Text Processing Using Support Vector Machine for Scientific Research Paper Content Classification

Hasanuddin Al-Habib[1(✉)], Elly Matul Imah[1], Riskyana Dewi Intan Puspitasari[1], and Binar Kurnia Prahani[2]

[1] Data Science Department, Universitas Negeri Surabaya, Surabaya, Indonesia
{hasanuddinhabib,ellymatul,riskyanapuspitasari}@unesa.ac.id
[2] Physics Department, Universitas Negeri Surabaya, Surabaya, Indonesia
binarprahani@unesa.ac.id

Abstract. Research related to text processing is carried out to analyze information in text data which can then be used in strategies for the development of the field of science and technology, one of which is the text processing of scientific research paper. Classification of scientific research paper is carried out to obtain information related to the field of research based on the substance of the scientific articles, then to map them into categories to identify areas of excellence and to determine development strategies related to publications and research. In this research, text processing of scientific research paper is carried out by implementing the supervised machine learning model, Support vector machine (SVM), for classification algorithms which divide it's into several groups based on the substance of scientific research paper in each field or subject area. Feature extraction on the substance of scientific research paper becomes input for modeling with SVM. Furthermore, the concept of kernel functions on the support vector machine forms the classifier model and classifies scientific research paper text data which results in high accuracy.

Keywords: Text processing · scientific research paper · support vector machine · machine learning · classification · data text

1 Introduction

Research related to text processing is an important subject in the field of natural language processing and is developing along with the use of text-based digital technologies on various media platforms. Research of text processing is substantial because the information obtained can be used in science and technology development strategies, one of which is text processing in the field of scientific papers. Text classification is a text mining problem that usually deals with the content-based assignment of labels to documents, from single label classification problems (one text to one label) to multi-label cases (one text to multiple labels) [1]. Analysis related to this labeling result can be used as a source of information and scientific research may result in a new discovery, which may have

I. H. Agustin (Ed.): ICONNSMAL 2022, AISR 177, pp. 273–282, 2023.
https://doi.org/10.2991/978-94-6463-174-6_20

a significant impact on the development of science, technology, and innovation, and even be considered as 'breakthrough'[2]. Component of scientific research paper can be reviewed as the input of classification process. Subsections in scientific research papers are unstructured data that must be modelled. This data modeling can be done in various ways, it can be present in a recent survey of 12 machine learning text classifiers applied to a corpus of public spam [3].

Furthermore, scientific research paper is imbalanced data class. It refers to a disproportion in the number of examples belonging to each class of a dataset and is known to bias classifiers towards the most represented concepts. This situation is especially critical when minority class concepts are associated with a higher misclassification cost, such as the diagnosis of rare diseases. Although this is an important problem in isolation, its combination with other factors creates a much more difficult setting for classifiers [4]. Another imbalanced classification was done by using random forest method for modelling the text data [5], neural networks learning algorithm [6, 7], under-sampling with support vector for multi-class classification [8], and learning SVM with weighted maximum margin criterion [9]. Classification for imbalanced data still challenging, because currently the amount of data is also growing very rapidly and it is possible that the data is also imbalanced. Especially scientific article paper data and it is very possible that the challenge of classifying scientific article paper as text data is also very necessary.

Research related to text data classification in scientific articles was carried out using the BERT model for multi-classification classification. It frames the problem as text classification and conducts experiments to compare ensemble architectures, where the selection criteria are mapped to the components of the ensemble. Uncased Sci-BERT, which is a caseless BERT model pre-trained on a collection of scientific articles, as the core model in our study. This model was pre-trained on a random sample of 1.14 M papers from Semantic Scholar [10]. In addition, text classification research that utilizes the TF-IDF model has been carried out by modifying and adding a sliding window so as to obtain good accuracy [11]. Another research related to text classification by utilizing machine learning on English text data. In this study, the support vector machine method was implemented as a classifier model builder in 1033 text documents and resulted in a good accuracy [12]. Support Vector Machine (SVM) classifier also used in research about predict publication years based on the latent topic distribution.

Based on the research above, machine learning is a good method in the text classification process and provides high accuracy. So, to improve the accuracy obtained in previous studies, in this research, we apply the TF-IDF method as a feature extraction method and the Support Vector Machine method as a classifier. The composition of this paper includes a literature review, pre-processing method, features extraction, classification process in Sect. 2, and explanations of the experiments that have been carried out will be explained in Sect. 3. Then Sect. 4 discusses the conclusions.

2 Method

In this section we will discuss about the research flow to classify scientific research paper content based on Author Keywords. All stages are shown in Fig. 1, where the 'Author Keywords' as the input data was pre-processed using several steps to get the

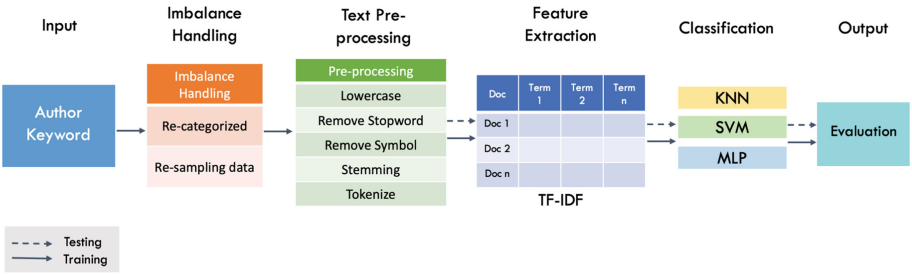


Fig. 1. Scientific research paper content classification research flow

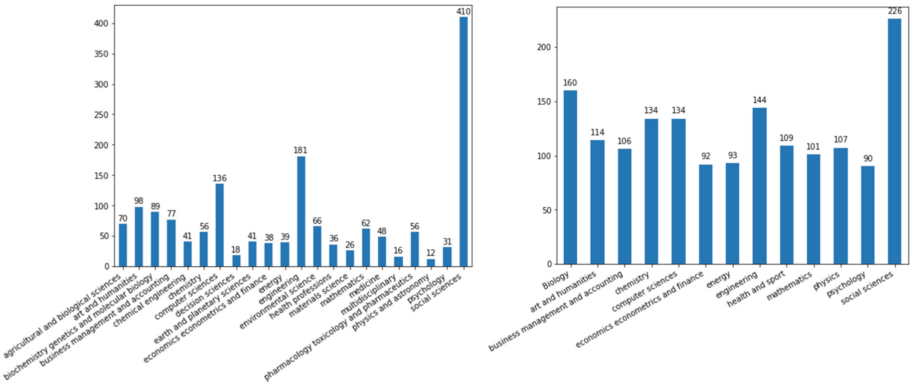


Fig. 2. (a) Data distribution with 22 categories (b) Data Distribution after re-categorization and undersampling with 13 categories

clean and consistent format of text. Feature extraction using TF-IDF was performed to extract essential information of data, then several machine learning algorithms was used to classify the author keyword into suitable categories. The performance of machine to classify the data are evaluated using metric such as accuracy, recall, precision, and F1-score.

A. Dataset

The data input used in this study was taken from the summary of the publication of the Surabaya State University journal on Scopus, which consists of several research scopes. The total number of research publications is *1647* journals, published from 2003 to 2022. Classification is done based on the 'Author Keywords' column, which is categorized into *22* categories shown in Fig. 2a. Based on the data distribution in Fig. 2a, it can be seen that the data is an imbalance, where the number of data for the Social Sciences category is much higher than the other categories, with 410 instances. In comparison, the average number for the other categories is 59 instances. In Fig. 2a, it shows that the category with the greatest number of data is Social Sciences. In contrast, the category with the lowest data is Physics and Astronomy, with only 12 instances.

Imbalance data can affect classification performance, so the machine will tend to learn in the class with the most data and ignore the minority class. We overcome data

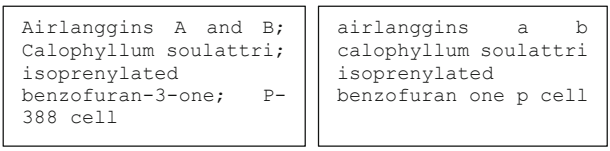


Fig. 3. (a) Author Keyword **before** pre-processing (b) Author Keyword **after** pre-processing

imbalance using re-categorization by regrouping similar categories. Re-categorization is done by regrouping several similar categories into relevant categories. From this re-categorization process, 13 categories were obtained, such as Biology, Art and Humanities, Business Management and Accounting, Chemistry, Computer sciences, Physics, Economics Econometrics and Finance, Energy, Engineering, Health and Sport, Mathematics, Psychology, and Social Sciences. After that, we resampled the data by removing data in the majority class (*under sampling*) and increasing the minority class (*over-sampling*) with a bias of 0.5. The result of data imbalanced handling is shown in Fig. 2b.

B. Pre-processing

Pre-processing of text data must be done to enhance classification performance by removing less meaningful information and changing some words into more proper terms, so it is ready to be processed by machines. In this study, the author's keywords string data was lowercase to synchronize the meaning of a word on the computer. We also removed stop words to get nouns and eliminated less important terms. In addition, we used to stem to get essential words and tokenized to split documents into several sets of meaningful pieces of words (Fig. 3).

C. Term Frequency-Inverse Document Frequency (TF-IDF)

A weighting method was proposed by George and Vinod [13, 14] during the feature extraction process, called term frequency-inverse document frequency (TF-IDF). The method was applied to reduce the dimensionality of the feature space and the noise contained within the dataset, therefore drastically improving the outcome of the classifier. This TF-IDF algorithm is a comprehensive assessment of the importance of a word for a text or class of texts. The TF (word frequency) is the frequency of occurrence of the word in that article or class of documents, which intuitively indicates the importance of the word for that article or class of documents; the IDF (inverse document frequency) characterizes the ability of the word to differentiate against the text classification [11].

The formula for the TF-IDF algorithm is shown in (1) - (3), where TF_{ij} denotes the TF value of the $i-th$ word for the $j-th$ document, IDFi denotes the IDF value of the $i-th$ word, and $TF-IDF_{ij}$ denotes the TF-IDF value of the $i-th$ word for the j-th document. D is the total number of all documents, D_{ij} is the $j-th$ document, nij is the number of occurrences of the $i-th$ word in the $j-th$ document, and Ni is the number of occurrences of the $i-th$ word in all documents.

$$TF_{ij} = n_{ij}/D_j \tag{1}$$

$$IDF_i = \lg\left(\frac{D}{N_i}\right) \tag{2}$$

$$TF - IDF_{ij} = TF_{ij} \bullet IDF_i \tag{3}$$

However, although the traditional TF-IDF algorithm can unsupervised find some keywords that have special significance and outstanding contribution to text classification, it also has certain shortcomings [11].

D. Support Vector Machine (SVM)

SVM is a useful model for pattern classification. The method is based on a statistical theory proposed by [15], and it can be applied to both linearly separable features and non-linearly separable features. It can classify linear data with the help of Maximum Margin Hyperplane (MMH), in which the distance is maximum between data points called support vectors. The two parallel lines separating the data are called positive and negative hyperplanes, as we can draw several [16]. In contrast, the negative hyperplane is drawn on the negative data point side It is better to draw the hyperplanes in such a way to get the maximum margin between positive and negative hyperplane [17]. For non-linear data, the kernel function can be used to form the multi-dimensional hyperplane for classification. There are multiple numbers of kernel functions available for classification purposes and it classify text data with high accuracy [18] There are different types of kernel functions available that we can use for classification purposes. This method has to find out the proper kernel function to classify data points appropriately. When the kernel function is used to classify the data points, it will transform one class of data to a higher dimension, and the decision surface is obtained to classify the data points [18].

Given the training set $(x_n, y_n), n = 1, \ldots, N$, where x_n is a vector containing the features associated with each instance n, and y_n is the class label for each instance n, the SVM classifier defines the "maximum-margin hyperplane" separating the classes. The hyperplane is defined so that the distance between the hyperplane and the nearest point x_n from either group is maximized. In classification problems, specifically regarding spam classification, classes are set as labels $+1$ and -1 retrospectively to spam and ham outcomes [15].

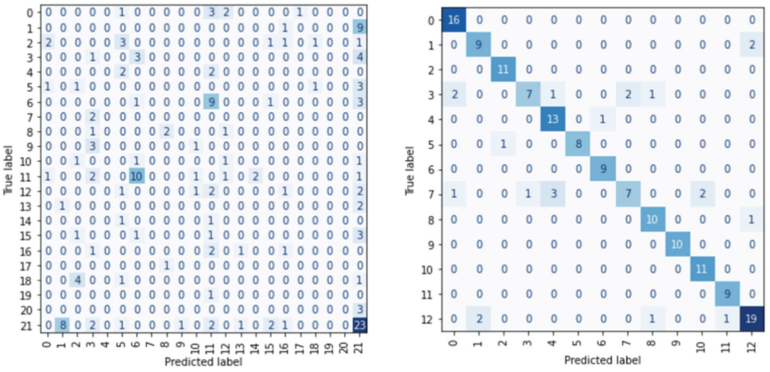
3 Result and Discussion

A. Result

In this section we will discuss about experimental results of classifying scientific research paper content based on Author Keywords using TF-IDF which are then classified using several machine learning algorithms. The classification results are then evaluated using accuracy, recall, precision, and F1-score metrics. In the previous section, it was discussed that the initial data to be used is imbalance data which has 22 categories where several categories have a high gap in the number of data with other categories. We compared the experimental results in the first experiment before and after the imbalanced handling was performed. The results of this comparison are shown in Table 1. In Table 1, it can be seen that the performance for each metric increased significantly after re-categorized and resampling the data. We also compared the confusion matrix results for predictions on test data using SVM on data before and after imbalanced handling, shown in Fig. 4. In Fig. 4, it can be seen that recognition for each category is very low in data before imbalance handling. After the imbalance handling was performed, the

Table 1. Classification results Before and After Imbalanced Handling

Imbalance Handling	Classifier	Precision	Recall	F1-Score	Accuracy
Before Imbalance Handling	KNN	0,078	0,064	0,067	0,170
	SVM	**0,065**	**0,066**	**0,064**	**0,170**
	MLP	0,024	0,063	0,030	0,206
After Imbalance Handling (Recategorization + Resampling)	KNN	0,869	0,863	0,865	0,855
	SVM	**0,869**	**0,863**	**0,865**	**0,862**
	MLP	0,870	0,862	0,865	0,819



(a) (b)

Fig. 4. (a) Confusion Matrix **before** imbalanced handling (b) Confusion Matrix **after** imbalanced handling

recognition of each category was quite good, as seen from the number of values on the diagonal matrix.

Table 1 compares classifier performance in classifying scientific paper content based on author keywords using the TF-IDF feature. The machine learning algorithms being compared are K-Nearest Neighborhood (KNN), SVM, and Multi-Layer Perceptron (MLP). Table 1 shows that SVM obtained better performance results than the other two classifiers, with an accuracy value of 0.862. In contrast, KNN only obtained an accuracy result of 0.855 and MLP with an accuracy of 0.815 (Table 2).

B. Discussion

In this paper, we classify scientific research paper content based on Author Keywords using the TF-IDF extraction feature, which is then classified using several machine learning algorithms. From the Tables and Figures in the Results section, it is found that proper data processing will improve classification performance significantly. Data processing in this context is not only at the text pre-processing stage but also understanding the underlying data distribution and looking for appropriate handling methods. From the

Table 2. Classification results use the TF- IDF feature

Classifier	Class	Precision	Recall	F1-Score	Accuracy
KNN	Biology	0.858	0.831	0.844	0.855
	Art and humanities	0.862	0.930	0.895	
	Business management and accounting	0.913	0.896	0.905	
	Chemistry	0.823	0.761	0.791	
	Computer sciences	0.812	0.806	0.809	
	Physics	0.914	0.897	0.906	
	Economics econometrics and finance	0.935	0.935	0.935	
	Energy	0.833	0.914	0.872	
	Engineering	0.802	0.674	0.732	
	Health and sport	0.869	0.853	0.861	
	Mathematics	0.930	0.921	0.925	
	Psychology	0.965	0.922	0.943	
	Social sciences	0.775	0.885	0.826	
SVM	Biology	0.858	0.831	0.844	**0.862**
	Art and humanities	0.881	0.912	0.897	
	Business management and accounting	0.912	0.877	0.894	
	Chemistry	0.840	0.746	0.791	
	Computer sciences	0.871	0.806	0.837	
	Physics	0.941	0.897	0.919	
	Economics econometrics and finance	0.935	0.935	0.935	
	Energy	0.850	0.914	0.881	
	Engineering	0.804	0.771	0.787	
	Health and sport	0.869	0.853	0.861	
	Mathematics	0.931	0.941	0.936	
	Psychology	0.964	0.900	0.931	
	Social sciences	0.760	0.898	0.824	
Backpropagation	Biology	0.858	0.831	0.844	0.819

(*continued*)

Table 2. (*continued*)

Classifier	Class	Precision	Recall	F1-Score	Accuracy
	Art and humanities	0.862	0.930	0.895	
	Business management and accounting	0.913	0.896	0.905	
	Chemistry	0.823	0.761	0.791	
	Computer sciences	0.812	0.806	0.809	
	Physics	0.914	0.897	0.906	
	Economics econometrics and finance	0.935	0.935	0.935	
	Energy	0.833	0.914	0.872	
	Engineering	0.802	0.674	0.732	
	Health and sport	0.869	0.853	0.861	
	Mathematics	0.930	0.921	0.925	
	Psychology	0.965	0.922	0.943	
	Social sciences	0.775	0.885	0.826	

initial data distribution in Fig. 1 (a), it can be seen that the data is imbalanced with a relatively high number of gaps for several data categories. The problem of data imbalance is quite problematic because the uneven distribution of data will decrease the machine's performance to acknowledge data patterns optimally. The machine will tend to focus on the majority data and find it challenging to recognize the minority data properly. In this study, we re-categorized data on data with similar categories and resampled the data. Resampling is done by increasing the number of minority data (*oversampling*) and removing some of the majority data (*undersampling*) with a bias of 0.5. After the imbalanced handling is performed, the classification performance increases significantly. It is proof, as seen from the value of each metric and each data category, as evidenced in Table 2.

4 Conclusion

Imbalance data is a kind of machine learning problems. Imbalance data for text classification can affect the classification performance, because the machine will tend to learn in the class with the most data and ignore the minority class. Handling the imbalanced data using resample by removing data in the majority class (under sampling) and increasing the minority class (oversampling) with a bias of 0.5 can increase the performance of classification. The accuracy increases more than 80%. The classification scientific research paper content based on Author Keywords using the TF-IDF extraction feature, is improved classification performance significantly. Data processing in this context is

not only at the text pre-processing stage but also understanding the underlying data distribution and looking for appropriate handling methods. The best classification result is SVM, which is obtained better performance results than the other two classifiers, with an accuracy value of 0.862.

Acknowledgment. This paper is part of the LPPM Competitive PNBP Research, Universitas Negeri Surabaya, Indonesia which is fully funded with a contract number [B/35071/UN38.9/K.04.00/2022].

References

1. R. Romero, P. Celard, J. M. Sorribes-Fdez, A. Seara Vieira, E. L. Iglesias, and L. Borrajo, "MobyDeep: A lightweight CNN architecture to configure models for text classification," *Knowl Based Syst*, vol. 257, p. 109914, 2022, https://doi.org/10.1016/j.knosys.2022.109914.
2. J. J. Winnink, R. J. W. Tijssen, and A. F. J. van Raan, "Searching for new breakthroughs in science: How effective are computerised detection algorithms?" *Technol Forecast Soc Change*, vol. 146, pp. 673–686, 2019, https://doi.org/10.1016/j.techfore.2018.05.018.
3. A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," *Expert Syst Appl*, vol. 201, p. 117193, 2022, https://doi.org/10.1016/j.eswa.2022.117193.
4. M. S. Santos *et al.*, "On the joint-effect of class imbalance and overlap: a critical review," *Artif Intell Rev*, vol. 55, no. 8, pp. 6207–6275, 2022, https://doi.org/10.1007/s10462-022-10150-3
5. Q. Gu, J. Tian, X. Li, and S. Jiang, "A novel Random Forest integrated model for imbalanced data classification problem," *Knowl Based Syst*, vol. 250, p. 109050, 2022, https://doi.org/10.1016/j.knosys.2022.109050.
6. Z. ao Huang, Y. Sang, Y. Sun, and J. Lv, "A neural network learning algorithm for highly imbalanced data classification," *Inf Sci (N Y)*, vol. 612, pp. 496–513, 2022, https://doi.org/10.1016/j.ins.2022.08.074.
7. S. Tyagi and S. Mittal, "Sampling approaches for imbalanced data classification problem in machine learning," in *Lecture Notes in Electrical Engineering*, 2020, vol. 597, pp. 209–221. https://doi.org/10.1007/978-3-030-29407-6_17
8. B. Krawczyk, C. Bellinger, R. Corizzo, and N. Japkowicz, "Undersampling with Support Vectors for Multi-Class Imbalanced Data Classification," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–7. https://doi.org/10.1109/IJCNN52387.2021.9533379.
9. Z. Zhao, P. Zhong, and Y. Zhao, "Learning SVM with weighted maximum margin criterion for classification of imbalanced data," *Math Comput Model*, vol. 54, no. 3, pp. 1093–1099, 2011, https://doi.org/10.1016/j.mcm.2010.11.040.
10. A. K. Ambalavanan and M. v Devarakonda, "Using the contextual language model BERT for multi-criteria classification of scientific articles," *J Biomed Inform*, vol. 112, p. 103578, 2020, https://doi.org/10.1016/j.jbi.2020.103578.
11. M. Liang and T. Niu, "Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs," *Procedia Comput Sci*, vol. 208, pp. 460–470, 2022, https://doi.org/10.1016/j.procs.2022.10.064.

12. X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, 2021, https://doi.org/10.1016/j.aej.2021.02.009.
13. P. George and P. Vinod, "Machine Learning Approach for Filtering Spam Emails," in *Proceedings of the 8th International Conference on Security of Information and Networks*, 2015, pp. 271–274. https://doi.org/10.1145/2799979.2800043.
14. P. George Princy and Vinod, "Composite Email Features for Spam Identification," in *Cyber Security*, 2018, pp. 281–289.
15. V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.
16. R. I. Kurnia, "Classification of User Comment Using Word2vec and SVM Classifier," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 643–648, Feb. 2020, https://doi.org/10.30534/ijatcse/2020/90912020
17. A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif Intell Rev*, vol. 52, no. 1, pp. 273–292, Jun. 2019, https://doi.org/10.1007/s10462-018-09677-1.
18. S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, pp. 238–248, Jan. 2022, https://doi.org/10.1016/j.susoc.2022.03.001.