Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Abstract Classification Using Support Vector Machine Algorithm (Case Study: Abstract in a Computer Science Journal)

To cite this article: F R Lumbanraja *et al* 2021 *J. Phys.: Conf. Ser.* **1751** 012042

Abstract Classification Using Support Vector Machine Algorithm (Case Study: Abstract in a Computer Science Journal)

F R Lumbanraja[1,*], E Fitri[2], Ardiansyah[3], A Junaidi[4], Rizky Prabowo[5]

[1,2,3,4,5] Department of Computer Science, Faculty of Computer Science, University of Lampung, Bandar Lampung, Indonesia

email: favorisen.lumbanraja@fmipa.unila.ac.id[1,*]

Abstract. Jurnal Komputasi is an online journal written by researchers and published by the Department of Computer Science, University of Lampung. Specific scientific information contained in journals is difficult to find because journals have not been structured and are classified into more specialized categories of computer science. Text mining can convert the shape of a journal into structured by homogeneous data form in it. 144 journal abstracts are collected into one corpus document in CSV format used as a research dataset. Journal abstract classification is done using one of the supervised machine learning methods, namely Support Vector Machine (SVM) so that the classification process is faster than the manual method. The TF-IDF technique is used to transform sentences in the abstract into vector so that they can be modelled with SVM. The classification model will be validated by applying the 10-fold cross validation technique. From these classifications a calculation of the resulting performance will be calculated based on the confusion matrix calculation of the resulting performance will be calculated based on the confusion matrix calculation and the use of 3 SVM kernels. The conclusion based on this research is that there are two factors that affect classification accuracy, that is the number of members between scientific classes that are not balanced and the number of features generated from text mining. The highest accuracy of testing result obtained on the use of 205 features and SVM Linear kernel with a value of 58,3%.

Keyword: Supervised Machine Learning, Text Mining, Support Vector Machine, TF-IDF, 10-Fold Cross Validation

1. Introduction

APJII as the manager of internet services in Indonesia, in 2017 noted that the growth of internet users in Indonesia has increased significantly with total internet users reaching 143.26 million people [2]. APJII also noted that in 2017, services accessed by the Indonesian public for the search engine category took the third highest position, reaching 74.84%. This fact shows that Indonesian people rely on the role of search engines to find the information or solutions needed. One of its uses is to find scientific literature such as journals for further research purposes and as a reading source.

The number of journals in Indonesia has increased from 23,876 journals in 2017 to 34,964 journals in 2018 [2]. This significant increase shows that the high interest in reading and the need for information which come from journals. Information on all articles in journals can generally be read briefly in the abstract section. Journal abstracts contain important information from a scientific article. Extracting information in a jurnal komputasi can be done by utilizing the abstract section of the journal using text mining techniques.

The University of Lampung, especially the Department of Computer Science, has an online journal that publishes various scientific articles written by researchers, both students and lecturers. This scientific article is called Jurnal Komputasi which is published every 6 months. Online Jurnal Komputasi are not yet structured and classified into more specific scientific categories**.**

In previous research, text classification and text mining were used to obtain implicit information in text data. This information is used as data in the text data retrieval system. The classification method applied is the Rocchio classification method by measuring the boundaries between classes to determine the centroid. The corpus data used are 150 undergraduate thesis abstract documents of the Computer Science Department, University of Lampung. These documents are classified into 12 scientific classes, namely Data Mining, Information Retrieval, Information Systems, Geographical Information Systems, Software Engineering, Cryptography, Computer Networks, Parallel Programming, Expert Systems, Digital Image Processing, Pattern Recognition, and Software Computing. There are two parameters used to measure the quality of information retrieval, namely precision and recall. From the test results of 30 test data documents, the accuracy of the recommendation results is 76.67%. Classification of training data in predetermined classes greatly affects the results of recommendations and search results [8].

Research related to classification has also been applied to the sentiment assessment of customer satisfaction levels for restaurants and food stalls in Tegal City. The data set used is the comments of visitors to the web rawa.co.id between 2017-2018 in Indonesian. Sentiment classification uses the Support Vector Machine (SVM) approach based on Information Gain (IG) feature selection and Chi Squared Statistics. Categories in the classification carried out are "good" and "average". The results of the classification experiment with the SVM model and IG feature selection resulted in an accuracy value of 72.45%. While the classification experiment with the SVM model and feature extraction Chi Squared Statistic resulted in an accuracy value of 70.09% [11].

The method of extracting information then becomes easier with CERMINE. In a high-quality document search system, access is required to obtain document metadata such as journal names, journal volumes, and number of pages. By simplifying procedures in supervised and unsupervised machine learning techniques. The PDF format is the most popular format for saving source documents [5].

Previous research also applied several methods that have the potential to improve the performance of top-level classifiers such as hierarchical multi-label classification, namely Convolutional Neural Network (CNN), distributed semantic models, and combined methods of lexical and semantic models. The term frequency of each feature is calculated using Word2Vec to construct the classifier. The dataset used is Indonesian news articles which are arranged into a dataset for multi-label classification. The dataset is then divided into two, first (dataset1) 677 Indonesian articles that have been classified manually and secondly 5713 Indonesian articles were added to the training data for the first dataset, resulting in a second dataset (dataset2). Based on the results, the combination method of lexical and semantics in feature engineering can improve the performance of the classification model. by increasing the false positive error can also improve the performance of hierarchical multi-label classification [9].

2. Data and Method

Abstract Data Collection

The data in this study were abstracts of the Jurnal Komputasi of the Computer Science Department of the University of Lampung in English with CSV format totaling 144 journal documents (volumes I-VI) published 2013-2018. For abstract data, manual classification was carried out by means of majority voting by 3 sources. The classification is based on 12 scientific categories according to the Dennings matrix. The last version released by Peter J. Dennings in 1999, Computer Science is divided into 12 sub-fields where previously it was divided into 9 sub-fields. The division is based on scientific reflection of three things, namely theory, abstraction, and design. The 12 sub-fields of classification are described in Table 1.

Table 1. Scientific Subjects in Denning Matrix.

Algorithms and Data Structures	Programming language	Architecture
Operating Systems and Networks	Software Engineering	Database and Information Retrieval System
Artificial Intelligence and Robotics	Graphic	Human Computer Interaction
Computational Science	Organizational Informatics	Bioinformatics

The results of manual classification can be seen in Table 2, where we collected 144 abstracts in Jurnal Komputasi.

Table 2. Frequency of Subjects in collected form 144 Jurnal Komputasi.

id	Scientific Class	Number of Documents
1	Algorithms and Data Structures	27
2	Programming language	0
3	Architecture	1
4	Operating Systems and Networks	5
5	*Software Engineering*	57
6	Database and Information Retrieval System	10
7	Artificial Intelligence and Robotics	15
8	Graphic	2

9	*Human Computer Interaction* (HCI)	2
10	Computational Science	1
11	*Organizational Informatics*	23
12	Bioinformatics	1
	Total	144
	Average	12
	Standard Deviation	16,2

The implementation of text mining in the Jurnal Komputasi abstract classification uses the Natural Language Toolkit (NLTK) while the classification modeling process to the evaluation stage uses the Scikit-Learn library available in Python. The flow of research carried out can be seen in Figure 1.

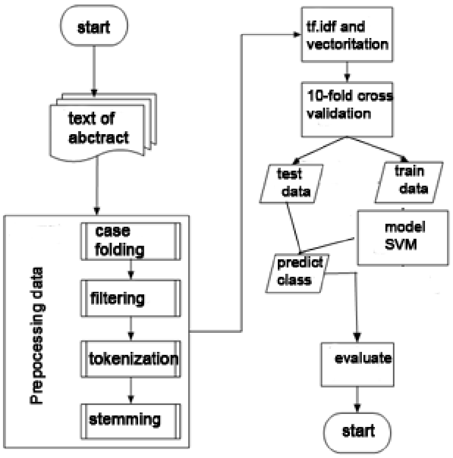


Figure 1. Follow chart of text classification in this research.

Based on the results of the abstract classification in Table 2, it is found that there is an imbalance in the number of members between classes. The Software Engineering class has 57 class members while the Programming Language class has no class members at all. Meanwhile, the Architecture, Computing, and Bioinformatics classes only have one class member. Based on the results of the abstract classification in Table 3, it is found that there is an imbalance in the number of members between classes. The average member of the scientific class from the classification results is 12. The standard deviation of the classification result data is 16.2, which means that the data points per class are close to the average value. But there is a high difference, for example the Software Engineering class has 57 class members while the Programming Language class has no class members at all. Meanwhile, the Architecture, Computing, and Bioinformatics classes only have one class member.

Pre-processed Data

Pre-processing abstract data is an implementation step of text mining techniques in extracting information. The abstract pre-processing stage is carried out with the aim of obtaining features in the abstract of the jurnal komputasi by homogenizing the abstract form. The first step of the abstract pre-processing is to run the lowercase function to homogenize abstract shapes by changing all capital letters in the abstract to lowercase.

The second step is to run the filtering process to filter out the words that will be used as features. The features produced in the pre-processing stage are words that contain information and are considered to represent the characteristics of a class. The filtering process consists of several sub-processes. First, run the remove number function to remove numbers from the abstract and change it to whitespace. The second subprocess of filtering is to run the remove whitespace function to remove excess space. The third subprocess is to run the remove punctuation function to remove punctuation. The last subprocess in filtering is to run the remove stop words function to remove words that have no meaning so as to produce a meaningful array of words. Stop words are a group of words that are not related to the main subject in question even though these words often appear in the data used [3]. Examples of words in English that are included in the stop words list are to be (is, am, are), this, of, many, and so on.

The third step of the abstract pre-processing is to run the tokenization function to cut each word in a sentence by using spaces as delimiters which will produce a token or term [6]. The final step in the abstract pre-processing is to run the stemming function to find the root word of each tokenized word by removing the prefix and ending (suffix) of the word. In the stemming process, words are grouped into several groups of words which have the same root word [1]. For example, drug, drugs, and drugged are grouped into the root word drug. The abstract pre-processing stage is simply described in Table 3.

Table 3. Illustration of Document Prepossessing.

Previous Stages	Result
Original Abstract	Library integrated service unit of Unila provides academic services for students through the website.
Lowercase	library integrated service unit of unila provides academic services for students through the website.
Filtering	library integrated service unit unila provides academic services students website
Tokenises	library integrated service unit unila provides academic services students website
Stemming	library integrate service unit unila provide academic servic student websit

The final result in the abstract pre-processing is tokens in the form of basic words which are then referred to as features that will be processed at a later stage.

Abstract Pre-processing Result Transformation

The document transformation is carried out to convert the token from a string type to a vector form using the TF-IDF weighting technique. The goal is to calculate the frequency of occurrence of words in abstract data so that it is known how important a word is in an abstract collection [6]. The level of importance increases when a word appears several times in a document but is offset by the frequency of occurrence of the word in the document [6].

Pre-processed data is converted into vector form by forming text data into a numeric matrix. The calculation of the matrix using TF-IDF is formulated in Equation 1.

$$TF\ x\ IDF = TF\ x\ log(\frac{n}{df}) \tag{1}$$

Where: *TF = word frequency*
df = document frequency
n = number of documents

First, calculate the TF or the occurrence of a word in an abstract data. Second, calculate the DF value or the amount of abstract data where the word is found. Third, calculate the TF x IDF value so that the weight value is obtained.

From the preprocessing stage, tokens are obtained as features in the form of strings. Furthermore, the document is transformed from a string type into a vector form using the TF-IDF weighting technique. In Table 4, an example of the features of the abstract preprocessing results has been given weight values from 3 documents. The greater the weight value obtained, the higher the level of document similarity to the word / feature.

Table 4. Illustration of weight value of terms.

id_doc	TFIDF			
	develop	print	research	utility
1	0	0	0.0542	0.1943
2	0	0	0.0249	0
3	0	0	0	0

In the Scikit-Learn library, the max_feature parameter is implemented to limit the number of words weighted in order to see the effect of word count on testing accuracy.

Modeling and Validation of Abstract Classification Models

The classification prediction model is made using the Support Vector Machine (SVM) approach. SVM has a basic principle of linear classifier, namely classification cases that can be separated linearly, however SVM has been developed to work on non-linear problems by incorporating the kernel concept in a high-dimensional workspace [4]. The maximum margin training algorithm finds the decision function for the x-dimensional pattern vector n belonging to one of the two classes A and B. The decision function must be linear in its parameter but not limited to the linear dependence x. These functions can be expressed either directly or in multiple spaces [4]. Kernel functions are used to map the original dimension (lower dimension) of a data set to a new dimension (a relatively higher dimension).

The prediction model from the Support Vector Machine approach is then validated using the K-Fold Cross Validation (K-Fold CV) technique to obtain the highest average accuracy value. K-Fold CV is a technique for estimating the performance of the training model that has been built and prevents overlapping of testing data [7]. K-Fold CV divides data into K equal pieces of data (Xi, i = 1,… K). At each fold or iteration, one K is used as a testing data set and combines the remaining K-1 as a training data set during K experiments [7]. In this study, the value of K = 10. A total of 144 abstract data were divided into 10 iterations so that 129 training data and 15 testing data were obtained for each iteration. The implementation of 10-Fold Cross Validation is illustrated in Figure 2.

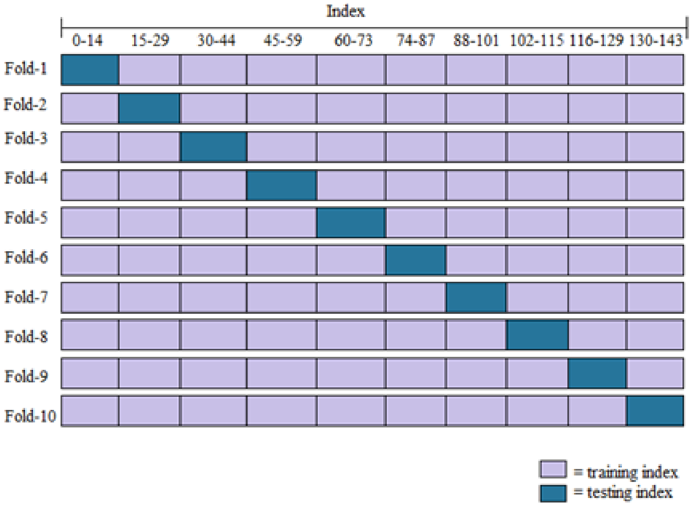


Figure 2. Illustration of 10-fold cross validation.

The distribution of training data and testing data in each iteration is in accordance with the NumPy index sequence because it uses the parameter Shuffle = False. Each iteration also does not represent all scientific classes because it uses the NumPy index sequence. NumPy is used because it has the ability to form N-dimensional array objects in the scientific computing process. Another advantage is the consumption of smaller memory and faster runtime [15].

Measuring Classification Performance

Classification system performance is measured to find out how well the system classifies data. In this study, the classification performance measurement method used is confusion matrix. The confusion matrix compares the results of the classification carried out by the system with the results of the voting classification. The basis for confusion matrix can be seen in Table 5.

Table 5. Confusion Matrix for Classification.

	Actual: Positive	Actual: Negative
Predicted: Positive	*True Positives*	*False Positives*

Predicted: Negative	*False Negatives*	*True Negatives*

1. **Accuracy**

Accuracy is the percentage of observations that are correctly classified [14]. The higher the accuracy results, the more effective the classification algorithm model being tested is [14]. Accuracy is formulated in Equation 6.

$$Accurancy \quad = \frac{TP+TN}{TP+FP+FN+TN} \qquad (6)$$

2. **Precision**

Precision is the percentage of observations that are correctly classified as positive data in the group that has tested positive by the classifier [14]. Precision is formulated in Equation 7.

$$Precision \quad = \frac{TP}{(TP+FP)} \qquad (7)$$

3. **Recall**

Recall is the percentage of observations labeled positive and classified correctly [14]. Recall is formulated in Equation 8.

$$Recall \quad = \frac{TP}{TP+FN} \qquad (8)$$

4. **F-Measure (F1-Score)**

F1-Score shows a balance between precision and recall. F1-Score is the harmonic mean of precision and recall with a value range of 0 (worst score) and 1 (best score) [14]. F1-Score is formulated in Equation 9.

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \qquad (9)$$

3. Result dan Discussion

The testing accuracy value is obtained from the average value of testing accuracy for 10 times iteration which is calculated using 3 SVM kernel functions, namely Linear, Polynomial, and RBF. Table 6 describes the testing accuracy values for each kernel function.

Table 6. Comparison accuracy of SVM classifer for each kernel.

Fold	Kernel		
	Linear	Polynomial	RBF
1	0.38	0.28	0.28
2	0.55	0.33	0.33
3	0.62	0.37	0.37
4	0.47	0.40	0.40
5	0.40	0.40	0.40
6	0.61	0.46	0.46
7	0.77	0.46	0.46
8	0.54	0.45	0.45
9	0.82	0.45	0.45

10	0.73	0.45	0.45
Average	0.59	0.41	0.41

The value for each performance measure is also obtained from each experiment in each fold so that 10 values are obtained which are then sought for the average. This average value then becomes the final value used to draw conclusions as a result of the assessment of the classification process carried out. Figure 3 represents the comparison of the average values obtained from the use of the three SVM kernel functions at each performance measure.

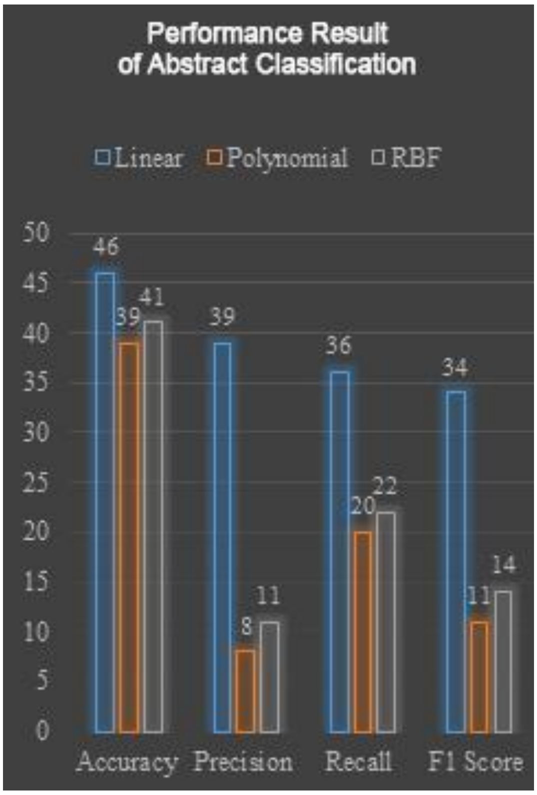


Figure 3. Comparison of classification performance using each kernels.

Based on the bar graph in Figure 3, it is found that the highest value on each performance measure is achieved using the Linear kernel function with details of the accuracy value of 46%, precision of 39%, F1 Score of 34%, and MCC of 30%. While the lowest value is obtained when using the Polynomial kernel function. The Linear kernel function yields the highest value compared to the other two kernel functions indicating that abstract data can be classified linearly even though it produces a small performance value. Figure 1 shows a bar graph of the percentage of testing accuracy values from the use of Linear, Polynomial, and RBF kernels. From the results of the analysis carried out during the study, there are several factors that affect the results of the classification model accuracy.

1. The number of class members in the dataset

The basis for the Dennings matrix grouping applied to the dataset in this study results in an unbalanced number of members between classes. This causes the classifier to be unable to recognize the classification pattern so that most of the features resulting from the text mining process are recognized as characteristics of the scientific class with the largest number of members. As a result, there was an error prediction in a class with few members.

2. Number of features of text mining results

In this study, an experiment to reduce the number of features was carried out by applying an interval of 100 features. The results of the experiment to reduce the number of features carried out show the effect of the number of features resulting from the feature extraction process on the value of testing accuracy. The results of this experiment can be seen in graphical form in Figure 4.

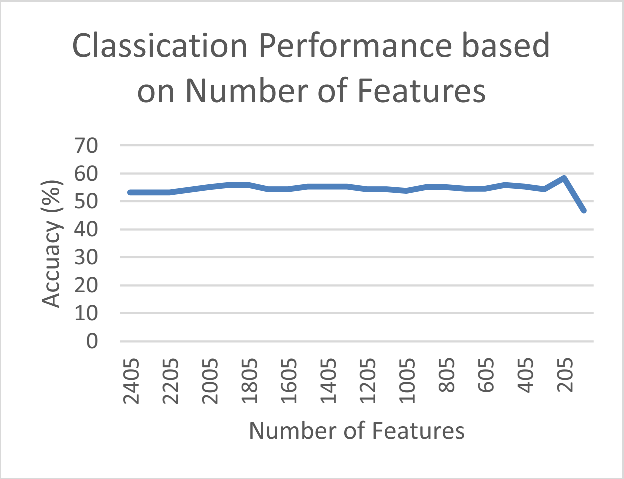
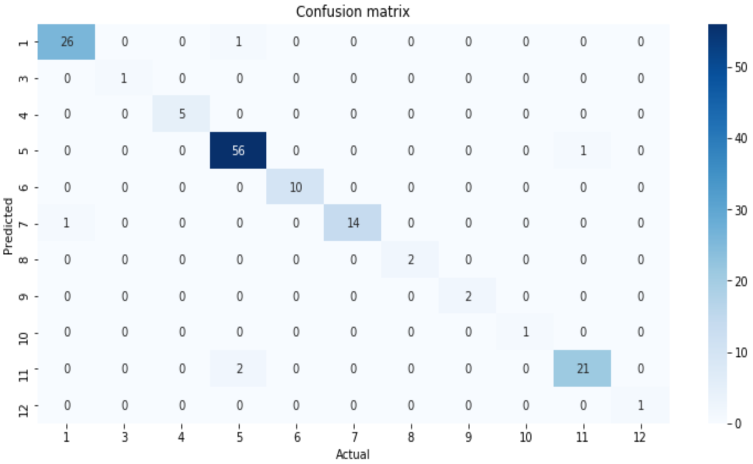


Figure 4. Classification performance based on number of features.

To see the effect of the number of features on the testing accuracy value, a feature value interval of 100 is applied. Based on Figure 1, the highest testing accuracy is obtained when using 205 features with an accuracy value percentage of 58.3%. However, when using 105 features, the accuracy value drops to 46.6%. This fact shows that the fewer or more features produced does not necessarily increase the value of testing accuracy. The comparison of the classification results carried out by the system with the voting classification results on the use of 205 features is represented in the form of confusion matrix in Figure 5.



	1	3	4	5	6	7	8	9	10	11	12
1	26	0	0	1	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0
4	0	0	5	0	0	0	0	0	0	0	0
5	0	0	0	56	0	0	0	0	0	1	0
6	0	0	0	0	10	0	0	0	0	0	0
7	1	0	0	0	0	14	0	0	0	0	0
8	0	0	0	0	0	0	2	0	0	0	0
9	0	0	0	0	0	0	0	2	0	0	0
10	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	2	0	0	0	0	0	21	0
12	0	0	0	0	0	0	0	0	0	0	1

Figure 5. Confusion matrix of 205 important features.

Based on Figure 5, misclassification occurs in 3 journal abstracts that are incorrectly predicted as Software Engineering class (5), 1 journal abstract is incorrectly predicted as Algorithm and Data Structure class (1), and 1 journal abstract is incorrectly predicted as Organizational Informatics class (11).

4. Conclusion

The use of the Linear kernel produces the greatest accuracy value, namely 58.3% in the Jurnal Komputasi abstract classification. There are two factors that affect classification accuracy. The first factor, namely the unbalanced number of members between scientific classes, causing the classifier to not experience a learning process for a class with few members. The second factor, namely the number of features generated from the abstract preprocessing stage. The highest accuracy is obtained when using 205 features, but the small number or the large number of features resulting from the feature extraction process does not necessarily improve classification accuracy.

REFERENCES

[1] Lutfi A A, Permanasari A E, and Fauziati S 2018 Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine *Journal of Information Systems Engineering and Business Intelligence* **4(1)**, p 57-64

[2] APJII 2018 Hasil Survey Penetrasi dan Perilaku Pengguna Internet Indonesia 2017 (Asosiasi Penyelenggara Jasa Internet Indonesia) https://apjii.or.id/content/read/39/342/Hasil-Survei-Penetrasi-dan-Perilaku-Pengguna-Internet-Indonesia-2017#.

[3] Setiawan A, Kurniawan E, and Handiwidjojo W 2013 Implementasi Stopword Removal untuk Pembangunan Aplikasi Alkitab Berbasis Windows 8 *Jurnal EKSIS* **6(2)** p 1-11

[4] Boser B E, Guyon I M, and Vapnik V N 1992 A Training Algorithm for Optional Margin Classifiers: https://www.svms.org

[5] Tkaczyk D, Szostek P, Fedoryszak M, Dendek P J, and Bolikowski 2015 CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature IJDAR **18** p.317-335

[6] Wisnu D and Hetami A 2015 Perancangan Information Retrieval (IR) untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris dengan Pembobotan Vector Space Model *Jurnal Ilmiah Teknologi dan Informasi ASIA* **9(1)** p 53-59

[7] Alpaydin E 2004 *Introduction to Machine Learning* (Massachusetts: MIT Press) p 609

[8] Lumbanraja F R 2013 Sistem Pencarian Data Teks dengan Menggunakan Metode Klasifikasi Rocchio (Studi Kasus:Dokumen Teks Skripsi) *Kumpulan Makalah Seminar Semirata* **1(1)** p 217-224

[9] Irsan I C and Khodra M L Hierarchical Multi-Label News Article Classification with Distributed Semantic Model Based Features *International Journal of Advances in Intelligent Informatics* **5(1)**, p 40-47

[10] Yusa M, Utami E, and Luthfi E T 2016 Analisis Komparatif Evaluasi Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes *Jurnal Buana Informatika* **7(4)** p 293-302,

[11] Somantri O and Apriliani D 2018 Support Vector Machine Berbasis Feature Selection untuk Sentiment Analysis Kepuasan Pelangga Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal *Jurnal Teknologi Informasi dan Ilmu Komputer* **5(5)** p 537-548

[12] PDII-LIPI (Pusat Dokumentasi dan Informasi Ilmiah), “Statistik Jumlah Artikel per Tahun,” Indonesian Scientific Journal Database, 2019. [Online]. Tersedia pada: http://isjd.pdii.lipi.go.id/index.php/public_no_login/dashboard. [Diakses pada 3 Januari 2019].

[13] P. J. Denning, “Computer Science: The Discipline,” Agustus 1997. [Online]. Tersedia pada: http://denninginstitute.com/pjd/PUBS/ENC/cs99.pdf. [Diakses pada 17 Oktober 2019].

[14] P. R. Nicolas, Scala For Machine Learning. UK: Packt Publishing Ltd, 2015, pp. 69-70. Y. A. Rohman, “Pengenalan NumPy, Pandas, Matplotlib,” Desember 2019. [Online]. Tersedia pada: https://medium.com/@yasirabd/pengenalan-numpy-pandas-matplotlib-b90bafd36c0