Subject Area Classification of Journal Articles Based on Metadata Using Bag of Words and Naïve Bayes

Manuscript received January 29, 2026; revised May 7, 2026

Ainunna'imah
Master's Program of Informatics,
Faculty of Industrial Technology,
Universitas Ahmad Dahlan,
Yogyakarta, Indonesia
2408048020@webmail.uad.ac.id

Herman Yuliansyah
Department of Informatics,
Faculty of Industrial Technology,
Universitas Ahmad Dahlan,
Yogyakarta, Indonesia
herman.yuliansyah@tif.uad.ac.id

Imam Riadi
Department of Information Systems,
Faculty of Applied Science and Technology,
Universitas Ahmad Dahlan,
Yogyakarta, Indonesia
imam.riadi@is.uad.ac.id

***Abstract*— The rapid growth of scientific publications poses challenges in grouping journal articles based on subject area, especially when using metadata such as titles, abstracts, and keywords. However, differences in feature representation and classification algorithms often result in varying performance, requiring comparative studies to determine the optimal model combination. This study compares four combinations of subject area classification models, namely TF-IDF + Naïve Bayes, TF-IDF + Support Vector Machine, Bag-of-Words + Support Vector Machine, and Bag-of-Words + Naïve Bayes. The research process included text preprocessing, feature extraction, and testing using an 80% training and 20% testing data split scheme in five scenarios. The evaluation was performed using confusion matrices, accuracy, precision, recall, and F1-score. The experimental results showed variations in performance between models, with an average F1-score of 0.8103 for TF-IDF + Naïve Bayes, 0.8494 for TF-IDF + Support Vector Machine, 0.8297 for Bag-of-Words + Support Vector Machine, and 0.8335 for Bag-of-Words + Naïve Bayes as the best performance. These findings indicate that a word frequency-based approach combined with Naïve Bayes is effective for classifying journal article subject areas based on metadata, although challenges remain in subject areas with semantic proximity.**

Keywords*— *Data Analysis, Feature Extraction, Performance Evaluation, Scientific Document Classification, Supervised Machine Learning

I. INTRODUCTION

Advances in science and technology have significantly driven the growth of scientific publications in various disciplines [1], [2]. Every year, the number of scientific articles published through journals, proceedings, and academic repositories continues to increase in line with the development of research activities in universities and research institutions [3], [4]. This situation poses new challenges in scientific information management, particularly in ensuring that every article can be found, accessed, and utilized appropriately by the academic community.

In the scientific publication ecosystem, grouping articles by subject area is fundamental. The subject area serves as the scientific identity of an article, indicating the field of study and the scope of research discussed [5]. Subject area classification serves as the basis for indexing systems that organize scientific databases, support information retrieval, and facilitate the mapping of research developments across various fields of science. Without accurate and consistent classification, scientific articles may be difficult for relevant readers to find, may not be optimally linked to other research in the same field, and may be underutilized in data-driven decision making.

Subject area classification also plays an important role in evaluating research performance at the individual, institutional, and national levels. Mapping publications by scientific field serves as a basis for formulating research policies, determining funding priorities, and analyzing academic productivity [6]. Errors in subject-area classification can distort bibliometric analysis and lead to inaccurate research policy formulation. However, the growing volume of interdisciplinary scientific publications has made manual classification by subject area increasingly complex and ineffective. The manual process is time-consuming, requires experts, and is prone to subjectivity and inconsistency, especially in multidisciplinary articles. With the increasing scale of scientific publication data, conventional approaches are no longer adequate to ensure speed, accuracy, and consistency of classification.

The metadata of a scientific article, consisting of the title, abstract, and keywords, is a concise summary of the research content and reflects the main focus of the article [7]. The metadata can be used as a textual representation to automatically build a subject-area classification system [8]. Text mining and machine learning approaches enable scientific documents to be grouped based on linguistic patterns and word characteristics that reflect their scientific field. In this context, an automatic classification system is urgently needed to support the management of modern scientific publications. The development of an accurate and reliable classification system is expected to assist in indexing, literature searches, research mapping, and data-driven policy-making [9]. Therefore, research is needed to examine classification approaches based on scientific article metadata to develop a model that effectively and efficiently groups articles into subject areas.

Scientific article classification has become an important research area for supporting literature management, knowledge organization, research field mapping, and publication recommendation systems. Numerous approaches

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002041

have been developed to address the growing volume of scientific publications, ranging from traditional machine learning algorithms to advanced deep learning techniques and large language models (LLMs). These methods aim to improve classification accuracy, scalability, and efficiency in processing large-scale scholarly datasets. Recent advances in natural language processing have further enhanced the capability of automated systems to identify, categorize, and recommend scientific articles across diverse academic disciplines.

Classic machine learning approaches are still widely used due to their efficiency and ease of implementation. The combination of Bag-of-Words (BoW) and TF-IDF-based text representation, combined with algorithms such as Naïve Bayes and Support Vector Machine (SVM), has been shown to deliver competitive performance across various scientific article classification scenarios [10], [11], [12]. Several studies show that SVMs are stable on high-dimensional data and can achieve high accuracy, especially in multi-label classification and large datasets [10]. However, the performance of classical methods is strongly influenced by data characteristics, class distributions, and feature schemes, so results are not consistent across studies [13], [14].

To improve performance, several studies have proposed ensemble and hybrid approaches that combine classical methods with transformer-based representations. These approaches have been shown to significantly improve precision and recall in scientific article classification [15]. However, the complexity of the system and greater computational requirements are major limitations in the widespread application of ensemble methods.

With the development of deep learning, transformer-based models such as BERT and its variants are widely used for scientific text classification. Several studies report that transformer models, especially those trained in a domain-specific manner, can outperform classical methods [16], [17], [18]. However, other studies show that this advantage is not always consistent, and under certain conditions, classical methods such as SVM with TF-IDF can still deliver comparable or better performance, especially on limited datasets and complex label structures [19], [20].

Based on this study, there is no clear conclusion regarding the most effective method for classifying scientific articles. In addition, there is still limited research directly comparing TF-IDF+Naïve Bayes and BoW+SVM using the same experimental setup and evaluation metrics, especially for the classification of journal articles based on metadata. Therefore, this study aims to fill this gap by structurally comparing the two classical machine learning approaches. This study focuses on the development and evaluation of a metadata-based scientific article field classification system by comparing the performance of various combinations of feature representations and algorithms, namely BoW with SVM, BoW with Naïve Bayes, TF-IDF with SVM, and TF-IDF with Naïve Bayes. The results of this study are expected to contribute to the development of an efficient and relevant scientific publication classification system to support academic information management in the digital era.

II. METHODS

This study was conducted using an experimental approach to compare the performance of two text classification methods, namely TF-IDF + Naïve Bayes and BoW + SVM, in classifying journal articles based on subject area. The research flow began with processing article metadata and ended with evaluating the classification model's performance, as shown in Fig. 1.

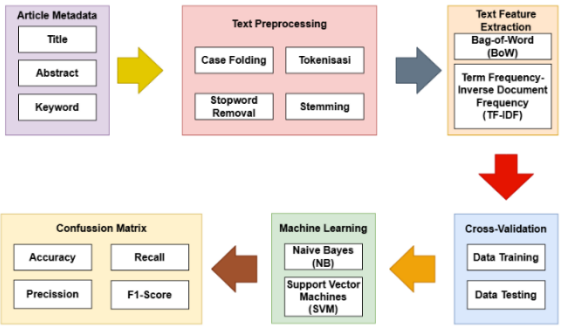


Fig. 1. Flowchart of the process of classifying journal article subject areas based on metadata with four combinations of TF-IDF and BoW models using Naïve Bayes and SVM

This research began with the processing of journal article metadata, such as titles, abstracts, and keywords, through the stages of text preprocessing, feature extraction using BoW and TF-IDF, data division with a ratio of 80% training data and 20% test data, as well as 5-fold cross validation, application of the Naïve Bayes and SVM algorithms, to evaluating model performance using a confusion matrix and accuracy, precision, recall, and F1-score metrics to determine the best method for classifying articles based on subject area.

A. Article Metadata

The data used in this study consisted of journal article metadata, including article titles, abstracts, and keywords. This metadata was selected because it effectively represents the content and main topics of the articles in a concise and informative manner. The class labels in this study correspond to the scientific fields of each journal article, namely Education, Social, Science, Health, Economics, and Engineering.

The data collection stage involved retrieving journal articles from the SINTA 1 portal and storing the metadata in CSV format for computational processing. To ensure data consistency, inclusion criteria were applied, namely articles indexed in SINTA 1, classified within the predefined subject areas (Education, Social, Science, Health, Economics, and Engineering), and containing complete metadata (title, abstract, and keywords). Meanwhile, exclusion criteria included articles outside SINTA 1, records with missing or incomplete metadata, and entries that did not match the defined subject categories or contained ambiguous labels. The inclusion and exclusion criteria applied in this study are summarized in Table I.

The final dataset consisted of 1,750 journal articles. An overview of the dataset distribution across subject areas is presented in Table II. The dataset was then split into training and test sets using an 80:20 ratio. To improve the reliability of the evaluation, a 5-fold cross-validation was applied to the training data, where each fold alternately served as training and a validation set.

TABLE I. INCLUSION AND EXCLUSION CRITERIA

Criteria Type	Description
Inclusion	Journal articles indexed in SINTA 1, classified into predefined subject areas (Education, Social, Science, Health, Economics, Engineering), and containing complete metadata (title, abstract, and keywords).
Exclusion	Articles not indexed in SINTA 1, records with incomplete metadata, and entries with ambiguous or incorrect subject labels.

TABLE II. RESEARCH OBJECT

Subject Area	Number of Data Sets	Journal Name
Education	250	Jurnal Pendidikan IPA Indonesia (Indonesian Journal of Science Education), Jurnal Ilmiah Ilmu Terapan Universitas Jambi, Jurnal Pendidikan Islam, EduLearn, IJERE
Social	500	Ahkam, Al-'Adalah, Syariah Jurnal Hukum dan Pemikiran, IJoLE, Jurnal Ilmiah Islam Futura, Al-Risalah, Mazahib Jurnal Pemikiran Hukum Islam, JAS, Jurnal Ilmiah Peuradeun, LAW REFORM, Jurnal Ilmu Sosial dan Ilmu Politik, Journal of Law, Journal of Legal Studies
Science	250	Indonesian Journal of Science and Technology, International Journal of Electrical and Computer Engineering, AJSE, Makara Journal of Science, Journal of the Indonesian Tropical Animal Agriculture
Health	250	Psikohumaniora, JHCLS, Bali Medical Journal, International Journal of Public Health Science, Belitung Nursing Journal
Economic	250	Gadjah Mada International Journal of Business, Journal of Indonesian Economy and Business, ETIKONOMI, Economic Journal of Emerging Markets, al-Uqud, Jurnal Akuntansi dan Keuangan
Engineering	250	IAES, JOIV, IJAIN, IJASEIT, JOIN
Total	**1750**	

B. Text Preprocessing

The text preprocessing stage aims to clean and normalize text data in order to improve the quality of the features produced [21]. The preprocessing steps applied include:

1. Case Folding, which converts all text to lowercase letters.
2. Tokenization, which breaks text into word units (tokens).
3. Stopword Removal, which removes common words that have no discriminatory meaning for classes, such as conjunctions and prepositions.
4. Stemming, which converts words to their base form by removing affixes.

All preprocessing steps are applied to the text metadata before the feature extraction.

C. Text Feature Extraction

After going through the preprocessing stage, the text data is converted into numerical representations using two feature extraction methods, namely:

1) Bag-of-Words (BoW)

The BoW model is an approach for representing text by using word frequencies in a document [22]. This model consists of two main components: a dictionary containing a collection of identified words and frequency values indicating how often each word appears in the document [23]. The BoW method represents documents as vectors of word frequencies, disregarding word order [24]. This approach produces high-dimensional features and is suitable for use with the SVM algorithm.

2) Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF is used to assign weights to each word based on its frequency of occurrence in a document and the frequency of that word across the entire document collection [25]. This method can suppress the influence of common words and highlight words that are more informative. Equation (1) defines term frequency (TF), Equation (2) describes inverse document frequency (IDF), and Equation (3) calculates the TF-IDF value based on the occurrence of a word in a document.

$$TF(t,d) = \frac{Number\ of\ occurrences\ of\ the\ word\ t\ in\ document\ d}{Total\ number\ of\ words\ in\ the\ document} \quad (1)$$

$$IDF(t) = log\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ the\ word\ t}\right) \quad (2)$$

$$TF - IDF(t,d) = TF(td) \times IDF(t) \quad (3)$$

D. Classification Algorithm

1) Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm that uses Bayes' theorem with the assumption of independence between features [26]. This algorithm is widely used in text classification because it is efficient in handling high-dimensional data and has stable performance in frequency-based feature representation [27]. In this study, Naïve Bayes is combined with TF-IDF feature representation and implemented using Multinomial Naïve Bayes, which is suitable for word-weight-based text data. The parameter used in this model is alpha = 1.0 for Laplace smoothing, the default value in the scikit-learn library. This study does not apply hyperparameter tuning to Naïve Bayes because the focus is on comparing the performance of two commonly used text classification approaches, not on model parameter optimization.

2) Support Vector Machine (SVM)

SVM is a classification algorithm that finds the optimal hyperplane to separate data into different classes. SVMs are known to be effective for handling high-dimensional data and multi-class classification problems [28]. In this study, SVM was implemented using LinearSVC with a linear kernel and combined with BoW feature representation. The C parameter was selected via GridSearchCV using a stratified 5-fold cross-validation scheme based on the best F1-macro score.

E. Performance Evaluation

Performance evaluation of the classification model was conducted to measure the model's ability to classify journal articles based on subject area [29]. The evaluation process began with the creation of a confusion matrix, which represents the number of correct and incorrect predictions in each class. Based on the confusion matrix, several key evaluation metrics were calculated, namely accuracy, precision, recall, and F1-score. The values of these evaluation metrics are then summarized in a classification report, which

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002041

serves as a basis for comparing the performance of the two classification methods tested: TF-IDF + Naïve Bayes and BoW + SVM. The calculation of each evaluation metric is based on Equations (4)–(7).

$$Accuracy = \frac{\sum_{i=1}^{n} \mathrm{TP}(C_i)}{\sum_{i=1}^{n} \sum_{j=1}^{n} C_{i,j}} \quad (4)$$

$$Precision(C_i) = \frac{TP(C_i)}{TP(C_i) + FP(C_i)} \quad (5)$$

$$\mathrm{Recall}(C_i) = \frac{\mathrm{TP}(C_i)}{\mathrm{TP}(C_i) + \mathrm{FN}(C_i)} \quad (6)$$

$$\mathrm{F1-Score}(C_i) = 2 \times \frac{\mathrm{Precision}(C_i) \times \mathrm{Recall}(C_i)}{\mathrm{Precision}(C_i) + \mathrm{Recall}(C_i)} \quad (7)$$

N is the number of classes, $TP(C_i), FP(C_i), \mathrm{FN}(C_i)$ are true positive, false positive, and false negative for class (C_i), respectively, while $C_{i,j}$ is the number of actual instances of class i that are predicted as class j.

III. Result and Discusssion

A. Data Collection

The data in this study was collected through a manual crawling process by taking scientific article metadata information and subject areas from the official Science and Technology Index (SINTA) website. The metadata collected in this study included journal names, article titles, abstracts, keywords, volumes, and publication years. The subject areas available on the SINTA platform were used as class labels in the scientific article classification process. SINTA was chosen as the data source based on its role as a national index that provides standardized journal and scientific article information in Indonesia, ensuring that the data obtained is credible and relevant to the research objectives. One of the data sets from the research is shown in Table III, which displays the scientific article metadata along with the subject area labels used as classes in the classification process.

Table III. Dataset Sample

Journal Names	Subject Area	Article titles	Abstracts	Keyword	Publication years	Volume
Jurnal Pendidikan IPA Indonesia	Education	Comparison of SWAT-based Ecohydrological Modeling in the Rawa Pening Catchment Area, Indonesia	The Soil and Water Assessment Tool (SWAT) are an ecohydrological model widely applied to assess water quality and watershed management. This tool also has the advantage of building watershed models even with limited monitoring data availability. The essential data required by this tool includes digital elevation models (DEM), land use maps, climate data, and soil data. Nonetheless, the availability of spatial data is still often a challenge in developing hydrological models, especially in developing countries such as Indonesia. This research will compare the accuracy of freely available data in Indonesia in facilitating the development of hydrological models from SWAT in the Rawa Pening catchment area. This research is crucial since Rawa Pening Lake is a priority lake for revitalization, so the research results will help provide suggestions regarding presenting data in SWAT modeling. This research compares SWAT models built from different land use and DEM (Digital Elevation Models) data. The land use data being compared is the result of processing from the Google Earth Engine (GEE) platform using machine learning with land use data from government agencies, namely the Ministry of Environment and Forestry, while the DEM data being compared is SRTM and DEMNAS data. The validation results using R, R2, RMSE, and NSE show that, in general, the model built from land use from GEE is the best compared to the other models. In modeling SWAT in Indonesia, we recommend using good-quality land-use data. Utilizing supervised classification through Random Forest (RF) algorithms within GEE can facilitate the acquisition of this data. To reduce computation time, the DEM can be SRTM with a small sacrifice of accuracy.	DEM; Google Earth Engine; land use; SWAT; streamflow	2024	Vol 13

B. Preprocessing

Case folding is performed to standardize all text to lowercase letters, so that differences between uppercase and lowercase letters do not affect text analysis. This process is important so that words that are the same but have variations in capitalization are still considered identical in subsequent processing stages. The results of applying case folding to the dataset are shown in Table IV.

Table IV.Case Folding

Doc	Title	Case folding results title	Abstract	Abstract of case folding results	Keyword	Keywords resulting from case folding
0	Comparison of SWAT-based Ecohydrological Model...	comparison of swatbased ecohydrological modeli...	The Soil and Water Assessment Tool (SWAT) is a...	the soil and water assessment tool swat is an ...	DEM; Google Earth Engine; land use; SWAT; stre...	dem google earth engine land use swat streamflow
…	…	…	…	…	…	…
1749	SENTIMENT ANALYSIS OF POST-COVID-19 INFLATION ...	sentiment analysis of postcovid inflation base...	The COVID-19 pandemic caused a crisis in globa...	the covid pandemic caused a crisis in global e...	COVID-19, Inflation, K-Nearest Neighbor, Senti...	covid inflation knearest neighbor sentiment an...

Tokenization is the process of breaking text into units of words or tokens, so that each word can be analyzed separately. This stage allows the model to understand the structure and frequency of words in a document. The results of applying tokenization to the dataset are shown in Table V.

TABLE V. TOKENIZATION

Doc	Case folding results title	Tokenization result title	Abstract of case folding results	Abstract of tokenization results	Keywords resulting from case folding	Tokenized keyword
0	comparison of swatbased ecohydrological modeli...	[comparison, of, swatbased, ecohydrological, m...	the soil and water assessment tool swat is an ...	[the, soil, and, water, assessment, tool, swat...	dem google earth engine land use swat streamflow	[dem, google, earth, engine, land, use, swat, ...
…	…	…	…	…	…	…
1749	sentiment analysis of postcovid inflation base...	[sentiment, analysis, of, postcovid, inflation...	the covid pandemic caused a crisis in global e...	[the, covid, pandemic, caused, a, crisis, in, ...	covid inflation knearest neighbor sentiment an...	[covid, inflation, knearest, neighbor, sentime...

Stopword removal is performed to remove common words that have no discriminatory meaning in the analysis, such as “and,” “or,” “which,” and so on. This process helps to focus on words that are more relevant for document classification. The results of applying stopword removal to the dataset are presented in Table VI.

TABLE VI. STOPWORD REMOVAL

Doc	Tokenization result title	Title after stopword removal	Abstract of tokenization results	Abstract of stopword removal results	Tokenized keyword	Keywords after stopword removal
0	[comparison, of, swatbased, ecohydrological, m...	[comparison, swatbased, ecohydrological, model...	[the, soil, and, water, assessment, tool, swat...	[soil, water, assessment, tool, swat, ecohydro...	[dem, google, earth, engine, land, use, swat, ...	[dem, google, earth, engine, land, use, swat, ...
…	…	…	…	…	…	…
1749	[sentiment, analysis, of, postcovid, inflation...	[sentiment, analysis, postcovid, inflation, tw...	[the, covid, pandemic, caused, a, crisis, in, ...	[covid, pandemic, caused, crisis, global, econ...	[covid, inflation, knearest, neighbor, sentime...	[covid, inflation, knearest, neighbor, sentime...

Stemming is a text processing stage that aims to convert each word to its root form. This process is important to simplify the variations of words that appear in the dataset so that each word can be analyzed consistently and reduce information redundancy. With the application of stemming, the classification model can focus on root words that have the main meaning. The results of applying stemming to the dataset are shown in Table VII.

TABLE VI. STEMMING

Doc	Title after stopword removal	Title of stemming results	Abstract of stopword removal results	Abstract of stemming results	Keywords after stopword removal	Stemming results keyword
0	[comparison, swatbased, ecohydrological, model...	[comparison, swatbas, ecohydrolog, model, rawa...	[soil, water, assessment, tool, swat, ecohydro...	[soil, water, assess, tool, swat, ecohydrolog,...	[dem, google, earth, engine, land, use, swat, ...	[dem, googl, earth, engin, land, use, swat, st...
…	…	…	…	…	…	…
1749	[sentiment, analysis, postcovid, inflation, tw...	[sentiment, analysi, postcovid, inflat, twitte...	[covid, pandemic, caused, crisis, global, econ...	[covid, pandem, caus, crisi, global, econom, g...	[covid, inflation, knearest, neighbor, sentime...	[covid, inflat, knearest, neighbor, sentiment,...

C. Feature Extraction

After the preprocessing stage is complete, the four metadata components of the article, such as title, abstract, keywords, and subject field as class labels, are combined into a single text document for each article. This combination aims to maximize the use of all contextual information from the article in the classification process, so that the model can utilize all relevant text representations to determine the subject field class. The processed text documents are then used to build TF-IDF and BoW features.

1) TF IDF

TF-IDF is a text feature representation method that assigns weights to each word based on two components: the frequency of the word's appearance in a document (Term Frequency, TF) and how unique the word is in the entire document collection (Inverse Document Frequency, IDF). In this way, words that appear frequently in one document but rarely appear in other documents will have a higher weight, thereby improving the model's ability to distinguish between classes. The feature extraction process using TF-IDF on the dataset is shown in Fig 3.

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002041

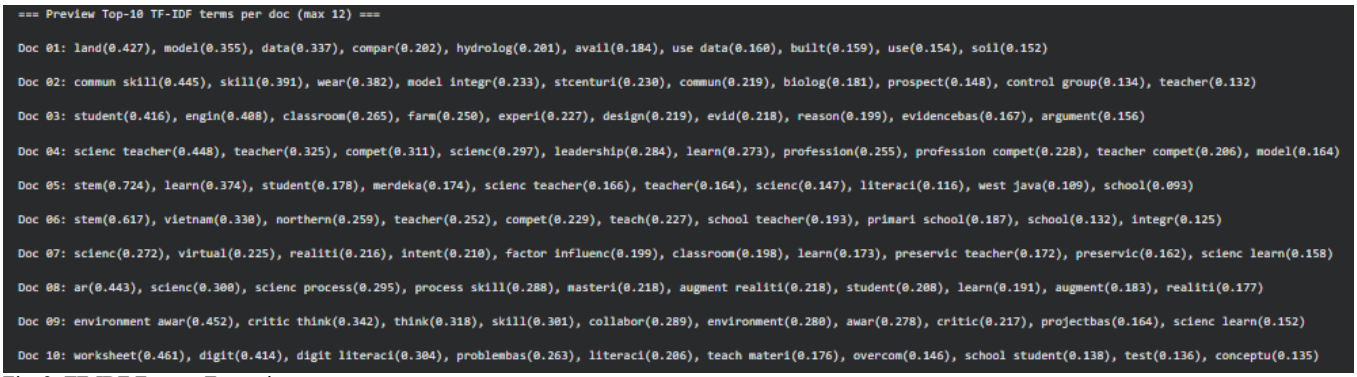


Fig. 2. TF-IDF Feature Extraction

2) Bag of Word

BoW is a text feature representation method that represents documents as a collection of unique words and their frequencies, without regard to word order. This method allows models to understand the distribution of words in documents and distinguish between classes based on the occurrence of certain words. The feature extraction process using BoW on the dataset is shown in Fig 4.

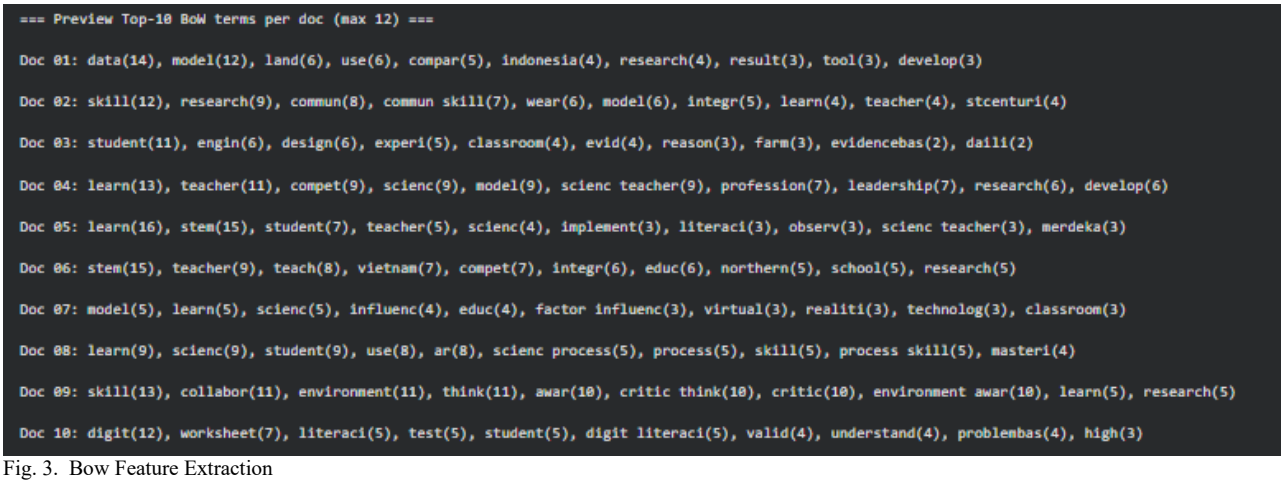


Fig. 3. Bow Feature Extraction

D. Evaluation

After feature extraction using TF-IDF and BoW, the next step is to build and test the classification model. The four model combinations used are TF-IDF + Naïve Bayes, TF-IDF + SVM, BoW + Naïve Bayes, and BoW + SVM. The evaluation was performed using 5-fold cross-validation with an 80:20 data split, and each model was tested five times with different random states. The metric values from the five experiments were then averaged to obtain the average performance of each model, making it easier to compare performance and determine the best model.

Model performance was analyzed using accuracy, precision, recall, and F1-score metrics, which provide an overview of each model's ability to classify articles into the appropriate subject area. This approach allows for a clear comparison between feature representations (TF-IDF vs. BoW) and classification algorithms (NB vs. SVM), and helps determine the most effective combination of models and features.

The results of the evaluation of the four classification models, namely TF-IDF + SVM, TF-IDF + Naïve Bayes, BoW + SVM, and BoW + Naïve Bayes, are presented in Table VIII, Table IX, Table X, and Table XI, respectively.

TABLE VII. CLASSIFICATION REPORT TF-IDF + NAÏVE BAYES

Experiment No.	Accuracy	Precission	Recall	F1-Score
1	0.7828	0.7917	0.7933	0.7845
2	0.8343	0.8288	0.8367	0.8315
3	0.7828	0.7978	0.7817	0.7825
4	0.8200	0.8468	0.8133	0.8266
5	0.8200	0.8468	0.8133	0.8266
Average	0.8080	0.8223	0.8076	0.8103

TABLE VIII. CLASSIFICATION REPORT TF-IDF + SVM

Experiment No.	Accuracy	Precission	Recall	F1-Score
1	0.8171	0.8184	0.8233	0.8175
2	0.8600	0.8549	0.8617	0.8574
3	0.8514	0.8581	0.8483	0.8514
4	0.8571	0.8579	0.8667	0.8605
5	0.8600	0.8705	0.8550	0.8603
Average	0.8491	0.8520	0.8510	0.8494

TABLE IX. CLASSIFICATION REPORT BOW + SVM

Experiment No.	Accuracy	Precission	Recall	F1-Score
1	0.8086	0.8050	0.8150	0.8080
2	0.8486	0.8447	0.8517	0.8470
3	0.8200	0.8239	0.8250	0.8217
4	0.8571	0.8592	0.8650	0.8613
5	0.8086	0.8172	0.8067	0.8108
Average	0.8286	0.8285	0.8326	0.8297

TABLE X.CLASSIFICATION REPORT BOW + NAÏVE BAYES

Experiment No.	Accuracy	Precission	Recall	F1-Score
1	0.7886	0.8021	0.8017	0.7921
2	0.8543	0.8538	0.8600	0.8536
3	0.8286	0.8399	0.8350	0.8310
4	0.8514	0.8582	0.8600	0.8549
5	0.8286	0.8424	0.8433	0.8362
Average	0.8303	0.8392	0.8400	0.8335

Based on the experimental results shown in Tables 2 to 5, the performance of each classification model combination can be analyzed as follows:

EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002041

1) Model Performance Based on Features and Algorithms

The combination of TF-IDF + SVM showed the best performance with an average F1-Score of 0.8494 (Table 9), while BoW + SVM came next with an F1-Score of 0.8297 (Table 10). Naïve Bayes-based models showed lower performance, with TF-IDF + NB achieving an F1-Score of 0.8103 (Table 8) and BoW + NB achieving 0.8335 (Table 11). These results show that TF-IDF feature representation is able to highlight the most relevant words for each class, so that the combination of TF-IDF and SVM provides the best performance on the research dataset. Meanwhile, BoW representation, which retains word frequency directly, still provides good results, but the average is slightly lower than TF-IDF + SVM.

2) Comparison of Feature Representations and Algorithms

Further analysis shows that the SVM algorithm excels in both feature representations, indicating good performance in handling high-dimensional text and multi-class classification (Table 9 and Table 10). Conversely, although Naïve Bayes is faster in training, it tends to be less stable when the distribution of words between classes is uneven, resulting in performance slightly below that of SVM (Table 8 and Table 11). This confirms that the right combination of algorithms and features is an important factor in achieving optimal classification performance.

3) Implications of Results

The results of this evaluation show that the combination of TF-IDF + SVM is the most effective model and feature representation for classifying scientific articles based on subject area in this research dataset. These findings emphasize the importance of selecting feature representations that highlight relevant words and algorithms that are capable of handling multi-class classification effectively. This analysis provides a strong basis for the application of classification models to scientific article datasets and can be used as a reference for further research on similar datasets.

E. Confusion Matrix Analysis

The confusion matrix is used to analyze the distribution of correct and incorrect predictions generated by the classification model for each field of study class. This analysis aims to identify patterns of classification errors that cannot be fully explained by numerical evaluation metrics such as accuracy or F1-score.

The confusion matrix used in this study is the result of aggregating five cross-validation experiments, thus providing a more stable and representative picture of model performance. Visualizations of the confusion matrix for each model combination are shown in Fig. 5 to Fig. 8, which represent the classification results for the TF-IDF + SVM, BoW + SVM, and BoW + Naïve Bayes models.

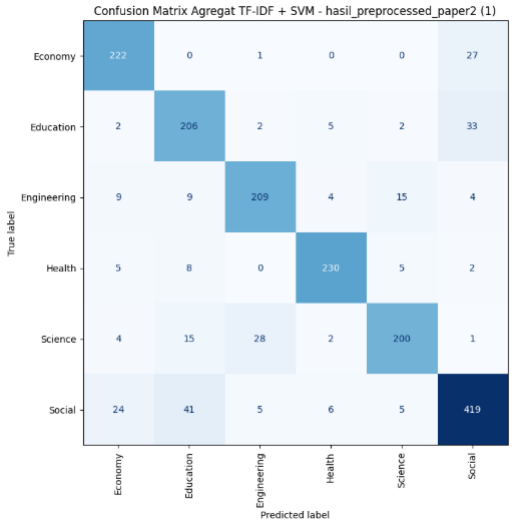


Fig. 4. Confusion Matrix Agregat TF IDF dan SVM

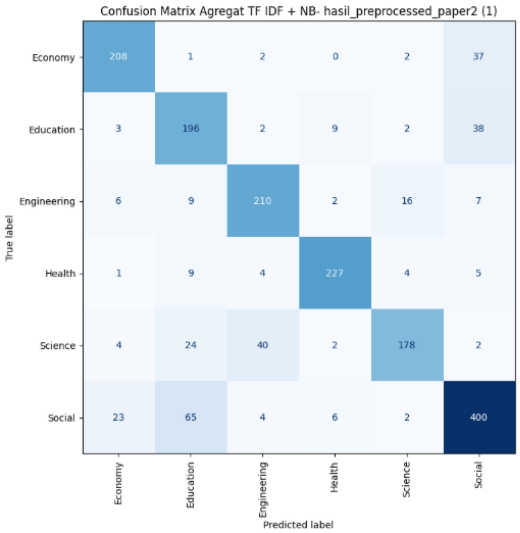


Fig. 5. Confusion Matrix Agregat TF IDF dan Naïve Bayes

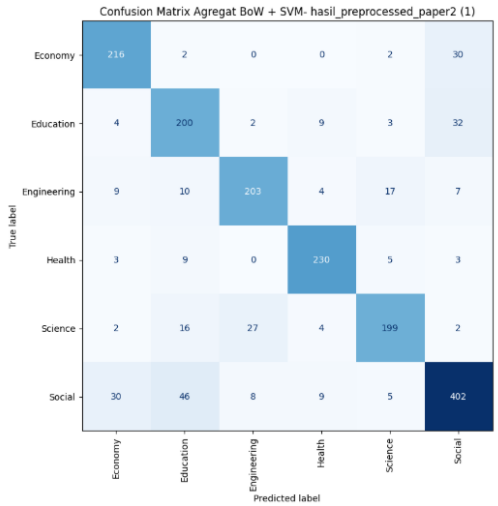


Fig. 6. Confusion Matrix Agregat BoW dan SVM

EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002041

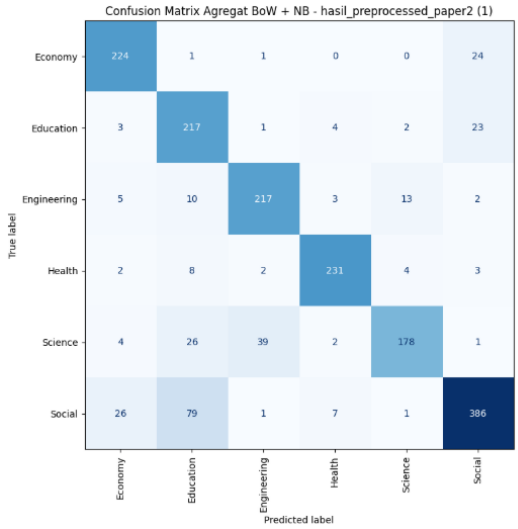


Fig. 7. Confusion Matrix Agregat BoW dan Naïve Bayes

Based on the results of the aggregate confusion matrices shown in Fig. 6 (TF-IDF + Naïve Bayes), Fig. 5 (TF-IDF + SVM), Fig. 7 (BoW + SVM), and Fig. 8 (BoW + Naïve Bayes), it can be observed that all models are generally capable of classifying journal articles effectively, particularly in the dominant class, namely the social class. However, misclassifications are still present, especially among classes with high semantic similarity.

In the TF-IDF + Naïve Bayes model (Fig. 6), misclassification occurs relatively frequently in the Science class, which is often predicted as Engineering, as well as between Education and Social classes. This indicates that Naïve Bayes has limitations in capturing complex dependencies between terms, particularly when dealing with classes that exhibit overlapping vocabulary distributions.The TF-IDF + SVM model (Fig. 5) shows an improvement in classification performance across most classes, particularly Health and Social. Nevertheless, misclassification patterns remain evident, especially in the Science class, which is frequently confused with Engineering and Education, as well as in the social class, which is still misclassified as Education and Economy.

For the BoW + SVM model (Fig. 7), the prediction distribution appears more balanced, with a relatively lower number of misclassifications compared to the TF-IDF-based models. Correct predictions are more dominant in the Social and Health classes, although some confusion between Science and Engineering classes still occurs, albeit in smaller proportions.

Meanwhile, the BoW + Naïve Bayes model (Fig. 8) achieves high correct classification rates in Education, Engineering, and Health classes. However, misclassifications remain prominent in the Science class, which is often predicted as Engineering, and in the social class, which is frequently classified as Education. This suggests that although BoW representation improves Naïve Bayes performance, the assumption of feature independence still limits its ability to distinguish closely related classes.

Overall, the confusion matrix analysis supports the findings from the numerical evaluation reported in the classification metrics. The observed error patterns indicate that classes with semantic proximity or interdisciplinary characteristics, such as Education–Social and Science–Engineering, remain challenging to distinguish. These findings are consistent with previous studies [14], [30], which also report difficulties in classifying overlapping textual categories.

In addition, a comparative analysis among models highlights different strengths. The BoW + Naïve Bayes model achieves the highest average F1-score, indicating that simple frequency-based representations are still highly effective in capturing general patterns in textual metadata. On the other hand, SVM-based models, particularly BoW + SVM and TF-IDF + SVM, demonstrate more stable classification behavior across classes and are more robust in reducing inter-class misclassification. These complementary characteristics align with previous findings [19], which suggest that no single model is universally optimal across all textual classification scenarios.

IV. Conclusion

Based on the experimental results, BoW + Naïve Bayes achieves the highest average F1-score (0.8335), followed by TF-IDF + SVM (0.8494), BoW + SVM (0.8297), and TF-IDF + Naïve Bayes (0.8103). These findings indicate that simpler feature representations such as Bag-of-Words combined with Naïve Bayes can still perform competitively for journal metadata classification. The results also highlight that model effectiveness depends on the interaction between feature representation and classifier characteristics. From a practical perspective, lightweight models can be considered efficient alternatives for text classification in resource-constrained environments.

Acknowledgement

This research is supported by Universitas Ahmad Dahlan.

References

[1] W. E. Savage and A. J. Olejniczak, “More journal articles and fewer books: Publication practices in the social sciences in the 2010’s,” *PLoS One*, vol. 17, no. 2, pp. 1–16, 2022, doi: 10.1371/journal.pone.0263410.

[2] J. J. Mallett, “The resilience of scientific publication : From elite ancient academies to open access,” *Learn. Publ.*, vol. 34, no. 1, pp. 49–56, 2021, doi: 10.1002/leap.1366.

[3] M. A. Hanson, P. G. Barreiro, P. Crosetto, and D. Brockington, “The strain on scientific publishing,” *Quant. Sci. Stud.*, vol. 5, no. 4, pp. 823–843, Nov. 2024, doi: 10.1162/qss_a_00327.

[4] M. Thelwall and P. Sud, “Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals,” *Quant. Sci. Stud.*, vol. 3, no. 1, pp. 37–50, Apr. 2022, doi: 10.1162/qss_a_00177.

[5] O. Fedoruk and S. Lecturer, “Society. Document. Communication,” *Soc. Doc. Commun.*, vol. 10, no. 1, pp. 55–64, 2025, doi: 10.69587/sdc/1.2025.55.

[6] M. Aria, C. Cuccurullo, L. D. Aniello, and M. Misuraca, “Comparative science mapping : a novel conceptual structure analysis with metadata,” *Scientometrics*, vol. 129, no. 11, pp. 7055–7081, 2024, doi: 10.1007/s11192-024-05161-6.

[7] P. Pottier *et al.*, “Title, abstract and keywords: a practical guide to maximize the visibility and impact of academic papers,” *Proc. R. Soc. B Biol. Sci.*, vol. 291, no. 2027, p. 20241222, Jul. 2024, doi: 10.1098/rspb.2024.1222.

[8] G. Mustafa, M. Usman, M. Afzal, A. Shahid, and A. Koubâa, “A Comprehensive Evaluation of Metadata-Based Features to Classify Research Paper’s Topics,” *IEEE Access*, vol. 9, pp. 133500–133509, 2021, doi: 10.1109/access.2021.3115148.

[9] W. I. Al-Obaydy, H. Hashim, Y. A. Najm, and A. Jalal, “Document classification using term frequency-inverse document frequency and K-means clustering,” *Indones. J. Electr. Eng. Comput. Sci.*, 2022, doi: 10.11591/ijeecs.v27.i3.pp1517-1524.

[10] R. C. Morales-hernández and D. Becerra-Alonso, “A Comparison of Multi-Label Text Classification Models in Research Articles Labeled With Sustainable Development Goals,” *IEEE Access*, vol. 10, no. November, 2022.

[11] S. Rahman, H. K. Shanto, U. A. Koana, and S. M. Danish, “Automated Research Article Classification and Recommendation Using NLP and Machine Learning,” *arXiv*, vol. 1, 2025.

[12] A. M. Chaid, Z. A. Abdulrazzaq, R. N. Sadoon, and M. A. Aljabery, "Comparative Analysis of Innovative Machine Learning Algorithms : Advancements in Natural Language Processing," *J. Inf. Syst. Eng. Manag.*, vol. 10, 2025.
[13] N. Susetyo, F. Putri, A. Prasetya, H. Ar, and A. Nur, "Classification of Engineering Journals Quartile using Various Supervised Learning Models," *Ilk. J. Ilm.*, vol. 15, no. 1, pp. 101–106, 2023.
[14] R. Guns, "Fine-grained classification of social science journal articles using textual data : A comparison of supervised machine learning approaches," *Quant. Sci. Stud.*, vol. 2, no. 1, 2021, doi: 10.1162/qss.
[15] M. Lu, L. Tang, and X. Zhou, "An ensemble approach for research article classi fi cation : a case study in arti fi cial intelligence," *PeerJ Comput. Sci.*, pp. 1–22, 2024, doi: 10.7717/peerj-cs.2521.
[16] L. P. F. Garcia, J. P. Mena-Chalco, R. L. Grando, W. M. C. da Silva, F. P. Guimarães, and V. de A. Jorge, "Hierarchical Article Classification: A Multi-Level Framework for Organizing Scholarly Literature," *IEEE Access*, vol. 13, no. May, pp. 102589–102601, 2025, doi: 10.1109/ACCESS.2025.3579232.
[17] T. Gupta, M. Zaki, N. M. A. Krishnan, and Mausam, "MatSciBERT: A materials domain language model for text mining and information extraction," *npj Comput. Mater.*, vol. 8, no. 1, p. 102, 2022, doi: 10.1038/s41524-022-00784-w.
[18] S. Aum and S. Choe, "srBERT : automatic article classification model for systematic review using BERT," *Syst. Rev.*, pp. 1–8, 2021, doi: 10.1186/s13643-021-01763-w.
[19] C. Arhiliuc, R. Guns, W. Daelemans, and T. C. E. Engels, "Journal article classification using abstracts: a comparison of classical and transformer-based machine learning methods," *Scientometrics*, vol. 130, no. 1, pp. 313–342, 2025, doi: 10.1007/s11192-024-05217-7.
[20] F. Aguilar-canto, C. Macias, A. E. Juárez, M. A. Cardoso-moreno, and H. Calvo, "Quartile Prediction and Journal Recommendation Using Deep Learning Models for Artificial Intelligence Articles," *J. Scientometr. Res.*, vol. 14, no. 1, pp. 373–382, 2025, doi: 10.5530/jscires.20251460.
[21] M. Sciences, "Exploring Preprocessing Techniques for Natural Language Text : A Comprehensive Study sing Python Code," *Int. J. Eng. Technol. Manag. Sci.*, vol. 5, no. 7, pp. 390–399, 2023, doi: 10.46647/ijetms.2023.v07i05.047.
[22] W. Costa and G. V. Pedrosa, *A Textual Representation Based on Bag-of-Concepts and Thesaurus for Legal Information Retrieval*. 2022. doi: 10.5753/kdmile.2022.227779.
[23] D. Sugiarto, E. Utami, and A. Yaqin, "Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng," vol. 12, no. 3, pp. 272–277, 2022.
[24] K. Juluru, H.-H. Shih, K. N. Keshava Murthy, and P. Elnajjar, "Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists," *RadioGraphics*, vol. 41, no. 5, pp. 1420–1426, Aug. 2021, doi: 10.1148/rg.2021210025.
[25] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, "Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports," *Math. Probl. Eng.*, vol. 2021, no. ii, 2021, doi: 10.1155/2021/6619088.
[26] R. Blanquero, E. Carrizosa, and P. Ramírez-cobo, "Computers and Operations Research Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, 2021, doi: 10.1016/j.cor.2021.105456.
[27] T. Simbolon, A. Prasetya, I. Ari, E. Zaeni, and A. Ritahani, "Text classification of traditional and national songs using naïve bayes algorithm," *Sci. Inf. Technol. Lett.*, vol. 3, no. 2, pp. 59–72, 2022.
[28] D. M. Abdullah and A. M. Abdulazeez, "Machine Learning Applications based on SVM Classification : A Review," *Qubahan Acad. J.*, pp. 81–90, 2021, doi: 10.48161/Issn.2709-8206.
[29] G. D. Bispo *et al.*, "Applied Sciences Automatic Literature Mapping Selection : Classification of Papers on Industry Productivity," *Appl. Sci.*, vol. 14, no. 9, p. 3679, 2024, doi: https://doi.org/10.3390/app14093679.
[30] A. Lyutov, Y. Uygun, and M. T. Hütt, "Machine learning misclassification networks reveal a citation advantage of interdisciplinary publications only in high-impact journals," *Sci. Rep.*, vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-72364-5.

 EISSN: 2961-872X ●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002041