March - April 2020
ISSN: 0193-4120 Page No. 8177 - 8184

Clustering social media user for grouping students in final project using K-Means Clustering and Support Vector Machine

[1]Patrick Adolf Telnoni, [2]Reza Budiawan, [3]Mutia Qana'a

[1,2,3] Diploma of Information System, School of Applied Science, Telkom University, Bandung, Indonesia
[1]patrick.telnoni@tass.telkomuniversity.ac.id, [2]rbudiawan@tass.telkomuniversity.ac.id,
[3]mutia@tass.telkomuniversity.ac.id

Article Info
Volume 83
Page Number: 8177 - 8184
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 09 April 2020*

Abstract

As social media has become inseparable attribute from a person, it is used in common to identify someone's behavior. This practice can be seen in Human Resource Development upon recruiting new employee or when an emigration officer checking someone in the airport or applying visa. This paper propose model to group students based on their preference of twitter account. The steps are by collecting dataset from news website and use each category to label the data. After dataset is collected, user classified according to the friend list and then cluster the users based on the classification result. Test shows that SVM has better accuracy compared to other algorithm, while the elbow method to determine number of cluster does not show best k number since graphic of the method shows exponential form. For future works, it is recommended to use silhouette method to determine k form clustering and Latent Dirichlet Allocation (LDA) to label dataset into multi-label data.

Index Terms; *social media, k-means clustering, svm, project grouping.*

I. INTRODUCTION

Students nowadays has been actively engaged in social media platform, the behavior, friend list or post from social media can be used to approximate personality. This paper propose to build machine learning model with SVM and KMeans clustering to form students group based on their social media friend list.

SVM is a highly used algorithm in machine learning, especially in classification case. In text analysis, SVM become favourite among other machine learning algorithm [3]. Compared to Naive Bayes [4] and Logistic Regression [5], SVM always come on top in term of accuracy score.

In [4], twitter data directly retrieved and feature were extracted using bag of words technique. in the F-Measure test, SVM score relatively high margin with naive bayes for each category classification. From 10% to nearly 30% margin. Svm not only limited to text classification, but also for malware detection [6], where using 271094 malware data, SVM scored 75% detection rate from the range between 74 - 83% where the data is at least 10000 per class.

Since in this paper, twitter is used for its easiness for its accessibility. Twitter itself has become favourite for text-based research, particularly for sentiment analysis, user interest [7], or user behaviour [8]. In [7], method such as bag of words, inverse document frequency or category extraction were not used, but use temporal dimension to determine user interest and future trends instead. [8] used Multi-Scale-Entropy (MSE) analysis to identify user behavior and separate them into different labels. In this research, the date of tweet posted was used as

Published by: The Mattingley Publishing Co., Inc.

feature. The dataset labelled manually into five labels (individual, news platform, advertising and promotion, robot, and other type account). Using date from the tweet data, MSE produced MSE vector and MSE slope and use it to classify using libsvm. The result is MSE Slope gave higher result (94%) compared to MSE Vector (64.2%) because of the noise from the vector itself.

On the other hand, K-Means clustering algorithm also favourite in doing grouping data. Garg and Rani used kmeans clustering to analyze and visualize twitter data to map distribution using users' geolocation. [9] Before k-means used for twitter, k-means has been used for many purpose such as document clustering [10], image clustering [11] and even text clustering [12]. In [10], it is mainly about comparison between K-Means and its variants such as heuristic K-Means and Fuzzy K-Means. In this research, methods for feature extraction are tested. The tested methods includes tf, *tf-idf*, and boolean. The result is *tf-idf*with stemming produce better result among other two, while heuristic k means provide better result than other flat clustering algorithm. In [11], performed image segmentation using K-Means and fuzzy K-Means and compared the result between them. What makes this research unique is the usage of fuzzy logic in K-means, where one data point can belong to more than one clusters rather single cluster. That way, fuzzy K-Means produce better speed in clustering image data. On [12], an improved K-Means was proposed by modifying initial cluster centre. The improvement was achieved by calculating density and average density of the data, eliminating isolated data and then select parameter with highest density as the initial cluster. The improved K-Means algorithm scored slightly higher result, from 1% to 3% margin with traditional K-Means.

SVM and K-Means also can be used concurrently. For example, these two algorithms can be used together in biomedical such as shown in [13]. The idea on [13] is to group brain MRI image into different labels. After that, images from each class are extracted into feature vector and then segmented again with SVM. This research shows that the modified algorithm scored up to 6% to 9% margin in 9% noise level compared to the common K-Means. Not limited to biomedical field, SVM and K Means also used for security purpose in cyber security for intrusion detection system [14]. In [14], the idea is quite similar with [13], where data of network traffic are separated into different groups with improved K-Means and use SVM to mark cluster with abnormal behavior for detailed classification as the final realization of the detection process.

The work here is quote similar with [13] and [14]. The striking difference is the built model use SVM first to classify twitter users and cluster them into different labels. This label will be used for profiling purpose for educational process, i.e creating cluster for study groups or extracurricular activities. SVM was chosen because it still scored higher than other classification algorithm such as Logistic regression although the margin is quite thin [5] [15]. Note that in this research, the model uses standard SVM and standard K-means, since the data for clustering is not huge.

II. METHODOLOGY

This section will explain about the frequently used term in this paper.

A. K Means clustering

K-Means clustering is unsupervised machine learning algorithm which is mainly used to cluster, classify, or group N object into k numbers of clusters or groups, with K and N are positive integers. Upon the grouping process, the algorithm calculate centroid from existing data point. K Means algorithm begin with determining *K* numbers of cluster will be formed. After the number of *K* is agreed upon, *K* number of centroids are randomly selected from the existing data points [1].

Each data point's proximity is calculated against each centroid, until all data point completely

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8177 - 8184

assigned to each cluster. The proximity commonly computed using eucledian distance. Centroid with the least euclidean distance will be assigned to the data point. Equation 1 and 2 denotes euclidean distance formula.

$$\mathrm{k}\mathbf{p}\mathrm{k} = \mathrm{d}(\mathbf{p},\mathbf{0}). \quad (1)$$

$$\|\mathbf{p}\| = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2)$$

After each cluster completely filled with data points, new centroids are calculated for each cluster by average each components from each data point. From this new centroid, every data point will be evaluated its proximity to the new centroid. In this step, there are data points which will move to another cluster. This change will trigger recalculation of new centroid. Calculation of new centroid formulated by equation3.

$$C_i = \frac{1}{N_i}\sum_{x \in x_i} x, i = 1, 2, .., k \quad (3)$$

After the new centroid is formed and no data point which change cluster, the iteration stops, thus provide final cluster.

B. Support Vector Machine

Support Vector Machine is supervised classifier algorithm which relies on hyperplane to separate data into distinguishable label often called linear separability [2]. SVM can be used for classification and regression, thus use Support Vector Classifier and Support Vector Regression for each case respectively. Mathematically, SVM hyperplane are denoted in equation 4.

$$h_{w,b}(x) = g(w^T x + b) \quad (4)$$

From equation 3, hyperplane equation is similar to what in Linear Regression, because single line is considered as hyperplane. This hyperplane used as boundary to separate data into different labels. In general, there are multiple hyperplane that can be formed to separate data.

What makes SVM different from typical Linear Regression is that SVM set maximum margin function denoted in equation5.

$$\delta = \frac{1}{\|\vec{w}\|} \quad (5)$$

where the $\|w\|$ is support vector, which the closest vector to the hyperplane. Margin between w~ and hyperplane are called maximum margin, denoted as δ.

Since the hyperplane distinguish data into two different label, meaning that for every data, a hypothesis in equation 6 is applied.

$$h(x_i) = \begin{cases} +1 & if \mathbf{w}.\mathbf{x} + b \geq 0 \\ -1 & if \mathbf{w}.\mathbf{x} + b < 0 \end{cases} \quad (6)$$

III. THE PROPOSED MODEL

This section will explain through process occurred in the model. This section include dataset, feature extraction, classification, dimensionality reduction, and finding optimum value of *k* with elbow method.

A. Dataset

Collecting the dataset begun by scrapping news website such as kompas.com and detik.com. Dataset from twitter did not directly used, because the twitter policy limits its data streaming up to 3200 data per 24 hours. Also, since most social media account today are heavily associated with hoax or any political issue so it will contain many rude words. Detik and kompas were chosen since those are the media which provide hoax debunking up until now. Not only that, the retrieved from twitter directly is unlabelled data. By scrapping the news website can get labelled data, because every news website contains sub-domain which post content according to the sub-domain's theme, such as sport, entertainment, politics etc.

The scrapper scrapped the website by using scrapping tools built with python's scrappy. The scrapper run the scrapper for 10 days and successfully retrieve over 9000 data with 19

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8177 - 8184

categories. Since not every content in the news website contain text, but also video, we filtered the news with video, resulting over 7500 data. Not every category has the same amount of data, so each category must be filtered which only contain less than 150 data and result with 14 categories as shown in Figure 1. This was conducted to prevent under sampling.

B. Feature extraction

Before train the text into SVM, the texts were transformed the document into numerical vector with fixed size. The text transformed using bag of words technique. Bag of words will count the frequency of occurrence of words while ignoring the order. Specifically, Term Frequency, Inverse Document Frequency, or commonly known as *tf-idf*was used. Stopword was not used since the dataset is based on Bahasa Indonesia. Using tf-idf, unigram, bigram will be acquired for every category. Unigram is a word which most associated to a category while bigram is two most associated words to a category. Unigram and bigram is measured using their respective frequency. Upon extracting unigram and bigram, stopwords are used to remove less meaningful words from the dataset. Since the dataset was scrapped from Indonesian news portal, Sastrawi python library was used for the stopwords. This is considered as noise removal.

Unigram and bigram represented in vector with the size of (7054 x 76504) which means each of 7054 news content are represented by 76504 features, representing the tf-idf score for different unigrams and bigrams. These vectors will be used to train document of tweet from an active account and determine which category fit the user most.

C. Classifying twitter account

After the vectors were formed, twitter api was used to retrieve twitter data. The retrieved data is not a trending topic nor an event. The model retrieve data from a user and classify the user according to the tweet. The model did not directly classify a user according to the latest post but classify them by using their friends. From an account, list it's 0 followers and for each follower, pick their latest tweet. Since twitter limits its user timeline data stream up to 3200 tweet per day [16], the model only streamed 20 latest post from a user. If it try to exceed 3,200, twitter api will return 401 response.

The model uses friends, which is the account followed by the target account, not by its follower. Friend was picked over follower because follower's account can be set to protected, preventing data streaming over twitter API. This is because the nature of twitter, where a user can see the complete information of other user which s/he follow but cannot see information about the follower who set their profile to be private. This protected/private account will prompt the 401-error message over the API.

After collecting the required number of tweet document, the model was trained with SVM algorithm using vector obtained from *tf-idf*transformation. After the vectors were trained, tweet from every respective user were concatenated to form a document. After document was formed, it was crosschecked against the trained model. The model will produce which category the document belongs to.

Each classification will be grouped according to its category. The grouping result will be inserted into vector to represent a user.

$$ProfFeynman\tilde{} = [6,0,3,21,21,2,7,3,0,6,31,2,3]$$

For example, one user named ProfFeynman from his friend list has 6 friends who heavily associated with politics, 3 friends associated with automotive, 0 friend associated with technology, 21 friend associated with lifestyle, and so on.

D. Dimensionality reduction

After the vector for each user were formed, it reduced into two dimensional data. Since there are 14 categories in the classification result, this means

Published by: The Mattingley Publishing Co., Inc.

that the representation from each user are in 14 dimensions. To ease the process of clustering and visualization, t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to trim it down into 2D data.

The t-SNE algorithm computes similarity measure between pairs of instances in high and low dimensional space. It then tries to optimize these two similarity measures using a cost function [17]. T-SNE used conditional probability to measure this similarity, as shown in equation 7.

$$p_{i|j} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\delta_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-\|x_i - x_k\|^2}{2\delta_i^2}\right)} \quad (7)$$

E. Elbow Method

After obtaining two dimensional data, clustering process can be started. But first, the number of optimal *k* must be computed. To determine number of *k*, elbow method was used. Elbow method are denoted with equation 7 [18].

$$\mathbf{SSE} = \sum_{k=1}^{k} \sum_{x \in x_i} \|X_i - C_k\|_2^2 \quad (8)$$

From equation 7, Sum Squared Error is used to calculated elbow method by summing of the average Euclidean Distance of every data point against the centroid. Upon sharp decrease of the SSE value and the angle start to align horizontally, k is determined. From k = 2 and the SSE value added on each iteration, where $k_n = k + 1$, the largest margin of SSE which is $SSE_{kn} - SSE_{kn-1}$ is the point in where the optimal k value is located.

To summarize the proposed model, the steps shown as follows:

- The scrapper collects data from news portal (detik and kompas) including the sub-domain.
- Every news content labelled according to the sub domain for each news. For example, news from sport sub-domain will be labelled with sport, news from automotive will be labelled as automotive.
- Collected data from news portal will be filtered by checking item per category. Category which contain less than 150 items will be dropped from the dataset.
- After the dataset is filtered, stopwords are removed and the cleaned dataset transformed into vector using *tf-idf*.
- Retrieve 60 accounts friend list from a designated account.
- From every retrieved account, collect 20 latest tweet text. Upon collecting the data, every tweet is filtered from noise character which heavily related to twitter i.e. 'RT', '@', URL such as 'http' or 'https'.
- Join cleaned tweets into one document. After that, the document will be classified using the trained model. The result of classification will be grouping according to each respective matching category.
- Each grouped category will be counted the sum for each user and appended into vector. Then each user has vector with the size of 1*14.
- Each vector will be reduced into two dimensional data using T-SNE.
- Optimal value of *k* will be calculated from two dimensional data using SSE.
- Group two dimensional data into *k* cluster from elbow method. Plot the result into two dimension chart.

Figure 1 shows complete diagram of the proposed model.

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8177 - 8184

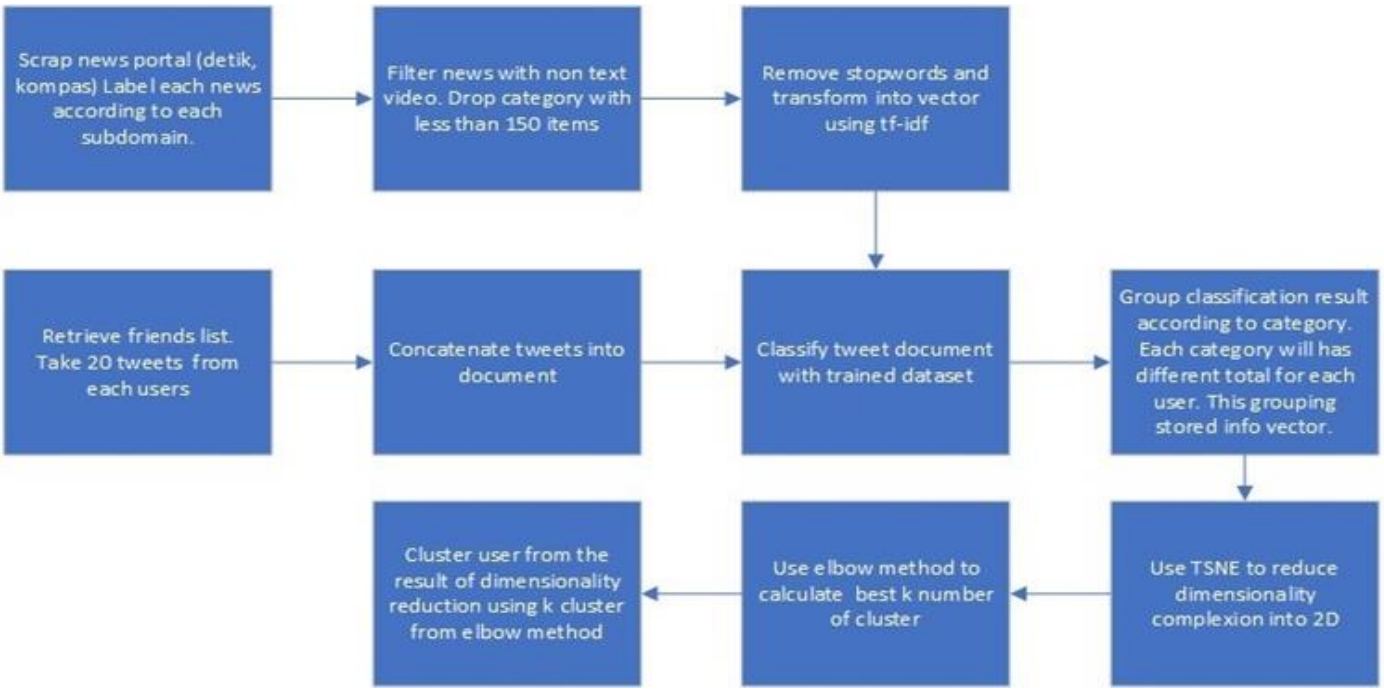

Fig. 1. Proposed Model

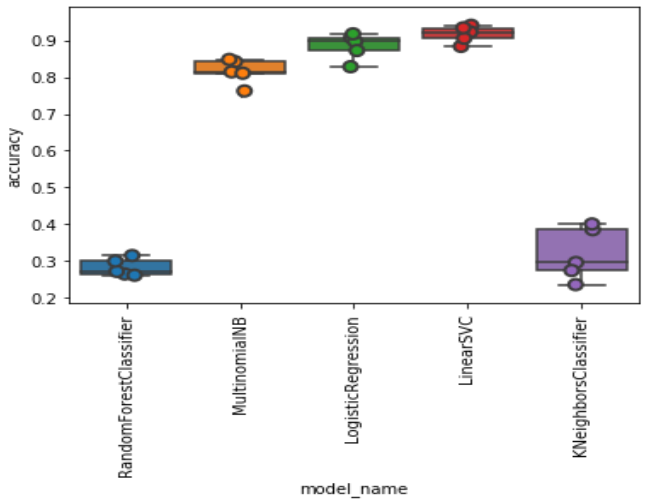

Fig. 2. Cross validation score for all tested algorithms

IV. RESULT AND DISCUSSION

The finding here will discuss only for the comparison of tested classification algorithms, the result of elbow function, and the graph of clustering.

Using SVM, the test compared the accuracy with other algorithm. The model tested against K Nearest Neighbor (KNN), Linear Support Vector Machine, Multinomial Naive Bayes, and Random Forest. This comparison are tested using cross validation score (CVS). CVS score for Logistic Regression (LR) is 0.884578, Multinomial Naive Bayes at 0.815395, KNN hit 0.318189, Random Forest at 0.282280, and SVM scored at 0.917324. This margin is typical to the [5] [15]. Figure 2 display accuracy chart of the tested algorithm with accuracy for each algorithm while table 1 display the numerical value of it.

Table I

Comparison Between Tested Algorithms

No	Algorithm	Accuracy Score
1	Random Forest	0.282280
2	Multinomial Naive Bayes	0.815395
3	Logistic Regression	0.884578
4	LinearSVM	0.917324
5	K Nearest Neighbor	0.318189

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8177 - 8184

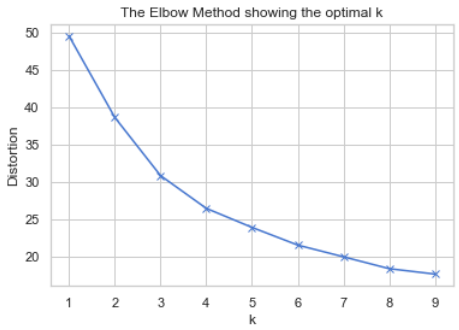

Fig. 3. Graphic of distortion for every number of k using elbow method

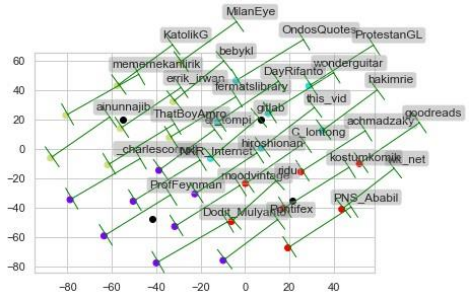

Fig. 4. Clustering result

While SVM came on top as expected, unique thing occurred during elbow test. While typical elbow graphic will show clear and sharp elbow, the test did not show dramatical result. Instead, result have slightly smooth exponential curve as shown in Figure 3. This could be the effect of large scale of two dimension value from T-SNE.

Nevertheless, 4 is picked between 3 or 4 as the *k* number to group the data. The plotted cluster shown in Figure 4.

From figure 4, desired group for students can be acquired. The data with least margin to each cluster will be used as the team leader.

Since this work focus on SVM and K-Means Clustering, this research did not consider other algorithms as the main classifier. From the dataset projection in 2 dimensions chart, found that it is quite difficult to determine which algorithms perfect for the dataset, since its projection in figure 4, which consist quarter of its feature, is not linearly separable. Back to the initial intention is to use other method which give identical accuracy to SVM as stated in [5] and [15], where the gap between SVM and LR were 2% and 3.33%. In this case, the gap still identical to the [5] and [15] since the data set is quite large, over 7000 records were used.

V. CONCLUSION AND FUTURE WORKS

From this research, it can be concluded that the model was successfully built to classify twitter user using SVM and dataset from news portals. While the SVM did not give surprising problem, the opposite happened with the elbow method. This paper recommends using silhouette function rather than elbow method to find optimal *k* value.

Also, it is better to classify dataset first into multi-label data, since one article from a news portal can associate with multiple topics.To do this, Latent Dirichlet Allocation (LDA) encouraged to use for future work.

REFERENCES

[1]. J. Wu, *Advances in k-means clustering. A data mining thinking*. 01 2012.

[2]. T. Joachims, *Support Vector Machines: Theory and Applications*. The Springer International Series in Engineering and Computer Science, Springer-US, 1 ed., 2002.

[3]. P. L. W. e. V. Kecman (auth.), *Learning to Classify Text Using Support Vector Machines*. Studies in Fuzziness and Soft Computing 177, Springer-Verlag Berlin Heidelberg, 1 ed., 2005.

[4]. I. Dilrukshi and K. De Zoysa, “Twitter news classification: Theoretical and practical comparison of svm against naive bayes algorithms,” in *2013 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 278–278, Dec 2013.

[5]. L. Sijia, T. Lan, Z. Yu, and Y. Xiuliang, “Comparison of the prediction effect between the logistic regressive model and svm model,” in *2010 2nd IEEE International Conference on Information and Financial Engineering*, pp. 316–318, Sep. 2010.

Published by: The Mattingley Publishing Co., Inc.

[6]. B. Sanjaa and E. Chuluun, "Malware detection using linear svm," in *Ifost*, vol. 2, pp. 136–138, June 2013.

[7]. R. Abbasi, G. Rehman, J. Lee, F. M. Riaz, and B. Luo, "Discovering temporal user interest on twitter by using semantic based dynamic interest finding model (tut)," in *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 743–747, Dec 2017.

[8]. S. He, H. Wang, and Z. H. Jiang, "Identifying user behavior on twitter based on multi-scale entropy," in *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pp. 381–384, Oct 2014.

[9]. N. Garg and R. Rani, "Analysis and visualization of twitter data using k-means clustering," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 670–675, June 2017.

[10]. V. K. Singh, N. Tiwari, and S. Garg, "Document clustering using kmeans, heuristic k-means and fuzzy c-means," in *2011 International Conference on Computational Intelligence and Communication Networks*, pp. 297–301, Oct 2011.

[11]. V. K. Dehariya, S. K. Shrivastava, and R. C. Jain, "Clustering of image data set using k-means and fuzzy k-means algorithms," in *2010 International Conference on Computational Intelligence and Communication Networks*, pp. 386–391, Nov 2010.

[12]. C. Xiong, Z. Hua, K. Lv, and X. Li, "An improved k-means text clustering algorithm by optimizing initial cluster centers," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pp. 265–268, Nov 2016.

[13]. J. Liu and L. Guo, "A new brain mri image segmentation strategy based on k-means clustering and svm," in *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol. 2, pp. 270– 273, Aug 2015.

[14]. Z. Xiaofeng and H. Xiaohong, "Research on intrusion detection based on improved combination of k-means and multi-level svm," in *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, pp. 2042–2045, Oct 2017.

[15]. A. Janghorbani, A. Arasteh, and M. H. Moradi, "Prediction of acute hypotension episodes using logistic regression model and support vector machine: A comparative study," in *2011 19th Iranian Conference on Electrical Engineering*, pp. 1–1, May 2011.

[16]. Twitter, "Get tweet timelines."

[17]. L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," 2008.

[18]. D. Marutho, S. Hendra Handaka, E. Wijaya, and M. , "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," pp. 533–538, 09 2018.