Automatic Labeling of Forums Using Bloom's Taxonomy

Vanessa Echeverría[1], Juan Carlos Gomez[2], and Marie-Francine Moens[2]

[1] Centro de Tecnologías de Información,
Escuela Superior Politécnica del Litoral,
Km 30.5 vía Perimetral, Guayaquil, Ecuador
vecheverria@cti.espol.edu.ec
[2] Department of Computer Science, KU Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
{juancarlos.gomez,sien.moens}@cs.kuleuven.be

Abstract. The labeling of discussion forums using the cognitive levels of Bloom's taxonomy is a time-consuming and very expensive task due to the big amount of information that needs to be labeled and the need of an expert in the educational field for applying the taxonomy according to the messages of the forums. In this paper we present a framework in order to automatically label messages from discussion forums using the categories of Bloom's taxonomy. Several models were created using three kind of machine learning approaches: linear, Rule-Based and combined classifiers. The models are evaluated using the accuracy, the F1-measure and the area under the ROC curve. Additionally, a statistical significance of the results is performed using a McNemar test in order to validate them. The results show that the combination of a linear classifier with a Rule-Based classifier yields very good and promising results for this difficult task.

Keywords: CSCL, Bloom's taxonomy, logistic regression classifier, Rule-Based classifier, combined classifiers.

1 Introduction

Discussion forums are considered as an application of Computer-Supported Collaborative Learning (CSCL). The goal of CSCL is to create a computer environment for assessing educational goals through a shared activity among participants of a discussion board [15]. The participants can construct their own knowledge through this social interaction by sharing ideas and negotiate their validity which leads to an active participation in this collaborative activity [8].

In order to assess the educational goals achieved in a learning environment, Bloom's taxonomy (BT) plays an important role. Benjamin Bloom created a categorization of learning objectives to evaluate learning outcomes [1], and former students of him modified this categorization to include the analysis of cognitive processes of participants [7]. The categories that Bloom's students proposed

H. Motoda et al. (Eds.): ADMA 2013, Part I, LNAI 8346, pp. 517–528, 2013.

are: *Remembering*, *Understanding*, *Analyzing*, *Applying*, *Evaluating* and *Creating*. Several studies have fostered BT to analyze the contributions of participants in the discussion forums at a cognitive level [13].

In a CSCL environment the information from forums needs to be labeled using BT. However, in large databases of information usually not all the data is labeled because the overall process is time-consuming, needing human resources with an appropiate background in education [3]. This is an important motivation for migrating from a manual to an automatic labeling by selecting automatically a level from the taxonomy and assign it to a discussion forum's response.

The automatic labeling of discussion forums with a cognitive level from BT could be considered in essence as a text categorization (TC) problem. TC is defined as a supervised task of assigning a value of true or false to a document with regard to the assignment of a certain category c_j, where $\mathbf{d}_i \in \mathbf{D}$ is the i-th element of the collection of documents $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, ...\mathbf{d}_n\}$. Thus, each document $\mathbf{d}_i$ is labeled with a category $c_j \in \mathbf{C}$, which is the j-th element of the set of categories $\mathbf{C} = \{c_1, c_2, c_3, ...c_k\}$ [12].

Several studies have been conducted for automatic labeling of questions using BT. These questions are used for designing an effective test in order to evaluate the skills of a participant using the taxonomy. Those studies yielded systems for labeling questions using Artificial Neural Networks (ANN) [14], a Rule-Based classifier [9] and a bank of words using weights for each category [2]. However, it is important to notice that those studies addressed the automatic labeling of questions, which is not equivalent to a text containing expressions and thoughts from a person. Additionally, a study by [10] implemented a system for automatic labeling of discussion forums, but a deeper analysis of the results was not performed. As far as we know, there is not a relevant study about the automatic labeling of discussion forums.

In this paper, we conduct an analysis of several models based on supervised machine learning methods in combination with the use of several features for labeling discussion forums using BT. Such models are compared using different evaluation measures and performing statistical significance tests.

The problem presented in this paper is treated as TC task, however it presents an important difference with common TC problems resulting from the complexity for modeling the data. The goal of most TC task is to find features that can be common for a category or topic (directories, news, emails, etc.). In our case, the categories to be assigned to a forum are not topics, but levels of a cognitive process. This makes this task harder than a common TC problem, since the data that we are dealing with are not simple documents containing information related to a given category or topic, but rather information about the cognitive level of persons, which is indeed topic independent.

Three approaches are used in order to have a wide coverage of results. The first one consists on training linear classifiers, in particular Support Vector Machine (SVM) and Logistic Regression (Logit), using the following features: words, verbs, bigrams and trigrams; weighting such features using tf-idf (term frequency - inverse document frequency) [11]. Linear classifiers are well known to have good

performance in TC tasks [4]. In this approach, we also perform a dimensionality reduction using Principal Component Analysis (PCA) [5] on the feature space and then use the reduced feature space with the linear classifiers. The second approach implements the creation of Rule-Based (RB) classifiers using the following features: verbs, bigrams and trigrams. Finally, the third approach comprises the combination of classifiers resulting from the first and second approaches [6].

For all the models, their performance is computed based on accuracy, F1-measure and Area Under the ROC Curve (AUC). From all the models created, at the end we select the best five models based on those measures for conducting a deeper analysis and a statistical comparison using the McNemar test.

The contributions of our work are: 1) the addressing of a novel and hard task that has not been explored in detail; 2) the good labeling results obtained; 3) the creation of highly effective but simple rules in a RB classifier; 4) the observation that a combination of classifiers, where the prior knowledge has been modeled, is a good approach to follow in order to enhance the general performance; 5) a deep experimental and error analysis of the performance of the different models, considering common measures and a statistical comparison of the results, and providing rationales for some methods performing better than others.

The remainder of this paper is organized as follows: in section 2, we provide the related work that has been carried out when implementing automatic classification of text using BT. Section 3 is dedicated to describe the framework used for finding a suitable model for this TC task, including a detailed description of the data collection and the experimental setup for each model. In Section 4 we provide the experiments and results together with a deeper analysis of them. Finally, in Section 5 we present the conclusions and future work.

2 Related Work

In the area of CSCL Several studies have been conducted for automatic labeling of examination questions from different areas (courses) using Bloom's taxonomy (BT). In [14], the aim of the study was to label a set of questions using Artificial Neural Networks (ANN). The model presented there used words to form an initial feature set, and then two feature reduction methods were performed: document frequency (DF) and category-frequency document-frequency (CF-DF), which introduces a discriminant value for features that appeared in few categories. Seven models were trained using a 3-layer feed forward neural network. The models were evaluated using precision, convergence time and error. Results showed that the use of all words reached the best performance according to the given measures.

The authors of [2] tackled the task of dealing with multiple keywords in one question. To solve the problem, they compared keywords extracted for each BT category. They stressed that the verbs in a question are the most important keywords to represent a category. To deal with shared keywords among categories, they gave weights to the keywords. Thus, in the test phase, when an unseen question was analyzed, they compared the verb with the database of keywords

constructed previously and the category with higher weight was the predicted one for that example. The evaluation measure used in this study was the number of correct matched items. Moreover, the results of this study showed that the category with the label "Knowledge", which corresponds to the first cognitive level on BT, achieved a good performance (57 percent), while for other categories, the performance was lower.

In [9] the authors determined the appropriate category for exam questions using a Rule-Based approach, employing keywords and verbs. Initially, they had a collection of questions that were manually categorized by a group of experts in the programming domain. Then, to beginning the analysis of the questions, they tag each question to obtain the question structure and then, find patterns according to the tags. They also mentioned that there were certain shared patterns among categories. Therefore, to overcome this problem, they gave weights to each category. Those weights were set by experts, meaning that they did not have a good method to determine the weights of the categories; which makes the analysis expert-dependent. Finally, they created rules based on the found patterns questions and the prediction is the category with the highest weight.

The closest study related to the automatic labeling of forums is presented in [10]. There, the authors trained a Bayesian classifier over a set of 420 documents in Spanish using document frequency for feature selection. The results of their study showed that their model reached an accuracy of 51 percent for the category "Understanding", while for the other categories the performance was lower.

Although most of the previous studies label questions using BT, they do not label large texts or responses from participants in order to identify the participant's cognitive level. Furthermore, we can observe from those studies a lack of a deeper analysis regarding the employed evaluation measures. Nevertheless, we observe that most of the studies show that there exists shared information among the categories of BT.

3 Framework

In this paper we propose a framework for performing the automatic labeling of forums using the cognitive levels of BT under several models. The framework is divided in three main parts: preprocessing, indexing and learning-testing phase. Figure 1 shows the framework used for creating our models.

3.1 Document Collection

The data used for this work was gathered from nine discussion forums about several topics[1]. The discussion forums were originally written in Spanish, then

[1] The topics related to the forums are from two courses of a Bachelor in Computer Science: Computer Graphics and Interactive Multimedia Applications; and from an Information Security course in a Master Program of an Ecuadorian University. The data gathered is property of the CTI-ESPOL.

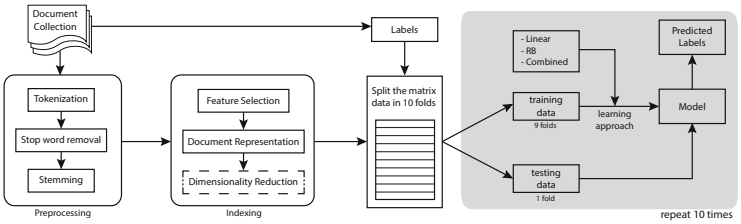

Fig. 1. Scheme for model creation

translated into English by people with an intermediate-advance knowledge of the language.

The manual tagging of forums was accomplished by three independent coders. Each entry of the forum (message) was labeled with one of the BT categories: Remembering, Understanding, Analyzing, Applying, Evaluating and Creating. An additional label *Uncodable* was used in order to track the messages that did not fit in the taxonomy.

After all coders finished the labeling of messages, a meeting between all coders was carried out and each individual result was discussed. The messages with disagreement were debated until an agreement was reached. The result of this activity is a set of messages with one category label for each message. At this point, each message in the set is treated as an individual document.

The dataset used here corresponds to a collection of 463 documents grouped in five categories (there are two categories from the taxonomy without messages). Table 1 shows the number of documents and the corresponding percentage for each category.

Table 1. Frequency of documents per category

Category	Frequency	Percent	Cum. Percent
Remembering	116	24.8	24.8
Understanding	186	39.7	64.5
Analyzing	46	9.8	74.4
Evaluating	27	5.8	80.1
Uncodable	93	19.9	100
Total	**468**	**100**	

3.2 Preprocessing and Indexing

In the preprocessing step, the document collection is analyzed document by document. Thus, we carried out a tokenization process for each document, resulting in a list of words. From this list we perform stemming and stop word removal. Hence, the result of the preprocessing part is a list of relevant words for each document.

The indexing step consists of selecting the features and representing them in machine readable data. Four types of features are identified and used: words, verbs, bigrams and trigrams. First, the verbs were identified from the returned list of words using a tagger. Then, the bigrams were created by joining each verb with the word that is to the left or to the right of it. Finally, the trigrams were created by joining the words that are to the left and right of each verb.

The document representation defines the input of the data for a given learning approach. The input for a linear classifier is a sparse vector of weights. Here we use tf-idf as a weighting factor for the vectors of features. The representation of a RB classifier is given by features related to a category; thus the words, verbs, bigrams and trigrams are used as the representation of a rule.

Dimensionality reduction is dedicated to reduce the number of features and for this work we use Principal Component Analysis (PCA), which performs a linear combination over the document representation matrix. Then, from the matrix of eigenvectors, a fixed number of components is selected and used to project the original matrix, giving a matrix with less but more relevant features. The selection of an optimal number of components q was done using a global grid search over the training set, varying the number of components from 2 to N, where N is the size of the document collection. Finally, each number of components is evaluated considering the classification performance and the best one is selected.

3.3 Training the Models

In this study, we used three types of learning approaches. First, two linear classifiers (SVM and Logit) [2]; second a RB classifier; and third the combination of the linear and RB classifiers. For each learning approach, we implement different models using variants of features or performing a dimensionality reduction.

Initially, the SVM and the Logit classifier were trained using words as features. Afterwards, only the Logit was considered for training models using verbs, bigrams and trigrams as features; and PCA with the same features. The selection of this classifier was given by analyzing the output of each classifier: the output of the Logit (a probability of membership for each category) was better suited for the combination of classifiers rather than the output of the SVM (a yes/no output). For this approach, we obtain nine trained models using the two classifiers with four types of features: SVM (words), Logit (words, verbs, bigrams, trigrams) and PCA Logit (words, verbs, bigrams, trigrams). Additionally, we accomplish for each model a global parameter optimization using a 3-fold cross validation together with a grid search over the training set. The best (regularization) parameter C of the SVM classifier was selected from the set of values {1e-6,1e-5,1e-4,1e-3,1e-2,1e-1,1,1e1}; while for the Logit classifier, the C, the tolerance (ε) and the norm were optimized using the range values [1 - 10],{1e-8,1e-7,1e-6,1e-5,1e-4,1e-3,1e- 2,1e-1,1,1e1,1e2} and {L1, L2} respectively.

[2] A non-linear version of SVM with a RBF kernel was also tested in the experiments, but compared to the linear SVM the performance was worse.

For the second approach, the RB classifiers, we select a set of rules for each category using verbs, bigrams and trigrams as features. First, the training set was separated by categories (giving five groups). Next, from each collection of documents per category we extract the verbs, bigrams and trigrams and compute a frequency distribution of the resulting list. Then, we select the verbs, bigrams and trigrams with a support threshold greater or equal than 1 (thr$\geq$1), 2 (thr$\geq$2) and 3 (thr$\geq$3) to compose the set of rules. However, for the verbs only a support threshold (thr$\geq$2) was selected because there were too many verbs with a threshold equal to 1 and this would affect the generalization of the classifier. In order to add meaning to the rule, the conditional probability that one verb v belongs to a category c_i given the category $P(v \mid c_i) = P(v, c_i)/p(c_i)$ was computed, where $P(v, c_i)$ corresponds to the likelihood of the verb in the category (frequency divided by the size of the list of verbs) and $p(c_i)$ was the likelihood of the category (frequency divided by the number of documents belonging to the category). This conditional probability corresponds to the confidence of the rule. Finally, by varying the support threshold for creating the rules and combining the different features we obtain a set of seven classifiers: RB (verbs thr$\geq$2); RB (bigrams thr$\geq$1; thr$\geq$2; thr$\geq$3) and RB (trigrams thr$\geq$1; thr$\geq$2; thr$\geq$3). Each classifier was represented by a list of rules where each item was associated with its confidence and frequency.

The last learning approach consists of a combination of the Logit classifier and the RB classifier. The idea behind combining models was to improve the general decision boundary by applying a linear combination over the decision surfaces constructed by the individual classifiers. The expected error of the combined classifier will have a lower bound given by the minimum error of the classifiers involved. The Logit classifier throws out a vector of probabilities for each test document (one probability per category). The RB classifier also yields a vector of weights, where an element of the vector corresponds to the sum of confidences regarding one category. The two classifiers can be combined by summing the outputs per category of both models and taking the highest probability value of the resulting output vector. The category associated with the highest value will be the predicted category corresponding to a test document. Thus, two Logit classifiers from the first group (Logit and PCA Logit classifiers, both with words as features) were used for being combined with the seven RB classifiers from the second group. The result of the classifiers of both group yields a total of 14 combined models: Logit (words) + RB, PCA Logit (words) + RB.

In order to have a more robust analysis of the models, each one of them has been tested using a 10-fold cross validation schema. In this schema, nine folds were used for training and one for testing, repeating the process 10 times (until each fold has been used for training and testing the model).

The performance of the models was compared using as evaluation measures the accuracy, F1-measure and the AUC. The final values of these measures were averaged over the 10 folds. Furthermore, the parameter optimization was done using a global approach, meaning that 10 different parameters of the learners

were given using the training data, and the parameter values with the best accuracy among the 10 learners were selected for training the final model.

In total, we train 30 classifiers: 9 linear, 7 RB and 14 combined. From these, the top five classifiers according to the accuracy, F1-measure and AUC were selected in order to develop a deeper performance and statistical analysis.

4 Experiments and Results

4.1 Experimental Evaluation Measures

Three groups of evaluations were conducted based on the three types of learning approaches. The first group corresponds to the linear classifiers which included the SVM and Logit classifiers together with the classifiers using PCA. The second group was composed of the RB classifiers using verbs, bigrams and trigrams as features. The third group included the combination of Logit classifiers with RB classifiers.

First Group: for this group, the SVM (words) classifier trained with the optimal parameter $C = 1$ reaches values of 0.6104, 0.3999 and 0.6815 for accuracy, F1-measure and AUC respectively. the Logit (words) classifier with optimal parameters ($C = 7$, L2 and $\varepsilon =$ 1e-8) reaches values of 0.6225, 0.4096 and 0.7023 for the same measures. When PCA is performed using the optimal number of components (q=39), the PCA Logit (words) classifier yields values of 0.6240, 0.4141 and 0.7122 for the measures, reaching a slightly better performance than the previous models. The performance of the others linear trained classifiers: SVM (words), Logit (verbs, bigrams, trigrams) and PCA Logit (verbs, bigrams, trigrams) is below the given results.

Second Group: for the second group, the accuracy, F1-measure, and AUC for the RB (trigrams thr$\geq$1) classifier are 0.6966, 0.6454 and 0.7838, respectively. The RB (bigrams thr$\geq$1) classifier reaches values of 0.6667, 0.5508 and 0.7494 for the same measures. The performance of the rest of the trained classifiers this group: RB (verbs thr$\geq$2), RB (bigrams thr$\geq$2; thr$\geq$3) and RB (trigrams thr$\geq$2; thr$\geq$3) is below the results of the two aforementioned classifiers.

Third Group: in this group, the Logit (words) + RB (trigrams thr$\geq$1) classifier reaches an accuracy, F1-measure and AUC of 0.7671, 0.6994 and 0.8230, respectively. The PCA Logit (words)+ RB (trigrams thr$\geq$1) obtains 0.7671, 0.7018 and 0.8246 for the same measures. The Logit (words) + RB (bigrams thr$\geq$1) classifier reaches a performance of 0.6838, 0.5703, 0.7542 for the measures. The results of the other combined classifiers: Logit (words) + RB (verbs thr$\geq$2), Logit (words) + RB (bigrams thr$\geq$2; thr$\geq$3), Logit (words) + RB (trigrams thr$\geq$2; thr$\geq$3), PCA Logit (words) + RB (verbs thr$\geq$2), PCA Logit (words) + RB (bigrams thr$\geq$2; thr$\geq$3) and PCA Logit (words) + RB (trigrams thr$\geq$2; thr$\geq$3) are lower than the ones obtained from the aforementioned classifiers.

4.2 Overall Results

After performing the three groups of experiments, the best five models were selected based on the highest values of accuracy, F1-measure and AUC: two of them correspond to linear classifiers (with and without PCA), one to the RB classifier and two to combined classifiers. These are shown in table 2.

Table 2. Summary of classifiers with best evaluation measures

Classifiers	Accuracy	F1-measure	AUC
Logit (words)	0.6225	0.4096	0.7023
PCA Logit (words q=39)	0.6240	0.4141	0.7122
RB (trigrams thr$\geq$1)	0.6966	0.6454	0.7838
Logit (words) + RB (trigrams thr$\geq$1)	0.7671	0.6994	0.8230
PCA Logit (words q=39) + RB (trigrams thr$\geq$1)	**0.7671**	**0.7018**	**0.8246**

Table 3. F1-measure and AUC per category

Category	Logit		PCA Logit		Trigrams		Logit + Trigrams		PCA Logit + Trigrams	
	F1-measure	AUC	F1-measure	AUC	F1-measure	AUC	F1-measure	AUC	F1-measure	AUC
Remembering	0.66	0.84	0.66	0.81	0.75	0.86	0.81	0.89	**0.82**	**0.89**
Understanding	0.65	0.65	0.62	0.66	0.73	0.78	**0.79**	**0.82**	0.78	0.82
Analyzing	0.00	0.00	0.00	0.00	0.60	0.72	**0.61**	**0.72**	**0.61**	**0.72**
Evaluating	0.00	0.00	0.00	0.00	0.53	0.68	**0.53**	**0.68**	**0.53**	**0.68**
Uncodable	0.73	0.90	0.76	0.86	0.62	0.77	0.76	0.83	**0.78**	**0.83**

Table 3 shows the results for F1-measure and AUC for the selected models regarding the individual categories. According to these results, the Logit (words) and PCA Logit (words) classifiers do not perform well in categories Analyzing and Evaluating. In contrast with the linear classifiers, the RB (trigrams thr$\geq$1) classifier reaches better performance for those categories. The results obtained from the combination of classifiers were better than the other ones. Here, the results of categories Remembering, Understanding and Uncodable increased considerably comparing with the results of individual classifiers; while the results for categories Analyzing and Evaluating remained the same as expected, mainly due to the misclassification coming from the linear classifiers.

In order to find the statistical significance of the results, the McNemar paired test was performed over the general results of the top five classifiers. In this test we applied the Bonferroni adjustment factor to calculate a new significance level α^* in order to avoid the multiplicity effect. This was calculated with the formula: $1 - (1 - \alpha^*)^n \leqslant \alpha$, where n was the number of learned models and α was the two tails significance level (0.05). Replacing the values and solving the equation, the new significance level for which the result of a paired test was compared is $\alpha^* = 0.0017083$. Thus, the paired test was performed between each pair of classifiers from table 2. Furthermore, The p-value resulting from the McNemar test was compared with the α^*. For this test, the null hypothesis $H_0 : p_a = p_b$ stands that there is not significant difference between two classifiers when $p \geqslant 0.00170832$. The p-value for each pair of classifiers is shown in table 4.

Table 4. Paired tests with their p-value

Classifier 1	Classifier 2	p-value
Logit (words)	PCA Logit (words q=39)	0.00005706
Logit (words)	RB (trigrams thr≥1)	0.00000000
Logit (words)	Logit (words) + RB (trigrams thr≥1)	0.00000024
Logit (words)	PCA Logit (words q=39) + RB (trigrams thr≥1)	0.00000002
PCA Logit (words q=39)	RB (trigrams thr≥1)	0.00000036
PCA Logit (words q=39)	Logit (words) + RB (trigrams thr≥1)	0.00128306
PCA Logit (words q=39)	PCA Logit (words q=39) + RB (trigrams thr≥1)	0.00060775
RB (trigrams thr≥1)	Logit (words) + RB (trigrams thr≥1)	0.00002019
RB (trigrams thr≥1)	PCA Logit (words q=39) + RB (trigrams thr≥1)	0.00030624
Logit (words) + RB (trigrams thr≥1)	PCA Logit (words q=39) + RB (trigrams thr≥1)	0.91383314

The McNemar test shows that the difference in performance between the Logit (words) + RB (trigrams thr≥1) classifier and the PCA Logit (words) + RB (trigrams thr≥1) is not significative (p=0.914). The test also shows that there is a statistical significance between the other results of the classifiers.

4.3 Analysis

Starting from the Logit classifier, the evaluation measures showed that while being a simple classifier it achieves a good performance, classifying correctly more than half of the test data. The F1-measure corroborates that there are many examples classified either as false positive or false negative. Moreover, the AUC measure shows a better value, but considering that the AUC is a weighted average per category, it could be the case that some categories with many examples helps to produce a high value for this measure.

In the same order of the Logit classifier is the PCA Logit classifier, which produces similar results. Eventually, if a choice of the learner needs to be done, it is important to consider the processing time of both classifiers. The training phase of the PCA Logit classifier needs to execute a grid search varying the number of components extracted from PCA in order to reach ist optimal performance, resulting in a higher computational cost. On the other hand, the test phase is performed faster because of the reduced number of features used. However, the general time spent for training and testing the Logit classifier is lower than the one of the PCA Logit classifier.

Equally important is the comparison of performance between the RB classifiers and Logit classifiers. The RB trigram learner has better performance than the Logit model. The reason of this behavior is given by the distribution of the data. Most of the data has many dependent features among categories. While the RB learner tries to manage the overlapping patterns in a better way allowing that a rule can belong to two or more classes, linear classifiers cannot deal with this type of data, and even the models using kernels did not succeed neither.

Finally, analyzing the results given by the combined classifiers, it can be easily observed how the performance increases when adding the output predictions of two classifiers. This clearly shows that the combination of classifiers based on probabilities helps to improve the accuracy by acting as a sum of weights where prior knowledge for categories has been modeled. The PCA Logit classifier

combined with the RB trigram classifier achieves the best performance among all the classifiers. However, one would prefer the simplicity of creating a Logit classifier using words as features rather than searching for the best number of components in PCA. This selection could be motivated by the Occam's Razzor principle, which states that the hypothesis selected should be the simplest one.

5 Conclusions

This paper has shown that the labeling of forums using Bloom's taxonomy can be done using a Rule-Based (RB) classifier in combination with a linear classifier. The evaluation measures showed that taking a Logit and a RB classifier as basis for combining them could yield a better performance of classifiers than the individual classifiers. The combination of the models is given by summing the outputs of each classifier and taking the highest value of the resulting vector of probabilities as the predicted label.

In addition, the results of this study show that RB classifiers are better suited than models for this particular dataset. This is mainly due to the distribution of data over categories. Linear classifiers act as discriminant functions but when many data overlap, the accuracy for such overlapping categories decreases and consequently the overall accuracy decreases. In contrast, RB classifiers can discriminate a category with rules belonging to more than one category.

Although the current study is based on a small and unbalanced dataset, the findings suggest that a RB classifier is a good model solution. An idea to overcome the problem of the unbalanced data could be to fulfill a stratified sampling over the training data and have an homogeneous distribution, but certainly in real life all categories are not equally probable to appear. Therefore, it is better to treat the problem as was initially proposed without making any special treatment for this particularity on the data.

Also, it should be noted that, despite of the use of SVM as a primary classifier for TC tasks, the present study achieved a better performance with logistic regression (Logit) and Rule-Based (RB) classifiers.

Finally, several limitations need to be considered. First, the number of samples for training and testing was small and this affected directly the performance of classifiers in categories with a low distribution over the whole dataset. Another limitation is the origin of the data. The original data was collected in Spanish and translated into English by different people. At the moment of translation some misspellings or bad interpretations could have been carried out.

A further work could improve the selection of relevant features for the Rule-Based classifiers using a topic modeling and discriminate words that are off-topic, meaning that those off-topic words are relevant for the category. Another approach that could be implemented is the use of algorithms like sequential pattern mining using projection databases to find frequent sequential patterns.

Acknowledgments. We would like to thank to the Centro de Tecnologías de Información and to Professor Katherine Chiluiza G. for the data provided for this study. This research was supported partially by the KU Leuven project RADICAL (GOA 12/003).

References

1. Bloom, B.S., Engelhart, M., Furst, E.J., Hill, W.H., Krathwohl, D.R.: Taxonomy of Educational Objectives: The Classification of Educational Goals. In: Handbook I: Cognitive Domain, vol. 19, Longman, Green (1956)
2. Chang, W., Chung, M.: Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items. In: 2009 Joint Conferences on Pervasive Computing (JCPC), pp. 727–734 (2009)
3. Chiluiza, K., Echeverria, V.: Cognitive and Meta-Cognitive Skills Measurement: What about the Task in Web 2.0 Environments? In: Proceedings of Society for Information Technology & Teacher Education International Conference, pp. 1685–1690 (2012)
4. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
5. Jolliffe, I.: Principal Component Analysis. Wiley Online Library (2005)
6. Kittler, J.: Combining classifiers: A theoretical framework. Pattern Analysis and Applications 1(1), 18–27 (1998)
7. Krathwohl, D.: A revision of Blooms Taxonomy: An overview. Theory into Practice 41(4), 212–218 (2002)
8. Miyake, N.: Computer supported collaborative learning. In: The SAGE Handbook of E-Learning Research, pp. 248–267. SAGE Publications Ltd. (2007)
9. Omar, N., Haris, S.S., Hassan, R., Arshad, H., Rahmat, M., Zainal, N.F., Zulkifli, R.: Automated Analysis of Exam Questions According to Bloom's Taxonomy. Procedia - Social and Behavioral Sciences 59, 297–303 (2012)
10. Pincay, J., Ochoa, X.: Automatic Classification of Answers to Discussion Forums According to the Cognitive Domain of Blooms Taxonomy using Text Mining and a Bayesian Classifier. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 626–634 (2013)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
12. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
13. Valcke, M., Wever, B.D., Zhu, C., Deed, C.: Supporting active cognitive processing in collaborative groups: The potential of Blooms taxonomy as a labeling tool. The Internet and Higher Education 12(3), 165–172 (2009)
14. Yusof, N., Hui, C.J.: Determination of Bloom's cognitive level of question items using artificial neural network. In: 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 866–870 (2010)
15. Zurita, G., Nussbaum, M.: Computer supported collaborative learning using wirelessly interconnected handheld computers. Computers & Education 42(3), 289–314 (2004)