ARTICLE

Exploring Final Project Trends Utilizing Nuclear Knowledge Taxonomy

An Approach Using Text Mining

Faizhal Arif Santosa

ABSTRACT

The National Nuclear Energy Agency of Indonesia (BATAN) taxonomy is a nuclear competence field organized into six categories. The Polytechnic Institute of Nuclear Technology, as an institution of nuclear education, faces a challenge in organizing student publications according to the fields in the BATAN taxonomy, especially in the library. The goal of this research is to determine the most efficient automatic document classification model using text mining to categorize student final project documents in Indonesian and monitor the development of the nuclear field in each category. The kNN algorithm is used to classify documents and identify the best model by comparing Cosine Similarity, Correlation Similarity, and Dice Similarity, along with vector creation binary term occurrence and TF-IDF. A total of 99 documents labeled as reference data were obtained from the BATAN repository, and 536 unlabeled final project documents were prepared for prediction. In this study, several text mining approaches such as stem, stop words filter, n-grams, and filter by length were utilized. The number of k is 4, with Cosine-binary being the best model with an accuracy value of 97 percent, and kNN works optimally when working with binary term occurrence in Indonesian language documents when compared to TF-IDF. Engineering of Nuclear Devices and Facilities is the most popular field among students, while Management is the least preferred. However, Isotopes and Radiation are the most prominent fields in Nuclear Technochemistry. Text mining can assist librarians in grouping documents based on specific criteria. There is also the possibility of observing the evolution of each existing category based on the increase of documents and the application of similar methods in various circumstances. Because of the curriculum and courses given, the growth of each discipline of nuclear science in the study program is different and varied.

INTRODUCTION

The National Nuclear Energy Agency of Indonesia (BATAN), now known as the Research Organization for Nuclear Energy (ORTN)—National Research and Innovation Agency (BRIN), in 2018 issued a decision regarding BATAN's six competencies: Isotopes and Radiation (IR), Nuclear Fuel Cycle and Advanced Materials (NFCAM), Engineering of Nuclear Devices and Facilities (ENDF), Nuclear Reactor (NR), Nuclear and Radiation Safety and Security (NRSS), and Management (Mgt). These areas of focus are also known as BATAN's knowledge taxonomy, which is used to support Nuclear Knowledge Management (NKM) and the grouping of explicit knowledge in repositories.[1]

The Polytechnic Institute of Nuclear Technology (PINT), which is under the auspices of BATAN and is now in one of the directorates of BRIN, can also utilize BATAN's knowledge taxonomy to classify students' final assignments. Every year the PINT Library accepts final assignments from

Faizhal Arif Santosa (faizhalarif@gmail.com) is Academic Librarian, Polytechnic Institute of Nuclear Technology, National Research and Innovation Agency. © 2022.

https://doi.org/10.6017/ital.v42i1.15603

students who have graduated from three study programs, namely Nuclear Technochemistry, Electronics Instrumentation, and Electromechanics. Over the past six years (2017 to 2022), 563 final assignments in Indonesian were collected and needed to be classified into the BATAN's knowledge taxonomy in order to see the document growth of each existing competency. However, it is quite time consuming for librarians to assign individual documents to the most appropriate taxonomy term. It is also possible to involve experts to determine the right group, which results in increased working time to complete a document. This obstacle arises because librarians do not have in-depth and detailed knowledge of the nuclear field so it is feared that grouping errors will occur.

In this study, the author tried to classify the collection of final project documents owned by the PINT Library based on BATAN's knowledge taxonomy. The author used text mining tools, choosing the k-nearest neighbors (kNN) algorithm for this study. Similar research also leads to trying to focus on automatic document classification of certain subjects,[2] which in this case is the subject of nuclear engineering. The hope is that users will find it easier to explore knowledge according to their area of interest through taxonomy grouping based on explicit knowledge,[3] in this case, PINT students' final project documents. Finding the trend of research conducted by students on each subject is also one of the goals of this research.

LITERATURE REVIEW

Text Mining in Libraries

The increasing number of publications currently makes it a challenge to classify and find out the growth and trends of a topic. Document classification is one of the jobs that is quite time consuming so document classification automation by utilizing text mining is very necessary.[4] The application and utilization of text mining itself is very broad. Several studies have demonstrated the usefulness of text mining in libraries.

Pong et al. from City University of Hong Kong conducted research to facilitate the classification process using machine learning.[5] This study aimed to streamline document categorization utilizing automatic document classification by using a system called the web-based automatic document classification system (WADCS) and claimed to be the pioneer of a comprehensive study of automatic document classification on a classification that is already popular in the world, namely the Library of Congress Classification (LCC) utilizing kNN and naive Bayes (NB). This research indicates that the machine-learning algorithm they used can be applied by the library for document classification.

Wagstaff and Liu utilized text mining to perform automatic classification to help make decisions to select candidate documents for weeding.[6] This study used data from Wesleyan University from 2011 to 2014 to predict which documents were eligible for weeding and which will be stored. Five classifier models, namely kNN, naive Bayes, decision tree, random forest, and support vector machines (SVM), were used to compare their performance. While this process may not replace librarians, this study can help librarians make better decisions and reduce their workload significantly.

Lamba and Madhusudhan applied the use of text mining to extract important topics which were published in the *DESIDOC Journal of Library and Information Technology* over a period of 38 years.[7] The Latent Dirichlet Allocation (LDA) method used in this study is able to find topics from within a collection of documents so that they can see how these topics develop over time. Because LDA is an algorithm for looking at topics from a group of words that appear together, the authors suggest that this study be expanded by utilizing articles that have been labeled using supervised classification.

kNN Classifier

Various studies try to find answers to the most appropriate method of grouping the collection of documents. The kNN and SVM algorithms were used as comparative methods in the document classification study.[8] However, there is no definite standard for the methods used in text mining.[9] Choosing the right technique in each phase of document classification can improve the performance of the text classifier, so, experts generally make adjustments to existing methods to get better results.[10]

Kim and Choi compared kNN, Maximum Entropy Model (MEM), and SVM to classify Japanese patent documents by focusing on the structure of patents.[11] Instead of comparing the entire text, specific components named semantic elements, such as purpose, background, and application fields, are compared from the training document. These semantically grouped components are the basis for patent categorization. In addition, the strategy used is the existence of cross-references from two semantic fields that are useful for determining the intentions of the patent writers who are still unsure or hidden. This strategy works well on kNN compared to MEM and SVM where SVM doesn't do very well when handling large data sets. However, research conducted by Alhaj et al. on Arabic documents showed that SVM can outperform kNN by implementing a stemming strategy.[12] Meanwhile, through the approach to the relationship between unstructured text documents, the study conducted by Mona et al. was able to increase the performance of kNN combined with TF-IDF by 5 percent.[13]

The kNN algorithm is one of the popular classifiers that categorizes new data based on the concept of similarity from the amount of data (determined by the specified "k" value) around it.[14] This method is believed to be able to group documents effectively because it is not limited to the number of vector sizes.[15] Wagstaff and Liu noted that one of the weaknesses of kNN is the long processing time when faced with large datasets, but kNN as a classifier is easy to apply.[16] In terms of measurement, previous experiments showed that kNN was not suitable when used with Euclidean distance.[17] Generally, similarity measures such as Cosine, Jaccard, and Dice were used in the kNN classifier.[18]

One of the problems in text classification is the number of attributes or dimensions so that many irrelevant attributes in the data set cause the classifier's performance to not run optimally.[19] For this reason, it is necessary to have a technique to increase effectiveness and reduce dimensions that are too large through the selection of features or terms,[20] such as within-document TF, weighting with one of the popular methods, namely TF-IDF (which sees how important a word is in a collection of corpus),[21] and binary representation which looks at the absence and presence of a concept in a document[22] by converting it to 0 and 1.[23]

Aims of the Study

University libraries have a vital role in managing internal publications to support the education ecosystem. In connection with the role of the PINT to support NKM and nuclear development, it is necessary to apply technology to help provide advice on certain classes of documents. In addition, in order to see scientific developments, generally experts conduct bibliometric studies which are limited to the title and abstract fields. Text mining provides an opportunity to dig deeper. Instead of just the title and abstract, this study used the full text of the final project collection. The trend of a subject will be seen from the growth and percentage of existing documents. So, the objectives of this study are to

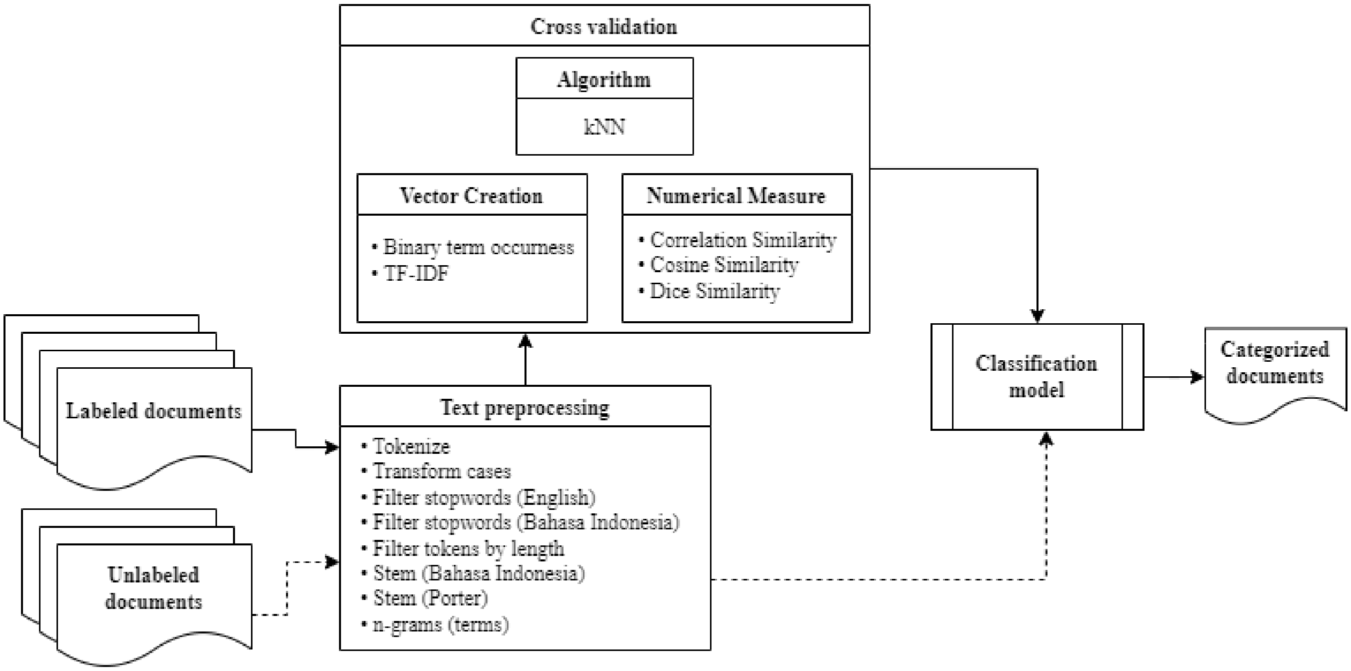
- explore the best kNN model to be applied to classify the final project;
- know the development of nuclear subjects based on BATAN's knowledge taxonomy; and
- know the development of nuclear subjects from each study program at the PINT.

METHODS

A total of 99 documents were taken from the BATAN repository and manually labeled as reference data. This study was conducted using RapidMiner Studio software. The first document processing method is to convert all words into lower case and divide the text into a collection of tokens. Filters on tokens are also applied based on the length of the token. In this case, the author applied a minimum of 3 characters and a maximum of 25 characters. Stop words were also applied to eliminate short words (e.g., "and," "the," and "are"), thereby reducing the vector size. English and Indonesian stop words were used for this study to overcome the use of English in the abstract section and Indonesian as the document language. The collection of words from Haryalesmana was chosen to be the stop words for Indonesian.[24] The stemming technique is applied to reduce dimensions that are useful for improving the function of the classification system[25] by changing word forms into basic word,[26] e.g., water, waters, watered, and watering into water. This analysis applies Wicaksana data to Indonesian stemming.[27]

Some words cannot be separated from other words because they form a meaning, e.g., nondestructive testing, biological radiation effects, structural chemical analysis, and water-cooled reactors. To overcome this case, the use of n-grams can help identify compound words that have a meaning so that the words are not reduced.[28] N-grams will record a number of "n" words that follow the previous word.[29] To accommodate these words, in this study, three words were assigned to n-grams.

Figure 1. Nuclear taxonomy classification framework.



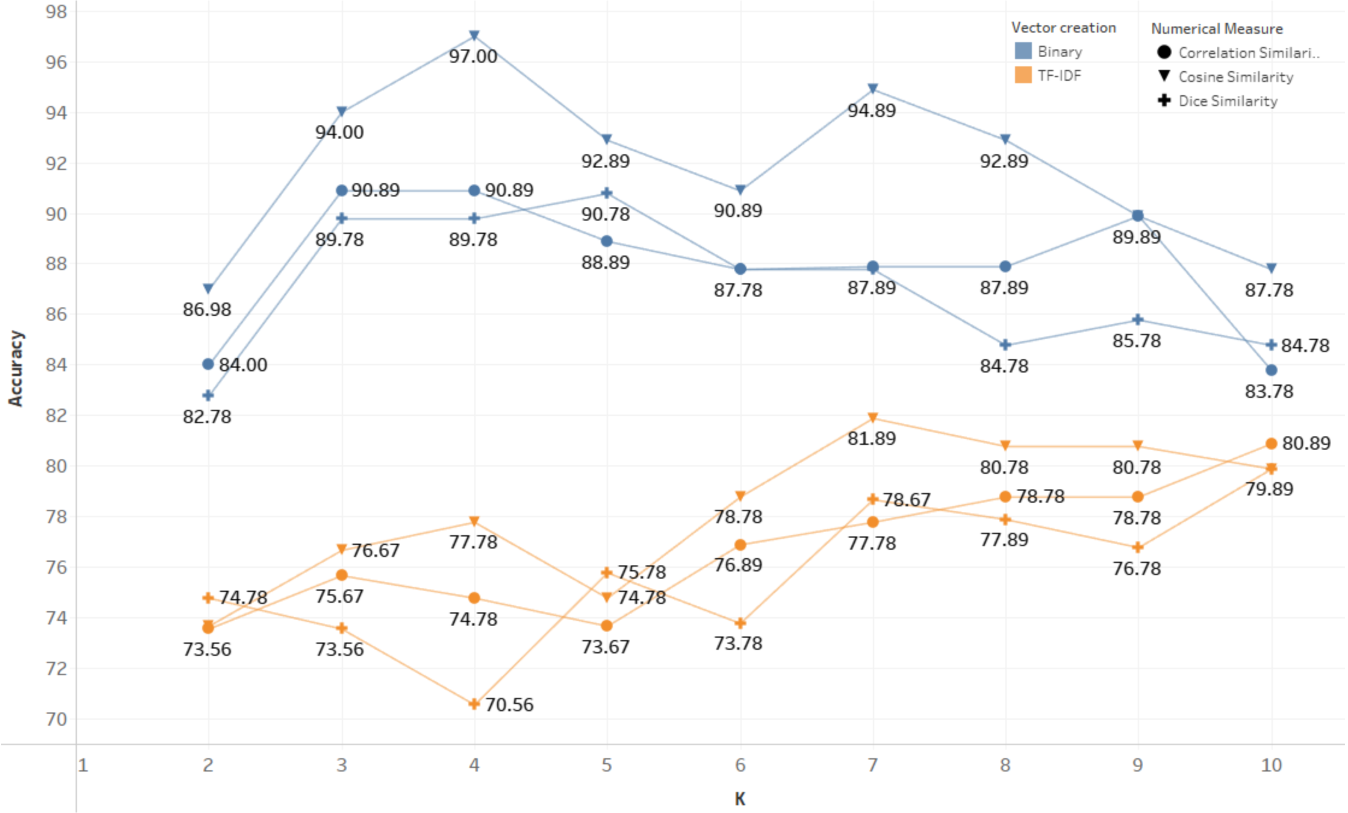
Vector creation in this study used TF-IDF and binary term occurrence and then compared them to determine the best performance. In the kNN method, it is necessary to determine the value of "k" manually, so a value of 2–10 was chosen by activating a weighted vote which is useful for weighing the contributions of neighbors in the vicinity. Weight voting indicates the use of multiple voting methods by assigning a weight to each neighbor depending on their distance from the unknown item.[30] The types of measurement chosen to get maximum results were Numerical Measure and tested Cosine Similarity, Correlation Similarity, and Dice Similarity. Meanwhile, to measure performance, the author used cross validation with a number of folds of 10. Then, using this set of procedures, documents from the BATAN repository are classified. The procedure that achieves the highest level of accuracy is then submitted as a model. This model was applied to 563 final project documents that have not been labeled so that each document has a label according to BATAN's knowledge taxonomy.

RESULTS

The experiment was carried out 54 times to determine the best kNN performance from the proposed approach, namely Cosine-binary, Correlation-binary, Dice-binary, Cosine–TF-IDF, Correlation–TF-IDF, and Dice–TF-IDF utilizing Cross validation. Cosine was still the most accurate in the TF-IDF vector creation process, with an accuracy of 81.89 percent on seven neighbors, and Dice reaches the lowest point when used on four neighbors. In contrast to Correlation and Dice, Cosine can perform well when creating binary vectors. Cosine on four neighbors had the best performance, with a 97 percent accuracy rate. The lowest accuracy occurred when the number of selected neighbors was two and the overall numerical measure had decreased in neighbors more than nine.

The classification model for unlabeled documents was determined to be the Cosine-binary method with four neighbors. The experiment found that this method did not successfully group three documents (for details of the Confusion Matrix, see appendix A). Even though document 7 ought to be on NFCAM, but with a lower score of 0.49921, it was predicted on the NRSS with a confidence value of 0.50079. Documents 86 and 93, which were supposed to be about ENDF, were unable to be foreseen. Document 93 was predicted on the NRSS with a confidence value of 0.50126 and document 86 was predicted on the NR with a value of 0.49936.

Figure 2. A comparison of the accuracy levels in the kNN method.



This study utilized 563 unlabeled documents that were divided into six years. There were 34 fewer documents in 2021 than there were in 2020, a significant drop from the previous year (see table 1). The number of documents then climbed again in 2022, reaching 98. RapidMiner's labeling process ran into issues when it got to the process document stage. To improve memory performance, the documents were split into three runs (2017–2018, 2019–2020, and 2021–2022) because the memory was not sufficient to execute a set of commands on document processing.

The results of the previous set of procedures were then exported as tabular data for further study. Every year, the evolution of each nuclear subject can be seen in the final project report (see fig. 3). During the test period, 282 documents (50.09%) of the total extant papers had an ENDF study, followed by IR with 95 documents (16.87%) and NFCAM with 69 documents (12.26%). While there were very little changes between NR and NRSS, NR contains 47 papers (8.35%) connected while NRSS had 45 documents (7.99%). Mgt was the subject with the fewest documents, with a total of 25 (4.44%) from 2017 to 2022.

Table 1. The PINT's final project documents growth from 2017 to 2022

Study Program	2017	2018	2019	2020	2021	2022	Grand total
Electromechanics	35	34	43	35	24	41	212
Electronics Instrumentation	27	34	38	38	22	28	187
Nuclear Technochemistry	31	31	26	27	20	29	164
Grand total	93	99	107	100	66	98	563

See appendix B for more information on the confidence value of each predicted document.

Of the 212 final project reports in the Electromechanics study program 63.68 percent (135 documents) were projected to be on the ENDF subject, followed by 17.92 percent (38 documents) on NFCAM, NRSS with 8.96 percent (19 documents), and NR 5.19 percent (11 documents). Meanwhile, IR had the fewest papers predicted, with 2.83 percent (6 documents) while Mgt had 1.42 percent (3 documents) predicted. Every year, ENDF was the most predicted subject in this study (see fig. 4).

Figure 3. Nuclear subject development by percentage each year.

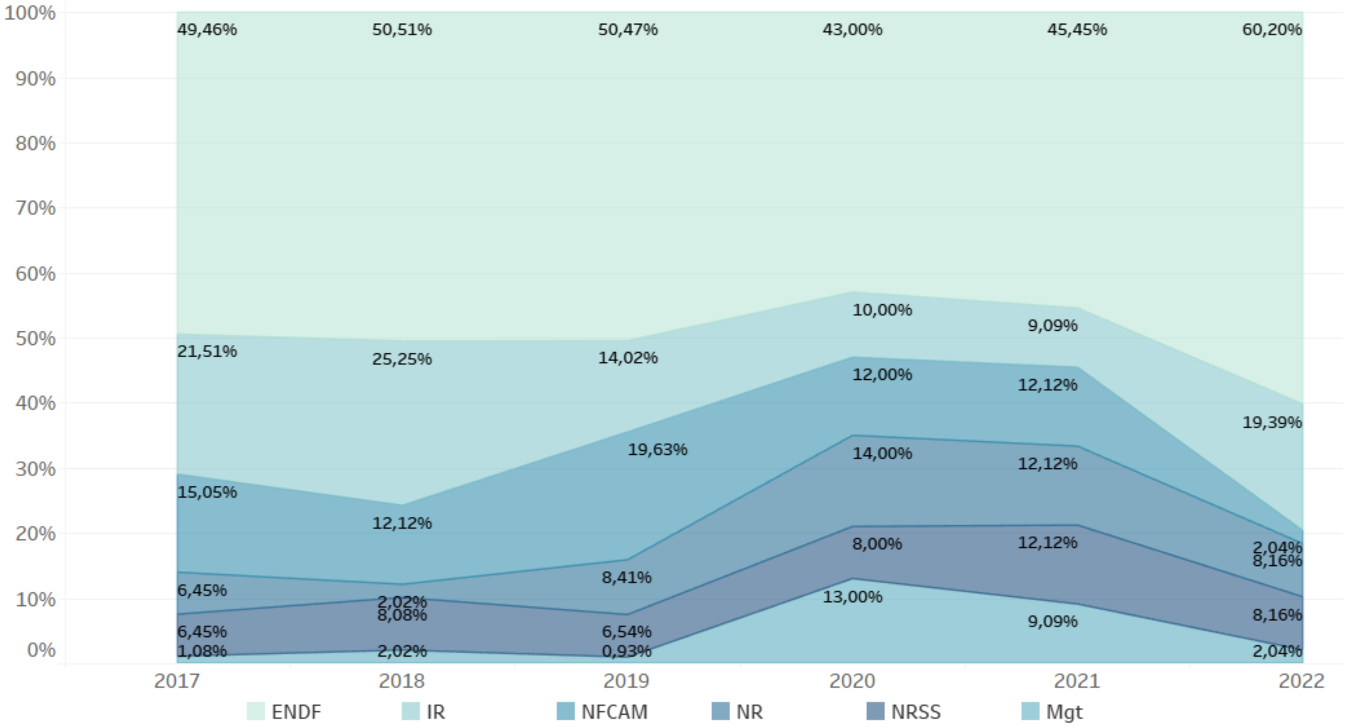
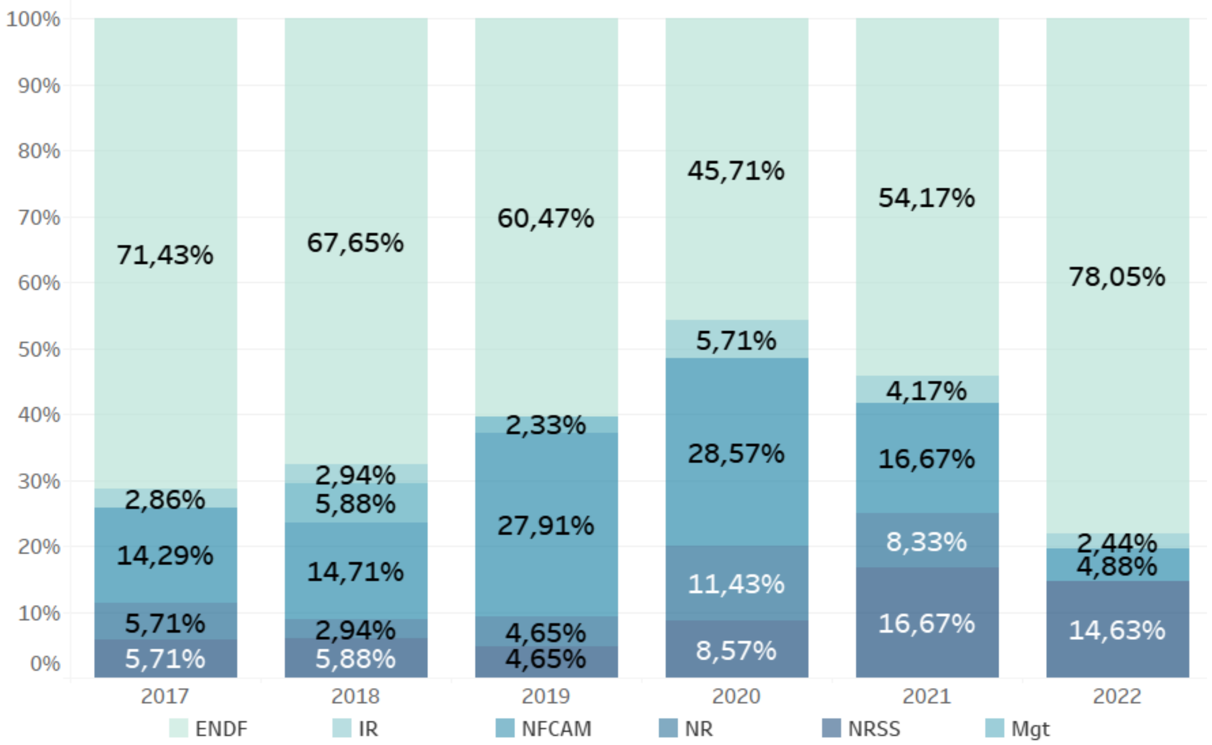


Figure 4. Nuclear subject development in Electromechanics by % each year.



The final project report on Instrumentation Electronics, which included 187 papers, was successfully predicted into five subjects. ENDF was projected to contain 141 documents (75.40%), NRSS was likely to contain 24 documents (12.83%), and NR was predicted to contain 14 documents (7.49%). Furthermore, only 7 documents (3.74%) on Mgt and 1 document (0.53%) on IR were predicted. NFCAM, on the other hand, is not mentioned in any of the Electronics Instrumentation publications (see fig. 5).

Final processing was performed on a collection of Nuclear Technochemistry documents. One hundred sixty-four documents are predicted at IR of 53.66 percent (88 documents), NFCAM of 18.90 percent (31 documents), NR of 13.41 percent (22 documents), Mgt of 9.15 percent (15 documents), ENDF of 3.66 percent (6 documents), and the remaining 1.22 percent (2 documents) were predicted on the NRSS. Subjects that were popular each year vary (see fig. 6) when compared to Electromechanics and Instrumentation Electronics, where ENDF was the most popular topic in these two study programs.

Figure 5. Nuclear subject development in Electronics Instrumentation by % each year.

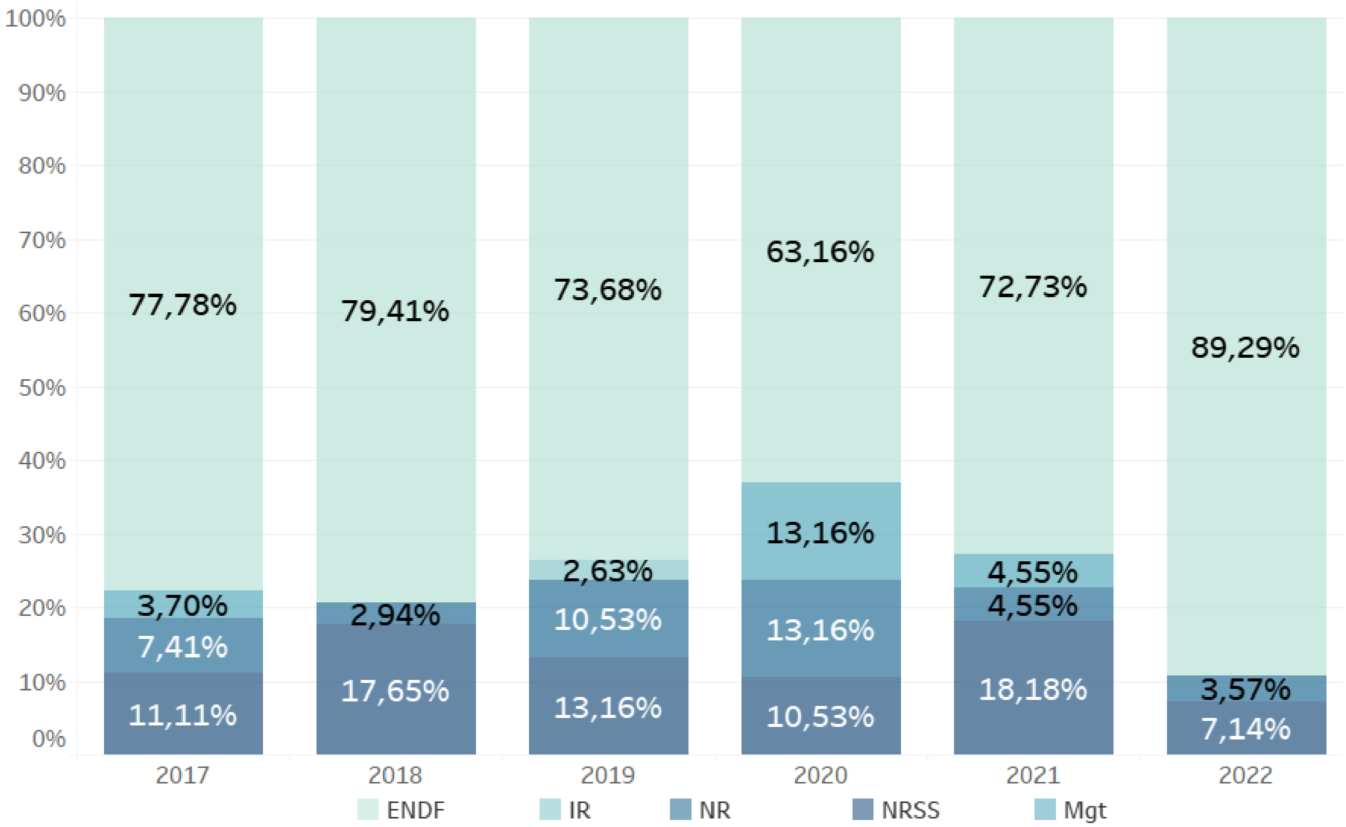
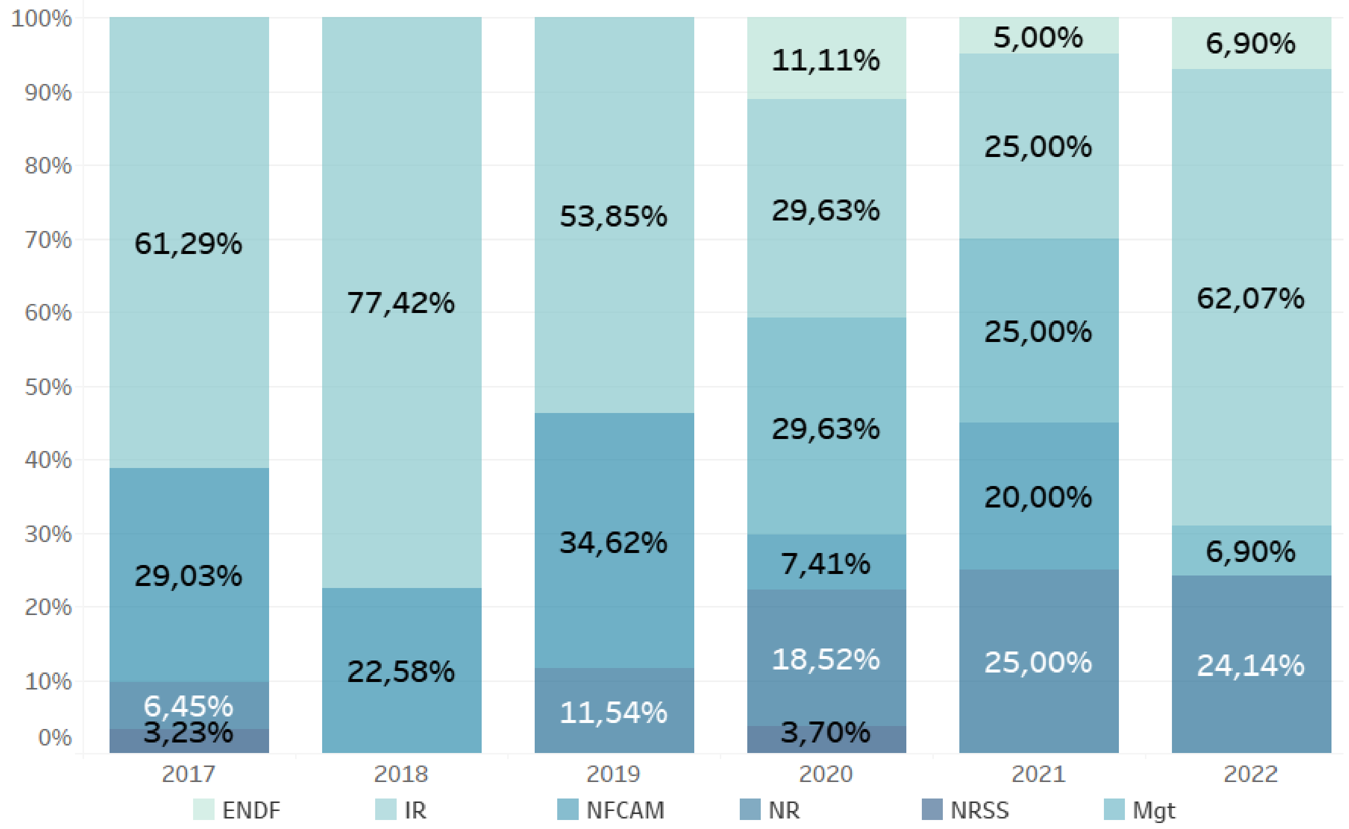


Figure 6. Nuclear subject development in Nuclear Technochemistry by % each year.



DISCUSSION

The study found that implementing kNN with Cosine Similarity in association with vector construction=Binary and k=4 resulted in the highest accuracy results of 97 percent. In general, this strategy outperformed in every class examined, and it can only be balanced on one occasion, notably at k=9 by utilizing Correlation Similarity. When compared to the use of TF-IDF, the results likewise indicated that binary term occurrence always functioned well. TF-IDF was only able to achieve its highest accuracy of 81.89 percent when k was 7 using Correlation Similarity. Cosine similarity also seemed to work efficiently on every vector creation, both when using binary and TF-IDF (in classes numbering 2, 5, and 10 the use of TF-IDF was not optimal), compared to numerical measures of Correlation Similarity and Dice Similarity. Cosine similarity evaluates the similarity of documents, and a high similarity score indicates that the documents are quite similar.[31]

Nuclear Field Growth

In general, aside from the ENDF field, which is steady and increasing, other subjects endure annual changes in development. For the past six years, ENDF has been the most popular subject among students. The ENDF reached the highest percentage rate in 2022, with 59 documents predicted on this subject. Students preferred engineering final project reports on mechanics and structures, electromechanics, control systems, nuclear instrumentation, or nuclear facility process technology. Research conducted by Wang et al. also suggests that the current popular topic of research on nuclear power is modeling and simulation.[32]

The ENDF document's average confidence value was 0.6499916, with a median value of 0.7490455. The two documents with the lowest confidence in the ENDF were document numbers 233 and 597. Document 233 had a confidence value of 0.25105 and was predicted in the other three subject areas (NRSS, NR, Mgt) with close values. Likewise, the 597 documents predicted in the ENDF with a confidence value of 0.25156 were higher than the NRSS, NFCAM, and IR subjects, but with a not too significant difference. Both of these documents can be investigated further and directly evaluated by the librarian in order to obtain a more precise field. The majority of the final project reports projected in the ENDF have confidence levels around 0.50, and some even higher at 0.75. This study also reveals that 11 documents in the ENDF category have a confidence value of 1. With lower NRSS confidence values, 239 ENDF documents connected to the NRSS field. This relationship demonstrates a good tendency among students conducting nuclear engineering related to the NRSS discipline.

Though it differs significantly from ENDF, IR is becoming a prominent field. The final project report for IR was developed in 2017–2018, but it shrank again from 2019 to 2021, then increased in 2022. In comparison to other fields, IR has the highest minimal confidence score of 0.4987, with many documents lying within the 0.5 and 0.75 range. Meanwhile, the confidence value for 26 documents predicted by IR is 1. The NFCAM subject area is a prediction that appears frequently in IR predictions but has a lower level of confidence. There are 54 documents indicating the existence of research that involves isotopes and radiation in nuclear materials, nuclear excavations, radioactive waste, structures, or advanced materials.

NFCAM is inversely proportional to the conditions that occur in ENDF. After increasing in 2019, this subject faced a reversal over the next three years, with only two documents classified in this subject through 2022. Students are still uncommonly interested in nuclear minerals, nuclear fuel, radioactive waste, structural materials, and advanced materials. Six projected documents in this field have confidence levels of 1, while many more have confidence levels between 0.50 and 0.75. The IR field is also expected to appear alongside the NFCAM field publications.

There were also ups and downs in NR and NRSS. Twenty-five of the 47 documents identified on the NR were also predicted with a lower value in the NRSS field. This demonstrates that students explored the relationship between the subject of reactor research and safety and security in various documents. Meanwhile, only eight of the 46 NRSS papers are unrelated to the ENDF field. This demonstrates that students who study nuclear safety and security tend to perform engineering to address situations involving nuclear safety and security. Documents in these two fields are usually concentrated in the 0.5 confidence value range in both NR and NRSS.

Mgt is one of the least studied topics among students. Human resources, organization, management, program planning, auditing, quality systems, informatics utilization, or cooperation are more commonly associated with the Mgt field. The Mgt increased in 2020, although it became the field with the fewest documents on earlier occasions (2017 to 2019 and 2021 to 2022). In terms of confidence value, 21 Mgt documents have a value greater than 0.5, with eight documents worth 1. With 10 documents, the ENDF is the most often discussed study area with Mgt.

Progression in Each Study Program

Even if they are still within the purview of nuclear science, the growth of the nuclear field in each study program differs depending on the curriculum. Students are influenced by knowledge, and more specifically the process of learning and comprehending (whether theoretical or more practical).[33]

ENDF is still the most popular field in Electromechanics and Electronics Instrumentation study programs. These two study programs offer courses in ENDF topic areas such as mechanical, civil and architectural, electromechanical, electrical, control systems, and radiation detection for nuclear devices. Furthermore, the Electronics Instrumentation study program offers courses on nuclear electronics, signal processing techniques, and practical work on interface and data acquisition techniques, all of which are part of the ENDF nuclear instrumentation group.

Apart from ENDF, the fields of NFCAM and NRSS have been present in Electromechanics for a period of six years. While Mgt is currently a less appealing topic, there have been no final project reports relating to Mgt in the most recent three years.

In Electronics Instrumentation, the absence of a field occurs in NFCAM. The findings of the predictions demonstrate that none of the documents predicted on NFCAM were proper. Meanwhile, only 10 documents that intersect with NFCAM which have lower confidence in the range of values from 0.247 to 0.251. Nuclear minerals, nuclear fuel, structural materials and advanced materials, and radioactive waste were not studied in depth in this study program, illustrating why NFCAM is not predicted in instrumentation electronics.

In contrast to other study programs, IR is the most predictable field in the final project report in Nuclear Technochemistry. In this investigation, Nuclear Technochemistry owns 88 of the 95 documents examined. This study program includes IR specializations such as the use of isotopes and radiation in agriculture, health, and industry. Radioisotope production becomes another discipline that specializes in the creation of isotopes and radiation sources, which explains why IR is so popular among Nuclear Technochemistry students. The NFCAM field was not present in 2022, despite the fact that it had been the topic of several students' studies throughout the preceding five years. While the ENDF and Mgt fields have only been present in the last three years, there were no predictable papers in the previous three years.

CONCLUSION

The trend of research activities carried out by students from one study program to the next appears to vary although they are both within the scope of the nuclear field. For example, the field of ENDF is quite popular among Electromechanics and Electronics Instrumentation students but not for Nuclear Technochemistry students because ENDF only appeared three years ago and the number of documents is still modest. However, ENDF deserves to be a field that needs attention. Nuclear Technochemistry students with radiochemistry learning experiences demonstrate that the IR field is linear and interesting to them. Due to a paucity of publications, the low proportion in certain categories, e.g., Mgt, shows a potential to further investigate this field.

This study demonstrates an opportunity to use text mining to assist librarians in performing automatic document classification based on specific subjects. The best model in this study is produced by combining kNN with Cosine similarity and binary term occurrence. The model used can help improve the quality of decisions made to accurately and efficiently categorize documents. To determine a more specific classification, pay close attention to documents that have a low level of confidence and intersect with other issues. This study is limited to the kNN method and documents from the BATAN repository, as well as final project documents for PINT students. Large-scale testing can be conducted, for instance, in the International Atomic Energy Agency's (IAEA) nuclear repository known as the International Nuclear Information System (INIS) Repository, or in other databases with the complexity of categorizing documents throughout many languages.

DATA ACCESSIBILITY

Datasets and data analysis code for RapidMiner have been uploaded to the RIN Dataverse: https://hdl.handle.net/20.500.12690/RIN/ASRGVO.

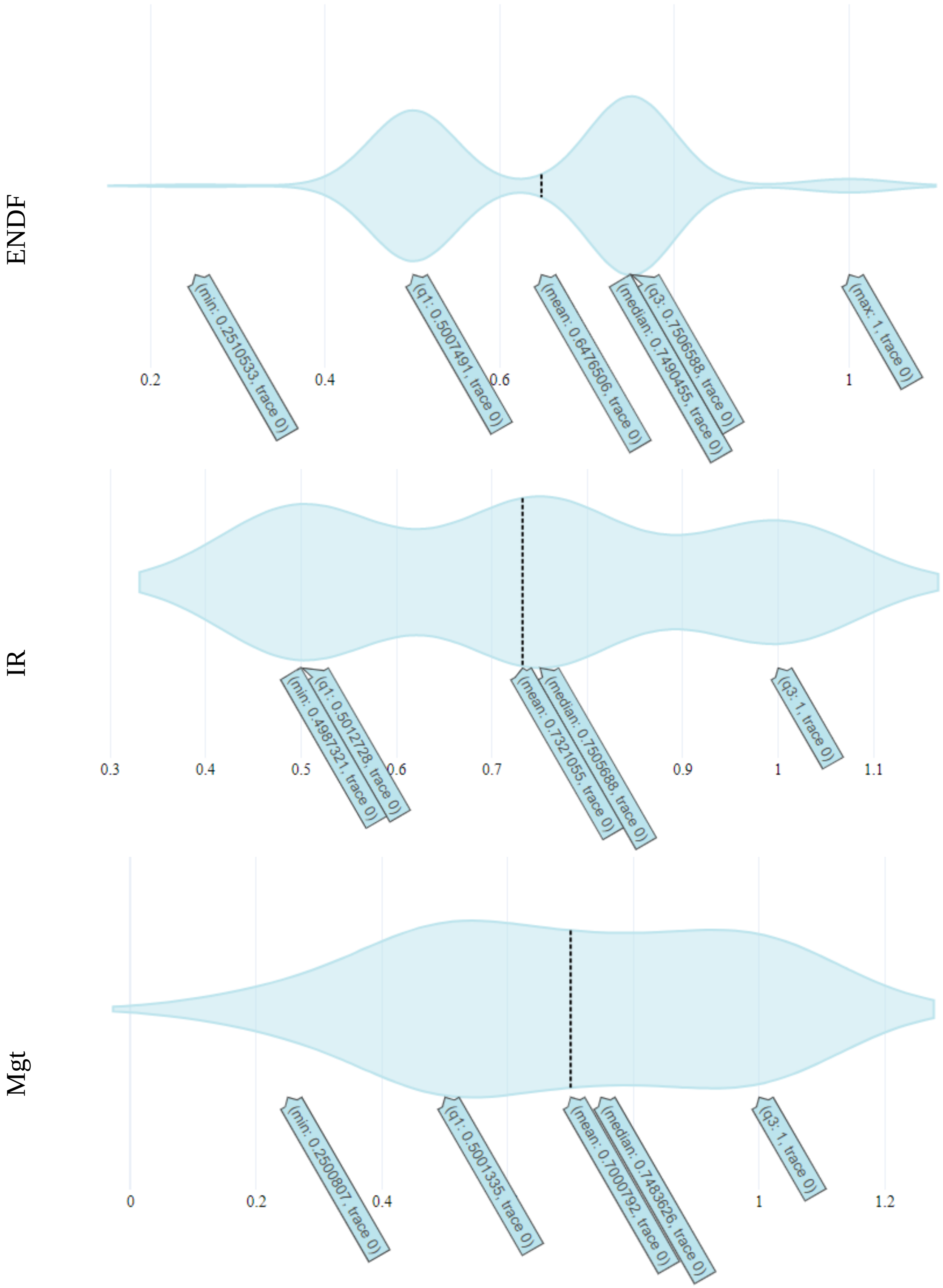
Data visualization can be accessed through Tableau Public: https://public.tableau.com/app/profile/faizhal.arif/viz/FinalProjectTrendsUtilizingNuclearKnowledgeTaxonomy/Story1

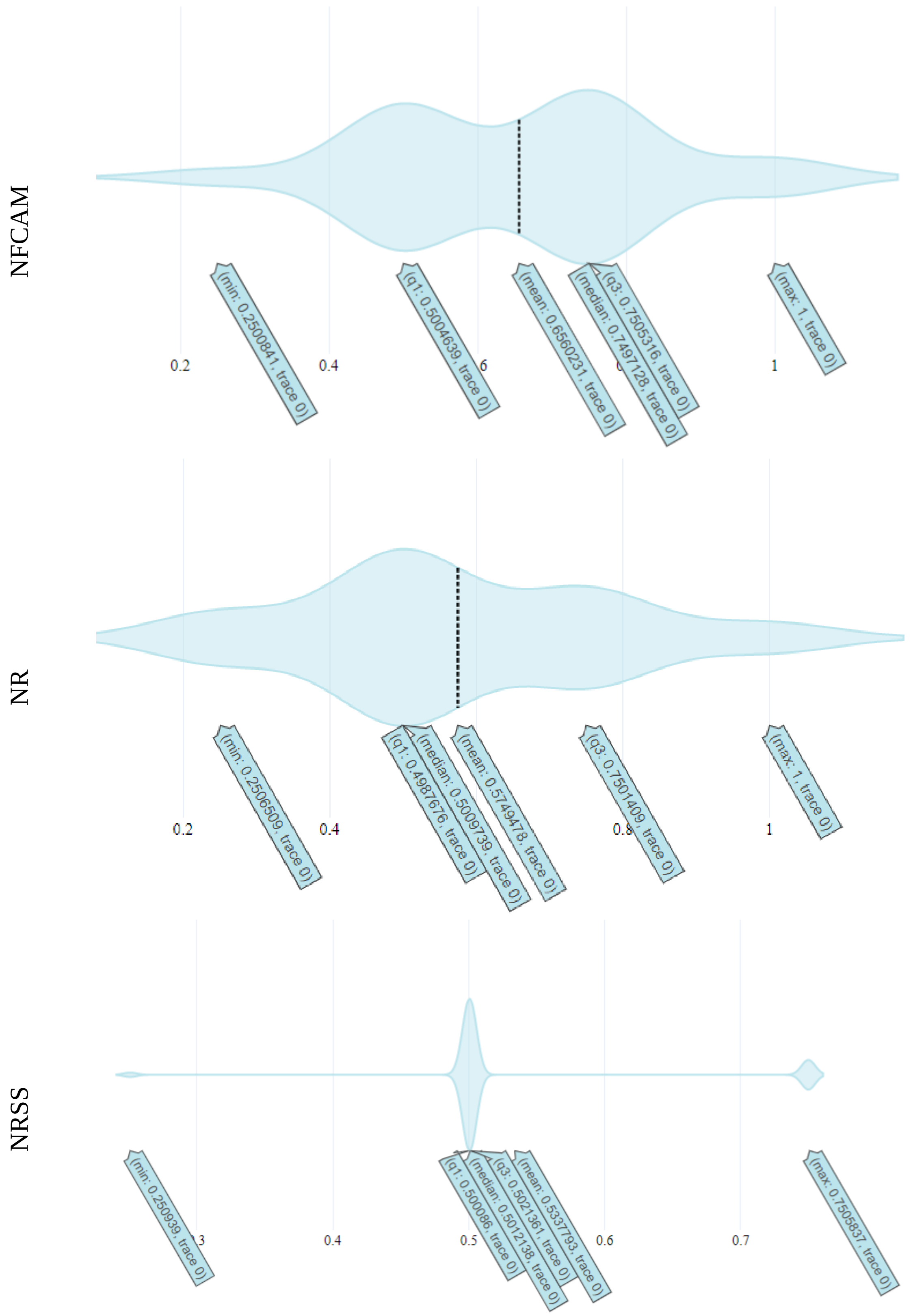
APPENDIX A: CONFUSION MATRIX OF 10-FOLD CROSS VALIDATION

Accuracy: 97.00% +/- 4.83% (micro average: 96.97%)

	True NFCAM	**True IR**	**True NRSS**	**True Mgt**	**True NR**	**True ENDF**	**Class precision**
Pred. NFCAM	13	0	0	0	0	0	100.00%
Pred. IR	0	18	0	0	0	0	100.00%
Pred. NRSS	1	0	20	0	0	1	90.91%
Pred. Mgt	0	0	0	19	0	0	100.00%
Pred. NR	0	0	0	0	13	1	92.86%
Pred. ENDF	0	0	0	0	0	13	100.00%
Class recall	92.86%	100.00%	100.00%	100.00%	100.00%	86.67%	

APPENDIX B: THE CONFIDENCE VALUE OF EACH FIELD




NFCAM
(min: 0.2500841, trace 0)
(q1: 0.5004639, trace 0)
(mean: 0.6560231, trace 0)
(median: 0.7497128, trace 0)
(q3: 0.7505316, trace 0)
(max: 1, trace 0)
0.2
0.4
6
0
1
NR
(min: 0.2506509, trace 0)
(q1: 0.4987676, trace 0)
(median: 0.5009739, trace 0)
(mean: 0.5749478, trace 0)
(q3: 0.7501409, trace 0)
(max: 1, trace 0)
0.2
0.4
0.8
1
NRSS
(min: 0.250939, trace 0)
(q1: 0.500086, trace 0)
(median: 0.5012138, trace 0)
(q3: 0.5021361, trace 0)
(mean: 0.5337793, trace 0)
(max: 0.7505837, trace 0)
.3
0.4
0.5
0.6
0.7

ENDNOTES

[1] Budi Prasetyo and Anggiana Rohandi Yusuf, “Pengelolaan Pengetahuan Eksplisit Berbasis Teknologi Informasi di BATAN,” in *Prosiding Seminar Nasional SDM teknologi Nuklir* (Seminar Nasional SDM Teknologi Nuklir, Yogyakarta: Sekolah Tinggi Teknologi Nuklir, 2018), 126–32, https://inis.iaea.org/collection/NCLCollectionStore/_Public/50/062/50062856.pdf?r=1.

[2] Joanna Yi-Hang Pong et al., “A Comparative Study of Two Automatic Document Classification Methods in a Library Setting,” *Journal of Information Science* 34, no. 2 (April 2008): 213–30, https://doi.org/10.1177/0165551507082592.

[3] Prasetyo and Yusuf, “Pengelolaan Pengetahuan Eksplisit.”

[4] Jae-Ho Kim and Key-Sun Choi, “Patent Document Categorization Based on Semantic Structural Information,” *Information Processing & Management* 43, no. 5 (September 2007): 1200–15, https://doi.org/10.1016/j.ipm.2007.02.002; Pong et al., “A Comparative Study”; Khusbu Thakur and Vinit Kumar, “Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools,” *New Review of Academic Librarianship* (May 12, 2021): 1–25, https://doi.org/10.1080/13614533.2021.1918190.

[5] Pong et al., “A Comparative Study.”

[6] Kiri L. Wagstaff and Geoffrey Z. Liu, “Automated Classification to Improve the Efficiency of Weeding Library Collections,” *The Journal of Academic Librarianship* 44, no. 2 (March 2018): 238–47, https://doi.org/10.1016/j.acalib.2018.02.001.

[7] Manika Lamba and Margam Madhusudhan, “Mapping of Topics in DESIDOC Journal of Library and Information Technology, India: A Study,” *Scientometrics* 120, no. 2 (August 2019): 477–505, https://doi.org/10.1007/s11192-019-03137-5.

[8] Fábio Figueiredo et al., “Word Co-Occurrence Features for Text Classification,” *Information Systems* 36, no. 5 (July 2011): 843–58, https://doi.org/10.1016/j.is.2011.02.002; Yen-Hsien Lee et al., “Use of a Domain-Specific Ontology to Support Automated Document Categorization at the Concept Level: Method Development and Evaluation,” *Expert Systems with Applications* 174 (July 2021): 114681, https://doi.org/10.1016/j.eswa.2021.114681; Yousif A. Alhaj et al., “A Study of the Effects of Stemming Strategies on Arabic Document Classification,” *IEEE Access* 7 (2019): 32664–71, https://doi.org/10.1109/ACCESS.2019.2903331.

[9] David Antons et al., “The Application of Text Mining Methods in Innovation Research: Current State, Evolution Patterns, and Development Priorities,” *R&D Management* 50, no. 3 (June 2020): 329–51, https://doi.org/10.1111/radm.12408; Muhammad Arshad et al., “Next Generation Data Analytics: Text Mining in Library Practice and Research,” *Library Philosophy and Practice* (2020): 1–12.

[10] Mowafy Mona, Rezk Amira, and Hazem M. El-Bakry, “An Efficient Classification Model for Unstructured Text Document,” *American Journal of Computer Science and Information Technology* 06, no. 01 (2018), https://doi.org/10.21767/2349-3917.100016.

[11] Kim and Choi, "Patent Document Categorization."

[12] Alhaj et al., "A Study of the Effects of Stemming Strategies."

[13] Mona, Amira, and El-Bakry, "An Efficient Classification Model."

[14] Thakur and Kumar, "Application of Text Mining Techniques."

[15] Kim and Choi, "Patent Document Categorization."

[16] Wagstaff and Liu, "Automated Classification."

[17] Najat Ali, Daniel Neagu, and Paul Trundle, "Evaluation of K-Nearest Neighbour Classifier Performance for Heterogeneous Data Sets," *SN Applied Sciences* 1, no. 12 (December 2019): 1559, https://doi.org/10.1007/s42452-019-1356-9.

[18] Roiss Alhutaish and Nazlia Omar, "Arabic Text Classification Using K-Nearest Neighbour Algorithm," *The International Arab Journal of Information Technology* 12, no. 2 (2015): 190–95.

[19] Mona, Amira, and El-Bakry, "An Efficient Classification Model."

[20] Guozhong Feng et al., "A Probabilistic Model Derived Term Weighting Scheme for Text Classification," *Pattern Recognition Letters* 110 (July 2018): 23–29, https://doi.org/10.1016/j.patrec.2018.03.003.

[21] Snezhana Sulova et al., "Using Text Mining to Classify Research Papers," in *17th International Multidisciplinary Scientific GeoConference SGEM 2017*, vol. 17, International Multidisciplinary Scientific GeoConference-SGEM (17th International Multidisciplinary Scientific GeoConference SGEM, Sofia: Surveying Geology & Mining Ecology Management (SGEM), 2017), 647–54, https://doi.org/10.5593/sgem2017/21/S07.083.

[22] Lee et al., "Use of a Domain-Specific Ontology."

[23] Man Lan et al., "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, no. 4 (April 2009): 721–35, https://doi.org/10.1109/TPAMI.2008.110.

[24] Devid Haryalesmana, "Masdevid/ID-Stop words," 2019, https://github.com/masdevid/ID-Stop words.

[25] Alhaj et al., "A Study of the Effects of Stemming Strategies."

[26] Pong et al., "A Comparative Study."

[27] Ananta Pandu Wicaksana, "Nolimitid/Nolimit-Kamus," 2015, https://github.com/nolimitid/nolimit-kamus.

[28] Antons et al., "The Application of Text Mining Methods."

[29] Kanish Shah et al., "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research* 5, no. 1 (December 2020): 12, https://doi.org/10.1007/s41133-020-00032-0.

[30] Judit Tamas and Zsolt Toth, "Classification-Based Symbolic Indoor Positioning over the Miskolc IIS Data-Set," *Journal of Location Based Services* 12, no. 1 (January 2, 2018): 2–18, https://doi.org/10.1080/17489725.2018.1455992.

[31] Hanan Aljuaid et al., "Important Citation Identification Using Sentiment Analysis of In-Text Citations," *Telematics and Informatics* 56 (January 2021): 101492, https://doi.org/10.1016/j.tele.2020.101492.

[32] Qiang Wang, Rongrong Li, and Gang He, "Research Status of Nuclear Power: A Review," *Renewable and Sustainable Energy Reviews* 90 (July 2018): 90–96, https://doi.org/10.1016/j.rser.2018.03.044.

[33] Ronald Barnett, "Knowing and Becoming in the Higher Education Curriculum," *Studies in Higher Education* 34, no. 4 (June 2009): 429–40, https://doi.org/10.1080/03075070902771978.