Automatic Classification of Answers to Discussion Forums According to the Cognitive Domain of Bloom's Taxonomy using Text Mining and a Bayesian Classifier

Jhonny Pincay
Centro de Tecnologías de Información
Escuela Superior Politécnica del Litoral
Ecuador
jpincay@cti.espol.edu.ec

Xavier Ochoa
Centro de Tecnologías de Información
Escuela Superior Politécnica del Litoral
Ecuador
xavier@cti.espol.edu.ec

Abstract: This paper presents the implementation of an automatic classification system to categorize student's answers in discussion forums according to Bloom's Taxonomy. Several works had been completed in this area, in this occasion the effectiveness of a Bayesian classifier solving this problem is analyzed alongside with the usage of text mining techniques and a dataset previously classified by experts. The possibility of considering the automatic classification system as a human coder is also explored. Several tests were conducted with the objective of obtaining metrics whose values allowed to rate the performance of the classification and the quality of the results.

Introduction

Bloom's Taxonomy is a cognitive skills taxonomy that has been widely used for learning objectives measurement and assessment (Khairuddin & Hashim, 2008). It is an essential concept that guides educators in writing learning objectives, preparing curricula and creating assessment (Yahya & Osman, 2011). Proposed in 1956 by Benjamin Bloom, this taxonomy divides learning objectives into three domains: cognitive, psychomotor and affective (Chang & Chen, 2009). The cognitive domain is related with intellectual knowledge, skills and abilities of an individual, which are classified from a low level that implies just recalling to high levels that denotes the creation of new knowledge and complex and abstract mental levels. The psychomotor domain is related to manual or physical skills which involve fine motor skills or some act that requires a neuromuscular coordination; the affective domain focuses on the growth in feelings or emotional areas, attitudes and the degree of acceptance or rejection (Hui, 2009).

Anderson and Krahwohl modified Bloom's cognition Taxonomy in 2001 being the biggest change the switch from the original unidirectional construction to knowledge with cognitive process bilateral construction; these changes were later known as Revised Bloom's Taxonomy (Chang & Chen, 2009). There are six categories or levels in the cognitive domain of the revised version of the taxonomy, which are listed from the simplest to the most complex (Churches, 2008), (Forehand, 2010): Remembering, Understanding, Applying, Analyzing, Evaluating and Creating.

The benefits of including the use of Bloom's Taxonomy in the educational and teaching practices are undeniable, however the process of categorizing questions, opinions and arguments is tedious, very time consuming and prone to mistakes, considering that is a process done manually. To make a proper classification under this taxonomy training and experience are required, if this is not met it will result in erroneous categorization and incorrect identification of the student level, even for educators with a certain grade of experience in the categorization this process is not free of errors and takes lot of time (Yahya & Osman, 2011), these are few reasons why most of the times the classification is avoided. Also to have an accurate classification, the participation of more than one person in the process is necessary and if the conclusions differ from one another, a discussion is needed to reach a final decision.

The reasons presented above have motivated to look for methods to automate the classification

process; thereby multiple implementation techniques had been proposed to solve this problem: text mining (Yahya & Osman, 2011), natural language processing (Sebastiani, 2002), intelligent systems with neural networks (Yusof & Hui, 2010), etc. Unfortunately, most of the proposed solutions are neither precise enough, nor available for users and in the end are not real solutions for the issues presented (Hui, 2009), (Yusof & Hui, 2010).

The task of automatic classification of questions, opinion and contributions can be casted as a text classification problem. Text classification is also known as text categorization and is the activity of labeling natural language text with thematic categories from a predefined set. Text mining is being used to denote the tasks of analyzing large quantities of text, detecting usage patterns and extract probably useful information that is not obvious at first instance. Under these criteria, text classification is an instance of text mining (Sebastiani, 2002). This paper proposes the implementation of an automatic classification system for contributions in discussion forums, employing text mining techniques and the use of a Bayesian Classifier; we also want to measure how accurate this classifier is in the specific task of assigning a category of Bloom's taxonomy to some text. The other purpose of this study is to determinate if the results provided by the classification system, are good enough to replace human coders.

The structure of the paper is as follows: related work section presents similar systems and some research conducted in this area, system architecture section introduces the components used to build the proposed system of classification, the methods employed to evaluate the effectiveness of the system are described in the section system evaluation, results obtained are presented in finding and results section, the paper finalizes with some conclusions.

Related Works

Having a system that automatically classifies text, opinions, questions, etc. according to Bloom's Taxonomy is a topic that has several proposed solutions and approaches. One of them is the presented by (Yusof & Hui, 2010). They implemented a system capable of classifying questions used to construct exams and evaluation instruments, this solution used a neural network as a classifier and was trained employing a learning algorithm of conjugate scale. One of the issues presented in this attempt was the poor scalability of the neural network, so it was necessary to apply several methods of feature reduction. The conclusions of this work were that considering the particular characteristics of their training set, the most effective method for feature reduction was the document frequency method because it kept the accuracy stable while the rate of convergence improved. Regarding to the accuracy of the classification, it was of 65.9% using a training set of 192 sample questions and getting a total of 605 characteristics.

Another approach was presented by (Yahya & Osman, 2011) the purpose of this study was similar to the mentioned above, but in this case instead of using a neural network they used a support vector machine – SVM. They evidenced that using a SVM it is possible to obtain satisfactory results regarding to accuracy and precision, however the fact of not having a good training set provoked a low result for recall and because of this they were unable to obtain more conclusive results. The training set used by them consisted of 190 questions which were uniformly distributed between the different levels of the taxonomy, although it was not a large number of examples they obtained an accuracy of 87% visibly better than 65.9% which was the result reached using a neural network as classifier.

In a similar way (Chang & Chen, 2009) proposed a system to automatically analyze the quality of question items used in exams based on Bloom's Taxonomy, the classification was performed by an expert system based on rulers where the definition of the rules was the most important and difficult task. The best result they reached was 51% of accuracy and they concluded that if the words or verbs used on each level are not in the knowledge base, the expert system cannot perform the inference with success.

Another study on question classification is presented in (Chang & Chung, 2009); in this work an online system was implemented. This system allowed teachers to enter questions they want to include in some test and the system returned the level of the taxonomy where the question was. The system architecture did not consider the inclusion of an automatic classifier, the determination of the Bloom's Taxonomy depended on words and verbs associated to each level that were stored on a database and to get a conclusion was necessary to compare each word of the question with the stored words, and the category with most words of the questions was the selected. The best result obtained was 75% of accuracy for the level of knowledge, while for the other levels was very low since the effectiveness was about 20%.

The solutions exposed above have in common that of all them use a similar step of preprocess of text and feature reduction, on the other hand the main differences are the classification method employed. Comparing these few studies is evident that the use of learning algorithms gives better results that using

comparison against a database. It should also be emphasized that none of the proposed solutions achieves 100% effectiveness, the main causes are the nonexistence of a fully effective classifier and not having larger and better training sets. This factors happens to be highly influential in the percentage of effectiveness achieved by the classifiers, if the examples of the dataset are not uniformly distributed through the levels of the taxonomy, the results will be good in some cases, regular in others and also could be totally bad (Yahya & Osman, 2011).

In this work the use of a Bayesian classifier specifically Naïve Bayes to obtain the classification model used to perform the prediction of the category is proposed; some tests are also conducted with aims to determinate if the automatic classification system can replace human coders. An aspect that should be emphasized is that the existent solutions are all implemented in English and the algorithms used in the preprocess steps are optimized for this language, so implementing a system of this type for items written in Spanish constitutes a significant contribution to this research line.

System Architecture

The automatic inference system consists of two main components:

- Learning Component
- Classification Component

To carry out the prediction of the label, the inference engine needs a classification model, this model is provided by the learning component using text mining processes, a learning algorithm and a training set. The input of the classification component is some text whose label is unknown; this text is preprocessed and the rules of the model previously obtained are applied and finally the most probable category of the text is returned. Both the learning component and the classification component were implemented using RapidMiner, a software for data mining and analysis developed in Java programming language. It allows the creation of simple and complex processes chaining a variety of operators of extraction, processing and display of data through a highly intuitive graphical environment (Jungermann, 2009).

Learning Component

The learning component consists of several subcomponents which are shown in Fig. 1; the input of this component is a training set that is going to be used to build a classification model. The training set is constituted by answers of students to discussion forums which had been categorized according to the Bloom's taxonomy by experts. The elements of the training set must pass through the subcomponent of text preprocessing and vector creation which is responsible for representing the problem so it can be understood by the learning algorithm. A vector representation is the way the text is represented, these vectors consist of features extracted from each category and each feature has an associated value that indicates how influential the feature for a given category is.

The elements of the training set go through several stages until get its vector representation, it is necessary to remove any information that is not relevant and influential that is why the text preprocessing is indispensable. The steps performed by the text preprocessing subcomponent are tokenizing, delete stop words and stemming; through the tokenization each word of a text is isolated, later stop words such as interjections, conjunctions and auxiliary verbs that do not affect the message to be transmitted, are deleted; finally, the root of the remaining words are extracted using a stemming algorithm that in our case was Snowball. The final form of the words is called stem. At this point there are several stems per category of the taxonomy, these stems are the components of the feature vectors; because the number of features of the vectors was large and thus affect the effectiveness and performance of the system, a process of feature reduction was applied. The feature reduction method used was document frequency – DF since for this type of problem it gives best results in terms of effectiveness and convergence times (Yusof & Hui, 2010); once the most representative features were selected, the vectors are ready to be the input of the learning algorithm. Fig. 2 shows the elements of the subcomponent of text preprocessing and vector creation.

Figure 1: Schema of the structure of the learning component

Figure 2: Schema of the structure of the subcomponent of text preprocessing and vector creation

The learning algorithm or classifier is the subcomponent that produces the model necessary to execute the prediction of the category; despite a large variety of text classifiers, Naïve Bayes was chosen because of its simplicity and effectiveness (Ikonomakis, Kotsiantis, & Tampakas, 2005). This is a probabilistic classifier which applies Bayes' theorem to determinate the probability that a document represented by a vector of terms belong to a category c_i. The Bayes' theorem is given by the equation presented below:

The event space is the space of documents, is thus the probability that a randomly picked document has vector representation and that a randomly picked document belongs to c_i. The estimation of is problematic since the number of possible vectors is too high, this in order to alleviate this problem the assumption that any two coordinates of the document vector are statistically independent. Probabilistic classifiers that use this assumption are known as Naïve Bayes (Sebastiani, 2002). In the end a probabilistic model is obtained which consists of the features extracted from the training set and the associated probabilities of each feature for each category which together are used when determining the level to which a text belongs (Ikonomakis, Kotsiantis, & Tampakas, 2005).

Classification Component

Figure 3: Schema of the structure of the classification component

The classification component is responsible for determining what level within the taxonomy of some text. In the same way as the elements of the training set, is necessary to apply preprocessing to the text received as input because it contains irrelevant words, misspelled words and punctuation, in this way we obtained the main features of the text to be classified and thus is prepared to apply the classification model.

Determining whether or not some text belongs to a level is done by applying the learned probabilistic model, this implies that given the features of the text is calculated the probability of belonging to a category or not is calculated. The returned result is the level of taxonomy most likely belongs to the text. Fig. 3 shows the schema of the structure of the classification component.

System Evaluation

Definition of the Training Set

The training set used to obtain the classification model consists of answers given by students of Escuela Superior Politécnica del Litoral – ESPOL to different discussion topics raised according to a particular subject; these responses were collected from 2010 through 2012 and were labeled by different experts. First, the answers to the discussion forums collected were labeled each by three different coders, who categorize them according to their personal criteria. Then they compared the labels assigned by each of them and discussed the reasons why they assigned that label in order to reach a consensus. Finally, the coders assigned the level of Bloom's Taxonomy. It should be noted that the responses were taken from Spanish to English because the coders did not speak Spanish. We consider this factor did not influence the quality of the results because the translations were done trying to preserve its message and the main ideas. Another aspect that should be emphasized is that the coders were different each year, but the number of them and the procedure to proceed to assign the level of Bloom's taxonomy to answers was the same.

The training subset number one was collected in 2010. A case study related to computer security was raised to a group of graduate students of ESPOL, the students were separated into five groups of four members and each group had to provide a solution to the problems presented in the case study and conclusions needed to be reached within 48 hours. One example of a translated post labeled as Understanding is "Type of support in an Open Source solution? In any case, our approach must be taken in consider support, in the case of CISCO, the product has a support office exclusive local consultations for any level and physical level. Of course, we have to investigate what is the support that provides open source companies?"

The training subset number two was obtained in 2011 and was responses of students of the subject Software Engineering to a discussion forum which should deliver as a result, a document of risks assessment of the projects they were developing at that time. The activity was done by for thirty-five students that were divided into seven groups of five members. A particular feature of that year's activity was that students were trained on how to classify their responses according to Bloom's taxonomy and they had to label each contribution according to what they learned. Subsequently labels assigned by students were compared with those assigned by expert coders, in order to reach consensus and determine the final category. An example of a post labeled as Remembering is "Another risk to take in consideration is a case in which a member of a group leaves the project, this person could have a critic role in the software development, and this is a personal risk."

In 2012 discussions were held on the subject interactive multimedia applications; there were two discussions one of them was decide the best web framework given certain conditions and the second was to decide what the best framework for mobile application development is. This time, the students had to take different roles within the discussion and they were divided into four groups from six to eight members. Each team had to get a final document outlining their proposals and explaining the reasons why they chose a particular platform. One example of a post labeled as Analyzing is "In my opinion regarding the frameworks, I think a lot depends on the type of website that we do, that is, if it's a blog is the best option is Wordpress to manage it very efficiently although this does not leave very useful for many other types of websites in other cases a framework web as Ruby on Rails is more convenient because of its flexibility", it should be noted that for this subset in particular the labels assigned per each coder are available.

The final training set is formed from the joining of the training subsets of the years 2010, 2011 and 2012, being finally established as shown in Tab. 1; some issues that need to be emphasized is that for the level of Creating in all years and Applying in the years 2010 and 2012, there are not examples because none of the student's answers reached that level of the taxonomy and although data was collected from three different years, the number of responses is not uniformly distributed among all levels and in many cases this factor influences the performance of automatic classifiers

Bloom's Taxonomy Level	Number of answers year 2010	Number of answers year 2011	Number of answers year 2012	Number of answers of the final training set
Creating	0	0	0	0
Evaluating	8	2	7	17
Analyzing	9	22	32	63
Applying	0	7	0	7
Understanding	13	46	155	214
Remembering	11	30	78	119
Total	41	107	271	420

Table 1: Number of answers per level of the three training subsets and of the final training set.

Definition of Metrics and Testing Set

An evaluation of the system was done with the objective of determining the effectiveness of the system performing the classification of the responses to the discussion forums. The way in which the effectiveness was measured was through several measures and metrics that are presented in a contingency table which usually consist of the following values (Yahya & Osman, 2011):

- **A:** The number of responses that the system correctly assigned to the category, i.e. the true positives.
- **B:** The number of responses that the system incorrectly assigned to the category, these are false positives.
- **C:** The number of responses belonging to a class, but the system assigns another, i.e. false negatives.
- **D:** The number of responses that the system does not properly assign to the category, which are the true negatives

The measurements commonly used to determine the effectiveness of a classifier include (Hui, 2009):

- **Precision:** Is the probability that a document or response is classified under a category, this decision is correct. Its value is in the range zero to one, one being the best value.

- **Recall:** Is the probability that if a random document ought to be classified under, this decision is taken.

- **Accuracy:** This is the most commonly metric used to measure the effectiveness of a classifier; however accuracy values are less hesitant to variations in the number of true positives than the measures of recognition and accuracy.

- **:** This is the harmonic mean of recognition and precision. The value of β depends on the importance given to the precision and the recall, if the precision is considered more important than recall then the value of β should be zero. If the recall is more important than the precision, the value of β should be taken to infinity, finally if the precision and recall are equally important the value of β must equal one. The equation for a value of β =1, is defined as follows:

To determine the values of A, B, C and D counting with a testing set was necessary. The testing set consists of elements that were randomly chosen from the total of examples available of all the years. We already know the level of Bloom's Taxonomy they belonged and were not part of the training set; the test measured that the classifier categorizes the examples of the testing set and then, the response returned by the classifier is compared to the true level of the answer, thus determining the true positives and negatives and false positives and negatives. Tab. 2 shows the number of examples per level of the testing set used.

Bloom's Taxonomy Level	Number of Answers
Creating	0
Evaluating	7
Analyzing	27
Applying	3
Understanding	92
Remembering	51
Total	180

Table 2: Number of answers per level of the testing set

It is also necessary to determine whether the automatic classifier can act as one human coder, considering that the classification process usually is performed by several people and the chosen decision is the result of consensus. That is why a test of inter-rater reliability with the Krippendorff's alpha statistic which indicates the degree of agreement between coders was applied. The alpha value should be between 0.75 and 1 to conclude that the coders agree on most of the labels assigned, while if the value is less than 0.75, it means that the decisions issued by the coders did not match and the level result consensus is not reliable.

Findings and Results

The result of adding the values of true positives, false positives, false negatives and true negatives, are presented in Tab. 3, the results are presented for each level of the taxonomy except for the level of Creating because of the lack of examples in this category in the datasets used.

Bloom's Taxonomy Level	A	B	C	D

Evaluating	7	37	0	135
Analyzing	16	13	11	140
Applying	3	4	0	173
Understanding	50	7	42	81
Remembering	28	15	23	114

Table 3: Results of the count of the values of A, B, C, and D after the application of the classification model to the testing set

It is possible to define the values of A and D as good results whereas B and C are errors. A and D values indicate that the classification is working properly while B and C are indicators that the predictions performed by the classifier are not accurate; A represents the true positives and C the false positives, this implies that adding these values the result must be equal to the total of elements of the training set in an specific level while the result of adding the values of A, B, C, and D per level must be equal to the total of elements of the testing set.

The resulting values of the metrics from the application of the classification model to each of the elements of the testing set are shown in Tab. 4. These values were calculated using the equations given previously and using the values of A, B, C and D showed in Tab. 3.

The results of the four metrics showed in Tab. 4 presented significant variances between them; regarding to precision the highest value is 0.88 for the level of Understanding while the lowest value is 0.16 for Evaluating, this means that the 88% of the times the classification to the level of Understanding is done correctly while in the level of Evaluating the prediction is wrong most of the times. The big differences among the values surely were caused because the elements of the training set were not uniformly distributed through the levels of the taxonomy, so the highest values of precision are for the levels of Understanding, Remembering and Analyzing which are the levels with more examples. The results for accuracy evidence the same pattern as the values of precision; the best value is for Understanding and the worst for Evaluating, in general the results of accuracy are lower than precision and not satisfactory.

Bloom's Taxonomy Level	Precision	Recall	Accuracy	
Evaluating	0.16	1.00	0.04	0.27
Analyzing	0.55	0.59	0.15	0.57
Applying	0.43	1.00	0.02	0.60
Understanding	0.88	0.54	0.51	0.67
Remembering	0.65	0.55	0.28	0.60
Average	0.53	0.74	0.2	0.54

Table 4: Classification effectiveness for each level of Bloom's Taxonomy

Recall values were in general around 50%, this means that a random text external to the dataset is classified correctly only half of the times. As the values of recall and precision were low, the results of F – measure were low too; this value indicates that in general the effectiveness of the classifier was of 67% for Understanding and 27% for Evaluating, once again these results turned out this way because of the structure and distribution of the training set. Comparing the best result obtained which was precision for the level of Understanding with the values obtained for similar systems, it should be emphasized that 88% is the best among the revised solutions, since the result for the system proposed by (Chang & Chung, 2009) was 26% for the equivalent level of comprehension and 50% for the SVM classifier of (Yahya & Osman, 2011). If we consider the precision reached for the level of Analyzing, it was lower than the value obtained by the SVM classifier which was 75% and higher than 32% that was the value reached by the proposal of (Chang & Chung, 2009). Overall, the effectiveness reached by the Naïve Bayes classifier is inferior to the obtained by the SVM classifier, but it showed a better performance than a neural network classifier (Yusof & Hui, 2010) and an expert system (Chang & Chen, 2009) in the levels that were trained using a higher number of examples. Even though some results were good, the systems still needs improvement and the use of a better dataset for training.

Regarding the inter-rater reliability test, it was completed but only using the elements of the answers collected in 2012, because to perform this test is necessary to know the assigned label of each coder; the objective of this proof was to determine if the inference system can act as an additional human coder, so several tests were performed and the results obtained per each experiment are shown in Tab. 5.

Test	Krippendorff's Alpha value
Three humans reliability	0.7189
Humans and automatic system reliability	0.53
Reliabitily withouth human one	0.579
Reliabitily withouth human two	0.418
Reliabitily withouth human three	0.418

Table 5: Results of the inter-reliability test

The Krippendorff's alpha value for the labels given by the three human coders was 0.718 which indicates that they have a high degree of agreement between their decisions; adding a fourth coder being this the automatic classification system the result was 0.53, this value suggests that the conclusions reached by the coders differ a lot between them and the coder that is introducing noise is the automatic system; the last three values shown in Tab. 5 confirm this, since excluding the human coder from the test the values remain low.

Conclusions

This paper proposes the implementation of an automatic classification system to contributions in discussion forums according to Bloom's Taxonomy, employing text mining techniques, the use of a Bayesian Classifier and a pre-classified data set of answers. The results obtained indicate that using the proposed architecture is possible but the results are highly dependent of the quality of the training set used to generate the classification model. The value of the metrics suggested that the Naïve Bayes classifier performs relatively well when there are a considerable number of examples to train. In general, the fewer number of examples are available, the less accurate the prediction is; this is an issue presented in most of the works in this field and some important questions still remain unanswered regarding to the improvement of the performance and results of the classification: what are the characteristic needed to be a good training set?, what is the ultimate classifier for this kind of problems?, what is the ultimate method for feature reduction?.

Despite the general low results obtained, it should be emphasized that the system provides a considerably good precision when it comes to the levels of Understanding, Remembering and Analyzing, so the results returned by the system could help a teacher to have a quick and approximate vision of the level reached by their students and whether they are achieving or not some learning objective, also this could help to the educators to improve the way they are teaching to a particular group of students. Another important point is that this entire implementation is for texts in Spanish, almost all the implementation of similar systems are for items in English and considering that Spanish is one of the languages with most native speakers in the world, makes that this approach constitutes a significant contribution.

Future work will focus on trying to obtain better results by getting and experimenting with larger datasets and different classifiers; also the feature reduction and transformation is a step in the process that needs to be carefully analyzed and selected. .Finally, if better results are obtained the values of the Krippendorff's alpha statistic will become subsequently higher and this will let to conclude that the system effectively can replace human coders.

References

Chang, W., & Chung, M. (2009). *Automatic Applying Bloom's Taxonomy to Classify and Analysis the Cognition Level of English Question Items.* IEEE.

Chang, Y., & Chen, H. (2009). An Automatic Inference System for the Quality Analysis of Test Items Based on The Bloom's Revised Taxonomy. *Proceedings of the Eight International Conference on Machine Learning and Cybernetics*, (pp. 2852-2856). Boading.

Churches, A. (2008). *Bloom's Taxonomy Blooms Digitally.* Retrieved August 25, 2012, from techLearning: http://www.techlearning.com/article/blooms-taxonomy-blooms-digitally/44988

Forehand, M. (2010). *Bloom's Taxonomy from Emerging Perspectives on Learning, Teaching and Technology.* The University of Georgia.

Hui, C. (2009). *Feature Reduction for neural Network in Determining the Bloom's Cognitive Level of Question Items.* Universiti Teknologi Malaysia.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification using Machine Learning Techniques. *WSEAS Transactions on Computers*, 966-974.

Jungermann, F. (2009). Information Extraction with RapidMiner. *Symposium Sprachtechnologie and eHumanities*, (pp. 50-61). Duisburg.

Khairuddin, N., & Hashim, K. (2008). Application of Bloom's Taxonomy in Software Engineering Assessments. *Proceedings of the 8th International Conference on APPLIED COMPUTER SCIENCE*, (pp. 66-69).

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 1-47.

Yahya, A., & Osman, A. (2011). *Automatic Classification of Questions into Bloom's Cognitive Levels using Support Vector Machines.* Najran: Najran University.

Yusof, N., & Hui, C. J. (2010). *Determination of Bloom's Cognitive Level of Question Items using Artificial Neural Network.* IEEE.