-_Term_Weighting_TF-IDF_ICF_Term_Distribution_Centroid_Text.pdf

by Student17 1

Submission date: 28-Apr-2024 08:58PM (UTC+0700)

Submission ID: 2225858076

File name: -_Term_Weighting_TF-IDF_ICF_Term_Distribution_Centroid_Text.pdf (377.66K)

Word count: 4965

Character count: 25064

Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 8, No. 4, November, Pp. 781-788

Combination of term weighting with class distribution and centroid-based approach for document classification

Christian Sri Kusuma Aditya*[1], Fauzi Dwi Setiawan Sumadi[1]
Universitas Muhammadiyah Malang, Malang, Indonesia[1]

Article Info

Keywords:
Term Weighting, TF-IDF, ICF, Term Distribution, Centroid, Text

Article history:
Received: July 13, 2023
Accepted: September11, 2023
Published: November 30, 2023

Cite:
C. Sri Kusuma Aditya and F. D. S. Sumadi, "Combination of Term Weighting with Class Distribution and Centroid-based Approach for Document Classification", KINETIK, vol. 8, no. 4`, Nov. 2023.
https://doi.org/10.22219/kinetik.v8i4.1793

*Corresponding author.
Christian Sri Kusuma Aditya
E-mail address:
christianskaditya@umm.ac.id

Abstract

A text retrieval system requires a method that is able to return a number of documents with high relevance upon user requests. One of the important stages in the text representation process is the weighting process. The use of Term Frequency (TF) considers the number of word occurrences in each document, while Inverse Document Frequency (IDF) considers the wide distribution of words throughout the document collection. However, the TF-IDF weighting cannot represent the distribution of words to documents with many classes or categories. The more unequal the distribution of words in each category, the more important the word features should be. This study developed a new term weighting method where weighting is carried out based on the frequency of occurrence of terms in each class which is integrated with the distribution of centroid-based terms which can minimize intra-cluster similarity and maximize inter-cluster variance. The ICF.TDCB term weighting method has been able to provide the best results in its application to SVM modeling with a dataset of 931 online news documents. The results show that SVM modeling had accuracy of 0.723, outperforming the use of other term weightings such as TF.IDF, ICF & TDCB.

1. Introduction

The need for information is pivotal and inevitable. Information in the form of news can be obtained not only from newspaper articles but also from online news articles. The popularity of Indonesian online news sites is currently increasing the volume of news available. Thus, a classification according to predetermined categories is necessary to make it easier for readers to choose the news they want to read. A survey conducted by UNESCO (United Nations Educational, Scientific, and Cultural Organization) in 2021, shows a pattern that continues to decline for news readers in conventional media, namely printed news, while on the contrary, since 2010, news searches have been through internet media, especially online news portals, continues to increase every year [1].

News grouping can be done in two ways, manually and automatically. Grouping news documents manually is very dependent on human ability and accuracy so that errors can occur in grouping these documents. Therefore, an automation in grouping news documents that have many similarities is needed so that the document search process becomes more optimal.

Several previous studies have carried out classification of text documents in classifying fake news. Trial comparisons are carried out by adding various selection feature techniques to get the model with the best results. Perfomance of Gaussian Naïve Bayesian improved significantly on best features selected by Chi-square as compared to other features selection techniques [2].

In another study aimed at classifying topic online news [3], various classification methods have been used such as Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF) and other methods, where in the text weighting phase, Term Frequency (TF) and Inverse Document Frequency (IDF) are employed. By using a real-world dataset, this research showed the significance of the hyperparameter tuning and its effect on the model's performance as it achieved 20.81% accuracy improvement for the SVM.

$TF.IDF$ weighting is a weighting that is frequently used in various kinds of Information Retrieval System development problems [4][5]. TF is used to measure the number of terms in a document. Meanwhile IDF is used to measure the informativeness of a term in a collection of documents [6]. $TF.IDF$ weighting based solely on the frequency of occurrence of terms in documents is not enough to determine the index of a document. Accurate index determination also depends on the informative value of the term for the class or cluster [7][8]. Terms that appear frequently in many classes or clusters should not be important terms even if they have a high $TF.IDF$ value. Research [9] adds Inverse Class Frequency (ICF) weighting to pay attention to the appearance of terms in a collection of categories or classes. The terms that appear very rarely are the most important terms. From these terms, documents can be grouped into topics according to these terms.

To optimize the term weighting process for online news documents where the collection already has a label, a centroid-based term distribution technique [10] was carried out to apply the concept of intra-cluster, inter-cluster and entire document weighting so as to increase the weight of discriminatory terms. This method is able to form a better weight representation or has a sense of each class [11].

Based on the description above, this study proposes a model for classifying new text documents by weighting based on the frequency of occurrence of terms in each class (ICF) and integrating them by minimizing intra-cluster variance or similarity and maximizing inter-cluster variance using centroid-based term weighting ($TDCB$). Thus, this modeling design is able to be more representative for a collection of news documents that have many classes and is able to increase the values of precision, recall and accuracy which are higher when compared to several other existing term weighting methods.

2. Research Method

Figure 1 explains in detail the main processes in this study, consisting of 4 stages, namely crawling stage, preprocessing stage, feature extraction stage and modeling stage, and prediction stage.

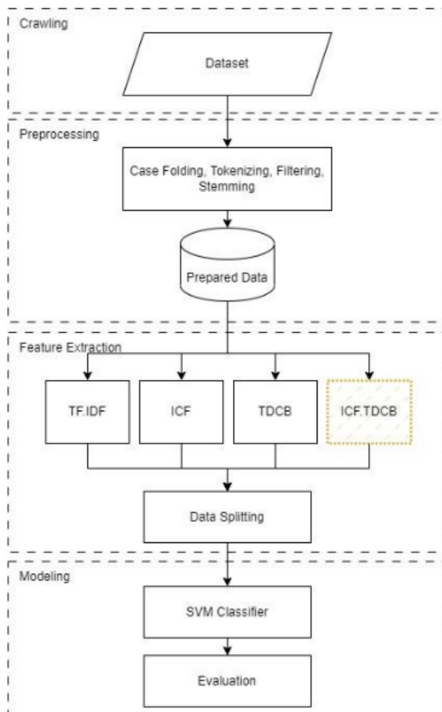

Figure 1. The Example of Figures that can be Seen Clearly

In the crawling process, or retrieval of news documents using Python's API, there are 4 categories, namely "Politics", "Sports", "Health", and "Technology". The total data for all news documents for training data is 931 with an average number of news for each category of 233. The process of obtaining news documents was started in January - April 2022 for all categories. More details about the number of per-category datasets used are shown in Table 1.

Table 1. An Example of Table Caption

Category	Amount of data
Politcs	252
Sports	224
Health	236
Technology	219
Number of News	**931**

The first stage after the dataset is obtained, the text document is preprocessed with the sequence of stages in the form of case folding, tokenization, filtering and stemming. The results of the preprocessing stage are a collection of words that have changed form to become basic words. The stemming algorithm used in this study is Nazief-Adriani [12][13]. An example of the results of text preprocessing of news documents for each category is shown in Figure 2.

Document	Category
['pemerintah', 'kota', 'malang', 'tengah', 'dorong', 'maju', 'sektor', 'ekonomi', 'salah', 'UMKM', 'langsung', 'maju', 'had	Politics
['lapor', 'kirim', 'data', 'aplikasi', 'tim', 'medis', 'pmi', 'kota', 'malang', 'datang', 'lokasi' ,'darurat', 'korban', 'operator', .	Politics
['ott, 'kpk', 'duga', 'tindak', 'pidana', 'korupsi', 'terima', 'perkara', 'pn', 'surabaya', 'aman', 'waktu', sikap', 'tangkap', ...]	Politics
['pss', 'sleman', 'joko', 'masalah', 'mental', 'eksekusi', 'bola', 'penting', 'evaluasi', 'kalah', 'lawan', 'persebaya', 'skor',	Sports
['getuk', 'pssi', 'regulasi', 'liga', 'degradasi', 'pola', 'kompetisi', 'normal', 'profesionalitas', 'mantan', 'latih', 'arema', 'ja	Sports
['dalam', 'sepak', 'bola', 'nomor', 'punggung', 'anggap', 'penting', 'banyak', 'klub', 'seluruh', 'dunia', 'putus', 'pensiun',	Sports
['bulan', 'puasa', 'umat', 'islam', 'syawal', 'konsumsi', 'kolestrol', 'tubuh', 'tingkat', 'risiko', 'jantung', 'stroke', 'kadar', '	Health
['kurma, 'pilih', 'rekomendasi', 'makan', 'sehat', 'lebaran', 'kandung', 'nutrisi', 'penting', 'tubuh', 'vitamin', 'protein', '	Health
['sun', 'damaged', 'skin', 'dapat', 'alam', 'orang', 'papar', 'sinar', 'uv', 'cara', 'terus', 'tahu', 'gejala', 'belum', 'lambat', 'a	Health
['akun', 'platform', 'intelijen', 'darkweb', 'ungkap', 'informasi', 'kait', 'identitas', 'hacker', 'bjorka', 'sebut', 'kelola', 'sit	Technology
['login', 'sekaligus', 'kata', 'sandi', 'jalur', 'bjorka', 'retas', 'situs', 'breakout', 'rooms', 'boleh', 'bilang', 'jadi', 'salah', 'sa	Technology
['aktif', 'tiap', 'kelompok', 'milik', 'akun', 'zoom', 'partisipasi', 'ruang', 'kerja', 'cara', 'gratis', 'masuk', 'profil', 'bagi', 'atu	Technology

Figure 2. The Example of Figures that can be Seen Clearly

After carrying out the preprocessing stage, the next step is to perform feature extraction from text documents by weighting the frequency of word occurrences.

TF (Term Frequency) is the simplest method of weighting terms. Each term is assumed to have a proportional importance to the number of occurrences of the term in the document. The following is the calculation of the weight of term t in document d in Equation 1, where $f(d,t)$ is the frequency of occurrence of term t in document d.

$$TF(d,t) = f(d,t) \tag{1}$$

If TF pays attention to the appearance of the term in the document, then IDF (Inverse Document Frequency) pays attention to the appearance of the term feature in the document set [14]. The background of this weighting is the term feature that rarely appears in the document set that is very valuable. The importance of each term is assumed to have the opposite proportion to the number of documents containing the term. Terms that frequently appear in one document but rarely appear in the entire dataset will be given a higher weight value (IDF indication) [15]. Equation 2 is the IDF factor of term t, where Nd is the total number of documents, $df(t)$ the number of documents containing term t.

$$IDF(t) = 1 + log(\frac{Nd}{df(t)}) \tag{2}$$

Multiplication between TF and IDF can produce better performance. The weight combination of term t in document d is explained in Equation 3, where $TF(d,t)$ is the frequency of the term, and $IDF(t)$ is the inverse of the occurrence of the term in the document.

$$TF.IDF(d,t) = TF(d,t) \times IDF(t) \tag{3}$$

ICF (Inverse Class Frequency) is a term weighting method that takes into class distribution. The term's frequency value has the opposite proportion to the number of classes that contain the term. In ICF, terms that are found in many classes cannot provide good differentiating values, which causes the function to give low values to terms that appear in many classes. Equation 4 is the ICF factor of term t, where Nc is the number of classes and $cf(t)$ is the number of documents containing term t [16].

$$IDF(t) = 1 + log(\frac{Nc}{cf(t)}) \tag{4}$$

The $TDCB$ (Term Distribution on Centroid Based) method uses the concept of term distribution based on intra-class, inter-class and the entire collection of documents to increase the weight of discriminatory terms. Each term has a weight according to the frequency of the document (intra-class information) and a discriminatory factor that is inversely proportional to the number of classes or clusters that contain the term (inter-class information) [17]. The distribution concept is based on the principle of minimizing intra-cluster variance or similarity and maximizing inter-cluster variance, so that data with similar characteristics are grouped in the same cluster and data with different characteristics are grouped into the other clusters. Object similarity is assessed based on the attribute value of an object [18].

Three types of information are defined as weighted representations for this method, including Equation 5 for $icsd$ (inter-class standard deviation), Equation 6 for csd (class standard deviation) and Equation 7 for sd (standard deviation), tf_{ijk} is the frequency of the document term ti dj in class C_k.

$$icsd_i = \frac{\Sigma_k \left[\bar{tf}_{ik} - \frac{\Sigma_k \bar{tf}_{ik}}{|c|}\right]^2}{|c|} \tag{5}$$

$$csd_{ik} = \sqrt{\frac{\Sigma_{d_j \epsilon C_k}\left[tf_{ijk} - \bar{tf}_{ik}\right]^2}{|C_k|}} \tag{6}$$

$$sd_i = \sqrt{\frac{\Sigma_k \Sigma_{d_j \epsilon C_k}\left[tf_{ijk} - \frac{\Sigma_k \Sigma_{d_j \epsilon C_k} tf_{ijk}}{\Sigma_k |C_k|}\right]^2}{\Sigma_k |C_k|}} \tag{7}$$

Where,

$$\bar{tf}_{ik} = \frac{\Sigma_{d_j \epsilon C_k} tf_{ijk}}{|C_k|} \tag{8}$$

Equation 8 for $\bar{tf}_{ik}$ is the average term frequency in all documents belonging to the C_k class category. $|c|$ is the number of classes and $|C_k|$ is the number of documents in class category C_k.

Each of the $icsd$, csd, and sd values is used to calculate the TDF weight value in Equation 9, and combine it with the $TF.IDF$ as in Equation 10 [19].

$$\mathrm{TDF}_{ik} = \mathrm{icsd}_i^{\alpha} \; x \; \mathrm{csd}_{ik}^{\beta} \; x \; \mathrm{sd}_i^{\gamma} \tag{9}$$

$$w_{ik} = \mathrm{tf}_{ik} \; x \; \mathrm{idf_i} \; x \; \mathrm{TDF}_{ik} \tag{10}$$

The parameters α, β, and γ are the weights of each information ($icsd$, csd, sd) where the weights with positive numbers play a role in increasing the information value, and conversely the weights with negative numbers decrease the information value. The greater the value of the parameters given, the greater the contribution for weighting either to increase (promoter) or decrease (demoter) the value of the information [20].

A term with a high $icsd$ value represents that term has discriminatory power against each class. The $icsd$ information allows a term to exist in almost all classes but the frequency for each class is different. In this case, the difference between the $TF.IDF$ weighting method provides less information.

The csd value is term information that appears throughout the document but is limited to the same class, therefore the csd for each term varies within each class. A term with a high csd value represents a high frequency number in all documents in one class, but inversely with the number of documents. There are two factors that make the csd score low, including the appearance of terms that are almost the same in all documents in one class or these terms rarely appear in that class [21].

A term with a high sd value is almost the same as the form of representation of csd. The difference is where a term has a high frequency number in the entire document collection, not each class, but also inversely proportional to the number of documents.

The terms weighting in this study used a new approach by integrating the frequency of occurrence of words in each class and the distribution of centroids and as explained in the previous section. The $TF.IDF$ weighting method only considers the distribution of words in the document as a whole without considering documents in a particular class. To optimize the term weighting process for online news documents that have many categories, the ICF method is added by calculating the occurrence of word features in a collection of categories or classes. The word features that appear very rarely in all existing categories are the most important word features.

$$w_{ik} = \mathrm{tf}_{ik}\ x\ \mathrm{idf}_{\mathrm{i}}\ x\ \mathrm{icf}_{\mathrm{i}}\ x\ \mathrm{TDF}_{ik} \tag{11}$$

The concept of word feature distribution for each class is continued by minimizing intra-cluster variance or similarity and maximizing inter-cluster variance, so that data with similar characteristics are grouped in the same cluster and data with different characteristics are grouped into the other clusters. Overall, the proposed weighting model can be seen in Equation 11.

3. Results and Discussion

The trial scenario is the stage aimed to test the readiness of the system. The first trial scenario is to compare the results of the term weighting values obtained where the resulting data from the term index process can be seen in Table 2.

Table 2. Term Weighting Representation

Term	Doc_Id	TF	TF.IDF	ICF	TDCB	ICF.TDCB
lapor	1	4	0.431	0.664	1.211	0.804104
	20	3	0.317	0.548	1.273	0.697604
	23	2	0.223	0.332	1.677	0.556764
	25	1	0.135	0.316	0.673	0.212668
aktif	4	4	0.721	0.528	1.444	0.762432
	12	2	0.605	0.564	1.313	0.740532
	15	3	0.412	0.346	1.673	0.578858
	32	1	0.231	0.482	0.781	0.376442
makan	9	1	0.259	0.453	0.834	0.377802
	12	1	0.259	0.453	0.732	0.331596
	14	4	1	1.112	1.451	1.613512
	19	2	0.508	0.606	1.334	0.808404
darurat	21	5	0.769	0.79	1.294	1.02226
	25	2	0.203	0.276	1.448	0.399648
	28	1	0.121	0.338	1.742	0.588796
	67	1	0.121	0.238	1.563	0.371994
kelompok	56	3	0.892	0.713	1.063	0.757919
	64	2	0.435	0.642	1.142	0.733164
	67	2	0.435	0.842	1.022	0.860524
	89	3	0.892	0.913	1.008	0.920304

The second trial scenario is a comparison of the average precision, recall, and accuracy results on the composition of the comparison between the amount of training data and testing data for each term weighting representation of the Support Vector Machine (SVM) classifier [22][23]. The test scenario is carried out to find out the best distribution ratio of train data and test data by using 5 ratios, consisting of 50:50, 60:40, 70:30, 80:20, 90:10 in the form of training data: data testing. After getting the best data sharing ratio, the initial stages of the data were weighted using the $TF.IDF$, ICF, $TDCB$ methods and the weighting method proposed in this study, $ICF.TDCB$. Furthermore, the classification process for test data and train data uses the SVM method using linear kernel parameters [24], constant C = 1 and degree d = 1. The results of the comparison of each term weighting to SVM can be seen in Table 3, Table 4, Table 5 and Table 6.

Table 3. TF.IDF Results

Composition	Precision	*Recall*	*Accuracy*
50:50	0.521	0.531	0.592
60:40	0.563	0.592	0.593
70:30	0.586	0.602	0.612
80:20	0.621	0.613	0.622
90:10	0.645	0.641	0.632

Table 4. ICF Results

Composition	Precision	*Recall*	*Accuracy*
50:50	0.569	0.551	0.596
60:40	0.572	0.596	0.598
70:30	0.589	0.616	0.619
80:20	0.633	0.643	0.654
90:10	0.685	0.678	0.662

Table 5. TDCB results

Composition	Precision	*Recall*	*Accuracy*
50:50	0.563	0.547	0.583
60:40	0.592	0.581	0.593
70:30	0.620	0.631	0.623
80:20	0.635	0.639	0.661
90:10	0.697	0.682	0.693

Table 6. ICF.TDCB Results

Composition	Precision	*Recall*	*Accuracy*
50:50	0.565	0.551	0.582
60:40	0.602	0.586	0.603
70:30	0.650	0.651	0.643
80:20	0.653	0.663	0.692
90:10	0.725	0.719	0.715

Based on the test results, the value of precision, recall and accuracy for $ICF.TDCB$ is better than the $TF.IDF$, ICF and $TDCB$ methods. ICF has advantages compared to $TF.IDF$ as it can pay attention to the appearance of terms in a collection of categories/classes. Terms that rarely appear in many classes are valuable terms for classification. The importance of each term is assumed to have the opposite proportion to the number of classes containing the term. For $TDCB$, the precision, recall and accuracy values are not too significantly different or almost the same when compared to ICF. The concept of distribution in $TDCB$ minimizes intra-cluster variance and maximizes inter-cluster variance, so that it has discriminatory power against each class. A term that has a high frequency number in the entire document collection, not each class, but also inversely proportional to the number of documents, this also provides added value information to the $TDCB$ term weighting.

Meanwhile, for modeling using the term weighting $ICF.TDCB$, it is sufficient to provide the difference in the addition of precision, recall and accuracy values in each composition, especially in the composition of the dataset division 70 : 30, 80 : 20, 90 : 10. If you look at the term weighting values, it looks like as shown in Table 3. ICF lowers several values from the term weighting generated by $TDCB$, where $TDCB$ takes into account information on the appearance of terms that are almost the same in all documents in one class or the term rarely appears in that class.

The third trial scenario is a comparison of the average precision, recall, and accuracy results, using a composition of 90:10 which gets the best results in trial scenario 1, for each representation of the term weighting in the SVM classifier using a comparison of the linear kernel and RBF (Radial Basis Function) [25], parameter values constant C and degree (d).

Table 7. Comparison of Linear Kernel Accuracy

	d							
	1				2			
C	**TF.IDF**	**ICF**	**TDCB**	**ICF.TDCB**	**TF.IDF**	**ICF**	**TDCB**	**ICF.TDCB**
0.1	0.632	0.662	0.693	0.715	0.632	0.656	0.690	0.705
1	0.632	0.662	0.693	0.715	0.632	0.656	0.689	0.704
10	0.629	0.661	0.692	0.713	0.629	0.661	0.692	0.713

Table 8. Comparison of RBF Kernel Accuracy

	d							
	1				2			
C	**TF.IDF**	**ICF**	**TDCB**	**ICF.TDCB**	**TF.IDF**	**ICF**	**TDCB**	**ICF.TDCB**
0.1	0.636	0.672	0.698	0.715	0.632	0.662	0.693	0.703

1	0.636	0.671	0.698	0.723	0.632	0.662	0.693	0.711
10	0.629	0.661	0.692	0.713	0.629	0.661	0.692	0.713

From the comparison between Table 7 and Table, 8 it can be seen that the use of the RBF kernel is slightly better when compared to the linear kernel. While the comparison of the good values in the linear kernel and RBF both show the same results where degree = 1 shows better results. And for the value of the constant C, it does not show too different results or it can be said that this research has no effect. The graph of the test results comparing the accuracy value with degree = 1 can be seen in Figure 3, and degree = 2 in Figure 4.

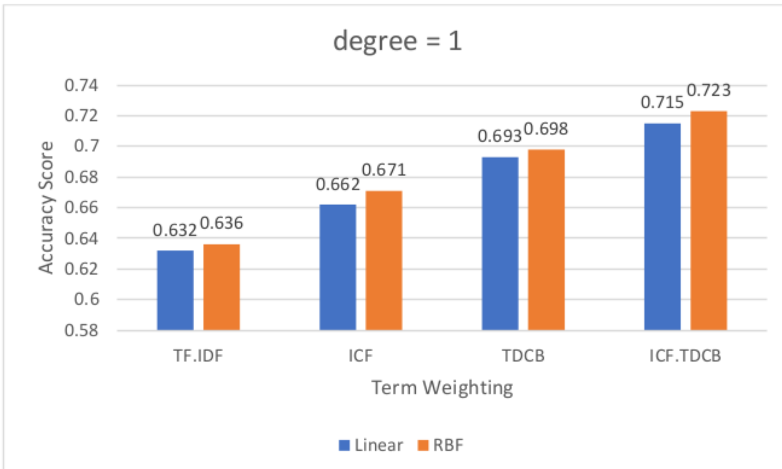

Figure 3. Graph of Comparison of Accuracy Degree = 1

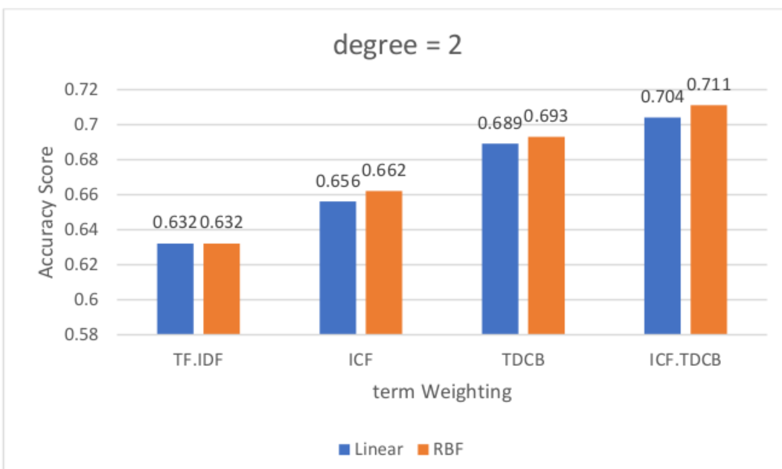

Figure 4. Graph of Comparison of Accuracy Degree = 2

The test results as shown in Figure 3 and Figure 4 show that the use of the $ICF.TDCB$ term weighting with a degree = 1 value is able to show better evaluation results when compared to other weighting terms with an accuracy value of 0.715 when using linear SVM kernel modeling and an accuracy value of 0.723 when using RBF kernels. On average, the RBF kernel has a better accuracy value for the overall term weighting used when compared to the linear kernel, due to its ability to separate data nonlinearly where text datasets tend to have high dimensions.

4. Conclusion

The conclusion of this study is that the term weighting method $ICF.TDCB$ has been able to provide the best results in its application to SVM modeling with a dataset of 931 online news documents. The results obtained in SVM modeling had accuracy of 0.723, outperforming the use of other term weightings such as $TF.IDF$, ICF & $TDCB$. Suggestions for further research are to conduct research on comparisons with more term weighting, deep learning modeling, and the application of feature selection in reducing the dimensions of datasets that are quite high.

Acknowledgement

This manuscript is based on research supported by the Universitas Muhammadiyah Malang under the Grant Number E.2.a/238/BAA-UMM/III/2020. The authors would also express their gratitude for the UMM Informatics Laboratory, who have supported the implementation of this research.

References

[1] https://news.un.org/en/story/2022/03/1113702
[2] Fayaz, M., Khan, A., Bilal, M., & Khan, S. U. (2022). Machine learning for fake news classification with optimal feature selection. Soft Computing, 26(16), 7763-7771. https://doi.org/10.1007/s00500-022-06773-x
[3] Daud, S., Ullah, M., Rehman, A., Saba, T., Damaševičius, R., & Sattar, A. (2023). Topic classification of online news articles using optimized machine learning models. Computers, 12(1), 16. https://doi.org/10.3390/computers12010016
[4] Alodadi, Mohammad, and Vandana P. Janeja. "Similarity in patient support forums using tf-idf and cosine similarity metrics." 2015 International Conference on Healthcare Informatics. IEEE, 2015. https://doi.org/10.1109/ICHI.2015.99
[5] Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25-29. https://doi.org/10.5120/ijca2018917395
[6] Guo, Aizhang, and Tao Yang. "Research and improvement of feature words weight based on TFIDF algorithm." 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference. IEEE, 2016. https://doi.org/10.1109/ITNEC.2016.7560393
[7] Uysal, Alper Kursat. "An improved global feature selection scheme for text classification." Expert systems with Applications 43 (2016): 82-92. https://doi.org/10.1016/j.eswa.2015.08.050
[8] Domeniconi, Giacomo, et al. "A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf. idf." International Conference on Data Management Technologies and Applications. Springer, Cham, 2015. https://doi.org/10.1007/978-3-319-30162-4_4
[9] Puspaningrum, Alifia, Daniel Siahaan, and Chastine Fatichah. "Mobile App Review Labeling Using LDA Similarity and Term Frequency-Inverse Cluster Frequency (TF-ICF)." 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE). IEEE, 2018. https://doi.org/10.1109/ICITEED.2018.8534785
[10] Lertnattee, Verayuth, and Thanaruk Theeramunkong. "Effect of term distributions on centroid-based text categorization." Information Sciences 158 (2004): 89-115. https://doi.org/10.1016/j.ins.2003.07.007
[11] Nguyen, T. T., Chang, K., & Hui, S. C. (2013). Supervised term weighting centroid-based classifiers for text categorization. Knowledge and information systems, 35, 61-85. https://doi.org/10.1007/s10115-012-0559-9
[12] Slamet, Cepi, et al. "Automated text summarization for indonesian article using vector space model." IOP Conference Series: Materials Science and Engineering. Vol. 288. No. 1. IOP Publishing, 2018. https://doi.org/10.1088/1757-899X/288/1/012037
[13] Wahyudi, Dwi, Teguh Susyanto, and Didik Nugroho. "Implementasi Dan Analisis Algoritma Stemming Nazief & Adriani Dan Porter Pada Dokumen Berbahasa Indonesia." Jurnal Ilmiah SINUS 15.2 (2017): 49-56. http://dx.doi.org/10.30646/sinus.v15i2.305
[14] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794
[15] Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. Human-centric Computing and Information Sciences, 9, 1-21. https://doi.org/10.1186/s13673-019-0192-7
[16] Takçı, H., & Güngör, T. (2012). A high performance centroid-based classification approach for language identification. Pattern Recognition Letters, 33(16), 2077-2084. https://doi.org/10.1016/j.patrec.2012.06.012
[17] Lertnattee, V., & Theeramunkong, T. (2004, October). Analysis of inverse class frequency in centroid-based text classification. In IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004. (Vol. 2, pp. 1171-1176). IEEE. https://doi.org/10.1109/ISCIT.2004.1413903
[18] Cieza, A., Fayed, N., Bickenbach, J., & Prodinger, B. (2019). Refinements of the ICF Linking Rules to strengthen their potential for establishing comparability of health information. Disability and rehabilitation, 41(5), 574-583. https://doi.org/10.3109/09638288.2016.1145258
[19] Lertnattee, V., & Theeramunkong, T. (2004). Effect of term distributions on centroid-based text categorization. Information Sciences, 158, 89-115. https://doi.org/10.1016/j.ins.2003.07.007
[20] Liu, C., Wang, W., Tu, G., Xiang, Y., Wang, S., & Lv, F. (2017). A new Centroid-Based Classification model for text categorization. Knowledge-Based Systems, 136, 15-26. https://doi.org/10.1016/j.knosys.2017.08.020
[21] Guan, H., Zhou, J., & Guo, M. (2009, April). A class-feature-centroid classifier for text categorization. In Proceedings of the 18th international conference on World wide web (pp. 201-210). https://doi.org/10.1145/1526709.1526737
[22] Huang, W., Liu, H., Zhang, Y., Mi, R., Tong, C., Xiao, W., & Shuai, B. (2021). Railway dangerous goods transportation system risk identification: Comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM. Applied Soft Computing, 109, 107541. https://doi.org/10.1016/j.asoc.2021.107541
[23] Dai, T. T., & Dong, Y. S. (2020, April). Introduction of SVM related theory and its application research. In 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) (pp. 230-233). IEEE. https://doi.org/10.1109/AEMCSE50948.2020.00056
[24] Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. Artificial Intelligence Review, 52(2), 803-855. https://doi.org/10.1007/s10462-018-9614-6
[25] Ring, M., & Eskofier, B. M. (2016). An approximation of the Gaussian RBF kernel for efficient classification with SVMs. Pattern Recognition Letters, 84, 107-113. https://doi.org/10.1016/j.patrec.2016.08.013

-_Term_Weighting_TF-IDF_ICF_Term_Distribution_Centroid_Text.pdf

ORIGINALITY REPORT

18%	14%	15%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to CSU, Chico Student Paper	4%
2	ijai.iaescore.com Internet Source	1%
3	d-nb.info Internet Source	1%
4	Lecture Notes in Computer Science, 2010. Publication	1%
5	ojs3.unpatti.ac.id Internet Source	1%
6	Rajendra Kumar Roul, Jajati Keshari Sahoo, Kushagr Arora. "Modified TF-IDF Term Weighting Strategies for Text Categorization", 2017 14th IEEE India Council International Conference (INDICON), 2017 Publication	1%
7	link.springer.com Internet Source	1%

8	www.mdpi.com Internet Source	1 %
9	vdoc.pub Internet Source	1 %
10	download.garuda.kemdikbud.go.id Internet Source	1 %
11	www.sciencegate.app Internet Source	1 %
12	ospro.usk.ac.id Internet Source	1 %
13	mafiadoc.com Internet Source	<1 %
14	www.researchsquare.com Internet Source	<1 %
15	archive.org Internet Source	<1 %
16	weblab.ing.unimore.it Internet Source	<1 %
17	repository.futminna.edu.ng:8080 Internet Source	<1 %
18	Dandan WANG, Qingcai CHEN, Xiaolong WANG. "A Framework of Centroid-Based Methods for Text Categorization", IEICE	<1 %

Transactions on Information and Systems, 2014
Publication

19 garuda.kemdikbud.go.id <1 %
Internet Source

20 "Database Systems for Advanced Applications", Springer Nature, 2017 <1 %
Publication

21 beei.org <1 %
Internet Source

22 Lertnattee, V.. "Effect of term distributions on centroid-based text categorization", Information Sciences, 200401 <1 %
Publication

23 www.coursehero.com <1 %
Internet Source

24 Alifia Puspaningrum, Daniel Siahaan, Chastine Fatichah. "Mobile App Review Labeling Using LDA Similarity and Term Frequency-Inverse Cluster Frequency (TF-ICF)", 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), 2018 <1 %
Publication

25 N. Venkata Sailaja, L. Padmasree, N. Mangathayaru. "Incremental learning for text categorization using rough set boundary <1 %

	based optimized Support Vector Neural Network", Data Technologies and Applications, 2020 Publication	
26	citeseerx.ist.psu.edu Internet Source	<1%
27	Chuan Liu, Wenyong Wang, Guanghui Tu, Yu Xiang, Siyang Wang, Fengmao Lv. "A new Centroid-Based Classification model for text categorization", Knowledge-Based Systems, 2017 Publication	<1%
28	docshare.tips Internet Source	<1%
29	ijcsi.org Internet Source	<1%
30	mimoza.marmara.edu.tr Internet Source	<1%
31	ses.library.usyd.edu.au Internet Source	<1%
32	www.frontiersin.org Internet Source	<1%
33	Anisa Dewi Anggraeni, Muhammad Farhansyah, Muhammad Risky Pratama Hermawan, Galih Wasis Wicaksono, Christian Sri Kusuma Aditya. "Comparison of	<1%

Classification Methods on Twitter Sentiment Analysis of PDAM Tugu Tirta Kota Malang", JUITA : Jurnal Informatika, 2023
Publication

34	Ba-Dung Le, Guanhua Wang, Mehwish Nasim, Muhammad Ali Babar. "Gathering Cyber Threat Intelligence from Twitter Using Novelty Classification", 2019 International Conference on Cyberworlds (CW), 2019 Publication	<1 %
35	Shahzada Daud, Muti Ullah, Amjad Rehman, Tanzila Saba, Robertas Damaševičius, Abdul Sattar. "Topic Classification of Online News Articles Using Optimized Machine Learning Models", Computers, 2023 Publication	<1 %
36	Willy, Erwin B. Setiawan, Fida N. Nugraha. "Implementation of Decision Tree C4.5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on Social Media Twitter", 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2019 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		