TWITTER BUZZER DETECTION SYSTEM USING TWEET SIMILARITY FEATURE AND SUPPORT VECTOR MACHINE

Ahmad Mustofa[1)], Fitrah Maharani Humaira[2)], Myrna Ermawati[3)], Peni Sriwahyu Natasari[4)], Akhmad Arif Kurdianto[5)], Aries Alfian Prasetyo[6)], and A Labib Fardany Faisal[7)]

[1, 5,6,7)]Teknik Listrik Industri, Politeknik Negeri Madura
Jl. Raya Camplong KM.4, Taddan, Camplong, Sampang Madura
[2)] Departemen Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya
Jl. Raya ITS, Keputih, Sukolilo Surabaya
[3)] Hexagon Mining Balikpapan
Sudirman Office Tower Lt.10 Jl. Jenderal Sudirman No.7, Klandasan Ilir, Balikpapan, Kalimantan Timur
[4)] Data Science Indonesia
Jl. Mampang Prapatan Raya Kav. 100, Pancoran, Duren Tiga, Jakarta
e-mail: mustofaahmad@poltera.ac.id[1)], fitrah@pens.ac.id[2)], myrnaerma.nina@gmail.com[3)], penitasari93@gmail.com[4)], aak@poltera.ac.id[5)], aries@poltera.ac.id[6)], a.labib@poltera.ac.id[7)]

ABSTRAK

Penggunaan media sosial, seperti twitter, sebagai sumber informasi oleh masyarakat semakin meningkat dalam beberapa tahun terakhir. Kondisi ini berpotensi memberikan banyak dampak positif dan negatif bagi masyarakat. Salah satu dampak negatif yang dapat terjadi adalah opini masyarakat menjadi mudah digiring menuju opini tertentu yang menguntungkan beberapa oknum. Penggiringan opini biasanya dilakukan oleh beberapa akun buzzer yang bermaksud untuk mempengaruhi masyarakat agar setuju pada opini tertentu. Dalam upaya untuk mengurangi dampak negatif yang ditimbulkan oleh akun buzzer, maka dibutuhkan sebuah sistem yang dapat mengenali apakah sebuah akun merupakan akun buzzer atau tidak. Ciri utama akun buzzer adalah mengunggah konten yang sama berulang kali dalam jangka waktu tertentu. Penelitian ini mencoba membangun sebuah sistem pendeteksi buzzer dengan menggunakan pengolahan teks dan metode klasifikasi. Kemiripan tweet yang didapatkan dengan menerapkan Cosine Similarity pada nilai Term Frequency - Inverse Document Frequency (TF-IDF) dari tweet digunakan sebagai fitur utama untuk sistem deteksi buzzer. Selain itu, fitur lain seperti jumlah followers, jumlah following, intensitas tweet, rasio retweet, dan rasio tweet yang mengandung link juga digunakan sebagai fitur tambahan dalam penelitian ini. Penelitian ini menggunakan fitur-fitur tersebut sebagai masukan pada model Support Vector Machine untuk menentukan apakah suatu akun termasuk buzzer atau tidak. Sistem ini memiliki hasil yang menjanjikan dengan akurasi 89%, presisi 86,67%, recall 70,91%, dan skor F1 78%.

Kata Kunci: *sosial media, deteksi buzzer, pengolahan text, support vector machine.*

ABSTRACT

Over the past few years, more people have turned to social media sites like Twitter as a source of news. This phenomenon may incur many positive and negative effects on society. One of the negative impacts is that public opinion becomes easier to be influenced towards certain opinions that benefit some person. Buzzer accounts usually try to convince people to agree on certain opinions. A system that can determine if an account is a buzzer account or not is required to reduce the negative impact caused by buzzer accounts. The main characteristic of a buzzer account is repeatedly uploading the same content within a certain period. This study uses text processing and classification methods to build a buzzer detection system. The tweet similarity, obtained by applying Cosine Similarity to the Term Frequency - Inverse Document Frequency (TF-IDF), is used as the main feature for the buzzer detection system. In addition, other features such as the number of followers, the number of followings, the intensity of tweets, the ratio of retweets, and the ratio of tweets that contain links are also used as additional features in this study. This study uses these features as input to the Support Vector Machine model to determine whether an account is a buzzer. This system has promising results with 89% accuracy, 86.67% precision, 70.91% recall, and 78% F1 score.

Keywords: *social media, buzzer detection, text processing, support vector machine.*

I. Introduction

THE significant increase in the use of social media in Indonesia is a good opportunity to seek out a wide variety of information. Social media makes it easier for people to express their opinions and feelings, thus the data become abundant and become one of the primary data resources. By processing those data, public opinion can be disclosed and concluded easily in real time. For example, tweets data from Twitter was utilized to make a prediction system for the winner of the Indonesian presidential election in 2014 and 2018 [1] [2]. Twitter data is also used for demographic analysis of Candidates Supporter during Indonesian Presidential Election in 2019 [3] and to analyze public sentiment toward candidates of US Presidential Election in 2012 [4]. The use of twitter data is not only in political domains, but others also use twitter data to analyze customer's perception about some products, brands, or events by applying sentiment analysis to user's tweets [5] [6] [7] [8] [9].

However, at the same time, the abundant information available online which can be accessed freely is risky for those who cannot distinguish which information is fake or genuine. Especially, if the society has a big event such as a public election, people are very sensitive with other opinions especially affected by false information. Buzzer account usually sway public opinion by spreading false information, provoking, insulting, and harassing other accounts. Thus, to get the neutral public opinion, buzzer detection is usually embedded in a system that utilizes data from social media.

From the research conducted by Suciati et al., several classifiers are used to categorize whether a Twitter account is a buzzer or non-buzzer. The research used eleven basic features such as most active days, most active hours, and frequency of most used hashtag resulting the performance of 67.9% [10]. Another promising feature is text-similarity which was used in bot detection system by implementing Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity, having best accuracy of 86% [11]. TF-IDF is commonly used in processing a text-based data, like in [12], TF-IDF is used to make a rumor detection system on twitter data. Other research in text-based data analysis often use cosine similarity together with TF-IDF to calculate similarity between 2 texts [13] [14] [15] [16] [17]. Cosine similarity and TF-IDF can also be used to classify text-based data [18] [19]. An Arabic text classification system is build using TF-IDF, cosine similarity, and some classifier methods in [20] having best result obtained by Support Vector Machine (SVM) classifier. To classify the buzzer and non-buzzer account, Support Vector Machine (SVM) was used in research conducted by Panatra et al. so that neutral information can be maintained for the 2019 presidential election in Indonesia [21]. The use of SVM in a binary class classification system also produces a fairly good performance in detecting glaucoma disease from retinal fundus of eye represented in images [22]. SVM also has a good performance by having 81% accuracy in classifying spam opinions on Steam Review [23]. Others also use SVM for object detection having result 82% accuracy [24].

In this research, we use SVM as the classifier to differentiate buzzer and non-buzzer accounts. We use the tweet similarity of each user, number of followers, number of followings, intensity of tweets, ratio of retweets, and ratio of tweets that contain links as features fed to the SVM classifier. The rest of this paper is arranged as follows: In section 2, we explain the methods used in this research. Then in section 3, we describe the result and analysis of our experiment. Finally, we conclude our work in section 4.

II. Research Method

A. Data Collection

The first task in our research was collecting relevant and appropriate amount of observation data for our classification model approach. We use Twitter public Application Programming Interface (API) to collect data needed in this research. First, we need to collect many twitters' users account whose nationality is Indonesian. Since twitter public API didn't have endpoint which can collect users account data directly, we used tweets collection by keywords endpoint using "Indonesia" as keyword. We also gave limitation to collect tweet that were tweeted by users located in Indonesia. We got 10000 tweets; each tweet contains user's id information. Then Then, we extracted 1096 distinct users from those tweet data. Endpoint of twitter public API used in this research is provided in Table. 1.

Table 1 Feature Description

No	Endpoint	Description
1	tweets/search/all	Endpoint to retrieve tweets with specific keyword and location of user when the tweet is posted
2	users/:id/timelines	Endpoint to retrieve tweets of specific user_id

To detect buzzer, we performed observation on our data and found that Twitter buzzers has specific characteristics such as high frequency of tweets in a short period since buzzer accounts actively post tweets to sway other users' opinion. We add those characteristics as features in our dataset. Later, we will explain these features in next section. To get these features, we collected more tweet from Twitter user accounts. For each user, we retrieved user's tweet in the last 3 month. If the user's tweet in the last 3 month is less than 10, we continue to retrieve user's tweet up to 10 tweets.

To apply supervised machine learning model to detect Twitter buzzer account, we need to obtain labeled data for each user accounts. Labeling data to provide dataset for our classification model approach usually requires human effort. In the next step, we observed Twitter users and manually labeled those accounts as ‘buzzer’ or ‘non-buzzer’. Twitter user that continuously promoting or discrediting the same things multiple times are labeled as ‘buzzer’. We involved three annotators to label those Twitter user by manual observation. Final label of each user is decided using a voting system. We found that 395 from the total of 1096 users from our data are labeled as buzzer.

B. *Buzzer Classifier Features*

To differentiate buzzers among normal Twitter users, we implemented a machine learning approach to build our detection model. We defined a buzzer detection problem as a classification problem that will be able to classify a Twitter user as a buzzer or not. In this research, we used statistical property of tweet similarity as the main feature together with other features like follower count and tweet per day. Table. 2 and Table. 3 displays the list of Twitter user-based features and tweet-based features for our classification approach.

Table 2 List of Proposed Features

No	Feature	Characteristic
1	Follower count	User based features
2	Following count	
3	Tweet per day	
4	Retweet ratio	Tweet based features
5	Link ratio	
6	Tweets similarity	

Table 3 Feature Description

No	Feature	Description
1	Follower count	number of users who follow each user
2	Following count	number of users followed by each user
3	Tweet per day	number of tweets tweeted by each user divided by the age of each account
4	Retweet ratio	number of retweets divided by the total number of tweets
5	Link ratio	number of tweets that contain links divided by the total number of tweets
6	Tweets similarity	Statistical properties of tweets similarity

User-based features like follower count and following count is retrieved directly from data provided by twitter public API. Twitter public API also provide accounts’ creation time and total number of tweets. These two data used to calculate tweet per day feature.

C. *Tweet Based Feature*

To obtain tweet-based feature used in this research such as retweet ratio, link ratio, and tweets similarity, we need to retrieve tweets data from each user and perform several text preprocessing steps as follows:

1. Tweet Retrieval

For each user, we use twitter public API to retrieve user’s tweet in the last 3 months. If the user’s tweet in the last 3 months is less than 10, we continue to retrieve user’s tweet up to 10 tweets.

Tweets data retrieved from twitter public API contains information whether a tweet is a retweet or not. For each user, we count how many tweets is a retweet and divide the result with total tweet retrieved to get retweet ratio feature.

Tweet content is also provided by twitter public API which value is used to calculate link ratio feature. For each user, we count number of tweets that contain links and divide the result with total number of tweets retrieved.

2. Tweet Preprocessing

After user’s tweets retrieved, we applied a filtering process, the stop word removal, to remove irrelevant words or characters. This preprocessing stage removed all words in stop words lists from tweets, including Twitter-specific words like RT. We also performed non-letter character removal, link removal, and mention removal in this preprocessing phase.

3. Tweet Vectorizing

In the next preprocessing stage, we need to vectorize each tweet to represent tweet in a vector space and calculate its similarity. We use a term frequency – inverse document frequency (TF-IDF) method to vectorize each tweet. The vectorization step is processed locally for each user.

This vectorizer takes tweets and tokenize them to obtain terms, and then computing term frequency – inverse document frequency (TF-IDF) as weighting method. The term frequency (TF) value is obtained by counting the number of occurrences of a term in one tweet, while the inverse document frequency value is obtained by (1) [25].

$$\mathrm{idf}_i = \log_2 \frac{N_t}{\mathrm{df}_i} \tag{1}$$

4. Tweet Similarity Feature

To extract tweet similarity feature of each user, we need to build a matrix of tweet similarity. Suppose we have a user with 8 tweets, the matrix of tweet similarity of this user will look like Fig. 1.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
t_1		$s_{1,2}$	$s_{1,3}$	$s_{1,4}$	$s_{1,5}$	$s_{1,6}$	$s_{1,7}$	$s_{1,8}$
t_2	$s_{2,1}$		$s_{2,3}$	$s_{2,4}$	$s_{2,5}$	$s_{2,6}$	$s_{2,7}$	$s_{2,8}$
t_3	$s_{3,1}$	$s_{3,2}$		$s_{3,4}$	$s_{3,5}$	$s_{3,6}$	$s_{3,7}$	$s_{3,8}$
t_4	$s_{4,1}$	$s_{4,2}$	$s_{4,3}$		$s_{4,5}$	$s_{4,6}$	$s_{4,7}$	$s_{4,8}$
t_5	$s_{5,1}$	$s_{5,2}$	$s_{5,3}$	$s_{5,4}$		$s_{5,6}$	$s_{5,7}$	$s_{5,8}$
t_6	$s_{6,1}$	$s_{6,2}$	$s_{6,3}$	$s_{6,4}$	$s_{6,5}$		$s_{6,7}$	$s_{6,8}$
t_7	$s_{7,1}$	$s_{7,2}$	$s_{7,3}$	$s_{7,4}$	$s_{7,5}$	$s_{7,6}$		$s_{7,8}$
t_8	$s_{8,1}$	$s_{8,2}$	$s_{8,3}$	$s_{8,4}$	$s_{8,5}$	$s_{8,6}$	$s_{8,7}$	

Fig. 1 Matrix of Tweet Similarity

Similarity between tweet is calculated using cosine similarity in (2) [26] [27].

$$s_{i,j} = \frac{\sum_{k=1}^{n} w_{i,k} \times w_{j,k}}{\sqrt{\sum_{k=1}^{n} {w_{i,k}}^2} \times \sqrt{\sum_{k=1}^{n} {w_{j,k}}^2}} \tag{2}$$

After matrix of tweet similarity is formed, we can extract the similarity feature of the matrix using (3) and (4):

$$f_1 = \frac{\sum_{i=1}^{N} \sum_{k=1,k\neq i}^{N} s_{i,k}}{N(N-1)} \tag{3}$$

$$f_2 = \sqrt{\frac{\sum_{i=1}^{N} \sum_{k=1,k\neq i}^{N} (s_{i,k} - f_1)^2}{N-1}} \tag{4}$$

D. *Classifier: SVM*

Support Vector Machine (SVM) was used as the classifier to differentiate buzzer and non-buzzer accounts. Our research used six features, as listed before, as features for the binary classification using SVM.

After feature extraction process is done, we combine all features extracted to represent each user in the buzzer detection system. Then all the user training data used to train classification model. Finally, we used all the user testing data to validate the trained model. Fig. 2 describe our proposed buzzer detection system.

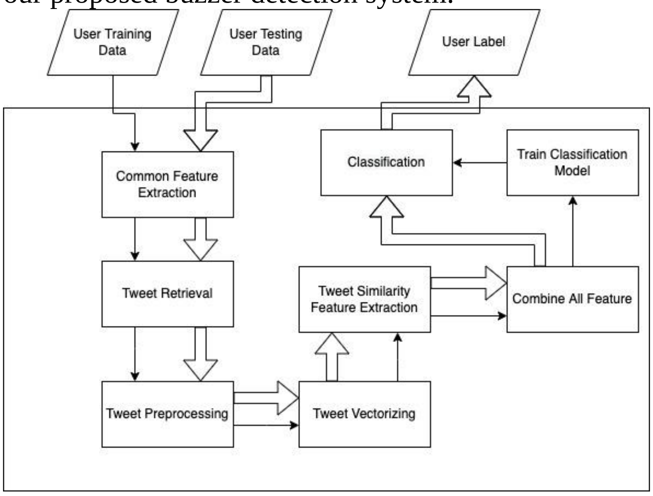

Fig. 2 Proposed Buzzer Detection System

III. Result and Discussion

A. *Dataset*

As the result of tweet data collection and user account extraction, we have 1096 distinct user accounts that had been labeled manually. Buzzer characteristic features from these user account data and the label are then used as our research dataset. To build a classification model for buzzer detection, we split our dataset into training data and testing data. Total dataset for training is 896 user account data that will be used to train our model. The rest of the dataset will be used as testing data to validate our model. Fig. 3 describe the distribution of our data.

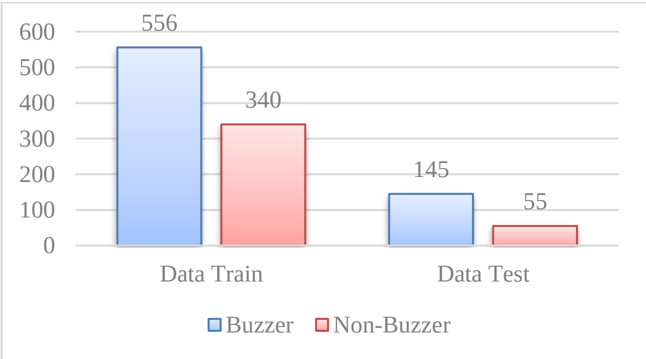

Fig. 3 User Data Distribution

B. *Result*

We use Support Vector Machine (SVM) as the classification method in our proposed buzzer detection system as SVM works well on a binary classification case. In this research, each user is represented by a 1x7 vector which will be used as features in SVM. Before training our model, we analyze the value range of each non-statistical feature to decide whether a normalization process is necessary or not.

Table 4 Value Range

No	Feature	Value Range
1	Follower count	0-289588
2	Following count	0-6849
3	Tweet per day	0.0005-83.333
4	Retweet ratio	0-0.92
5	Link ratio	0-1

From Table. 3, we can see that the value range is variative. We decide to do normalization process for feature 1-4 so that we have same value range for all 5 features (0-1).

We experiment using 3 different kernels on SVM such as linear, radial basis function (rbf), and polynomial kernel. The result of our experiment is shown in Table. 4.

Table 5 Experiment Result

No	Kernel	Accuracy (%)	Recall (%)	F1 Score (%)
1	Linear	86.5	60	70.97
2	Polynomial	87.5	65.45	74.23
3	RBF	89	70.91	78

Considering our research plan, we need a buzzer detection system that has a good recall and f1 score. From

Table. 4, we achieve the best result when using RBF kernel.

IV. Conclusion

Data from social media can be used to achieve many purposed, like predicting election result, analyzing product review, etc. Buzzer detection is needed to get an objective and valid result. In the next step of our research plan, we will build a system that rely on people's opinion in social media, so we need a good buzzer detection system.

Our experiment shows that our proposed buzzer detection system can distinguish buzzer user from normal user well. Experiments using other classification method can be done to get a better result. Another approach that we assume can improve the performance of the buzzer detection system is implementing similarity threshold on the tweet similarity feature extraction.

Notation

N_t : the total number of tweets
df_i : number of tweets containing term-i
n : number of unique terms in the bag of words (BOW) of the corresponding user
$s_{i,j}$: similarity of tweet i and tweet j
$w_{i,k}$: the weight of term k in tweet i
$w_{j,k}$: the weight of term k in tweet j
N : number of tweets of the corresponding user

Acknowledgment

We would like to thank Direktorat Riset dan Pengabdian Kepada Masyarakat Republik Indonesia for funding this research, and Pusat Penelitian dan Pengabdian kepada Masyarakat Politeknik Negeri Madura for the support to this research.

References

[1] M. Ibrahim, O. Abdillah, A. F. Wicaksono dan M. Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in A Twitter Nation," dalam International Conference on Data Mining Workshops, 2015. Available: https://doi.org/10.1109/ICDMW.2015.113

[2] "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," Journal of Big Data, vol. 5, no. 51, pp. 1-10, 2018. Available: https://doi.org/10.1186/s40537-018-0164-1

[3] T. A. Arafat, I. Budi, R. Mahendra dan D. A. Salehah, "Demographic Analysis of Candidates Supporter in Twitter During Indonesian Presidential Election 2019," dalam 2020 International Conference on ICT for Smart Society (ICISS), Bandung, 2020. Available: https://doi.org/10.1109/ICISS50791.2020.9307598

[4] H. Wang, D. Can, A. Kazemzadeh, F. Bar dan S. Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle," dalam ACL '12: Proceedings of the ACL 2012 System Demonstrations, Jeju Island, 2012. Available: https://dl.acm.org/doi/abs/10.5555/2390470.2390490

[5] A. Sarlan, C. Nadam dan S. Basri, "Twitter Sentiment Analysis," dalam Proceedings of the 6th International Conference on Information Technology and Multimedia, Putrajaya, 2014. Available: https://doi.org/10.1109/ICIMU.2014.7066632

[6] M. Ghiassi, J. Skinner dan D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," Expert Systems with Applications, vol. 40, no. 16, pp. 6266-6288, 2013. Available: https://doi.org/10.1016/j.eswa.2013.05.057

[7] Z. Jianqiang, G. Xiaolin dan Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," IEEE Access, vol. 6, pp. 23253-23260, 2018. Available: https://doi.org/10.1109/ACCESS.2017.2776930

[8] P. Ray dan A. Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," dalam 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), Pune, 2017. Available: https://doi.org/10.1109/ICDMAI.2017.8073512

[9] N. Yadav, O. Kudale, S. Gupta, A. Rao dan A. Shitole, "Twitter Sentiment Analysis Using Machine Learning For Product Evaluation," dalam 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2020. Available: https://doi.org/10.1109/ICICT48043.2020.9112381

[10] A. Suciati, A. Wibisono dan P. Mursanto, "Twitter Buzzer Detection for Indonesian Presidential Election," dalam 3rd International Conference on Informatics and Computational Sciences, Semarang, 2019. Available: https://doi.org/10.1109/ICICoS48119.2019.8982529

[11] M. Kantepe dan M. C. Ganiz, "Preprocessing Framework for Twitter Bot Detection," dalam International Conference on Computer Science and Engineering, 2017. Available: https://doi.org/10.1109/UBMK.2017.8093483

[12] Y. M. Vianny dan E. B. Setiawan, "Implementation of Rumor Detection on Twitter Using J48 Algorithm," Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 4, no. 5, pp. 775-781, 2020. Available: https://doi.org/10.29207/resti.v4i5.2059

[13] B. Li dan L. Han, "Distance Weighted Cosine Similarity Measure for Text Classification," Intelligent Data Engineering and Automated Learning – IDEAL 2013, vol. 8206, pp. 611-618, 2013. Available: https://doi.org/10.1007/978-3-642-41278-3_74

[14] R. A. Lahitani, A. E. Permanasari dan N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," dalam 2016 4th International Conference on Cyber and IT Service Management, 2016, Bandung. Available: https://doi.org/10.1109/CITSM.2016.7577578

[15] F. Rahutomo, T. Kitasuka dan M. Aritsugi, "Semantic Cosine Similarity," dalam The 7th International Student Conference on Advanced Science and Technology ICAST, Seoul, 2012.

[16] D. Gunawan, C. A. Sembiring dan M. A. Budiman, "The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents," Journal of Physics: Conference Series, vol. 978, 2018. Available: https://doi.org/10.1088/1742-6596/978/1/012120

[17] C.-H. HUANG, J. YIN dan F. HOU, "A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method," Chinese Journal of Computers, no. 5, p. 856—864, 2011. Available: https://doi.org/10.3724/SP.J.1016.2011.00856

[18] K. Park, J. S. Hong dan W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," Applied Artificial Intelligence, vol. 34, no. 5, pp. 396-411, 2020. Available: https://doi.org/10.1080/08839514.2020.1723868

[19] P. Y. Ristanti, A. P. Wibawa dan U. Pujianto, "Cosine Similarity for Title and Abstract of Economic Journal Classification," dalam 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, 2019. Available: https://doi.org/10.1109/ICSITech46713.2019.8987547

[20] F. S. Al-Anzi dan D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," Journal of King Saud University - Computer and Information Sciences, vol. 29, no. 2, pp. 189-195, 2017. Available: https://doi.org/10.1016/j.jksuci.2016.04.001

[21] J. A. Panatra, F. B. Chandra, W. Darmawan, H. L. H. S. Warnars, W. H. Utomo dan T. Matsuo, "Buzzer Detection to Maintain Information Neutrality in 2019 Indonesia Presidential Election," dalam 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), Toyama, 2019. Available: https://doi.org/10.1109/IIAI-AAI.2019.00177

[22] A. Mustofa, H. Tjandrasa dan B. Amaliah, "Deteksi Penyakit Glaukoma pada Citra Fundus Retina Mata Menggunakan Adaptive Thresholding dan Support Vector Machine," Jurnal Teknik ITS, vol. 5, no. 2, pp. A572-A575, 2016. Available: http://dx.doi.org/10.12962/j23373539.v5i2.18929

[23] R. Taqiuddin, F. A. Bachtiar and W. Purnomo, "Opinion Spam Classification on Steam Review using Support Vector Machine with Lexicon-Based Features," KINETIK: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, vol. 6, no. 4, 2021. Available: https://doi.org/10.22219/kinetik.v6i4.1323

[24] M. I. Rosadi, "DETEKSI DAN PENGHITUNGAN MANUSIA PADA VIDEO MENGGUNAKAN METODE HOG DAN SVM," Nusantara Journal of Computers and Its Application, vol. 1, no. 1, 2016. Available: https://journal.csnu.or.id/index.php/njca/article/view/15

[25] B. Trstenjak, S. Mikac dan D. Donko, "KNN with TF-IDF Based Framework for Text Categorization," Procedia Engineering, no. 69, pp. 1356-1364, 2014. Available: https://doi.org/10.1016/j.proeng.2014.03.129

[26] G. Salton dan M. J. McGill, Introduction to Modern Information Retrieval, New York, 1986.

[27] A. Singhal, "Modern Information Retrieval: A Brief Overview," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001.