2015 International Conference on Information Technology Systems and Innovation (ICITSI)
Bandung – Bali, November 16 – 19, 2015
ISBN: 978-1-4673-6664-9

Automatic Indonesia's Questions Classification Based On Bloom's Taxonomy Using Natural Language Processing

A Preliminary Study

Selvia Ferdiana Kusuma
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mailbox.selvia@gmail.com

Daniel Siahaan
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya Indonesia
daniel@if.its.ac.id

Umi Laili Yuhana
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya Indonesia
Yuhana@if.its.ac.id

***Abstract*— Identification of students' cognitive ability should be done to know students' understanding towards what have been taught. The identification result will be the benchmark to choose the basis of assessment. The identification process of cognitive ability can be done by giving questions in certain difficulties levels. The appropriateness of difficulty levels can be made based on bloom taxonomy introduced by Benjamin Bloom in 1956 and revised by Lorin Anderson Krathwohl in 1994. There are 6 levels in bloom taxonomy, namely remembering, understanding, applying, analyzing, evaluating and creating. However, the questions classification process based on bloom taxonomy is not easy when it is done manually. Classification process needs long time if there are many questions items. Besides, the different perception in classification make manual classification process is varied from one to another. This research suggests a method that produces automation classification of Indonesian language question items based on new bloom taxonomy levels. The method includes indentifying the question items' characteristic of nature language used. The identification is done based on lexical feature extraction and syntactic feature extraction. The features extraction output is classified by using algorithm of Support Vector Machine (SVM). The dataset used for the test is the question items from many lessons in elementary school. This research showed that the method suggested can be used to classify Indonesian language question items well.**

Keywords*— *question classification, bloom taxonomy, nature language processing, SVM.

I. INTRODUCTION

Identification of students' cognitive ability should be done to ensure students' understanding towards the lesson which has been taught. The identification result can be a benchmark of basis assessment. The identification process of students' cognitive ability can be done by giving questions in certain difficulty levels. The appropriateness of difficulty levels can be made by using bloom taxonomy introduced by Banjamin Bloom in 1956 and revised by Lorin Anderson Kratwohl in 1994. There are 6 levels in the taxonomy, namely remembering, understanding, applying, analyzing, evaluating and creating. However, exercises classification process based on bloom taxonomy is not easy when it is done manually. Classification process needs long time if there are many questions' items. Besides, the different perception in classification make manual classification process is varied from one to another.

In order to solve the problem, the automatism in classification process can be done. The automatism of classification process can be done by using nature language processing, then categorizing based on the possessive content [*1*]. There are some research examine the automatism of questions' classification as level of bloom taxonomy. The statistical approach used such as TF-IDF or Ngram is not enough to be used in questions' classification [*2*]. Statitical approach needs many data to be used for well accuracy [*3*]. Another approach which can be done is rule utilization made based on feature extraction of every question. However, based on [*4*] the usage of rule to this classification is not efficient because many more rules needed to enhance the accuracy result.

This research suggests a new approach for classification automatism of Indonesian language questions by using identification from lexical and syntactical feature for each question. Many previous research are only tested for English questions, hence, there is not any similar research done for Indonesian language questions. Thus, the researcher should make the dataset to prove the hypothesis. Feature extraction result of the questions will be classified by using SVM algorithm. SVM algorithm is chose because it is one of algorithm which is most used for text classification. This algorithm is appropriate to classify many data dimension.

This paper is organized as follows. In section 2 we will describe the literature study. In section 3 we will introduce the implementation of our approach. In section 4 we will evaluate the result. In section 5 we will describe the conclusion of this

method and future work.

II. Literature Study

A. Related Work

Yusof & Hui [5] use Artificial Neural Network (ANN) to classify questions. This research focuses on Document Frequency (DF) feature and Category Frequency-Document Frequency (CF-DF). However, the classification accuracy of this research is not good enough, in which it is around 65,9%.

Yahya & Osman [*1*] classify the question based on bloom taxonomy by using Support Vector Machine. The classification done provide the good accuration, in which it is 87,4%. However, this research did not use semantic feature or syntactic feature. The accuracy is based on Bag od Word (BOW) feature.

Haris & Omar [*6*] classify questions based on bloom taxonomy as the rules made. The rule and attern are made based on POS tagging in practice data. The accuration result is 77%. There should be many more rules to enhance the accuration result.

Omar et al. [*7*] classify question based on bloom taxonomy as the rule and weighting to handle overlapping of keyword. However, the weighting of keyword done is really influenced by expert.

B. Bloom Taxonomy

Bloom's taxonomy was introduced by Benjamin Bloom, Englehart, Furst, Hill and Krathwohl in 1950's. Bloom taxonomy is hierarchy structure used to identify the ability of someone from the lowest level to the highest level [*8*]. This concept is divided into three domains of intellectual ability, namely cognitive, affective, and psychomotor. Cognitive aspect is related to the knowledge and intellectual ability development [*1*]. Cognitive aspect in bloom taxonomy is divided into 6 levels, namely knowledge, comprehension, application, analysis, synthesis and evaluation. In 1994, Lorin Anderson Krathwohl and psychology experts of cognitivism reform the bloom taxonomy, so it is appropriate with the advanced era. The clear explanation of the aspect of cognitive domains describe are as follows [*9*]:

1) Remembering

The ability to mention the information/knowledge in memory is called remembering. For example, mention the definition of taxonomy.

2) Understanding

Understanding is the ability to understand the instruction and assert the meaning/idea/concept which has been taught in spoken, written, or even graph/diagram. For example: Summarizing the material by using own words.

3) Applying

Applying is the ability to do something and apply concept in certain situations. For example: doing income process remittance as the system defined.

4) Analyzing

Analyzing is the ability to separate concept into some components and relate each other to get the understanding of the whole concept. For example: analyzing the cause of main cost disposal increasing in the financial report by separating it based on every component.

5) Evaluating

Evaluating is the ability to decide degree based on certain values, criteria or standards. For example: comparing the students answer with the answer key.

6) Creating

Creating is the ability to integrate the elements into new form which is coherent, or making something original. For example: making curriculum by integrating opinion and material from many sources.

C. Question Classification

Question classification has similar function with text classification. Question classification is done to know the information of the question. The classification process is done based on the similarity from the main document/question. Different from the application of text classification focusing on document level, question classification focuses on the sentence and word [*10*]. The classification process of the question has its own difficulty. First, there is only little information in a question, so the information is not effective yet to classify the question. Second, stopwords used in text classification process and question classification is different. The words such as *apa, dimana, kapan* will be deleted in text classification. However, it is very important in question classification because it interprets information about the question [*10*]. Performance from the question classification is influenced by linguistics analysis (semantic, syntactic, and morphology) [*11*]. The question classification can be divided into two−based on rules or learning model [*12*].

D. Word Identification in Indonesian Language

Word is a syntactical in a sentence. Word in Indonesian Language is not always recognized directly, because sometimes the word has affix. In order to know the labeling, there is deleting process of affix. Affix is the additional which attach in word and make new meaning [*13*]. Affix is divided into 3 kinds namely prefix, suffix, and confix. Table I shows word types in Indonesian Language.

TABLE I. Word Types in Indonesian Language

Symbol	Word Types	Description
JJ	Adjectiva	Adjective; word which give explanation about something
RB	Adverbia	Adverb
AR	Artikula	Article
CC	Konjugtor Koordinatif	Connection word to relate clause in the compound sentence
CS	Konjugtor Subordinatif	Connection word in complex sentence

MD	Modal	Modal auxiliary
PR	Pronomina	Possessive pronoun; word which is used to substitute word or gerund
WH	Kata Tanya	Word which is used to ask something
NN	Nomina	Noun; word which mentions noun or geround
CD	Numeralia	Numerical word; word which asserts amount of noun or group
IN	Preposisi	Preposition; word which links words or sentence parts.
UH	Interjeksi	interjection
RP	Partikel	Particle
VB	Verba	Verb; word which has meaning doing activities
AUX	Kata bantu	Auxiliary
FW	Kata asing	Foreign word
PU	Tanda baca	Punctuation
SYM	Simbol Matematika	Math symbol
X	Tidak dikenali	Word which cannot be predicted.

E. Nature Language Processing

Natural Language Processing (NLP) is process of nature language identification done, in order to make human communicate with computer by using human language [*13*]. The process of nature language identification includes tokenizing, stemming, and POS tagging. Tokenizing is process used to scatter input string as the word arranging [14]. Stemming is losing affix in a word [14]. Stemming in Indonesian language is better based on dictionary, in order to producing fewer mistakes than based on rules [13]. POS Taging aims labeling every word in a sentence, for instance, *'Arinta membeli apel di pasar'* →|Arinta/X| |membeli/VB| |apel/NN| di/IN| pasar/NN|.

F. Feature Extraction

Feature is a differentiate characteristic used to classify a question. The feature to classify question can be distinguished into three types, namely lexical feature, syntactical feature, and semantics feature [*12*].

1) Lexical Feature

Lexical feature is feature from a question which is extracted from a word context of the question [*15*]. The example of lexical feature extraction is showed in Table II.

TABLE II. LEXICAL FEATURE

Feature Space	Extraction Result
Unigram	{(Berapa,1)(rupiahkah,1)(nilai,1)(uang,1)(2,1)(lembar,1)(lima,1)(ribuan,1)(?,1)}
Bigram	{(Berapa-rupiahkah,1)(rupiahkah-nilai,1)(nilai-uang,1)(uang-2,1)(2-lembar,1)(lembar-lima,1)(lima-ribuan,1)(ribuan-?,1)}
Trigram	{(Berapa-rupiahkah-nilai,1)(nilai-uang-2,1)(2-lembar-lima,1)(lima-ribuan-?,1)}
Wh-word	{(Berapa,1)}
Word-shape	{(Huruf kecil,6)(gabungan,1)(digit,1)(simbol,1)}
Question length	{(panjangkata, 8)}

TABLE III. SYNTACTICAL FEATURE

Feature Space	Extraction Result
Tag unigram	{(Berapa_wh,1)(rupiahkah_nn,1)(nilai_nn,1)(uang_nn,1)(2,num,1)(lembar_nn,1)(lima_num,1) (ribuan_nn,1)(?_pu,1)}
Pos tagging	{(wh,1)(nn,5)(x,1)(pu,1)}
Headword	{(uang,1)}

The example of question used for extraction feature is *'Berapa rupiahkah nilai uang 2 lembar lima ribuan?'*

2) Syntactical Feature

Syntactical feature is feature of question which is extracted from syntax to the question [*12*]. The example of syntactical feature extraction is showed in Table III. The example of question used to feature extraction s *'Berapa rupiahkah nilai uang 2 lembar lima ribuan?'*

III. METHODOLOGY

In this section we will describe our approach in detail. There are five main steps in this study namely the construction of dataset, keyword defining, stopwords defining, classification process and process of classification performance calculation.

A. Dataset

This research is the first research which classifies the question in Indonesian language based on level of bloom taxonomy, so there is not any dataset which can be used. The researcher should make the data set needed. The dataset was made from the questions from five lessons in the elementary school. All of the lessons were Indonesian language, Mathematics, Science, Social, and Civic. The amounts of all dataset were 130 questions. The questions which will be used as training data were classified by seven expert teachers who have experience of teaching in ten years or more. The seven exerts included teacher class and headmaster. Thus, it can be ensured that the questions are appropriate with the bloom taxonomy levels.

B. Keyword

Keyword defining is important to the question classification success. The process of question classification can be done if the keyword used can represent taxonomy level indentified. This research used keyword existed in the research [*9*]. However, there were some keywords which is less specific because it is also in another bloom taxonomy level.

The unspecific keywords will be discussed by the expert to be distinguished in certain levels.

C. Stopwords

One of processes done before question identification was deleting words which were not descriptive (*stopwords*). It is because not all words in a sentence have important meaning. The stopwords deleting done to decrease the use of word dimension−the option was done toward relevant words to interpret the main idea of text. Stopwords in Indonesian Language were provided, although they should be modified as the need. The stopwords should be modified to do question classification, because there some words including in stopwords all at once the keyword in the classification process. For instace, the word *buat*, *gunakan* and *sebutkan.* Those words is consideres as stopwords, although they are keywords used in introduction level in bloom taxonomy. The solution was by deleting the list of words in stopwords which were similar with keywords used in every level of bloom taxonomy.

D. Classification Process

Classification process was done based on the extraction result of lexical and syntactical feature in every question. SVM algorithm was chose as algorithm to classify this problem. Based on the previous research, SVM algorithm is the best algorithm to classify text [*12*] [*16*]. The classification in this research was done by using tools of MATLAB R2010b. Before doing the classification process, there was some process which should be done. The classification process was showed in Figure 1 and the explanation of the process was explained as follows:

1) Preprocessing

Preprocessing was the beginning step in classification. The aim of preprocessing was to interpret a sentence or a document became a feature vector by distracting text into words [*17*]. This step was very affected the classification process. The preprocessing steps included tokenizing, stopwords removal, stemming, and POS tagging. Stopwords removal was done before stemming process, it is because to decrease word complexity which should be processed.

2) Extraction Feature

In order to classify in a question sentence, it needed extraction feature. The purpose was to get unique information which can be used to do classification. The features which can be used in this research were explained in Table IV. The question example that will be extracted was showed in Table V. Extraction process of lexical feature was done by counting question words amount and words amount used in a question. The question words used were 5W+1H. Lexical feature extraction was showed in Table VI.

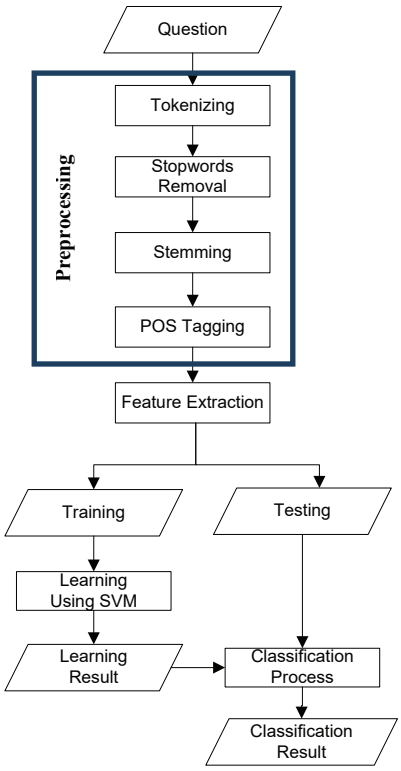

Fig. 1. Classification Process

TABLE IV. CLASSIFICATION FEATURE

Feature	Description	Reason
Lexical	Type of Question Words	As one of the characteristics that can be used to classify question [*12*]
	Question length	
Syntactical	Number of Verb	Characteristics that can be used to identify the level of the taxonomy bloom [*6*]
	Number of Adjective	
	Number of Noun	
	Number of other words	
	Keyword	Keywords can interpret a topic [*18*]

TABLE V. THE QUESTION WHICH WILL BE EXTRACTED

No	Question Sentence	POS *tangging* Result
1	Sebutkan 3 jenis bangun datar!	Sebut/vb jenis/nn bangun/vb datar/jj
2	Bandingkan kedua angka berikut, mana yang lebih besar 5 atau 7?	Banding/nn angka/x
3	Urutkan bilangan berikut 3,5,6,2!	Urut/vb bilangan/x

4	Bagaimana cara mengukur luas bangun persegi?	Bagaimana/wh ukur/nn luas/jj bangun/vb persegi/x
5	Kesimpulan dari cerita tersebut adalah?	Simpul/nn cerita/nn

TABLE VI. LEXICAL FEATURE EXTRACTION

Description	P1	P2	P3	P4	P5
What	0	0	0	0	0
Where	0	0	0	0	0
Who	0	0	0	0	0
When	0	0	0	0	0
Why	0	0	0	0	0
How	0	0	0	1	0
Question length	4	2	2	5	2

TABLE VII. KEYWORD FEATURE EXTRACTION

Bloom Taxonomy Level	P1	P2	P3	P4	P5
Remembering	1	0	0	0	0
Understanding	0	1	0	0	0
Applying	0	0	1	0	0
Analyzing	0	0	0	1	0
Evaluating	0	0	0	0	1
Creating	0	0	0	0	0

TABLE VIII. POS TAGGING FEATURE EXTRACTION

Description	P1	P2	P3	P4	P5
Number of Verb	2	0	1	1	0
Number of Adjective	1	0	0	1	0
Number of Noun	1	1	0	1	2
Number of other words	0	1	1	1	0

The process of feature extraction of keywords was done by counting all frequency of keywords in questions, in order to map as the appropriate level of bloom taxonomy. Then, the information were arranged in the table, in order to made it easy to be classified. The example of keywords feature extraction result through

TABLE IX. PERFORMANCE RESULT

Kernel SVM	Testing Accuracy					Accuracy Average
	I	II	III	IV	V	
Linier	89%	94%	83%	87%	90%	88,6%
Polinomial Orde 1	83%	87%	89%	82%	83%	84,8%
Polinomial Orde 2	65%	72%	72%	79%	78%	73,2%
Polinomial Orde 3	73%	76%	73%	75%	80%	75,4%
RBF	66%	63%	69%	67%	65%	66,0%

preprocessing was showed in Table VII. P1 to P5 is considered as disextraction question.

Extraction feature process by using POS tagging was done by counting words amount of verb, adjective, noun, and symbol used in a question. The example of POS tagging feature result for the same question with Table V was shown in Table VIII.

E. Evaluation Process Method

Evaluation process was done by comparing the question classification result from suggested method with the classification of the expert. The comparison will produce an accuracy result which can be used as the benchmark to see the successful level of this method. The rule used to do the evaluation was shown as follow:

$$\text{Accuracy} = \frac{\text{Data Amount Classified}}{\text{Data Amount}} \quad (1)$$

IV. RESULT

Process of testing was done by using technique 1/3 holdout cross validation in 130 questions in dataset. Testing data will be chosen randomly as much as 1/3 of the training data. Table IX showed the classification result which was done by using SVM algorithm in different kernel. Experiments on each kernel will be repeated 5 times then find the average value, this is done to ensure that the results were correct because they have the same relative value when using different testing data. This experiment shows that the linear kernel gives best results of the classification question.

V. CONCLUSION

The testing done showed that the method can be used to classify question based on new bloom taxonomy automatically. The used of lexical and syntactical feature were success to classify the questions with better accuracy result than the previous research. The successfulness of questions classification in many lessons of Elementary School proved that lexical and syntactical feature can be used in other case studies without doing recollection keywords or rule in the classification process. Besides, this research showed that classification accuracy result used linear kernel having the best accuracy value than using polynomial kernel and RBF kernel. For future work, this research will be developed to used of semantics feature.

ACKNOWLEDGMENT

The great gratitude to all teachers and Headmaster of SDN Dono III Sendang who helped this research.

REFERENCES

[1] Anwar Ali Yahya and Addin Osman, "Automatic Classification Of Questions Into Bloom's Cognitive Levels Using Support Vector Machine," *IEEE*, 2011.

[2] Xuan Hieu Phan, Le Minh Nguyen, and Susumu Horiguchi, "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections," in *ACM*, 2008, pp. 91-100.

[3] Jun Wang, Lei Li, and Fuji Ren, "An Improved of Keywords Extraction Based on Short Technology Text," in *International Conference on Natural Language Processing and Knoledge Engineering (NLP-KE)*, August 2010, pp. 1-6.

[4] Dhuha Abdulhadi Abduljabbar and Nazlia Omar, "Exam Question Classification Based On Bloom's Taxonomy Cognitive LEvel Using Classifiers Combination," *JATIT*, vol. 78, pp. 447-455, August 2015.

[5] Norazah Yusof and Chai Jing Hui, "Determination of Blooms Cognitive Level of Questions Item Using Artificial Neural Network," in *International Conference on Intelligent System Design and Applicatons*, October 2010, pp. 866-870.

[6] Syahidah Sufi Haris and Nazlia Omar, "A Rule-based Approach in Bloom's Taxonomy Question Classification Through Natural Language Processing," in *IEEE*, 2012, pp. 410-414.

[7] Nazlia Omar et al., "Automated Analysis of Exam Questions According to Bloom's Taxonomy," in *IEEE*, 2012, pp. 297-303.

[8] Nyoman Sukajaya, Mauridhi Hery Purnomo, and I Ketut Eddy Purnama, "Intelligent Classification of Learner's Cognitive Domain using Bayes Net, Naive Bayes, and J48 Utilizing Bloom's Taxonomy-based Serious Game," *International Journal Of Emerging Technologies in Learning (IJET)*, pp. 46-52, 2015.

[9] Retno Utari, "Taksonomi Bloom Apa dan Bagaimana Mengunakannya?," 2011.

[10] Anbuselvan Sangodiah, Rohiza Ahmad , and Wan Fatimah Wan Ahmad , "A Review in Feature Extraction Approach in Question Classification Using Support Vector Machine," *IEEE*, pp. 536-541, 2014.

[11] Ali Harb, Michael Beigbeder, and Jean Jacques Girardot, "Evaluation of Question Classification Systems Using Differing Feature," *IEEE*, 2009.

[12] Babak Loni, *Enhanced Question Classification With Optimal Combination of Features*.: Delft University of Technology, 2011.

[13] Rosa A. Sukamto and Dwi H. Widyantoro, "Indonesian Parsing using Collins's Parser," 2009.

[14] Bonifacius Vicky Indriyono, Ema Utami, and Andi Suyanto, "Pemanfaatan Algoritma Porter Stemmer Untuk Bahasa Indonesia Dalam Proses Klasifikasi Jenis Buku," *Jurnal Buana Informatika*, pp. 301-310, 2015.

[15] Somnath Banerjee and Sivaji Bandyopadhyay, "Bengali Question Classification: Towards Developing QA System," in *SANLP*, Mumbai, 2012, pp. 25-40.

[16] Abdiansyah and Anny K Sari, "Survey: Question Classification untuk Question Answering System," in *SNATI*, Yogyakarta, 2015, pp. 1-9.

[17] V Srividhya and R Anitha, "Evaluating Preprocessing Techniques in Text Categorization," *International Journal of Computer Science and Application Issue 2010*, pp. 49-51, 2010.

[18] Ning Kang, Carlotta Domeniconi, and Daniel Barbara, "Categorization and Keyword Identification of Unlabeled Documents," *IEEE*, 2005.