ICIC Express Letters
Volume **14**, Number **5**, May **2020**

ICIC International ©2020 ISSN 1881-803X
pp. **423–430**

CLASSIFICATION OF MATHEMATICS QUESTION USING NATURAL LANGUAGE PROCESSING AND TEXT MINING

Jafar Sadik, Alston Evan Wijaya and Antoni Wibowo

Department of Computer Science, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ jafar.sadik001; alston.wijaya }@binus.ac.id; anwibowo@binus.edu

Received September 2019; accepted December 2019

Abstract. *Mathematics is one of the most difficult subjects in the world. In Indonesia, one of the requirements to graduate from high school is to pass national mathematics examination. Many schools provided question bank to facilitate students in practicing mathematics. Unfortunately, the mathematics problems in question bank are not classified well, since the one who input the mathematics problem is not always a mathematics teacher. Hence, the problems should be labeled manually so administrator is able to know which topic and subtopic that problem had. It made students difficult to find the precise problem of math based on topic or subtopic. This research presents the classification of mathematic problem written with LaTeX based on topic and subtopic by using rule-based natural language processing, modified Term-Frequency – Inverse Document Frequency (TF-IDF) and Naive Bayes as a text mining algorithm. 780 mathematics problems are used in this research consisting of 18 topics and 52 subtopics from senior high school curriculum. The accuracy of this text mining is 91%. With this research question bank will be classified automatically that simplify administrator to input the mathematics problem into database and make it easier to students to find mathematics problem based on topic and subtopic.*

Keywords: Mathematics, Natural language processing, Text mining, TF-IDF

1. **Introduction.** Math is one of the main subjects in Indonesia. Therefore, math is one of requirements to pass the high school. However, Indonesian students still lack understanding of math problems. In fact, Indonesia is ranked last based on the results of the math test conducted by the Program for International Student Assessment (PISA) in 2012 [1] and is ranked 63rd out of 69 countries in 2015 [2]. In order to improve students' ability to work on math problems, some schools create question banks and their solutions on school websites that students can access anytime and anywhere. Unfortunately, students still have difficulty in finding mathematical questions related to a particular topic because the problem is not well classified based on topic or subtopic. This is because the administrators who fill the questions are not mathematics teachers, so they are confused in determining topics and subtopics if they are not labeled beforehand by the question maker. Therefore, there is the initiative of the author to create a system that can classify math questions so that the administrator can input them without having to know the topics and subtopics of the math problem. This research was conducted by combining Natural Language Processing (NLP) and text mining. Text mining, also known as text data mining or knowledge discovery from textual databases, refers generally to the process of extracting interesting and nontrivial patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases [3]. Text mining with classification techniques, divided into 8 stages, 1) tokenization, 2) stemming, 3) stop word removal, 4) weightening tokens, 5)

DOI: 10.24507/icicel.14.05.423

feature selection, 6) labeling, 7) classification and 8) evaluation [4]. There are various ways of weighting tokens, and one of the most widely used is TF-IDF [5-7]. The value of TF-IDF is a value that measures how important a word is to a document. The higher the TF-IDF value, the more significant the word affects the document [8]. TF-IDF is the product of the term frequency and inverse document frequency, where the term frequency refers to the frequency of occurrence of a term on one document and the inverse document frequency refers to how relevant a term is to all documents in the dataset [9]. To determine the value of TF-IDF, use the following formulae:

$$tf_{t,d} = \frac{f_{t,d}}{n_d} \tag{1}$$

$$idf_t = \log\left(\frac{N}{df_t}\right) \tag{2}$$

$$W_{t,d} = tf_{t,d} \times idf_t \tag{3}$$

where $tf_{t,d}$ is the value of the term-frequency of term t in document d, idf_t is the value of inverse-document-frequency term t, $f_{t,d}$ is the frequency of occurrence of term t in document d, df_t is the number of documents in the dataset that contains term t, n_d is the total number of terms in the document d, N is the total number of documents in the dataset and $W_{t,d}$ is the value of TF-IDF in term t of the document d.

After getting the value of TF-IDF from each term, a feature selection will be conducted to reduce the number of tokens that have been weighted. One of ways to do a feature selection is by simply only taking the term with the highest weight in some parts [4]. Term that has been weighted and filtered with feature selection, will be stored in the dataset with labels on each term.

NLP is one branch of the field of artificial intelligence that allows computers to be able to understand human language [10]. NLP performs a number of tasks, including word tokenization, identifying stop-words and named entity recognition which are tasked with grouping a number of words into one category [11]. NLP is divided into 3 types, namely machine-learning NLP, rule-based NLP and hybrid NLP [12]. NLP machine-learning is NLP that uses statistical data to process data. Rule-based NLP is NLP that uses manually written rules to process. Hybrid NLP is a combination of machine-learning NLP and rule-based NLP.

This study uses the NLP method to convert mathematical equations written in the form of LaTeX into Indonesian which will then be classified using the text mining method.

Research relating to NLP mostly processes alphabetical words in everyday language into various outputs. In research conducted by Zhuang et al. [13], NLP was used to convert Chinese words into alphabet (Pinyin) based on stroke. According to him, NLP for Chinese character reader algorithm is more difficult than Latin text, because one word in Mandarin is represented by one character. However, there are some mandarin characters that have meaning only when combined with other characters. There are 31 strokes that can be used to form one Chinese character.

In research conducted by Arora et al. [14], NLP is used to check the suitability of the sentence with the fixed template. In that study, each sentence requirement will be matched with the RUPP template. RUPP template envisages six slots: 1) an optional condition at the beginning; 2) the system name; 3) a modal (shall/should/will) specifying how important the requirement is; 4) the required processing functionality; this slot can admit three different forms based on the manner in which the functionality is to be rendered; 5) the object for which the functionality is needed; and 6) optional additional details about the object. The research emphasizes its contribution to the text-chunk stage, which is identifying sentence segments without using complicated and inefficient analysis. Example of text-chunks is classify the word belongs to noun phrase, adjective-phrase or verb-phrase.

Research conducted by Deshmukh et al. [15] develops a system called Sia. Sia is an interactive medical assistant that uses natural language processing method to process the symptoms of a disease spoken by humans into a text that contains the possibility of the disease he suffers. Broadly speaking, this research is divided into two parts, pre-processing and question-answering. In the pre-processing section, Sia will receive input from the user in the form of sound, then Sia will convert the sound into text. After that, the text extracted from the sound will be processed with Stanford's Part-of-Speech tagger which will convert the words in the text into basic words. The next stage is to implement a medical term identifier that converts the everyday words of human language into medical languages so that they can be matched with medical databases. The second part of this research is question-answering, which is answering questions from users based on a list of questions that are already available in the database.

In research conducted by Maldonado et al. [16], NLP is used to check for technical debt in the program source code. The study investigated the efficiency of using NLP to detect the two most common types of technical-debt, namely design-debt and requirement-debt. The study analyzed ten open-source projects from various application domains such as Ant, ArgoUML, Columba, EMF, Hibernate, JEdit, JFreeChart, JMeter, JRuby and SQuirrelSQL. Comments from the open-source project will be extracted and classified. Then, using a dataset that has been classified, the project will be classified using Stanford Classifier to identify design-debt and requirement-debt.

As far as this paper is written, NLP research has not been found to process mathematical notations written in LaTeX into everyday language.

The results of this study are a machine learning to classify mathematical questions that make administrators may input questions easier because the questions will be classified automatically. Thus, students will be faster, more precise and more accurate in conducting questions based on specific topics or subtopics.

The rest of the paper is structured as follows: Section 2 discusses the methodology about NLP and text mining followed by finding in Section 3 followed by a conclusion in Section 4.

2. **Methodology.**

This research is exploratory research, highlighting how natural language processing can be used to transform mathematics notation into everyday language and how text-mining can be used to predict the topic and subtopic of a question. We analyze, there are some variables that influence the significance of the word towards the topic.

Since the classification uses text mining technique, the relevance of word towards topic and the relevance of topic towards word are very significant point. For example, the word "probability" may exist more in topic of **Probability** than other topics. Hence, the word "probability" has high relevance towards topic of **Probability**. Another example is the word "variable_x" often exists in topic of **Trigonometry**. Nonetheless, the word "variable_x" has low relevance towards topic of **Trigonometry**, due to the fact that its existence often occurs in other topics. This research uses Indonesian Senior High School Curriculum as topics and subtopics. The dataset consists of 780 questions divided into 18 topics and 52 subtopics. The mathematical questions in this study are categorized based on topics in the high school KTSP curriculum. High school math topics based on the KTSP curriculum are 1) Roots, exponent and logarithms, 2) Quadratic Equations, 3) Mathematical Logic, 4) Basic Trigonometry, 5) Probability 6) Advanced Trigonometry, 7) Circle equations, 8) Polynomial, 9) Functions Composition, 10) Limit, 11) Derivatives, 12) Integral, 13) Linear Program, 14) Matrix, 15) Vector, 16) Geometry Transformation, 17) Exponents and Advance Logarithms and 18) Sequences and Series. Collection of questions is taken from formal textbook. Each subtopic of each topic above has 15 questions.

3. Finding and Result.

3.1. Natural language processing algorithm.

In this study, NLP is used for changing mathematics formula into Indonesian language. NLP that is used in this study is rule-based NLP where text will be analyzed by matching the text with the pattern [12]. The inputted data is a text that contains mathematical notation written by LaTeX. LaTeX is a programming language used to write a mathematical formula like notation and equation [17]. The following is an example of the data used in this study.

Table 1. Dataset example used in this research

Question	Subtopic
\(\int{x^{2}-5x+4}\;\; \partial{x}=\)<\span>...	Algebra Integral
\(\sum_{i=1}^{5}i^{2}=\)<\span>...	Sigma Notation
The roots of the quadratic equation \(x^{2}-x+2=0\)<\span> is \(x_{1}\)<\span> and \(x_{2}\)<\span>, The new quadratic equation whose roots \(\frac{x_{1}}{x_{2}}\)<\span> and \(\frac{x_{2}}{x_{1}}\)<\span> is...	Quadratic Equation
The inverse of \begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix} is...	Matrix Inverse

The first step in converting the LaTeX equation to text is to cut terms by components. Components will be categorized into 6: 1) Numbers, 2) Variables, 3) Operations, 4) Symbols, 5) Subscript and Supscript and 6) Functions, as in Table 2 below.

Table 2. Example of components

Notation	Category	NLP Output
$=$	Symbol	sama_dengan
a	Variable	variabel_a
32	Number	angka
$-$	Operation	operasi_a
a^3	Variable and Supscript	variabel_a_pangkat_polinom
$\sin x$	Function	sinus trigo nometri variabel_x
$\hat{j} + \hat{k}$	Function and Variable	vektor_j operasi_a vektor_k
$\frac{a}{b}$	Function and Variable	membagi variabel_a variabel_b

Once the terms are categorized, terms will be recombined and will be categorized again based on exponent as follows: 1) quadratic exponent, 2) polynomial exponent, 3) negative exponent, 4) function exponent, 5) decimal exponent.

3.2. Weighting algorithm.

In order to give weight to the terms, a labeled-question sentence will be separated into LaTeX and non-LaTeX. If the text contains LaTeX notation, it will be processed with NLP. Otherwise, the text will be cleaned by stemming. There have been many previous studies that have created libraries for stemming in Indonesian

language, one of which is the Sastrawi library [18]. The stemming process is done by eliminating affixes from a word according to Indonesian grammar and removing stop word from the question. Hence, a question only consists of basic words and mathematical terms as a result of NLP. Each of these terms will be calculated based on the relevance of the class towards a terms and the relevance of the terms toward the class.

Table 3. Relevance of word towards topic

No.	Word	Topic	$P_{(w\|t)}$	$P_{(t\|w)}$
1	akar_sqrt	Roots, exponent and logarithm	0.240701	0.288889
2	akar_sqrt	Integral	0.185185	0.123457
3	akar_sqrt	Sequence and Series	0.056666	0.00422535
4	akar_sqrt	Advanced Trigonometry	0.0185185	0.1

In Table 3, it can be seen that 24% of the questions in the dataset that contain the word "akar_sqrt" are the topic of the **Roots, exponent and logarithm** and 28% of the questions in the topic of **Roots, exponent and logarithm** contain the word "akar_sqrt". From the table it can also be seen that 1.8% of the questions in the dataset containing the word "akar_sqrt" are the topic of **Advanced Trigonometry** and 10% of the questions in the **Advanced Trigonometry** topic contain the word "akar_sqrt".

After getting these two values, the unique words will be sorted ascendent based on the sum of: question that contains the word on another topic divided by number of questions on that topic. In this research, it is called k. Significant point will be calculated using the formula from the modified TF-IDF. Calculating significant point will take place from the top to the bottom. The sequence constant (u) for the first data is 0.1. If the value of k in the next word is larger than the previous word, then the sequence constant will increase by 0.5. Otherwise, then the sequence constant does not increase. By taking an empirical approach, the following results of the evaluation of machine learning by tuning in the exponents of each component are shown in Table 4.

Table 4. Empirical approach to get the best significant point

Component	Exponent								
$(P_{(w\|t)})$	2	3	4	5	6	7	8	10	15
$(P_{(t\|w)})$	1	2	3	4	5	6	7	9	14
$(1 + k_{(w\|\sim t)})$	2	3	4	5	6	7	8	10	15
Topic Classifaction Accuracy	98.08	100	98.08	98.08	96.15	96.15	96.15	94.24	92.31
Subtopic Classifaction Accuracy	96.15	96.15	92.31	92.31	92.31	90.38	90.38	90.38	86.54

Thus, to get a significant point with the best accuracy, the following formula is used:

$$SP_{w|t} = \frac{(P_{(w|t)})^3 (P_{(t|w)})^2}{(1 + k_{(w|\sim t)})^3 u} \tag{4}$$

where $SP_{w|t}$ is the significant point of word w towards topic t, $P_{(w|t)}$ is the relevance of the word w to the topic t, $P_{(t|w)}$ is the relevance of topic t to the word w, $k_{(w|\sim t)}$ is the sum of: number of questions containing the word w other than the topic t divided by the number of questions on that topic, and u is a sequence constant. Table 5 contains words that are sorted by k on the topic "Function Composition".

After being classified by topic, the question will be classified according to subtopics. For weighting based on subtopics, a question will be compared with other subtopics in the same topic.

From Table 6, it was known that 36% of the questions on the topic of "Derivatives" that contains the word "derivative", were subtopics of "Algebraic Derivatives". 44% of

TABLE 5. Significant point of terms in "Function Composition" topic

Terms	Topic	$P_{(w\|t)}$	$P_{(t\|w)}$	k	u	Significant Point
fungsi_h	Function Composition	1	0.0333333	0	0.1	0.011111089
angka_fungsi_f	Function Composition	1	0.0333333	0	0.1	0.011111089
variabel_h	Function Composition	1	0.0333333	0	0.1	0.011111089
domain	Function Composition	1	0.0666667	0	0.1	0.044444489
fungsi_g	Function Composition	0.943768	0.6	0.06666	0.6	0.472847377

TABLE 6. Relevance of word towards subtopic in the same topic

Word	Subtopic	$P_{(w\|t)}$	$P_{(t\|w)}$
turunan	Formal Derivative	0.368421	0.44444467
turunan	Second Derivative	0.315789	0.4
turunan	Derivative of trigonometry function	0.315789	0.4

the questions in the subtopics of "Algebra Derivatives" contain the word "derivative". From Table 6 it can also be seen that 31% of the questions on the topic of derivatives that contains the word "derivative", are subtopics of Trigonmetry Derivatives. And 40% of the questions in the subtopics of Trigonometry Derivatives contain the word "derivative".

After getting both values, next step is to find significant point on the subtopic, in the same way as finding for a significant point on the topic.

3.3. **Classification algorithm.** This study uses the Naive Bayes method. Naive Bayes was chosen because it is robust to noise in input data, relatively easy to implement, faster to predict classes than many other classification algorithms and can be easily trained using a small data set [19].

To implement classification, a question sentence will be separated between LaTeX and non-LaTeX. If the text contains LaTeX notation, it will be processed with NLP. Otherwise, the text will be cleaned by stemming. The stemming process is done by eliminating affixes from a word according to Indonesian grammar and removing stop word from the question. Hence, a question only consists of basic words and mathematical terms as a result of NLP. The next stage is to collect the potential topic of each term by seeing the significant word of the terms toward the topic. The significant word of each term will be summed up related to the topic. Thus, the results of the total signifficant word will be taken as the topic value as shown in the following formula:

$$P_t = \sum SP_{w|t} \tag{5}$$

where P_t is a weight value of topic t and $SP_{w|t}$ is the significant point of word w towards topic t.

For example on question below:

$$\text{Inverse of } \begin{pmatrix} 1 & 3 & 5 \\ -4 & 5 & -2 \\ 0 & 2 & 4 \end{pmatrix} \text{ is}$$

the first step to do is turning the question into human language by using stemming and NLP.

After the question has been processed with NLP and stemming, the terms will be separated by spaces and find the total of significant word as can be seen in Table 8.

From Table 8, it can be seen that the probability of the question belonging to the topic of the Matrix is 97.86%, Function Composition is 2.14%, Geometry Transformation 0%

Table 7. Stemming and NLP process of question

Raw Question	NLP and Stemming Output
Inverse of $\begin{pmatrix} 1 & 3 & 5 \\ -4 & 5 & -2 \\ 0 & 2 & 4 \end{pmatrix}$ is	Inverse matrix order_3

Table 8. Significant word of each terms based on topic

Terms	Matrix	Function Composition	Geometry Transformation	Vector
invers	0.00108247	0.0113469	0	0
matriks	0.0388183	0	0.00000413031	0.000000337563
ordo_3	0.478516	0	0	0
Total	0.51841677	0.0113469	0.00000413031	0.000000337563
Percentage	97.86%	2.14%	0%	0%

Table 9. Significant word of each term based on subtopics in "Matrix" topic

Terms	Matrix Inverse	Determinant of a Matrix	Matrix Operation
invers	**1.95112**	0	0
matriks	**0.000404948**	0.000092824	0.000298666
ordo_3	0.000000784452	**0.000389099**	0.000000297713
Total	**1.951525732452**	0.000481923	0.000298963713
Presentage	99.96%	0.02%	0.02%

and Vector 0%. These questions will be classified based on the greatest probability. In Table 8, the biggest probability is the topic of the Matrix, which is equal to 97.86%. Thus, the question is classified as the topic of the Matrix.

After the question has been classified to the topic, the next stage is to classify the question based on subtopics on that topic. The result can be seen in Table 9.

From Table 9, it can be seen that the probability of the question belonging to subtopic of "Matrix Inverse" is 99.96%.

3.4. **Result.** The NLP used in this research cannot be compared to other algorithms, since there is no NLP method that converts LaTeX script to human language. However, the weighting algorithm in this research can be compared to another well-known weighting algorithm like TF-IDF. In order to acquire optimal predicition in text mining, selecting weighting algorithm is the most important step. Proper weighting algorithm will lead to better classification. To evaluate the performance of the text mining, this study uses the k-fold cross validation method with $k = 15$. Use 780 data divided into 52 subtopics and 18 topics. Accuracy by using TF-IDF is 80.39% for classification based on topic and 68.21% for classification based on subtopics. While the accuracy of using modified TF-IDF is 97.18% for classification based on topic and 91.66% for classification based on subtopics. From these results, it can be seen that the weighting of the modified TF-IDF has a higher accuracy than using the TF-IDF. This is because the TF-IDF only calculates the frequency of words for one document and calculates the frequency of occurrence of the word on the topic in the entire document. While modified TF-IDF calculates how the word affects a topic and calculates how influential the word is on a topic.

4. **Conclusions.** Text mining with modified TF-IDF and natural language processing with rule-based NLP can be used to ease administrator of question-bank to input a question that makes students search the question more precise based on topic and subtopic. The relationship between weighting algorithm and classification step is very critical. The performance of text mining with modified TF-IDF is better than text-mining with general TF-IDF. There is an urgent need for academic institution to use this research as a tag generator for question bank on their academic institution. For future work, this research can be improved by classifying mathematics question that contains image or table.

Acknowledgment. We thank Najla Maulachaela and Rizal Faiz for assistance with giving an idea about the research and for comments that greatly improved the manuscript.

REFERENCES

[1] E. Pisani, *Indonesian Kids don't Know How Stupid They are*, http://indonesiaetc.com/indonesian-kids-dont-know-how-stupid-they-are/, 2013.
[2] H. Iswadi, *A Little from Newly Released PISA 2015 Result*, http://www.ubaya.ac.id/2014/con ten-t/articles_detail/230/Sekelumit-dari-Hasil-PISA-2015-yang-Baru-Dirilis.html, 2016.
[3] A. Akilan, Text mining: Challenges and future directions, *The 2nd International Conference on Electronics and Communication Systems (ICECS)*, pp.1679-1684, 2015.
[4] A. R. C and Y. Lukito, Classification of the sentiment of political commentary from Facebook pages uses naive bayes, *JUISI*, vol.2, no.2, pp.26-34, 2016.
[5] I. Yahav, O. Shehory and D. Schwatz, Comments mining with TF-IDF: The inherent bias and its removal, *IEEE Trans. Knowledge and Data Engineering*, no.14, pp.437-450, 2015.
[6] D. Surian, S. Seneviratne, A. Seneviratne and S. Chawla, App miscategorization detection: A case study on Google play, *IEEE Trans. Knowledge and Data Engineering*, pp.1591-1604, 2017.
[7] B. Shi, G. Poghosyan, G. Ifrim and N. Hurley, Understand short texts by harvesting and analyzing semantic knowledge, *IEEE Trans. Knowledge and Data Engineering*, pp.499-512, 2017.
[8] K. Sato, J. Wang and Z. Cheng, Credibility evaluation of Twitter-based event detection by a mixing analysis of heterogeneous data, *IEEE Access*, vol.7, pp.1095-1106, 2018.
[9] Z. Zhu, J. Liang, D. Li, H. Yu and G. Liu, Hot topic detection based on a refined TF-IDF algorithm, *IEEE Access*, vol.7, pp.26996-27007, 2019.
[10] S. A. Al-Ghamdi, J. Khabti and H. S. Al-khalifa, Exploring NLP web APIs for building arabic, *The 12th International Conference on Digital Information Management*, pp.175-178, 2017.
[11] R. Talib, M. K. Hanif, S. Ayesha and F. Fatima, Text mining: Techniques, applications and issues, *International Journal of Advanced Computer Science and Applications*, pp.414-418, 2016.
[12] E. Pons, L. M. M. Braun, M. G. M. Hunink and J. A. Kors, Natural language processing in radiology: A systematic review, *RSNA*, pp.329-343, 2016.
[13] H. Zhuang, C. Wang, C. Li, Y. Li, Q. Wang and X. Zhou, Chinese language processing based on stroke representation and multidimensional representation, *IEEE Access*, vol.6, pp.41928-41941, 2018.
[14] C. Arora, M. Sabetzadeh, L. Briand and F. Zimmer, Automated checking of conformance to requirements templates using natural language processing, *IEEE Trans. Software Engineering*, pp.944-968, 2015.
[15] S. Deshmukh, R. Balani, V. Rohane and A. Singh, Sia: An interactive medical assistant using natural language processing, *The 3rd International Conference on Science Technology Engineering & Management (ICONSTEM)*, pp.368-373, 2016.
[16] E. D. S. Maldonado, E. Shihab and N. Tsantalis, Using natural language processing to automatically detect self-admitted technical debt, *IEEE Trans. Software Engineering*, vol.43, pp.1044-1062, 2017.
[17] M. Cantiello, D. Ginev, A. Pepe and Authorea Help, How to write mathematical equations, expressions, and symbols with LaTeX: A cheatsheet, *AUTHOREA*, https://www.authorea.com/users/77723/articles/110898-how-to-write-mathematical-equations-expressions-and-symbols-with-latex-a-cheatsheet, Accessed on 20 April 2019.
[18] N. Yusliani, R. Primartha and M. D. Marieska, Multiprocessing stemming: A case study of Indonesian stemming, *International Journal of Computer Appllication (0975-8887)*, pp.15-19, 2019.
[19] A. Moldagulova and R. B. Sulaiman, Document classification based on KNN algorithm by term vector space reduction, *International Conference on Control, Automation and Systems*, Daegwallyeong, South Korea, 2018.