Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 661−666
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1435

Cyberbullying Detection on Twitter using Support Vector Machine Classification Method

Ni Luh Putu Mawar Silveria Putri Waisnawa*, Dade Nurjanah, Hani Nurrahmi

Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1,*]mawarsilveria@students.telkomuniversity.ac.id, [2]dadenurjanah@telkomuniversity.ac.id,
[3]haninurrahmi@telkomuniversity.ac.id
Correspondence Author Email: mawarsilveria@students.telkomuniversity.ac.id
Submitted: **21/03/2022**; Accepted: **26/03/2022**; Published: **31/03/2022**

Abstract−Bullying is when someone or a group of individuals is continuously attacked. Because of the advancement of the internet, it has become very easy for society to engage in harmful acts of bullying by attacking a person or group of people who can hurt the victim, this is known as cyberbullying. Twitter is a social media platform that may be used by the society to share information and can also be used to perpetrate cyberbullying actions by sending messages (tweets) that addressed to the victims. This final project was developing a system to detect cyberbullying on Twitter. The system uses the Support Vector Machine method to classify whether the tweets that are shared include cyberbullying or not. In addition, this research also uses Term Frequency-Inverse Document Frequency (TF-IDF) and N-gram feature extraction for data that has gone through the pre-processing stage. In collecting data, the author crawled tweets based on the keywords 'jelek', 'bodoh', 'goblok', 'brengsek', 'bangsat', 'memalukan', 'laknat', 'bacot' and 'pelacur'. The best performance results of the research is 76.2% accuracy, 73.2% precision, 78.2% recall and 75.6% F1-Score generated by the RBF kernel with a total of n=1.

Keywords: Cyberbullying; Twitter; Support Vector Machine; N-Gram; TF-IDF

1. INTRODUCTION

Bullying can be defined as an aggressive activity or behavior that is intentionally carried out by a group of people or a person repeatedly time by time against a victim [1]. With the advancement of internet, term of cyberbullying appears which is defined as the behavior of a person or group intentionally and repeatedly doing actions that hurt others through computers, cell phones, and other electronic devices [1]. Cyberbullying have various motivations, sometimes just for fun, seeking attention, anger, frustration and wanting revenge. The development of the internet presents social media that can be used by its users to share, participate and create virtual world spaces. In this digital era, it is very difficult for people not to use social media, this is because social media can make it easier for people to find information and also share information. However, many users use social media presence for negative things, for example doing cyberbullying With the existence of social media, the number of cyberbullying can also develop unconsciously. Cyberbullying can be easily found due to the freedom of social media users in submitting or sharing their reviews to individuals, groups and the public. It's rare for social media users to think about the impact of a shared review or content including cyberbullying as an action that could hurt or disturb others and harm their mental health.

In 2016 Indonesian Twitter users occupied the top three in the world with 24.3 million users [2]. The number of Twitter users is affect to the number of positive and negative messages shared by them. Cyberbullying is one of the negative messages found on Twitter. Not many Twitter users are aware that reviews or content that is shared and directed at individuals or groups is an act of bullying that can hurt and even cause the victim of bullying to hurt himself. This can be caused by users who are awareless on sharing reviews or content. Because of the lack of awareness on sharing reviews or bullying content, a system is needed that can classify the reviews or content including cyberbullying or not, and also this classification is carried out to prevent cyberbullying actions that can cause victims to feel afraid, stressed, anxious, lack self-confidence to depression and cause the desire to hurt and suicide.

For classification there are several methods that can be used to classify the data, including Support Vector Machine (SVM) and Naïve Bayes. In research [3] regarding sentiment analysis on cyberbullying on Twitter, the SVM method produces a better performance than Naïve Bayes. SVM produces an accuracy of 71.25% and Naïve Bayes produces an accuracy of 52.70%. Another study regarding sentiment analysis on Cyber-Aggressive Comments [4] using the logistic regression method produces an accuracy of 73.76% and SVM produces an accuracy of 77.65%, in this research SVM is better than logisic regression.

Support Vector Machine is a technique to make a prediction, either in the case of classification or regression [5]. In a research by Mifta Sintaha and Moin Mostakim regarding cyberbullying cases using the Naïve Bayes method and the Support Vector Machine with Term Frequency-Inverse Document Frequency (TF-IF) method as a feature extraction. The accuracy of the model obtained from this study is Naïve Bayes of 73.03% and the accuracy of the Support Vector Machine is 89.54% [6]. In research [7] using the Support Vector Machine algorithm and using N-gram feature extraction produces an accuracy of 92.75%.

In a research by R.R. Dalvi, S. Baliram Chavan, and A. Halbe [3] about cyberbullying on Twitter uses support vector machine method and TF-IDF as a feature extraction. The research produces performance 71.25% of accuracy, 71% of precision, 71% of recall and 70% of F1-Score. In this research, cyberbullying detection on Twitter will use the support vector machine and TF-IDF methods, also the n-gram will be used as a feature extraction with the number of n=1,2,3,4,5 and four different kernels of support vector machine namely linear kernel, radial basis function (RBF)

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 661−666
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1435

kernel, sigmoid kernel and polynomial kernel. The purpose of using the different number of n and different kernel of SVM is for knowing the effect of different number of n on n-gram and the effect of different kernel on support vector machine to produce a better performance than previous research.

2. RESEARCH METHODOLOGY

The stages of this reaserach shows at Figure 1.

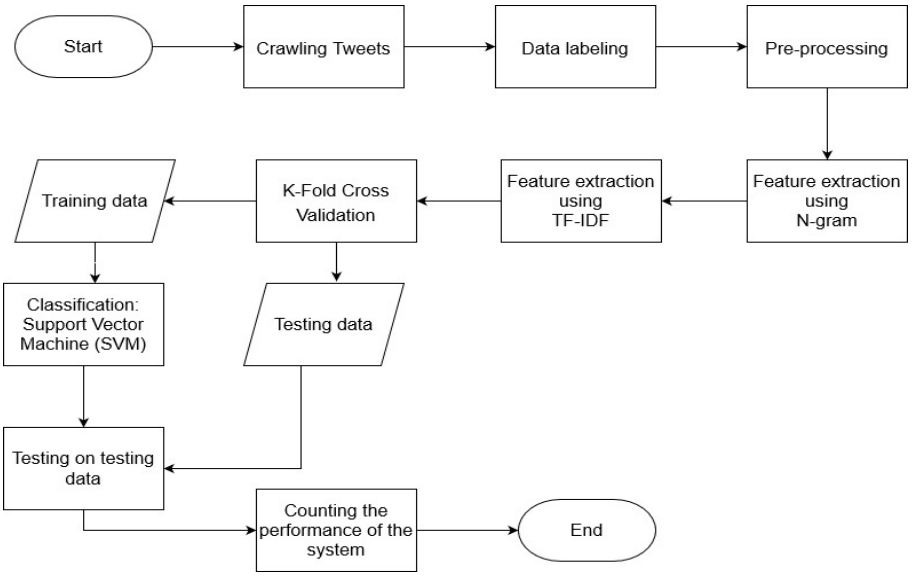

Figure 1. The Stages of The Research

2.1 Crawling Tweets

Data collection (crawling) is done by taking data from Twitter with a total of 10,000 data, 80% as training data and 20% as testing data. The data crawling process is carried out using the Application Program Interface (API) provided by Twitter. The keywords that used are ‘jelek’, ‘bodoh’, ‘goblok’, ‘brengsek’, ‘bangsat’, ‘memalukan’, ‘laknat’, ‘bacot’ and ‘pelacur. Based on [1] examples of cyberbullying actions are flaming (burning) which is sending text messages whose contents are angry and frontal words, and also harassment (disturbance) the messages containing interference with email, sms, or text messages on social networking. Therefore, a survey was conducted to 122 people regarding the keywords used to find out the word is a word full of anger and can hurt other people's feelings or not. The process of selecting keywords for data crawling is done by distributing survey questions in the form of Google Forms to respondents, and from the survey results it is determined that more than 50% of respondents stated that these keywords are words that are full of anger or disturb and hurt others. Figure 2 is the result of the survey.

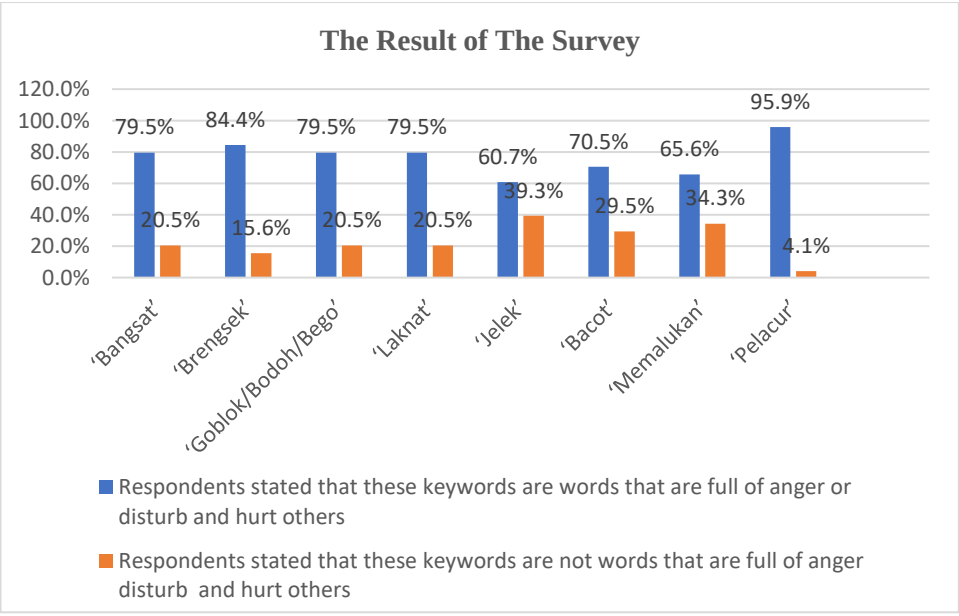

Figure 2. The Result of The Survey that used as *Keyword*

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 661−666
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1435

2.2 Data Labeling

In the process of making datasets for the classification system needs a mechanism so that the datasets that have been collected have the correct class label [8]. In this study, the data labeling stage was carried out by sharing the data that had been collected, after the data was collected the data was distributed to five participants to give label to the data in the form of a number 0 (non-cyberbullying) or 1 (cyberbullying). After labeling the data, a polling is conducted on the labels, the labels used for each data are the results of the polling from the five participants.

2.3 Data Pre-processing

Data pre-processing is the stage of preparing data to be processed by the classification system with the purpose to improve the quality and produce an efficient data when processed by the classification system such as removing punctuation marks, eliminating characters other than letters, eliminating spaces, deleting words that are considered not have an important influence, converts capital letters to lowercase, removes the prefix or suffix from a word and also breaks sentences into words. The steps in pre-processing are as belows:

1. Data Cleaning
   Data cleaning is process that is applied to remove noise and fix inconsistencies in data. Data in the real world tends to be noisy and inconsistent. In tweet cleaning data, URLs, numbers, symbols, and attributes that contain missing values or are empty are deleted.
2. Case Folding
   Case folding is the process of changing capital letters to lowercase on the data. This is done so that all letters in the data are uniform.
3. Tokenizing
   Tokenizing is the process of separating words separated by spaces. This is done to facilitate the next step.
4. Stop Words
   Stop words is the deletion of words in tweets that still contain words that are considered not to have an important influence in determining classification such as conjunctions.
5. Stemming
   Stemming is the process of removing words and turning them into basic words. This can be done by removing the prefix or suffix from a word.

2.4 K-fold Cross Validation

K-fold cross validation is the division of data into k datasets of the same size. Purpose of using k-fold cross validation is to eliminate bias in the data. Training and testing were carried out k times. In the first experiment, the S1 subset was treated as test data and the other subset was treated as training data, in the second experiment the S1, S3,...,Sk subsets became training data and S2 became test data, and so on [9]. In this experiment, number of k that used is k=5.

2.5 N-gram

N-gram is a document processing method that is usually used in spelling correction, word prediction and other processing. An N-gram is a series of tokens of length n. The value of n only refers to the number of tokens. Each data consists of n-grams per character such as unigram (one character), bigram (two characters) and trigram (three characters). With that, the N-gram method is used as a method to find the characteristics of each document which will produce a language model [10] and to take 5 character pieces of n letters from a word that is read continuously from the source text to the end. from the document[11].

2.6 Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF weighting is a calculation that describes how important a word (term) is in a document and corpus. This process is used to evaluate the weight of the relevance of the term from a document to all documents in the corpus [8]. Intuitively, this calculation determines how relevant a given word is in a particular document. Common words in one or a small group of documents tend to have higher TF-IDF scores than common words such as articles and prepositions [11]. The frequency with which a word occurs in a given document indicates how important it is in the document. The frequency of documents containing the word indicates how common the word is. The word weight is greater if it appears frequently in a document and is smaller if it appears in many documents [12]. The TF-IDF algorithm uses a formula to calculate the weight (W) of each document against keywords with the formula:

$$W_{ij} = t f_{ij} * Idf_j, \text{ where } Idf_j = (\log\left(\frac{N}{df}\right)) \quad (1)$$

With W_{ij} is the weight of the i document to the j word, tf_{ij} is the number of words searched for in a document, Idf_j is the Inverse Document Frequency, N is the total document, and df is the number of documents that contain the searched word.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 661−666
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1435

2.7 Support Vector Machine (SVM)

After performing the feature selection stage, a classification process will be carried out using the support vector machine algorithm. Support Vector Machine is a supervised learning method. SVM can be used for regression and classification [3]. The use of the SVM algorithm which aims to classify text using term index weights as a feature was pioneered by Thorsten Joachim. SVM learning has been popularized since 1992 by Boser, Guyon, and Vapnik [13].

SVM can solve problems linearly and non-linearly. Solving non-linear problems using the kernel concept in a high-dimensional workspace, by looking for a hyperplane that can maximize the margin between data classes. Hyperplane is useful in separating 2 groups of class +1 and class -1 where each class has its own pattern. In making decisions using the SVM method, kernel functions are used. The kernels used in this experiment are linear, polynomial, radial basis function, sigmoid. Here are the mathematic formula of the kernel functions used in the SVM method:

1. Linear Kernel
   The linear kernel is the simplest kernel function. Linear kernel is used when the analyzed data is linearly separated. The linear kernel is suitable when there are many features. The following is the equation of the linear kernel [14].

   $$\text{Linear Kernel} = x^T x \quad (2)$$

2. *Polynomial* Kernel
   Polynomial Kernel is a kernel function that is used when the data are not linearly separated. The polynomial kernel is very suitable for problems where all training data are normalized [14].

   $$\text{Polynomial Kernel} = (x^T x + 1)^p \quad (3)$$

3. RBF kernel
   The RBF kernel is a kernel function that is used when the data is not linearly separated. RBF has two parameters, namely Gamma and Cost (C) to produce an optimal classification and avoid misclassification. Gamma is used to measure how far the influence of a sample of training data is. A low gamma value means “far” and a high value means “near” [14]

   $$\text{RBF Kernel} = \exp(\gamma||x - x||^2) \quad (4)$$

4. Sigmoid kernel
   The equation of the sigmoid kernel is as follows.

   $$\text{Sigmod Kernel} = \tanh(\beta_0 x^T x_i + \beta_1) \quad (5)$$

2.8 System Performance

System performance on text classifiers can be done using a confusion matrix to measure the performance of a system. In this study the confusion matrix is used to determine the performance of the system by calculating precision, recall, accuracy and F1-Score. Precision is the ratio of a positive correct prediction to the overall positive predicted result. Equation 7 is a formula of calculating precision.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad (6)$$

Recall is a comparison of true positive predictions with all true positive data. Equation 8 is formula of calculating recall.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad (7)$$

Accuracy is a comparison of correct predictions (positive and negative) with the entire data. Equation 9 is a formula for calculating accuracy.

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{True Positive+False Positive+False Negative+True Negative}} \quad (8)$$

F1-Score is a weighted comparison of the average precision and recall. Equation 10 is formula of calculating F1-Score.

$$\text{F1-Score} = 2*\frac{\text{Recall*Precision}}{\text{Recall+Precision}} \quad (9)$$

3. RESULTS AND DISCUSSION

3.1 Results

This reserach is done by doing five experiments on each kernel with a different number of n to get the best results. Research [15] conducted a classification using N-grams with the number of n=1,2, 3 and the combination of n=1,2

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 661−666
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1435

and 3 resulted in different accuracy and obtained the best accuracy with the number of n=1. Research [7] conducted SVM classification with linear kernel, RBF, polynomial and sigmoid give different accuracy with the highest accuracy of 97.81% with polynomial kernel. Based on the results of research conducted in [15] and [7], therefore a study will be conducted on the effect of the SVM kernel and the effect of n on n-grams, in this study an experiment will be conducted with n=1,2,3,4,5 and kernel SVM Linear, sigmoid, RBF and Polynomial to get the best results. The classification results for each experiment are as belows.

Table 1. Tweet Classification using Support Vector Machine and N-gram with n=1

	Accuracy	Precision	Recall	F1 Score
SVC Linear	75.2%	72.1%	77.5%	74.7%
SVC Sigmoid	74.8%	71.6%	77.4%	74.4%
SVC RBF	**76.2%**	73.2%	78.2%	75.6%
SVC Poly	69.2%	75.4%	51.6%	61.3%

Table 2. Tweet Classification using Support Vector Machine and N-gram with n=2

	Accuracy	Precision	Recall	F1 Score
SVC Linear	64.9%	66.1%	52.7%	58.7%
SVC Sigmoid	64.8%	68.5%	47.3%	56%
SVC RBF	64.0%	72.1%	39.1%	50.7%
SVC Poly	63.6%	72.33%	37.31%	49.23%

Table 3. Tweet Classification using Support Vector Machine and N-gram with n=3

	Accuracy	Precision	Recall	F1 Score
SVC Linear	55.9%	74.24%	10.35%	18.18%
SVC Sigmoid	55.85%	77.87%	9.3%	16.61%
SVC RBF	48.4%	47.64%	92.07%	62.79%
SVC Poly	55.3%	76.53%	7.92%	14.36%

Table 4. Tweet Classification using Support Vector Machine and N-gram with n=4

	Accuracy	Precision	Recall	F1 Score
SVC Linear	53.1%	75%	1.26%	2.49%
SVC Sigmoid	53.1%	75%	1.26%	2.49%
SVC RBF	47.65%	47.44%	99%	64.15%
SVC Poly	52.9%	69.23%	0.09%	1.87%

Table 5. Tweet Classification using Support Vector Machine and N-gram with n=5

	Accuracy	Precision	Recall	F1 Score
SVC Linear	52.8%	62.5%	0.52%	1.04%
SVC Sigmoid	52.8%	62.5%	0.52%	1.04%
SVC RBF	47.75%	47.5%	99.78%	64.3%
SVC Poly	52.8%	62.5%	0.52%	1.04%

Based on the table above, the best accuracy of 76.2% is generated by SVC RBF and n-grams with n=1 in table 1. The following are the classification results based on the best accuracy in table 6.

Table 6. Classification Results based on The Best Accuracy

Data Class	Cyberbullying Classification	Non-cyberbulying Classification
Cyberbullying	740	206
Non-cyberbulying	270	784

Performance calculations based on table 6 are as belows:

$$\text{Accuracy} = \frac{(740+784)}{(740+206+270+784)} = 0.761 = 76.2\%$$

$$\textit{Precision} = \frac{740}{(270+740)} = 0.732 = 73.2\%$$

$$\textit{Recall} = \frac{740}{(740+206)} = 0.782 = 78.2\%$$

$$\text{F1-}\textit{Score} = \frac{2*0.782*0.732}{(0.782+0.732)} = 0.756 = 75.6\%$$

3.2 Discussion

After this system is implemented on twitter data with a ratio of 20% of testing data and 80% of training data tested with a different number of n-grams produces different performance, the significant of the difference in performance

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 661−666
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1435

is caused by the use of the kernel on the support vector machine and also the number of n in n-grams. In table 1 there are the best results, the classification using a support vector machine with RBF kernel and the number of n in n-grams is 1 with an accuracy of 76.2%, precision 73.2%, recall 78.2% and F1-Score 75.6%. This study gives better performance than research [3] which conducted the same research, namely Detecting A Twitter Cyberbullying Using Machine Learning, resulting in 71.25% accuracy, 71% precision, 71% recall and 70% F1-Score. Based on the results of classifying the number of n=1 works better on each SVM kernel than the other n numbers, it can be seen in the table of results for each test that the smaller the number of n, the better performance will be.

4. CONCLUSIONS

Based on the research that has been done, it can be concluded that the classification of text about cyberbullying on Twitter using 10,000 tweet data produces the best performance with an accuracy of 76.2, precision 73.2%, Recall 78.2% and F1-Score 75.63% using the Support Vector Machine algorithm with the RBF kernel and N-gram feature extraction method with n=1. The diference of the performance is caused by the different kernel of SVM and the number of n in N-gram. Suggestions that can be given by researchers for future researchers are to use more datasets and use data sets other than Indonesian so that the system is better trained in classifying so that it can produce research with better performance results.

REFERENCES

[1] M. Rifauddin, "Fenomena Cyberbullying pada Remaja," *Khizanah Al-Hikmah J. Ilmu Perpust. Inf. Dan Kearsipan*, vol. 4, no. 1, pp. 35–44, Jun. 2016, doi: 10.24252/kah.v4i1a3.
[2] R. M. Huda, "Indonesia Pengguna Twitter Terbesar Ketiga Dunia," *Setara.net*, Jun. 11, 2017. https://setara.net/indonesia-pengguna-twitter-terbesar-ketiga-dunia/
[3] R. R. Dalvi, S. Baliram Chavan, and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, May 2020, pp. 297–301. doi: 10.1109/ICICCS48265.2020.9120893.
[4] V. S. Chavan and Shylaja S S, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Kochi, India, Aug. 2015, pp. 2354–2358. doi: 10.1109/ICACCI.2015.7275970.
[5] R. M. Kamal, "Analisis Sentimen Cyberbullying Pada Komentar Facebook Dengan Metode Klasifikasi Support Vector Machine," Universitas Komputer Indonesia, 2019.
[6] M. Sintaha and M. Mostakim, "An Empirical Study and Analysis of the Machine Learning Algorithms Used in Detecting Cyberbullying in Social Media," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, Dec. 2018, pp. 1–6. doi: 10.1109/ICCITECHN.2018.8631958.
[7] Noviantho, S. M. Isa and L. Ashianti, "Cyberbullying classification using text mining," 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 2017, pp. 241-246, doi: 10.1109/ICICOS.2017.8276369.
[8] A. Rachmat and Y. Lukito, "Sentipol: Dataset Sentimen Komentar Pada Kampanye Pemilu Presiden Indonesia 2014 Dari Facebook Page," *Konf. Nas. Teknol. Inf. Dan Komun. 2017*, pp. 218–228, 2016.
[9] M. Bramer, *Clustering*. Springer, 2007.
[10] R. S. Putra, "Klasifikasi dokumen menurut bahasa berbasis N-Gram," 2018.
[11] C. P. Medina and M. R. R. Ramon, "Using TF-IDF to Determine Word Relevance in Document Queries Juan," New Educ. Rev., vol. 42, no. 4, pp. 40– 51, 2015.
[12] R. Melita, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, 2018.
[13] Paramita, "Penerapan Support Vector Machine untuk Ekstraksi Informasi dari Dokumen Teks," Jun. 2008.
[14] A. Kesumawati, "Perbandingan Metode Support Vector Machine (SVM) Linear, Radial Basis Function (RBF), dan Polinomial Kernel dalam Klasifikasi Bidang Studi Lanjut Pilihan Alumni UII," 2018.
[15] S. Naz, A. Sharan, and N. Malik, "Sentiment classification on twitter data using support vector machine," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018, pp. 676–679.