International Journal Of Artificial Intelegence Research
Vol 3, No 3, December 2019, pp. 28-37
ISSN: 2579-7298

Journal Classification Using Cosine Similarity Method on Title and Abstract with Frequency-Based Stopword Removal

Piska Dwi Nurfadila[a,1,*], Aji Prasetya Wibawa[a,2], Ilham Ari Elbaith Zaeni [a,3], Andrew Nafalski [b,4]

[a] *Electrical Engineering Department, Universitas Negeri Malang, Jl. Semarang No.5, Kota Malang, Jawa Timur 65145 Indonesia*
[b] *University Of Soult Australia, Australia*
[1] *Piskadwi12@gmail.com*;* [2] *aji.prasetya.ft@um.ac.id;* [3] *Ilham.ari.ft@um.ac.id;* [4] *Andrew.nafalski @unisa.edu.au*
** corresponding author*

ARTICLE INFO

Article history:
Received June 2019
Revised September 2019
Accepted December 2019

Keywords:
Vector Space Model
Cosine Similarity
Stopword Removal
K-Fold Cross-Validation
Frequency

ABSTRACT

Classification of *Artikel jurnal ekonomi* has been done using the VSM (Vector Space Model) approach and the Cosine Similarity method. The results of previous studies are considered to be less optimal because Stopword Removal was carried out by using a dictionary of basic words (tala). We can assume that because only deleted words are basic words. This study focuses on the phase of word elimination by adding frequency-based stopword removal. Because the term with a certain frequency is assumed to be an insignificant word and will give less relevant results. The method performance in this study was tested using K-fold Cross-Validation. The performance of the cosine similarity method that has been added to stopword removal increased 2% compared to previous research.

I. Introduction

There are a lot of text mining methods that can be used to classify documents, including K-NN, Jaccard, and Cosine Similarity. Understanding the K-NN method can be seen in the literature [1]. while the understanding of the Jaccard method can be seen in the literature [2]. The third understanding of the Cosine Similarity method can be seen in the literature [3]. The cosine similarity method is a method that produces the highest performance value compared to the K-NN and Jaccard methods [4]. This can occur because the cosine similarity value between two vectors depends on the number of word frequencies of the test and training documents [5]. This method uses the normalization concept of vector length by comparing word frequency between two documents so that it can produce high accuracy values [6]. Before being classified, the document will go through the pre-processing stage.

Pre-processing is a stage that functions to change data so that it is more structured and ready to be processed to the next stage [7]. The benefits of applying the pre-processing stage, which will help reduce noise, improve classification performance, and speed up the classification process [8]. In general, the pre-processing stages carried out for text mining consist of changing capital letters to lowercase letters, eliminating numbers and punctuation, stopword removal, and stemming [9]. From several stages, the most regularly applied pre-processing is stopword removal [10]. This stage is done to eliminate words that do not affect the classification process [11] [12].

Besides using the tuning *Tala* dictionary, deletion of words can be done with: Term Frequency Filtering is done to reduce dimensions by reducing all words that have a specific frequency, Supervised Word Removal is done to control the stoplist because the creation of bigram causes invalid words to be entered in the stoplist list [13], and Deterministic Finite Automata (DFA) functions to improve the performance of an algorithm by detecting whether a word-finding includes a stopword or not [14].

This research focuses on the stopword removal pre-processing technique on the economic dictionary with frequency-based. The application of stopword removal based on the economic dictionary aims to eliminate important words related to the world economy. This stage is done so that the word is not registered in the frequency-based stopword removal dictionary anymore. Whereas, frequency-based stopword removal dictionary is made to make the term with certain frequency documents that appear in the document as a stopword dictionary because the term with a certain frequency is assumed to be not important words and it will provide less relevant results on the calculation

II. Method

The classification of *Artikel jurnal ekonomi* by adding frequency-based stopword removal was carried out in this study. The stages consisted of collecting datasets, pre-processing to change the data to be more structured, calculating VSM-based Cosine Similarity values, and testing algorithms using k-fold cross-validation.

A. Research Dataset

The dataset used in this study is *Artikel Jurnal Ekonomi Universitas Negeri Malang* of Indonesian language. The datasets were collected in March 2019 consisting a total of 126 data containing titles and abstracts. The datasets of *Artikel Jurnal Ekonomi* were grouped into four labels; (1) *Ekonomi Bisnis*, (2) *Pendidikan Akuntansi & Bisnis*, (3) *Pendidikan Bisnis & Manajemen*, and (4) *Pendidikan Akuntansi*.

Table 1 displays the proportion of label members or categories based on their fields.

Table 1. The initial proportion of the number of dataset- labeled members

Label	Number of Instances	Percentage (%)
Ekonomi Bisnis	29	23
Pendidikan Akuntansi & Bisnis	31	25
Pendidikan Bisnis & Manajemen	30	24
Pendidikan Akuntansi	36	28

As much as 126 *Artikel Jurnal Ekonomi* of Indonesian language were used as test and training documents. The following are examples of documents taken from each field of each 1 document. The table of sample documents can be seen in Table 2.

Table 2 displays an example document consisting of label attributes and title & abstract.

Table 2. The initial proportion of the number of datasets- labeled members

Label	Title and Abstract
Ekonomi Bisnis	*PENGARUH KUALITAS PRODUK DAN LAYANAN TERHADAP LOYALITAS PELANGGAN COFFEE SHOP Penelitian ini bertujuan untuk menganalisis pengaruh kualitas produk dan layanan terhadap loyalitas pelanggan, dan kepuasan pelanggan sebagai mediasi. Populasi yang digunakan dalam penelitian ini adalah pelanggan DW Coffee yang telah datang lebih dari satu kali pada rentang usia 20-30 tahun, dengan metode non-probability sampling sebanyak 100 orang.*
Pendidikan Akuntansi & Bisnis	*PENGEMBANGAN MULTIMEDIA INTERAKTIF UNTUK PERUSAHAAN JASA Penelitian pengembangan multimedia interaktif pada mata pelajaran akuntansi Pokok Bahasan Siklus Akuntansi Perusahan Jasa ini bertujuan untuk memaksimalkan media yang telah disediakan oleh sekolah, dan diharapkan dapat memotivasi siswa dalam belajar sehingga tujuan pembelajaran dapat tercapai.*

Label	Title and Abstract
Pendidikan Bisnis & Manajemen	*Pengaruh Penerapan Presensi Sidik Jari (Fingerprint) terhadap Kinerja Guru Melalui Motivasi Kerja di SMA Negeri 5 Malang Hasil penelitian ini menunjukkan (1) Penerapan presensi sidik jari (fingerprint) termasuk dalam kategori sangat baik, motivasi kerja guru termasuk dalam kategori tinggi, dan kinerja guru termasuk dalam kategori baik*
Pendidikan Akuntansi	*PENGARUH PERSEPSI SISWA TENTANG KOMPETENSI PROFESIONALISME GURU TERHADAP MOTIVASI BELAJAR DAN PRESTASI BELAJAR MATA DIKLAT AKUNTANSI Penelitian ini menguji pengaruh persepsi siswa tentang kompetensi profesionalisme guru terhadap motivasi belajar dan prestasi belajar mata dilat akuntansi. Penelitian ini termasuk penelitian kuantitatif dengan metode penjelas.*

B. Pre-processing

The function of pre-processing can be seen in the literature[15]. There were three (3) stages of pre-processing in this study; (1) case folding, (2) stopword removing based on economic dictionary, and (3) frequency-based stopword removing.

The first step was case folding, which was done by changing uppercase letters to lowercase letters [6]. And followed by removing punctuation characters (! @ # $% ^ & *> <?) and numbers (0123456789) in the document [8].

The second stage was a stopword removing based on the economic dictionary that served to eliminate words related to the word economy. Thus, important words related to the word economy were not included in the stage of making frequency-based stopword removal dictionaries. The Economic Dictionary used consisted of 331 words. The steps taken were:

- Tokenizing documents and dictionaries
- Matching all words in the documents with the words in the economic dictionary
    - If the word in the document is the same as the word in the economic dictionary, the word in the document will be deleted,
    - If the word in
    - the document is not the same as the economic dictionary, it can be assumed that the word will not affect the classification process.
- They are recombining the decapitated word into a complete sentence.

The third stage was frequency-based stopword removal, which functioned to delete words on the test document based on the frequency term. The important word related to economics had already been described in the previous process so that the word was not included in the list of frequency-based stopword removal dictionary. The steps taken were:

- Counting the number of the terms' occurrences of the training document.
- Building stopword removal dictionaries based on the frequency terms in training documents.
- Decapitating sentences in test documents and dictionaries based on tokenizing.
- They are matching the frequency-based stopword removal dictionary with terms contained in test documents.
    - If the term is the same as the frequency-based stopword removal dictionary, the term in the test document will be deleted.
    - The term is assumed to be an important word that will influence the classification process.
- They are recombining the decapitated words into a complete sentence.

C. VSM Approach

At this stage, the document was represented by a vector using the VSM approach. The definition of the VSM approach can be seen in the article [16]. The function of VSM is to convert documents into numbers so that we can calculate the weight [17]. Each different word term will be represented by (t_i), whereas $(w_i)\,d$ is the appropriate weight (t_i) in the document *d* [18]. With

VSM approach, the calculation of weight from each term in the training document and test documents was carried out using the TF-IDF weighting method. The TF-IDF has the main ideas that can be seen in the literature [19]. To determine the value of TF-IDF, two elements were used; TF and IDF. There were three (3) stages to determine the value of TF-IDF weighting, namely:

- Calculation of TF (Term Frequency)
  It was done to calculate the frequency of term *i* in document *j* [20]. The formula for calculating TF can be seen in the literature[15].
- IDF Document (Inverse Dokument Frequency)
  IDF reflected the distribution of terms contained in the literature [19].
- TF-IDF
  The TF-IDF value was obtained by combining both values of TF and IDF. The TF-IDF weighting scheme can be seen in the literature [21] [22].

Whereas, to classify documents using the Cosine Similarity method. Cosine Similarity method uses a calculation based on a vector space similarity measure. The similarity value between two documents stated in two vectors using keywords from a document [2]. The equation for calculating cosine similarity can be seen in the literature [23].

The output of the Cosine Similarity method is a similarity value with a range of zero to one. If the similarity value is closer to one, it means that the level of document similarity is high. Conversely, if the similarity value is close to zero, it means that the level of similarity between the two documents is low [24].

D. Testing Method

The stage of testing the Cosine Similarity method was carried out in two (2) stages, namely:

- K-Fold Cross Validation

  The definition of the K-Fold Cross Validation method can be seen in the article [25]. The way the Cross Validation method works was by dividing the data into almost the same set of k parts. In each repetition, a set of k was used as test data and the remainder was used as training data. The process was repeated as many as k until all the data alternately changed into random test data [26] [27]. The output of this step was a k estimate of the test error which was then averaged to get the estimated value of the expected testing error [28].

- Confusion Matrix

  The definition of confusion matrix can be seen in the literature [29]. At this stage there is an accuracy test of the algorithm used to classify the data. Accuracy test was done by using the confusion matrix method. Testing was done using equations:

  - Accuracy

    The value of the method accuracy was obtained by dividing the number of true documents to true value with the number of all classified documents [11] [30].

  - True Positive Rate (Recall)

    Recall was done through the calculation of the ratio of true positive. The recall calculation formula can be seen in the literature [15].

  - Presision

    Presision was calculated from the ratio of the amount of data in the true dataset that is true positive to the number of true positive data and the number of false negative data. The precision calculation formulas can be seen in the literature [15].

III. Result

In the tests that have been done by removing several different word frequencies, the comparison results of the number of words before, after stopword removal with *tala* dictionary and after

frequency-based stopword removal are obtained. A comparison of the number of words can be seen in Table 3.

Table 3 displays the comparison of the number of words before, after stopword removal with *tala* dictionary and after frequency-based stopword removal with a specified frequency limit of less than 60.

Table 3. Comparison of the Number of Words

Words	Word Numbers Before	Word Numbers After	
		With Tala dictionary	***With dictionary based frequency***
pengaruh kualitas produk dan layanan terhadap loyalitas pelanggan coffee shop penelitian ini bertujuan untuk menganalisis pengaruh kualitas produk dan layanan terhadap loyalitas pelanggan dan kepuasan pelanggan sebagai mediasi populasi yang digunakan dalam penelitian ini adalah pelanggan dw coffee yang telah datang lebih dari satu kali pada rentang usia tahun dengan metode non probability sampling sebanyak orang	56	32	25

Based on Table 3, the number of words before frequency-based stopword removal was 56 words. After stopword removal with *tala* dictionary the remaining 32 words. We can assume that because only deleted words are basic words. And after being executed with frequency-based stopword removal with a specified frequency limit, only 25 words remain. This can be assumed because the words that are displayed are not just basic words. Not only for less than 60 frequencies, testing was done by removing frequencies that are less than 30, 40, 50, and 70. Comparison of the remaining words from each frequency can be seen in Table 4.

Table 4 displays the comparison of the remaining words before and after stopword removal based on the specified frequency.

Table 4. Comparison of the Remaining Words Based On Frequency

Before	After				
	Frequency				
	30	**40**	**50**	**60**	**70**
56	32	27	27	25	24

Based on Table 4, the difference in the number of words remaining from each set frequency limit. This can be assumed because the number of dictionaries on each boundary frequency varies. So that it can affect the number of words remaining after being matched with a dictionary frequency based on the prescribed limits.

In addition to knowing the number of words remaining, the purpose of this study is to know the value of accuracy, precision, recall and the results. The results of the confusion matrix can be seen in Table 5.

Table 5 displays an example of confusion matrix by removing terms that have a frequency of less than 60.

Table 5. Sample of Confusiin Matrix Calculation

Predicted Class	Real Class			
	0	***1***	***2***	***3***
0	27	1	1	1
1	1	17	5	8

Predicted Class	Real Class			
	0	***1***	***2***	***3***
2	1	3	15	5
3	0	10	9	22

Not only term with frequency > 60 was removed, terms with frequencies less than 30, 40, 50, and 70 were deleted. The results comparison of accuracy, precision and recall can be seen in Table 6.

Table 6 displays the result of testing performance based on the frequency that was deleted.

Table 6. Test Result Based on the Frequency That Has Been Deleted

Deleted Frequency Limit (<)	Accuracy (%)	Precision (%)	Recall (%)
30	59,52	62,28	61,94
40	60,31	61,06	60,03
50	61,90	62,28	61,94
60	64,28	64,76	65,24
70	57,93	58,48	58,27

Based on Table 6, the highest number of accuracy results from deletion of frequency that is less than 60. Frequency less than 60 is used as a treshhold value. Whereas, accuracy decreases when terms with frequencies less than 30, 40, 50, and 70 are deleted. We can assume that when the term with frequency < 30 is deleted, the deleted term becomes too little so the accuracy value decreases. Meanwhile, we can assume that the accuracy result of term removal with frequency < 70 decreases because too many terms are deleted causing a decrease in accuracy value.

The results of this study can be compared with the values of accuracy, precision and recall testing of the cosine similarity method.

Table 7 displays the difference between the accuracy comparison of the Cosine Similarity method with stopword removal with *Tala* dictionary and the accuracy of the Cosine Similarity method that has been combined with frequency-based stopword removal.

Table 7. Result of Comparative Performance Accuracy, Precision, and Recall with Previous

	Cosine Similarity with stopword removal with *Tala* dictionary (%)	Cosine Similarity with Stopword Removal based frequency (%)
Accuracy	61,37	64,28
Precision	60,18	64,76
Recall	64,52	65,26

Based on Table 7, the value increases of accuracy, precision and recall are 2.91%, 4.58% and 0.74%. It seems that the increase in accuracy value is still not significant. This is because there are still too many words left after the frequency-based stopword removal stage that can affect the document classification process.

In addition to performance accuracy, precision, and recall, the execution time of the classification process was compared as well. Comparison of the execution time can be seen in Table 8.

Table 8 displays the comparison of the execution time of the Cosine Similarity method which has been combined with stopword removal based *Tala* dictionary with the Cosine Similarity method which has been combined with frequency-based stopword removal.

Table 8. The Result of the Execution Time With Previous Research

	Cosine Similarity With Stopword Removal Based Tala Dictionary (S)	**Cosine Similarity With Stopword Removal Based Frequency (S)**
Execution Time Pre-Processing	0,650	61,6266
Execution Time Classification	0,791	0,05033

Based on Table 8, the required execution time in pre-processing the Cosine Similarity method with stopword removal based *Tala* dictionary is faster; 0.650 s. It can happen because there are not many pre-processing steps in the basic method. Meanwhile, the Cosine Similarity method with frequency-based stopword removal requires an execution time of 61,6266 s. The execution time at the combined pre-processing stage is longer because more stages are carried out. However, the execution time in the classification of the combined Cosine Similarity method with frequency-based stopword removal is faster because the number of words is matched slightly. The execution time needed is only 0.05033 s so that it can speed up the classification process. Meanwhile, the execution time required for classification in the basic Cosine Similarity method is longer. It happens because the number of words that need to be matched are a lot. The execution time required is 0.791 s.

IV. Conclusion

This study concludes that adding frequency-based stopword removal can improve the performance of the Cosine Similarity algorithm. This study resulted in accuracy value of 64.28%. Compared with the previous research which produced accuracy value of 62.70%, the accuracy increase in this study was approximately 2%. Meanwhile, the execution time is needed when the classification process is faster, which is 0.05033 s. However, the results of this study are considered to be less than optimal. It happens because the term frequency is not evenly distributed so that an increase in the value of accuracy is still not optimal. Therefore, the researchers suggest adding stemming to future studies.

Acknowledgment

Thanks to Putri Yuni Ristanti for providing a dataset and research results to be developed.

References

[1] N. Chandra, S. K. Khatri, and S. Som, “Anti Social Comment Classification based on kNN Algorithm,” 2017.

[2] N. D. Nurdiana Ogie, Jumadi, “Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al- Qur ’ An,” vol. I, no. 1, pp. 59–63, 2016.

[3] G. Orellana, M. Orellana, V. Saquicela, F. Baculima, and N. Piedra, “A text mining methodology to discover syllabi similarities among higher education institutions,” *Proc. - 3rd Int. Conf. Inf. Syst. Comput. Sci. INCISCOS 2018*, vol. 2018–Decem, pp. 261–268, 2018.

[4] L. Zahrotun, “Comparison Jaccard similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method,” vol. 5, no. 1, pp. 11–18, 2016.

[5] A. I. Kadhim, Y. N. Cheah, N. H. Ahamed, and L. A. Salman, “Feature extraction for co-occurrence-based cosine similarity score of text documents,” *2014 IEEE Student Conf. Res. Dev. SCOReD 2014*, pp. 2–5, 2014.

[6] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro*, vol. 9, no. 1, pp. 18–23, 2017.

[7] D. S. Maylawati, W. B. Zulfikar, C. Slamet, M. A. Ramdhani, and Y. A. Gerhana, "An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media," *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, 2019.

[8] M. Mhatre, D. Phondekar, P. Kadam, A. Chawathe, and K. Ghag, "Dimensionality Reduction for Sentiment Analysis using Pre-processing Techniques," no. Iccmc, pp. 16–21, 2017.

[9] S. M. Babapour and M. Roostaee, "Web pages classification: An effective approach based on text mining techniques," *2017 IEEE 4th Int. Conf. Knowledge-Based Eng. Innov. KBEI 2017*, vol. 2018–Janua, pp. 0320–0323, 2018.

[10] K. Amarasinghe, M. Manic, and R. Hruska, "Optimal stop word selection for text mining in critical infrastructure domain," *Proc. - 2015 Resil. Week, RSW 2015*, pp. 179–184, 2015.

[11] R. Geetharamani, M. N. Kumar, and L. Balasubramanian, "Identification of emotions in text articles through data pre-processing and data mining techniques," *Proc. 2016 Int. Conf. Adv. Commun. Control Comput. Technol. ICACCCT 2016*, no. 978, pp. 611–615, 2017.

[12] M. Jahantigh, N. Daneshpour, M. Erfani, and N. Orojlou, "Presenting an improved combination for classification of Persian texts," *2016 8th Int. Conf. Inf. Knowl. Technol. IKT 2016*, pp. 234–240, 2016.

[13] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated Document Classification for News Article in Bahasa Indonesia based on Term Frequency Inverse Document Frequency (TF-IDF) Approach," pp. 0–3, 2014.

[14] K. S. Dar, A. Bin Shafat, and M. U. Hassan, "An efficient stop word elimination algorithm for Urdu language," *ECTI-CON 2017 - 2017 14th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol.*, pp. 911–914, 2017.

[15] A. Mishra and S. Vishwakarma, "Analysis of TF-IDF Model and its Variant for Document Retrieval," *Proc. - 2015 Int. Conf. Comput. Intell. Commun. Networks, CICN 2015*, pp. 772–776, 2016.

[16] C. Langcai, L. Zhihui, and L. Yuanfang, "Research of text clustering based on improved VSM by TF under the framework of Mahout," *Proc. 29th Chinese Control Decis. Conf. CCDC 2017*, pp. 6597–6600, 2017.

[17] V. Carrera-Trejo, G. Sidorov, S. Miranda-Jiménez, M. M. Ibarra, and R. C. Martínez, "Latent Dirichlet Allocation complement in the vector space model for Multi-Label Text Classification," *Int. J. Comb. Optim. Probl. Informatics*, vol. 6, no. 1, pp. 7–19, 2015.

[18] X. Liu, H. Xiong, and N. Shen, "A Hybrid Model of VSM and LDA for Text Clusteing," pp. 230–233, 2017.

[19] A. Guo and T. Yang, "Research and improvement of feature words weight based on TFIDF algorithm," *Proc. 2016 IEEE Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2016*, pp. 415–419, 2016.

[20] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014.

[21] R. Premalatha and S. Srinivasan, "Text processing in information retrieval system using vector space model," *2014 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2014*, no. 978, pp. 0–5, 2015.

[22] I. Yahav, O. Shehory, and D. Schwartz, "Comments Mining With TF-IDF: The Inherent Bias and Its Removal," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 437–450, 2019.

[23] M. E. Sulistyo, R. Saptono, A. Asshidiq, J. Informatika, and U. S. Maret, "Penilaian Ujian Bertipe Essay Menggunakan Metode Text Similarity," vol. 12, no. 02, pp. 146–158, 2015.

[24] M. Alodadi and V. P. Janeja, "Similarity in Patient Support Forums: Using TF-IDF and Cosine Similarity Metrics," *Proc. - 2015 IEEE Int. Conf. Healthc. Informatics, ICHI 2015*, pp. 521–522,

2015.

[25] I. K. Hadihardaja, M. Cahyono, and I. Soekarno, "A Study of Hold-Out and K-Fold Cross Validation for Accuracy of Groundwater Modeling in Tidal Lowland Reclamation Using Extreme Learning Machine," pp. 228–233, 2014.

[26] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," *Proc. - 6th Int. Adv. Comput. Conf. IACC 2016*, no. Cv, pp. 78–83, 2016.

[27] X. Meng, Q. Zhou, J. Hu, L. Shu, and P. Jiang, "A Global Support Vector Regression Based on Sorted K-Fold Method 1," pp. 2169–2173, 2017.

[28] S. Sci, M. Ljumovi, and R. B. Gmbh, "Estimating Expected Error Rates of Random Forest Classifiers : A Comparison of Cross-Validation and Bootstrap," pp. 212–215, 2015.

[29] J. L. García-balboa, M. V Alba-fernández, F. J. Ariza-lópez, and J. Rodríguez-avi, "Homogeneity Test For Confusion Matrices : A Method And An Example," pp. 1203–1205, 2018.

[30] N. R. Fatahillah, "Implementation Of Naive Bayes Classifier Algorithm On Social Media (Twitter) To The Teaching Of Indonesian Hate Speech," pp. 128–131, 2017.