters, $1 < K < M$.
b. Determine initial centroid $\boldsymbol{\mu}_k$ of each cluster, $k = 1, 2, \dots, K$.
c. Determine the value of $\left\|\boldsymbol{a}_m\text{-}\boldsymbol{\mu}_k\right\|^2$, the squared distance of each data into each centroid, $m = 1, 2, \dots, M$ and $k = 1, 2, \dots, K$.
d. Group data based on closest squared distance to a centroid.
e. Calculate the new centroid of each cluster using Equation 2.
f. Repeat steps c, d and e, the process stop (converge) when there is no data change or reaches the maximum number of iterations.

B. Singular Value Decomposition

Singular Value Decomposition (SVD) is defined as follows:

Definition B.1 [12] Let A be an $m \times n$ matrix. A factorization $A=U\Sigma V^t$ is said Singular Value Decomposition for A, where U is an $m \times m$ orthogonal matrix, Σ is an $m \times n$ pseudo-diagonal matrix whose elements nonnegative, and V is an $n \times n$ orthogonal matrix. The diagonal elements of the matrix Σ are called singular value of A.

SVD uses the properties of the symmetric matrix shown by the following theorems.

Theorem B.1 [13] Let A be $m \times n$ matrix,
(a) The matrices A^tA and AA^t are symmetric.
(b) Nullity(A) = nullity(A^tA).
(c) Rank(A) = rank(A^tA).
(d) The eigenvalues of A^tA and AA^t are real and nonnegative.
(e) The nonzero eigenvalues of A^tA and AA^t are the same.

Theorem B.2 [13] Matrix A is symmetric if and only if there exist a diagonal matrix D and an orthogonal matrix P with $A=PDP^t$.

1) SVD Algorithms

Suppose that matrix A is an $m \times n$ matrix. Then, matrix A^tA is a symmetric matrix (Theorem B.1), and by Theorem B.2, it can be obtained a factorization

$$A^tA=PDP^t,$$

where D is a diagonal matrix whose entries are the eigenvalues of A^tA, and P is an orthogonal matrix such that the i-th column of the matrix P is the eigenvector corresponding to the i-th eigenvalues on the diagonal matrix D [13].

According to Definition B.1, if given a matrix A then $A=U\Sigma V^t$ is the Singular Value Decomposition for A, where U and V is an orthogonal matrix, and Σ is a pseudodiagonal matrix. By Definition B.1, it can be formed symmetric matrix A^tA as follows:

$$A^tA=(U\Sigma V^t)^t(U\Sigma V^t)$$
$$A^tA=V\Sigma^tU^tU\Sigma V^t$$

Because U is an orthogonal matrix, therefore $U^tU=U^{-1}U=I$, thus

$$A^tA=V\Sigma^tI\Sigma V^t=V\Sigma^t\Sigma V^t$$
$$A^tA=V\Sigma^t\Sigma V^{-1}.$$

Because Σ is a pseudodiagonal matrix, matrix$\Sigma^t\Sigma$ is a diagonal matrix, where the diagonal elements of $\Sigma^t\Sigma$ are eigenvalues of matrix A^tA. If λ_i for i=1, 2,..., n is eigenvalue of A^tA, and $\sigma_i=\sqrt{\lambda_i}$ is singular value of A, then matrix Σ can be obtained by arranging the singular value from the largest to the smallest,

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$$

Thus,

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \sigma_n \\ 0 & \ldots & \ldots & 0 \\ \vdots & \ldots & \ldots & \vdots \\ 0 & \ldots & \ldots & 0 \end{bmatrix}.$$

Matrix V is an orthogonal matrix whose columns are the normalized eigenvectors of the matrix A^tA corresponding with singular values of Σ.

Matrix U can be formed in a similar way to V by forming a symmetric matrix AA^t as follows:

$$AA^t = (U\Sigma V^t)(U\Sigma V^t)^t$$
$$AA^t = U\Sigma V^t V \Sigma^t U^t$$
$$AA^t = U\Sigma I \Sigma^t U^t = U\Sigma\Sigma^t U^t.$$

The efficient SVD algorithms are discussed in more detail in [13].

2) Truncated SVD

Illustrations of matrix factorization using SVD can be seen from the following figure:

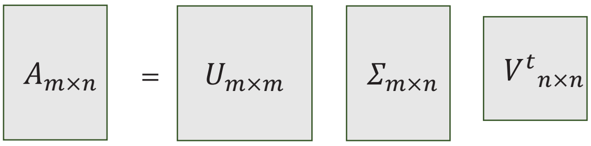

Fig. 1. Illustrations of matrix factorization using SVD

Singular value on matrix Σ are sorted from the largest to the smallest, then the best possible approximation to the matrix A can be formed by taking the first p rows and columns of matrix Σ. Taking p rows and p columns of the matrix Σ not only eliminates the zero vector, but also delete some singular values that are relatively small [13][14]. Fig. 2 illustrates the truncated SVD.

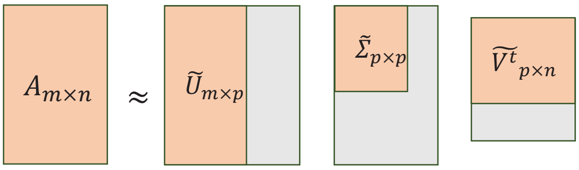

Fig. 2. Illustrations of truncated SVD

III. Combination of SVD and K-means Clustering Method for Topic Detection

Latent semantic analysis (LSA) is a technique of projecting the document to the pent-dimensional space of the data relating to latent semantic space [10][11]. LSA is an extension of the vector space model that uses SVD which can projects document into smaller dimensions than the original dimension. Because the singular values are ordered in decreasing order, it is possible to remove the smaller dimensions and still account for most of the variance. This means that LSA can be used as a method to reduce the dimension of document using truncated SVD algorithm.

Let A be a $m \times n$ *word-tweet* matrix, LSA decomposes the matrix A into a $m \times p$ orthogonal matrix $\tilde{U}$, a $p \times p$ matrix pseudo-diagonal $\tilde{\Sigma}$, and a $p \times n$ orthogonal matrix $\tilde{V}^t$. $\tilde{U}$ is a *word-latent semantic* matrix, $\tilde{\Sigma}$ is a *latent semantic-latent semantic* matrix, and $\tilde{V}^t$ is a *latent semantic-tweet*. Therefore, the $p \times n$ matrix $\check{\Sigma}\tilde{V}^t$ is a reduced form of matrix A because p can be set much smaller than m.

To detect topics from tweets, K-means clustering algorithm can be performed on the reduced form of the tweets, that is, the matrix $\check{\Sigma}\tilde{V}^t$. The reduced matrix is a *latent semantic-tweet* matrix, then the centroid of each cluster cannot be directly interpreted as a topic because the reduced matrix is not showing the relationship between words and tweets. Therefore, it is necessary to transform the discovered centroids into the original dimension of tweets, that is, the words. In general, the combination algorithm for topic detection from Twitter is described as follows:

a. Reduce dimension of a word-tweet matrix A using truncated SVD to form a reduced latent semantic-tweet matrix $\check{\Sigma}\tilde{V}^t$.
b. Cluster tweets in the reduced form $\check{\Sigma}\tilde{V}^t$ using K-means clustering.
c. Transform the discovered centroids into the original dimension of tweets, that is, the words.
d. The top or most heavily *weighted words* in each centroid describe a specific topic.

IV. Experimental Result

For the simulations, we use three Twitter datasets focused on three popular real-world events, *The US Super Tuesday*, *The US Presidential Elections*, and *The English FA Cup* [1]. For creating the word-tweet matrix, tweets are parsed, and vocabularies are created using a standard tokenization method. The non-alphabet characters are removed, and standard stop words are applied. Finally, the matrices are weighted by a term frequency-inverse document frequency (TFIDF) weighting scheme.

Each dataset have ground truth topics that will be used to evaluate the performance of topic detection methods. Each ground truth topic has three sets of words: required, optional and forbidden. These three sets of words are separated with a tab character from each other. The words in each set of words are separated with a semicolon and when there are alternatives words, these are placed in square brackets and separated with a space. This informations were discussed in an email from G. Petkos, received on June 29, 2015.

To evaluate the accuracy of a topic detection method, we use TopicEvaluator.jar provided by the dataset owners [1]. Firstly, the method needs to produce words that describe topics. These words are saved in a file with one topic per line.

These words are compared to the ground truth words by the toolkit using three scores:

- *Topic recall*
  Percentage of ground truth topics is successfully detected by a method.
- *Keyword precision*
  Percentage of correctly detected keywords out of the total number of keywords for the topics matches to some ground truth topic in the time slot under consideration.
- *Keyword recall*
  Percentage of correctly detected keyword over the total number of keywords of ground truth topics that have matched to some candidate topic in the time slot under consideration.

These scores are calculated for the *N* first topic generated by the method. In order to get a match, a topic must have all required words but none of the forbidden words. Optional words only count for keyword recall and keyword precision and not for topic recall. A word matches a ground truth keyword when their Levenshtein similarity is > 0.8. The topic detection methods examine each timeslot of dataset for which at least one topic is contained in the ground truth. In total, FA Cup has 13 one-minute slots, Super Tuesday has eight one-hour slots, and US Election has 26 ten-minute slots.

A. The Selection of the Optimal Dimension

For some real world application data, it is usually difficult to get the optimal number of dimension (*p*). The method that is usually used for a practical purpose is a heuristic approach. The optimal *p* has been chosen by running the combination truncated SVD and K-means several times and for multiple values of *p*. The *p* that results in the best performance is chosen as the optimal *p*.

Fig. 3 shows the heuristic approach to determine the optimal *p* for each dataset. The candidate values of *p* are 2, 3, 100, 200, and 300. From Fig. 3, we see that the optimum values of *p* for FA Cup, Super Tuesday, the US Election are 2, 200, 200, respectively.

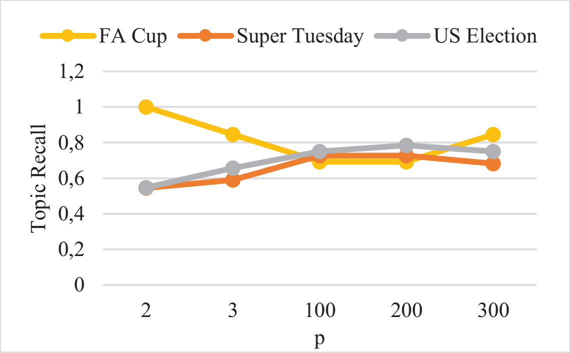

Fig 3. The heuristic selection of the optimal dimension (*p*)

B. The Comparison of the Performance

For comparison purpose, we set the number of topics to 100 for all dataset in the simulation. Besides accuracy, we also compare the computation time K-means clustering method with (both SVD and K-means) and without (K-means only) SVD. Table I shows the topic recall, keywords recall and keyword precision generated by the K-means clustering method and the combination method for the first 10 topics. The combination method gives better topic recall for the FA Cup and Super Tuesday, while for the US Election it produces worse performance. The K-means clustering method has better keywords recall and keywords precision except for the FA Cup.

TABLE I. THE COMPARISON OF ACCURACY FOR THE FIRST 10 TOPICS

Dataset	Method	SVD, *K-means*	*K-means Clustering*
FA Cup	*Topic Recall*	0.938	0.892
	Key. Precision	0.242	0.222
	Key. Recall	0.572	0.526
Super Tuesday	*Topic Recall*	0.236	0.236
	Key. Precision	0.367	0.400
	Key. Recall	0.595	0.696
US Elections	*Topic Recall*	0.216	0.231
	Key. Precision	0.300	0.325
	Key. Recall	0.551	0.575

Next, we examine the performance of methods for the first 100 topics. From Fig. 4, we see that there are no significant differences on the value of topic recall between the two methods. For the first 80 to 100 topics, the topic recalls generated by both methods reach a value of 1. This means that all topics of ground truth topics are successfully detected by these methods.

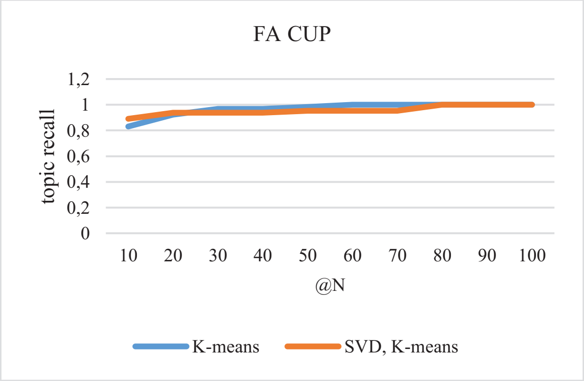

Fig 4. The comparison of topic recalls for the first N topics (@N) on FA Cup dataset

Overall the computing time of the combination method on FA Cup is faster than K-means method as shown in Fig. 5. This is due to the dimension reduction matrix that transforms tweet vectors into only two-dimensional vectors using the truncated SVD.

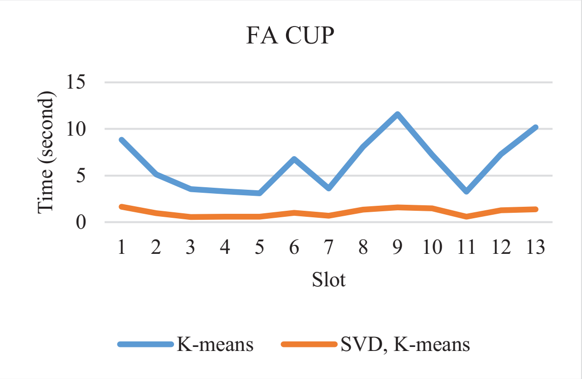

Fig 5. The comparison of computing time for each slot on FA Cup dataset

Fig. 6 shows that the combination method has higher topic recalls than K-means method on the Super Tuesday. By reducing the tweets vectors up to 200, the computing time of the combination method becomes two times faster than K-means clustering method as shown in Fig. 7.

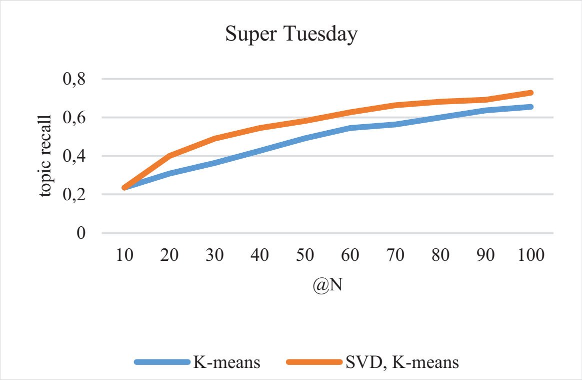

Fig 6. The comparison of topic recalls for the first N topics (@N) on Super Tuesday dataset

Fig. 8 and Fig. 9 show that the dimension reduction of tweet vectors to 200 causes a big impact on the computing time on US Election dataset. The computation time of the combination method reaches two times faster than the computing time of K-means clustering method. Meanwhile, the topic recalls generated by both methods are relatively similar.

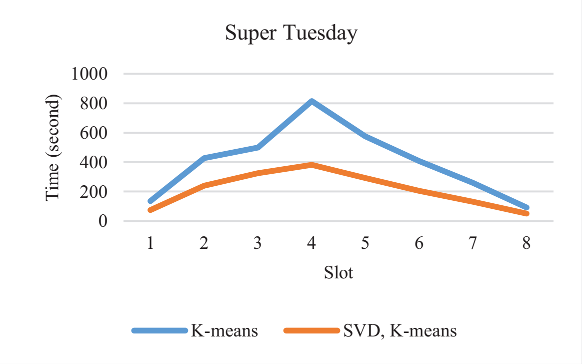

Fig 7. The comparison of computing time for each slot on Super Tuesday dataset

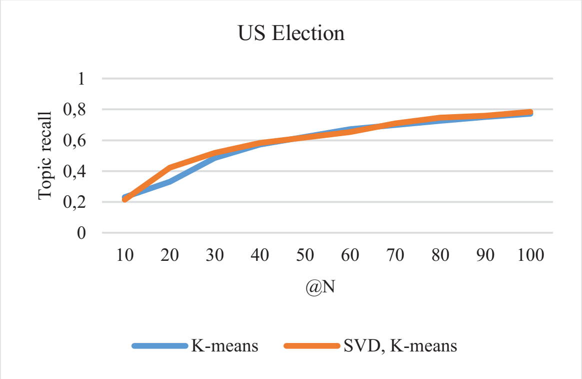

Fig 8. The comparison of topic recalls for the first N topics (@N) on US Election dataset

Overall, the results show that topic recall in FA Cup always higher than in political datasets for both methods. This result was influenced by the data used, as discussed in [1]. This is due to the nature of the target event. Users commenting on FA Cup much more consistent, their attention are focused on very narrow scope and for a limited time. But, the stories about the primaries in US are plenty and interleaving and more difficult to capture [1].

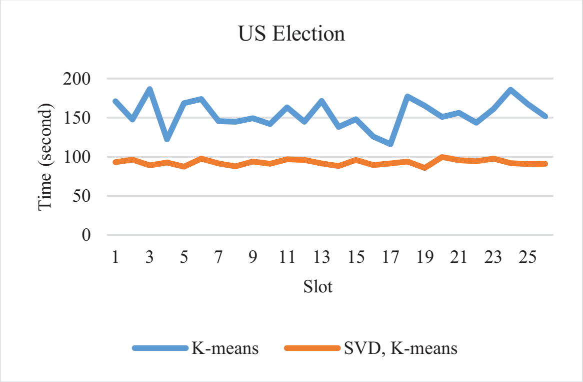

Fig 9. The comparison of computing time for each slot on US Election dataset

K-means is an iterative method where the results are highly dependent on the initial value of the centroids. This method converges to a local minimum rather than a global minimum. This condition makes the method may produce different result for every simulation. Therefore, we run the method several times and use average of the results for the comparison. In general, our simulations show that the combination method produces results that are not much different from the k-means method. Even the combination method yields better results in some cases. In term of the computing time, the combination method is better than the K-means method does.

V. Conclusion

In this paper, we analyze the performance of the combination method as a topic detection method on Twitter. Firstly, we use SVD to reduce the dimension of tweets. Next, K-means clustering method is performed on the reduced tweets. Our simulations show that the combination method gives comparative accuracy in term of topic recall, keyword precision and keyword recall. On the issue of computing time, the use of SVD to reduce the dimension of tweets improves the computing time significantly.

REFERENCES

[1] L. M., Aiello et al. "Sensing trending topics in Twitter". *IEEE Transactions on Multimedia*, 15(6), pp. 1520-9210, 2013.

[2] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer, 2002.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent dirichlet allocation”. Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[4] D. M. Blei. “Probabilistic topic models”. Communication of the ACM, vol. 55, no. 4, pp. 77–84, 2012.

[5] D. D. Lee and H. S. Seung. “Learning the parts of objects by nonnegative matrix factorization”. *Nature*, 401:788–791, 1999.

[6] H. Murfi. "Incorporating Semantic Metadata into Nonegative Matrix Factorization Based Topic Modeling and its Application to Main Topic Extraction". *International Journal of Intelligent Information Processing*, vol. 5 (2), pp. 29-38, 2014

[7] Y. -W. Seo, K. Sycara. "Text clustering for topic detection". Tech. report CMU-RI-TR-04-03, Robotics Institute, Carnegie Mellon University, 2004

[8] G. Petkos, S. Papadopoulos, Y. Kompatsiaris. "Two-level message clustering for topic detection in Twitter". *Proceeding of the SNOW 2014 Data Challenge*, Seoul, Korea, 2014

[9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] S. C. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. "Indexing by latent semantic analysis". *Journal of the American Society of Information Science,* 41(6), pp. 391-407, 1990.

[11] S. T. Dumais. "Latent semantic analysis". *Annual Review of Information Science and Technology*, 38 (1), pp. 188-230, 2005.

[12] B. Jacob. *Linear Algebra*. New York: W. H. Freeman and Company, 1990.

[13] R. L. Burden, J. D. Faires. *Numerical Analysis.* Brooks/Cole Cengage Learning, 2011.

[14] G. H. Golub, C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 2013.