Topic Classification Based on Scientific Article Structure: A Case Study at Can Tho University Journal of Science

Hai Thanh Nguyen, Tuyet Ngoc Huynh, Anh Duy Le, and Tran Thanh Dien(✉)

Can Tho University, Can Tho, Vietnam
{nthai.cit,ldanh,thanhdien}@ctu.edu.vn

Abstract. With a massive amount of stored articles, text-based topic classification plays a vital role in enhancing the document management efficiency of scientific journals. The articles can be found faster by filtering out the appropriate topic and speeding up to determine appropriate reviewers for the review phase. In addition, it can be beneficial to recommend related articles for the considered manuscript. However, fetching entire documents for the process can consume much time. Especially, Can Tho University Journal of Science (CTUJS) is a multidisciplinary journal with many topics. Therefore, it is necessary to evaluate various common structures in an article. Extracted sections can be short but efficient in determining the article's topic. In this study, we explore and analyze the paper structure of articles obtained from CTUJS for topic classification using Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB). The results show that Random Forest outperforms Naïve Bayes and SVM regarding performance and training time. As shown, the topic classification performance based on the section of "Method" can reach 0.53 compared to the whole content of the paper with 0.61 in accuracy.

Keywords: Paper Structure · Scientific journal · Topic classification

1 Introduction

Text classification has been the most widespread problem of NLP, and is often used for spam detection [1], topic classification, etc. In scientific journals, managing many articles requires a scientific and time-saving management process. Automatic classification of topics is quite essential to help manage articles easier. A faster search based on selected topics to filter and label topics before detecting similar text saves processing costs than scanning the entire data warehouse. Review is an indispensable process for a scientific article to be approved for publication. The classification of documents by topic helps the article be evaluated by the right reviewers in the main areas of the article, improving high-quality articles.

N. Thai-Nghe et al. (Eds.): ISDS 2023, CCIS 1949, pp. 208–215, 2024.
https://doi.org/10.1007/978-981-99-7649-2_16

The case study at CTUJS is a multidisciplinary journal that provides information on Can Tho University's scientific research and introduces domestic and foreign scientific research. Therefore, there are both English and Vietnamese scientific articles. In the previous work, there have been studies on automatic topic classification. However, the classification results in English documents are rather positive, and for Vietnamese, it is still limited, and no feasible results have been found. Therefore, this study proposes a method to classify topics based on the structure of scientific articles in Vietnamese and English.

2 Related Work

Text classification in many studies aimed to efficiently find the necessary documents and save time than searching for irrelevant data. The authors in [2] also proposed topic classification approaches through Support Vector Machines, Naïve Bayes, and k-Nearest Neighbors and preprocessed input data, extracting information, and vectorizing. According to [3], authors described some main steps in text classification, including document preprocessing, feature extraction/selection, model selection, training, and classifier testing. The text preprocessing stage converts the original textual data to a raw data structure, where the most significant text features that distinguish between text categories are identified [4]. The study in [5] presented a method to indicate similar Vietnamese documents from English Documents.

The work in [6] investigated the process of text classification, the process of different methods of weighing and measuring terms, and compares other classification techniques, including Naïve Bayes classifiers (NBC), SVM, and k-Nearest Neighbors (kNN), KNC [7] - a combination of KNN and other three classifiers (C4.5 algorithm, NBC and SVM), etc. According to the authors in [8], Naïve Bayes is quite effective in data mining tasks, but it gives pretty bad results when used for automatic text classification tasks.

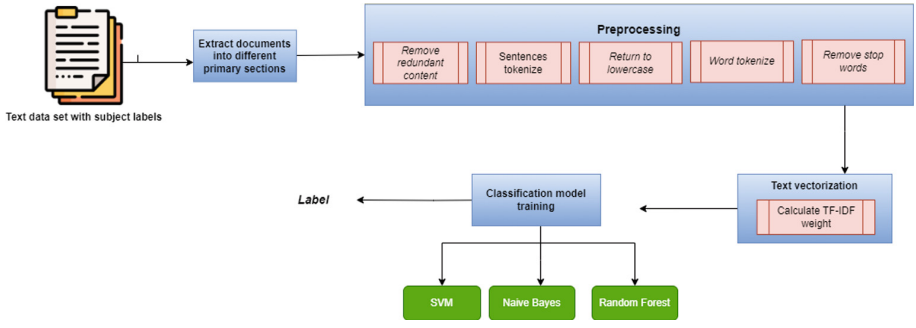


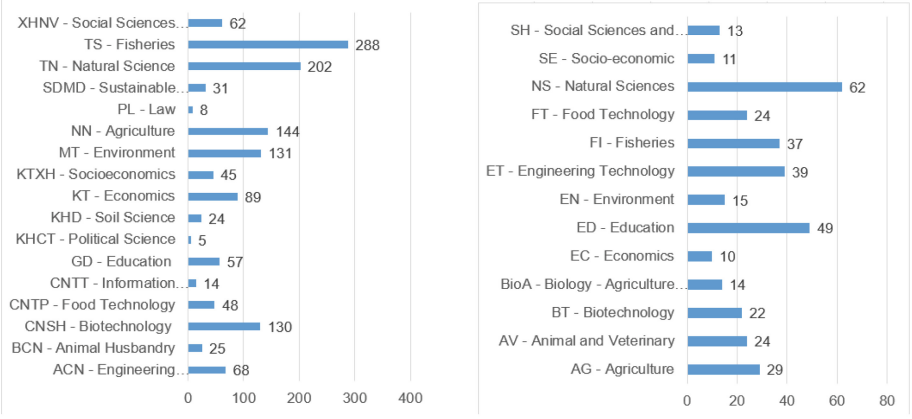
Fig. 1. The main steps in topic classification tasks based on article structure.

3 Methods

Figure 1 illustrates several of the main steps for topic classification based on article structure with the main steps as follows. Step 1 collects articles, each with a corresponding topic. Input data are Vietnamese documents presented based on the labeled scientific paper structure. In Step 2, we extract the documents with various primary sections of typical article structure such as Abstract, Introduction, Method, Results (Experiments), and Conclusion. Preprocessing data by sentence splitting, word splitting, removing redundant characters in sentences (such as punctuation marks, mathematical formulas, etc.), removing stop words, etc. are performed in Step 3. In Step 4, After preprocessing, we vectorize the data set corresponding to an article that will be a feature vector, including the weights of words. In step 5, we leverage some classification models. After preprocessing and vectorizing each document, we perform the classification tasks. In the next section, we will detail the steps above.

3.1 Data Collection

This is one of the essential process steps in the text classification model. It determines the complete elements of the text classification system. Experimental data were collected from the system of CTUJS[1]. The collected data of Vietnamese scientific articles include 1371 articles, divided into 17 different topics described in Fig. 2a.



(a) Vietnamese articles (b) English articles

Fig. 2. Topic distribution of Vietnamese and English articles

The English documents for classification include 348 articles, divided into 13 different topics described in Fig. 2b. In this study, the data collected was not

[1] https://ctujs.ctu.edu.vn/.

evenly distributed among the topics. The number of Law or Political Science papers in the Vietnamese data section is less than ten. In contrast, the number of papers on others, such as Fisheries or Agriculture, is much higher than others, including 144 articles on Agriculture and 288 articles on Fisheries. As for the English data section, topics such as Information Technology or Political Science have very few samples, with only 1–2 articles. The difference in the number of articles can lead to the situation that topics with too few articles will have low accuracy. In contrast, topics with a rich number of articles will give high accuracy. Therefore, to avoid reducing the classification efficiency, this study omitted topics with less than three articles.

3.2 Extracting Article Structure

As the guidelines from Springer Nature publisher [9], they described a standard structure of the body of research papers, including four primary sections: Introduction, Materials and Methods, Results, Discussion, and Conclusions. Many articles have followed these guidelines. Such structures are prevalent in most scientific journals. We briefly present "Materials and Methods" as "Method" and "Discussion and Conclusions" as "Conclusion" in the Experiment section. Besides, we also consider "Abstract" in classification tasks. For articles published in the CTUJS, although most of the articles have the standard structure as above, some papers may have sections with slight differences in names. In this study, we analyzed the published articles' structure in CTUJS and observed the positions of each section. The Abstract section is a summary of the article and is named "Abstract", placed at the top of the article. The introduction is usually located at the beginning of the paper, right after the Abstract, and usually gives an overview of the main problem that the research paper needs to solve. Methods are usually techniques and approaches to problem-solving, usually after the "Introduction". Results section is often called "Results" or "Experimental Results", presenting the results obtained after applying the proposed method, usually right after the Methods section. The Conclusion section, which is usually called "Conclusion" or "Discussion and Conclusion", is the conclusion of the article (before References). The content of this section is to discuss the results of the problem and summarize the main ideas of the study and future directions. Based on the above characteristics, we proceed to extract the sections for the classification task.

3.3 Data Preprocessing

Data preprocessing in natural language processing is an important step that affects the performance and accuracy of the classification model. In this problem, data (which are structured articles of scientific articles collected from CTUJS) will be preprocessed through the following steps:

– **Extract text:** We remove distracting content such as author information, title, etc. Only keep the main content, including the summary and main content of the article (from introduction to conclusion).

- **Split sentences:** After removing redundant content, proceed to split the document into a list of sentences, keeping only sentences with a length of more than four words (because often these sentences are meaningless).
- **Remove special characters and math formulas:** Keep only Alphabet characters, and remove special characters, math characters, and punctuation marks.
- **Convert content to lowercase:** Due to the nature of the computer processor, it will understand uppercase and lowercase letters differently, so it is necessary to convert content to lowercase to reduce the number of features, increase the accuracy in the classifying topics.
- **Split words**: use the Underthesea library[2] to separate words in Vietnamese text - this is a relatively popular Vietnamese processing support library.
- **Eliminate stop words (stop words):** Stop words appear frequently but are meaningless in a sentence. Removing the stop word increases the performance and reduces the number of features of the text topic classification model. The classification model will remove the 100 most frequently occurring words in the dataset to ensure classification accuracy.

3.4 Classification Algorithms

We conduct a topic classification test using three classification algorithms: SVM, Naive Bayes, and Random Forest. Naive Bayes classifier [10] is an algorithm belonging to the class of statistical algorithms. SVM [11] is a supervised learning algorithm that can be used for either classification or prediction tasks. In addition, we also perform the classification with Random Forest. Random forest [12] is a popular machine learning algorithm belonging to supervised learning techniques, widely used in the ML field's classification and regression problems. We use the default hyperparameters for the Naïve Bayes model, for SVM we train the model with C=1.0 and kernel=sigmoid, while for RF we run with n_estimators = 100. All three models are chosen max_df = 0.8

4 Experimental Results

The scenarios include a test script for word separation and topic classification of Vietnamese and English scientific articles using three techniques SVM, Naïve Bayes, and Random Forest.

4.1 Environmental Settings and Dataset

In both Vietnamese and English data sets, they are divided into two parts: 80% for the training set and 20% for the test set, using Stratified sampling. This study used the data of scientific articles in English and Vietnamese collected from CTUJS for classification. The collected Vietnamese scientific article data includes 1371 articles, divided into 17 different topics, and the English data for classification includes 348 articles, divided into 13 different topics. Detailed data distribution has been mentioned in Sect. 3.1 of this study.

[2] https://github.com/undertheseanlp/underthesea.

4.2 Experimental Results

We conduct problem classification on three proposed models on data from Vietnamese and English articles. Table 1 summarizes the three algorithms' classification performance results of Vietnamese/English articles. Although Naïve Bayes has the fastest training time, It has the lowest accuracy on both the training set and test set. The accuracy on the SVM and Random Forest test sets is quite different, 60% and 62%, respectively, but SVM has the disadvantage that it takes longer to train. In the case of English articles, Random Forest has the highest accuracy on the test set of the reviewed models (55%) and takes 0.92 s to train. SVM and Naïve Bayes have relatively low accuracy, and Naïve Bayes has the fastest training speed. In general, the accuracy of English data is relatively low, partly due to the limited number of samples.

Table 1. Topic classification performance with algorithms on Vietnamese articles

	Algorithm	Accuracy	Training time (seconds)
Vietnamese articles	Naïve Bayes	0.40	1.56
	SVM	0.60	14.12
	Random Forest	0.62	3.58
English articles	Naïve Bayes	0.41	0.45
	SVM	0.41	1.19
	Random Forest	0.55	0.92

In the second scenario, the data to experiment is to separate English scientific articles into five parts: Abstract, Introduction, Method, Result, and Conclusion to assess the classification ability based on each component of a text-structured article. The classification model used is Random Forest as reported in Table Table 1. Table 2 presents the performance of each topic based on different sections using the Random Forest algorithm. Fisheries and Education topics that have high classification efficiency in all sections, from 72–93%. The classification based on the whole paper gives the highest accuracy. However, it takes longer to train than other sections. If we ignore entire-paper-based classification because of the long time for processing, the classification based on the "Method" section can also be a much better alternative than the other sections.

4.3 Discussion

Through the experimental results above, it can be concluded that Random Forest outperforms the models considered in this study. Accuracy, F1 score in each topic, and training time proved the model's effectiveness. Moreover, under the same training conditions, with a limited number of samples such as datasets of English scientific papers, Naïve Bayes and SVM also give much worse classification performance than Random Forest. Random Forest is known as an ensemble

Table 2. Topic classification Performance Comparison in F1-scores on English articles using Random Forest. Some rows include Average (AVG) Accuracy (ACC) on all topics, macro AVG, and weighted AVG, respectively.

Topic	Abstract	Introduction	Method	Result	Conclusion	All sections
AG	0.00	0.29	0.00	0.00	0.67	0.60
AV	0.00	0.75	0.57	0.33	0.75	0.75
BT	0.40	0.40	0.75	0.33	0.40	0.75
BioA	0.00	0.00	0.00	0.00	0.00	0.50
EC	0.00	0.67	0.00	0.00	0.00	0.00
ED	0.82	0.83	0.76	0.78	0.72	0.80
EN	0.00	0.00	0.00	0.00	0.00	0.00
ET	0.18	0.17	0.50	0.25	0.20	0.46
FI	0.78	0.78	0.70	0.53	0.86	0.93
FT	0.29	0.33	0.67	0.33	0.33	0.33
NS	0.47	0.39	0.53	0.44	0.41	0.62
SE	0.00	0.00	0.00	0.00	0.00	0.00
SH	0.00	0.00	0.00	0.00	0.00	0.00
AVG ACC	0.44	0.49	0.53	0.41	0.49	0.61
Macro AVG	0.23	0.35	0.34	0.23	0.33	0.44
Weighted AVG	0.34	0.43	0.46	0.34	0.44	0.56
Training time (s)	0.52	0.47	0.58	0.60	0.38	0.80

approach [12] that can integrate and consider multiple decision trees to select the best one, while SVM and Naïve Bayes are single models. Therefore, in many cases, SVM and Naïve Bayes may not resolve complex relationships in the content of the articles as effectively as the ensemble-based techniques such as Random Forest. Therefore, Random Forest can be applied in severe data shortage conditions and still maintain accuracy. Moreover, "Method" is a crucial section in a typical research article. It may contain special terms which can be key to discriminating the research topic, so topic classification based on "Method" can obtain the best result among all considered sections. "Conclusion" can give more information on future directions compared to "Abstract", hence, it achieves better performance than "Abstract". In addition, "Introduction" which usually reveals the specific context of the research also exhibits an informative section to indicate the topic of the article.

5 Conclusion

In this study, we collected 1371 scientific papers in Vietnamese and 348 in English from the CTUJS to serve the work of topic classification. The results show that the Random Forest model has a better classification ability than the SVM or Naïve Bayes models, even for missing data. Extracting sections in articles gives very positive results, improving time efficiency in topic classification. Giving users more options is to choose one of the elements in an article for quick classification. Besides, training time is also an advantage of Random Forest. In

addition, we analyzed the potential benefits of each separated section in a typical research article structure to perform the topic classification. As shown from the experiments, the performances vary depending on the topic. The section "Method" can discriminate the topic of articles, while "Introduction" obtained approximate performance compared to "Conclusion".

References

1. Ghanem, R., Erbay, H.: Spam detection on social networks using deep contextualized word representation. Multimedia Tools Appl. **82**(3), 3697–3712 (2022). https://doi.org/10.1007/s11042-022-13397-8
2. Dien, T.T., Loc, B.H., Thai-Nghe, N.: Article classification using natural language processing and machine learning. In: 2019 International Conference on Advanced Computing and Applications (ACOMP). IEEE (2019). https://doi.org/10.1109/ACOMP.2019.00019
3. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49 (1999)
4. Kadhim, A.I.: An evaluation of preprocessing techniques for text classification. Int. J. Comput. Sci. Inf. Secur. (IJCSIS) **16**(6), 22–32 (2018)
5. Nguyen, H.T., Le, A.D., Thai-Nghe, N., Dien, T.T.: An approach for similarity vietnamese documents detection from English documents. In: Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications, pp. 574–587. Springer Nature Singapore (2022). https://doi.org/10.1007/978-981-19-8069-5_39
6. Kadhim, A.I.: Survey on supervised machine learning techniques for automatic text classification. Artif. Intell. Rev. **52**(1), 273–292 (2019). https://doi.org/10.1007/s10462-018-09677-1
7. Hao, P., Ying, D., Longyuan, T.: Application for web text categorization based on support vector machine. In: 2009 International Forum on Computer Science-Technology and Applications. IEEE (2009). https://doi.org/10.1109/ifcsta.2009.132
8. Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: Some effective techniques for naive bayes text classification. IEEE Trans. Knowl. Data Eng. **18**(11), 1457–1466 (2006). https://doi.org/10.1109/tkde.2006.180
9. Structuring your manuscript | Springer - International Publisher, https://www.springer.com/gp/authors-editors/authorandreviewertutorials/writing-a-journal-manuscript/author-academy/10534936
10. Webb, G.I., Keogh, E., Miikkulainen, R., Miikkulainen, R., Sebag, M.: Naïve bayes. In: Encyclopedia of Machine Learning, pp. 713–714. Springer, US (2011). https://doi.org/10.1007/978-0-387-30164-8_576
11. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
12. Ho, T.K.: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282. IEEE (1995)