Shared Nearest Neighbour in Text Mining for Classification Material in Online Learning Using Mobile Application

https://doi.org/10.3991/ijim.v16i04.28991

Irawan Dwi Wahyono[1(✉)], Djoko Saryono[1], Hari Putranto[1], Khoirudin Asfani[1], Harits Ar Rosyid[1], Sunarti[1], Mohd Murtadha Mohamad[2], Mohd Nihra Haruzuan Bin Mohamad Said[2], Gwo Jiun Horng[3], Jia-Shing Shih[3]
[1]Universitas Negeri Malang, East Java, Indonesia
[2]Universiti Teknologi Malaysia, Johor Bahru, Malaysias
[3]Southern Taiwan University of Science and Technology, Tainan, Taiwan
irawan.dwi.ft@um.ac.id

Abstract—There are many resources for media learning in online learning that all of the teachers made many media which it made a problem if there have the same subject and material. This problem made online learning having a big database and many materials made useless because the material has the same purpose. The big problem in overload database is that online learning can't be accessed by everyone. This research to fix this problem developed an algorithm in Artificial Intelligence for the classification of material in online learning with the same subject and purpose so that teachers can use already media. This algorithm is text mining and Shared Nearest Neighbour (SSN) that is embedded in the mobile application to display the classification and the location of searching media in database online learning. The testing in this research applied in 142 media with 130 data training and 12 data testing is the result of testing is 94.7% of the accuracy of the algorithm and The average of validation is 73.33%.

Keywords—text mining, classification, mobile application

1 Introduction

The effect of the pandemic era is that all learning uses online learning in web-based applications. This reason is that all of the face-to-face learning move to online learning to avoid the coronavirus in the class. The problem is that all teachers must make a media learning such as a video, text note, and animation and upload in online learning so it makes overload in database online learning [1–3]. If online learning is an overload in the database, it makes a big problem that it can't be accessed by all people. The solving this problem is that if there is the same material with the same subject or topic, another teacher doesn't need to upload it. Another problem is that how to know if the material already exists in online learning [2–3]. It can fix by another application to classify the material in online learning and the position of material in online learning so

the teacher can use the material for their teaching in online learning [1–3]. Clearly, the problem will be fixed by making an application to classify all of the material in online learning based on topic and subject with a specific category.

Another hand, There is much material in online learning with the same topic and same subject. It makes it useless and loading in the database. For instance, video of introduction of a network computer already exists in 10–11 file that is made database that is an overload and difficult to find the specific material with the result of time out. The problem will be worst if many people find and access with the same time that the result condition of online learning is down and can't be accessed [4–5]. Meanwhile, many research uses artificial intelligence (AI) to solve this problem but it needs more resources for online learning. Online learning has a limited resource, so this problem will be fixed by using AI that is a little resource such as machine learning. Machine learning had used in much online learning for assessing a student and grading the student. Machine learning can be integrated into online learning with the web-based application but now, all people use the mobile application in online learning [6–7]. Obviously, now, online learning needs machine learning in the mobile application to solve the many problems in a database online learning such as making classification of material online learning.

However, now online learning uses machine learning that is used for optimization or effective of usage online learning. This reason is that online learning has limited resources and is accessed by many peoples. For instance, all students use online learning in the morning in the pandemic era, so the database will overload because of the time and the total of accessing it [8–10]. Online learning needs more space or eliminated material in there if the material is the same as another material in the same subject or topic and make classification of the material. Meanwhile, there are many machine learning algorithm that is integrated into online learning but just for assessment user or grading the user in online learning [10–11]. To illustrate, a text-mining algorithm is used for assessment, or naïve Bayes is used to classifying the ability of users or students [11–13]. Two algorithms can use for classification material in the database in online learning by processing the title of the file. Needless to say, after the processing of classification, if the material already exists in there, the teacher doesn't make or upload another material in there.

The problem is fixed by using a machine learning algorithm that is mobile-based. The purpose of this research made a mobile application to classify the material online learning based on each category of subject. The application uses a machine-learning algorithm to make classification by using the title of the file in the material in database online learning. The algorithm is text mining and Shared Nearest Neighbour (SNN) Algorithm. The length of the title is processed by a text-mining algorithm and after that gives a weighting for each word in the title. Every title of file with weighting has a value that is calculated by SNN to get near the cluster for each category. The utilization of this is a mobile application to know the accuracy of the algorithm and the result of classification. At the end of this research will be tested in a real database of online learning to get a real validation of the result of the application.

2 Method

This study uses 2 methods, namely the Text Mining algorithm and the SNN Algorithm. Text mining performs the process of retrieval of training data, data testing, tokenizing, filtering, steaming, and cleaning, and then finally weighting the value of the IDF TF algorithm by grouping SNN based on the closeness of similarity values to classify each type of document contained in online learning. The processing in this research is shown in Figure 1.

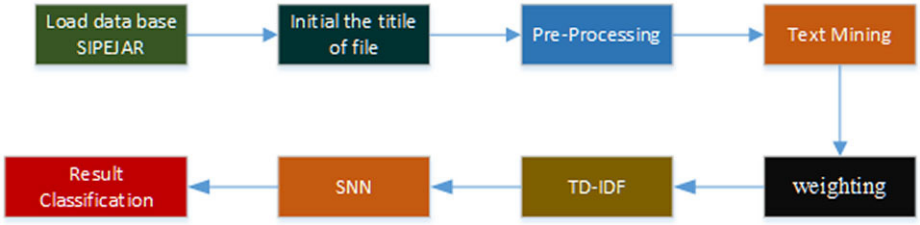

Fig. 1. The processing of this research

2.1 Text mining

The stages of text mining include the cleaning, tokenizing and filtering processes. In the cleaning process, words are truncated in file titles that exceed 12 words. So from the initial data in the database, if a title is found that has more than 12 words, the 13th word to the last word is omitted.

The tokenizing process in this study begins by taking the practical work title data in the database, from the practical work title data then the tokenizing process is carried out. The results of the tokenizing process are stored back in the database. In this study, the filtering process was carried out using a stop list model or eliminating words that were not important. First, the words that are considered unimportant are stored in the database, namely in, to, from, and, to, at, or. Once stored in the database, unimportant words will be called to match the words in each title. If one of the stop lists is found in the file title, the word will be deleted by the system. The results of this filtering process are then stored in a database. Then the weighting is based on the title match using the TF-IDF equation as in equation 1 [13–14] to produce several categories.

$$IDF = \log \frac{d}{df} \tag{1}$$

The description of equation 1 is that IDF is the value of Frequency Document Invers, df is the total of frequency document and d is the total of the document. The sample of TF-IDF is shown in Table 1 that F is the Name of File and item is component text in the title of the file. This sample of calculation in this research uses 10 titles of files in the database of material online learning.

Table 1. The result of TF-IDF on sample

Text in the Title of the File	TF					IDF
	F1	***F2***	***F3***	**….**	***F10***	***Log***
Network	0	0	1	….	0	1.5522
Security	1	1	1	…	1	1.4273
Technology	0	1	1	…	0	0.5980
…….	….	….	….	….	….	…..
RPL	1	0	1	….	1	2.0293

After all of the documents have value based on weighting using TF-IDF and have many categories of the file, the application will a clustering based on near of value each of file of media in online learning using SNN algorithm.

2.2 Shared nearest neighbour (SNN) algorithm

The Shared Nearest Neighbour (SNN) algorithm is a grouping process on high-dimensional data that has been developed [15–17]. The SNN algorithm requires 3 input parameters, namely, k which is the number of nearest neighbors, e which is the shared neighbor threshold value, and mint which is the minimum amount of data for each group.

Shared nearest neighbor algorithm (SNN) steps in this research is [15–17]

1. Calculating the similarity value from the existing data
2. Form a list of the k-nearest neighbors of each data point for k data
3. Forming a neighboring graph from a list of k nearest neighbors
4. Find the density for each data
5. Finding representative points
6. Form a group of these representative points

Meanwhile, to calculate the similarity distance between titles, the Euclidean equation is used. Euclidean equality is the determination of the square root of the difference between the coordinates of a pair of objects. The distance vectors x and y (x, y) is shown in equation 2 [15],[17].

$$sim(x, y) = d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

Where x and y are n-dimensional vectors.

For example, after calculating TF-IDF, 10 of the title of the file is processed in the SNN algorithm and the result of the example is shown in Table 2.

Table 2. The result of SNN in data set

Text in the Title of the File	d				
	F1	***F2***	***F3***	**…..**	***F10***
Network	0.011	0.2211	2.4095	…..	0
Security	0.3576	0.3576	0	….	2.0372
Technology	4.1183	0	0.5980	…..	0.5980
Network	0	0.5980	0	…..	0.5108
….	….	….	….	….	….
Total	9.1302	11.6856	22.4470	…..	2.2154
Distance vectors	3.0216	3.4184	1.4884	……	4.1349

Most of the styles are intuitive. However, we invite you to read carefully the brief description below.

3 Result and discussion

This section is about implementation and testing. The implementation uses a mobile application and it has a validation submission. The testing in this research makes testing for algorithm and validation. After testing, the analyses data will check the accuracy in algorithm and application.

3.1 Implementation

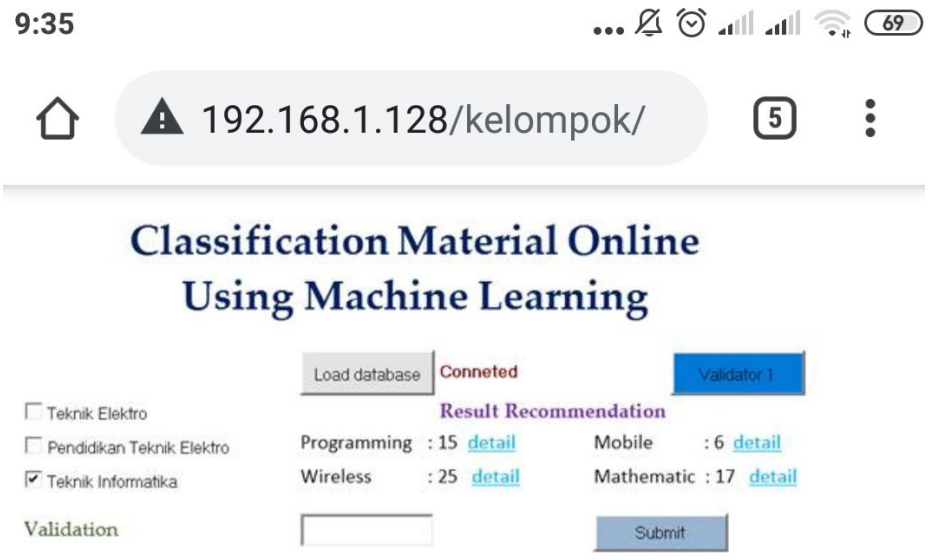

Fig. 2. The processing of this research

Implementation made in a mobile platform that is shown in Figure 2. The use of this application is

1. Users must log in as a validator or as a teacher or lecture.
2. Select one or more choosing the subject of material. This application has 3 option to choose: Teknik Elektro, Pendidikan Teknik Elektro, or Teknik Informatika material. This step makes auto-select the database based on the subject.
3. Load database that had done to initial and pre-processing step.
4. Afterload the database, the application shows the result of classification based on the category of subject.
5. Users can use an option detail in each category to show the species of material online based on a specific subject.
6. After the user shows the detail of the result of each category, the user can give validation for each category or all categories.

3.2 Testing

The testing in the application has 2 sections for checking accuracy. First is the testing of the algorithm for knowing the effective and valid algorithm.

1) Testing for algorithm

The application is embedded in online learning with the database of material and then it is tested in all of the systems to get accuracy by using a classification algorithm. The technical of testing is K-Fold Cross-validation to get the performance of this algorithm. In K-Fold Cross Validation, the data set of training divide into all of the multiple random values (k) without replacement where a multiply equal with the sum of k-1 as model training. Besides that one of the rest from multiple is used for testing. This step was repeated by all of k so the kind of model and the calculation of performance was the same repeated by all of k.

The total data set in this research for testing is 200 data with 10 data for each category. Selecting of sample is used by testing of data with the random method for each of category that is done for spreading of data rated in all of the categories. Data set is got from the labeling of all material in online learning in a specific subject that is electrical engineering subject.

The result of testing to get the performance of the application showed in data of qualitative that is presented by the implementation of the algorithm. The data of the result of performance is got from 10 times of testing using k-Fold Cross-Validation. The total sample is 200 of material in online learning in a specific subject and the result is showed in Table 3.

Table 3. The result of testing in the algorithm

Testing	Accuracy	Precision	Recall
1	94.29	82.50	71.74
2	94.76	85.09	73.25
3	92.80	79.60	64.18
4	93.29	76.34	66.96
5	94.92	83.65	72.53
6	94.34	83.21	70.61
7	93.50	75.05	66.90
8	94.92	84.01	73.24
9	94.23	79.52	71.29
10	93.16	76.92	66.22
Average	94.7	80.6	70.69

Based on Table 3, the testing for performance using Text mining and SNN algorithm with k-Fold Cross-validation is got the result that for the average of accuracy is 94.7%, the average of precision is 80,6% and lastly, the average of recall is 70.69%.

2) Testing for validation

The processing of validation is the same with testing in the algorithm that it has taken 10 times to test the validation. This testing took 3 validators to check the result of the classification of material in online learning. The validator use application that is showed in Figure 2. The validator is a teacher that teaches an electrical engineering subject. The teachers were checked all of the material that had been classified and they sent feedback by application. The format of feedback is valid or no. The result of validation is shown in Table 4.

Table 4. The result of the validation of the application

Testing	Validator 1	Validator 2	Validator 3
1	Valid	Valid	Valid
2	Valid	Valid	Valid
3	Valid	No	Valid
4	Valid	No	No
5	Valid	Valid	Valid
6	Valid	No	Valid
7	No	No	No
8	Valid	Valid	No
9	Valid	No	No
10	Valid	Valid	No
Average	90	60	60

Based on Table 4, the analysis is

1. Validator 1 was given a 9 valid status in 10 times of testing.
2. Validator 2 was given a 5 valid status in 10 times of testing. The validator gave no valid in testing 3, 4, 6, 7 and 9.
3. Validator 3 was given a 5 valid status in 10 times of testing. The validator gave no valid status in testing 5,7, 8, 9, and 10.

However, Validator 2 and Validator 3 given same the sum of valid status but they had a different number in testing given no valid status. The result of validation is that the average is 73.33%.

The result of Table 1 and Table 2 has a relationship about the result of validation and the result of the recall. If the value of recall is high in Table 1, all validators in Table 2 are given valid status in their feedbacks. All of the testing given a significant average both testing in algorithm and testing invalidation. The average rate for all testing is 83, 5% that this research success to classify the material on online learning based on a specific subject.

4 Conclusion

This research made a mobile application to classify the material online learning based on each category of subject. The application uses a machine-learning algorithm to make classification by using the title of the file in the material in database online learning. The algorithm is text mining and Shared Nearest Neighbour (SNN) Algorithm. The length of the title is processed by a text-mining algorithm and after that gives a weighting for each word in the title. Every title of file with weighting has a value that is calculated by SNN to get near the cluster for each category. The end of processing is that there are many categories of the subject with each of specific material online learning. Clearly, this application helps teachers or students to find material online learning based on specific subjects and topics in online learning material

5 Acknowledgment

This research funded by PNBP Universitas Negeri Malang, Indonesia in 2021.

6 References

[1] Martin, F., Sun, T., & Westine, C. D. (2020). A systematic review of research on online teaching and learning from 2009 to 2018. Computers & Education, 159, 104009. https://doi.org/10.1016/j.compedu.2020.104009

[2] Rasheed, R. A., Kamsin, A., & Abdullah, N. A. (2020). Challenges in the online component of blended learning: A systematic review. Computers & Education, 144, 103701. https://doi.org/10.1016/j.compedu.2019.103701

[3] Wahyono, I., Saryono, D., Asfani, K., Ashar, M., & Sunarti, S. (2020). Smart online courses using computational intelligence. International Journal of Interactive Mobile Technologies (iJIM), 14(12), 29–40. https://doi.org/10.3991/ijim.v14i12.1560

[4] Marcus, V. B., Atan, N. A., Yusof, S. M., & Tahir, L. (2020). A systematic review of e-service learning in higher education. International Journal of Interactive Mobile Technologies, 14(6), 4–14. https://doi.org/10.3991/ijim.v14i06.13395

[5] Hoi, S. C., Sahoo, D., Lu, J., & Zhao, P. (2021). Online learning: a comprehensive survey. Neurocomputing, 459, 249–289. https://doi.org/10.1016/j.neucom.2021.04.112

[6] Joy, J., & Pillai, R. V. G. (2021). Review and classification of content recommenders in e-learning environment. Journal of King Saud University-Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.06.009

[7] Wahyono, I. D., Fadlika, I., Asfani, K., Putranto, H., & Hammad, J. (2019, October). New Adaptive Intelligence Method for Personalized Adaptive Laboratories. In 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE) (Vol. 6, pp. 196–200). IEEE. https://doi.org/10.1109/ICEEIE47180.2019.8981477

[8] Zahour, O., Benlahmar, E. H., Eddaouim, A., & Hourrane, O. (2020). A comparative study of machine learning methods for automatic classification of academic and vocational guidance questions. International Journal of Interactive Mobile Technologies, 14(8), 43–60. https://doi.org/10.3991/ijim.v14i08.13005

[9] Luo, X. (2021). Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal, 60(3), 3401–3409. https://doi.org/10.1016/j.aej.2021.02.009

[10] Wahyono, I. D., Putranto, H., Asfani, K., & Afandi, A. N. (2019, September). VLC-UM: A Novel Virtual Laboratory using Machine Learning and Artificial Intelligence. In 2019 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 360–365). IEEE. https://doi.org/10.1109/ISEMANTIC.2019.8884288

[11] Cheng, M. Y., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. Automation in Construction, 118, 103265. https://doi.org/10.1016/j.autcon.2020.103265

[12] Baharudin, N. A., & Jantan, H. (2019). Mobile-based word matching detection using intelligent predictive algorithm. International Journal of Interactive Mobile Technologies, 13(9), 140–151. https://doi.org/10.3991/ijim.v13i09.10848

[13] Wahyono, I. D., Saryono, D., Ashar, M., & Asfani, K. (2019, September). Face Emotional Detection Using Computational Intelligence Based Ubiquitous Computing. In 2019 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 389–393). IEEE. https://doi.org/10.1109/ISEMANTIC.2019.8884320

[14] Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights, 1(1), 100008. https://doi.org/10.1016/j.jjimei.2021.100008

[15] Xie, X., Fu, Y., Jin, H., Zhao, Y., & Cao, W. (2020). A novel text mining approach for scholar information extraction from web content in Chinese. Future Generation Computer Systems, 111, 859–872. https://doi.org/10.1016/j.future.2019.08.033

[16] Liu, R., Wang, H., & Yu, X. (2018). Shared-nearest-neighbor-based clustering by fast search and find of density peaks. Information Sciences, 450, 200–226. https://doi.org/10.1016/j.ins.2018.03.031

[17] Wahyono, I. D., Ashar, M., Fadlika, I., Asfani, K., & Saryono, D. (2019, October). A New Computational Intelligence for Face Emotional Detection in Ubiquitous. In 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE) (Vol. 6, pp. 148–153). IEEE. https://doi.org/10.1109/ICEEIE47180.2019.8981420

7 Authors

Irawan Dwi Wahyono is a lecture on Department of Engineering in Universitas Negeri Malang, Indonesia (Email: irawan.dwi.ft@um.ac.id).

Djoko Saryono is a Professor on Department of Literature in Universitas Negeri Malang, Indonesia (Email: djoko.saryono.fs@um.ac.id).

Hari Putranto is a lecture on Department of Engineering in Universitas Negeri Malang, Indonesia (Email: Hari.putranto.ft@um.ac.id).

Khoirudin Asfani is a lecture on Department of Engineering in Universitas Negeri Malang, Indonesia (Email: khoirudin.asfani.ft@um.ac.id).

Harits Ar Rosyid is a lecture on Department of Engineering in Universitas Negeri Malang, Indonesia (Email: harits.ar.ft@um.ac.id).

Sunarti is a lecture on Department of Literature in Universitas Negeri Malang, Indonesia (Email: sunarti.fs@um.ac.id).

Mohd Murtadha Mohamad is a lecture on School of Computing in Universiti Teknologi Malaysia, Malaysia (Email: murtadha@utm.my).

Mohd Nihra Haruzuan Bin Mohamad Said is a lecture on Department of Educational Sciences, Mathematics and Creative Multimedia Universiti Teknologi Malaysia, Malaysia (Email: nihra@utm.my).

Gwo Jiun Horng is a lecture on Department of Computer Science and Information Engineering in Southern Taiwan University of Science and Technology, Taiwan (Email: grojium@stust.edu.tw).

Jia-Shing Shih is a lecture on Department of Electrical Engineering in Southern Taiwan University of Science and Technology, Taiwan (Email: jasonshih@stust.edu.tw).

Article submitted 2021-12-21. Resubmitted 2022-01-24. Final acceptance 2022-01-25. Final version published as submitted by the authors.