An Evaluation of Automatic Text Categorization in Online Discussion Analysis

Andrew Kwok-Fai Lui
School of Science and Technology
Open University of Hong Kong
andrew.lui@computer.org

Siu Cheung Li
Department of Educational Studies
Hong Kong Baptist University
sandyli@hkbu.edu.hk

Sheung On Choy
School of Science and Technology
Open University of Hong Kong
sochoy@ouhk.edu.hk

Abstract

Content analysis is often employed by teachers and research to analyse online discussion forums to serve various purposes such as assessment, evaluation, and educational research. Automating content analysis is desirable so that such analysis can be carried out efficiently on large amount of data. This paper evaluates text categorization and examines whether the attainable accuracy can satisfy the requirements of common content analysis tasks. It shows that even simple text categorization techniques can support tasks such as online learning progress monitoring. Methods of augmenting text categorization with other techniques are also discussed.

1. Introduction

Online asynchronous discussion forums are often archives of rich but unexploited information. From the perspective of content analysis, communication content may be put through an objective analysis process to infer the background and the outcome of the communication. In a teaching and learning context, the communication content affords the revelation of the learning process, the learners' characteristics, and the learning outcome. Online discussion is now regularly employed in teaching and learning, and using online discussion content as a basis for assessment, evaluation, and education research is common.

There is an increasing demand for automatic, efficient, and objective analysis of online discussion content. While the methodology of content analysis is well established, the application of this technique remains largely manual. A typical content analysis task requires teachers, educators, or researchers to assess each message, which is labour intensive. An automation of this task would enable the processing of even voluminous content in minutes rather than days, and employing content analysis in routine day-to-day teaching would become more viable.

The aim of this paper is to evaluate the application of text categorization to the analysis of online discussions in a few teaching and learning situations. Text categorization is an active research area, but its application to online discussion analysis is challenging. The content of online discussions is often incomplete, error-prone, and poorly structured. However, we suggest even simple automatic text categorization is sufficiently accurate for some selected purposes of content analysis where the accuracy requirement is not high. For example, a lower accuracy is acceptable in an evaluation of learning progress, while student assessment based on content analysis requires a high accuracy because of the student performance measurement at stake.

This paper studies the performance of automatic text categorization in three teaching and learning situations:

1. Assessment of students' online participation and contribution. Only messages containing academic relevant content contribute to the score. Irrelevant messages posted by students may be ignored in the assessment (i.e. not worth any mark).
2. Investigation of an aspect of online learning instigated by a research project. Messages are to be coded into categories according to the design of the research.
3. Evaluation and monitoring of learning progress. Messages are categorized into topics and the current theme of discussion can be deduced.

The remainder of this paper is structured as the following. The next section describes the background of online discussion content analysis and text mining. It is then followed by a description of the research

Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007)
0-7695-2916-X/07 $25.00 © 2007 IEEE

design and a presentation of the result. The paper concludes with a discussion of the findings and suggestions of future work.

2. Background

Content analysis is based on the tenet that the content of communication contains information about the background and the effect of communications. Such information can therefore be inferred objectively and consistently from communication content. Content analysis is often employed in formal social and educational research work, but it is also used extensively in various practical applications in online opinion analysis [9] and financial news analysis [4].

Studying the content of online discussion forums can serve a number of purposes. A common study purpose found in the literature is to understand various online teaching and learning issues through studying the online discussion content. A component of such study is to categorize segments of communication content according to the desired type of information [1]. For example, [2] investigated online social presence and the message categories include affective responses, interactive responses, and cohesive responses, among others; and [3] studied expertise presence and the message categories include knowledge seeking and knowledge contribution.

Another purpose of content analysis is to assist in online learning assessment where the quality of participation is considered salient [5]. Messages are categorized according to a quality measurement, which is often related to the amount of academic content contained.

Content analysis has also been applied in monitoring the learning progress of students. For example, [6] applied text categorization to identify the topic of discussion of each message, and from it inferred the learning progress of the cohort.

Manual content analysis is a labour-intensive task, but recent advances in information retrieval and text mining have demonstrated the promise of automatic content analysis. However, online discussion content presents major challenges to the existing information retrieval and text mining techniques:

- Discussion messages are usually short and the small word count reduces the possibility of the occurrences of features that distinguish a category.
- The authoring of messages is usually done in a less rigorous manner than formal articles, and the resulting spelling errors and grammar mistakes can cause problems.

Some trivial forms of automatic content analysis can be found in content and qualitative analysis software. These analysis tools are largely based on keyword matching that the categorization is based on the occurrence of specified keywords or phrases. On the other hand, current text categorization techniques can offer superior performance.

3. Research Design

In the research experiment, we will study three scenarios of content analysis, each of which is represented by a text categorization experiment. An algebraic vector-based text categorization technique will be applied to the three text categorization schemes and the performance will be studied.

3.1 Text Categorization

In a text categorization scheme, documents are to be categorized into a specific number of categories. In this research, each document represents a message in online discussion. For each category, a set of documents is specified as the representation of the category, and this set is known as the training set. The training sets of all the categories are then used to train a classifier or categorizer, which can then be used to category unseen or test documents.

There are various manners to encode a document for classifier training, and in this experiment we will use the vector-space model, coupled with latent semantic analysis (LSA) [8] or a Naïve Bayes (NB) classifier [12]. Each document is to be converted into a feature vector, and each element in the vector indicates the number of occurrence of a word or a bi-gram. During the conversion, the following techniques are applied to reduce the dimension and noise of the feature vectors: case normalization, Porter's stemming algorithm to normalize the various inflections of a word [7], and the removal of common words such as articles and personal pronouns. The feature vectors of the documents of all categories are then used to form a document-term matrix. Then the matrix is put through an entropy weighting process to amplify the information-content-wise more important terms, and then a latent semantic analysis process to reduce noise and to reveal latent relation between terms and documents. The resulting document-term matrix can then used to categorize a query document by finding the feature vector of the document/category that is most similar to the feature vector of the query document.

3.2 Experiments

Experiment 1 involves categorizing messages into two categories (code set 1):

1. Academic (A): the message containing academic content.
2. General (G): the message containing non-academic content such as that related to course administration and social conversations.

This experiment stems from the requirement of identifying messages of academic nature in the assessment of student online participation. Text categorization can automatically eliminate messages purposely posted by students to create a false impression of participation. Because assessment is involved, the requirement of text categorization accuracy is high.

Experiment 2 involves categorizing messages into two categories (code set 2):

1. Knowledge seeking (S): the message contains content that asks a question of academic nature.
2. Knowledge contributing (C): The message contains a response to a question of academic nature.

This experiment stems from the research work that measures the expertise presence [3].

Experiment 3 involves categorizing messages into five categories that represent the five topics covered in a multimedia and networking course (code set 3):

1. Networking (N).
2. Audio (A).
3. Image (I).
4. Video (V).
5. Java Programming (J).

This experiment stems from the requirement of identifying the topics of messages for the monitoring of the learning progress [6].

3.3 Procedure

An online discussion forum supporting an honour level course in advanced networking and multimedia is used in the research. The student population of 37 is well acclimatized with online discussion. A total of 322 messages have been posted.

The messages are first manually coded according to the categories employed in each of the three experiments. Only messages coded with 'A' in code set 1 are then coded with code sets 2 and 3 because only messages containing academic content are considered in experiments 2 and 3. In each experiment, a portion of messages is randomly selected as the training set, and the remaining messages are used as the test set, unless otherwise specified.

4. Results

The following table shows the number of coded messages in each category.

Table 1: Message Count of Various Categories

Code 1	*Count*
G	206
A	115

Code 2	*Count*
S	48
C	67

Code 3	*Count*
N	20
J	16
I	21
V	7
A	25

We applied both the LSA and the NB classifiers in the experiments, and their performances are similar. Only the results of the LSA experiments are shown because of the limited scope of this paper.

4.1 Experiment 1: Academic and General

The performance of the text categorizer against the percentage of messages used as the training set is shown in Figure 1. The accuracy reaches over 80% when half of the message is used as the training set. The accuracy is simply the proportion of all correctly classified messages of all types.

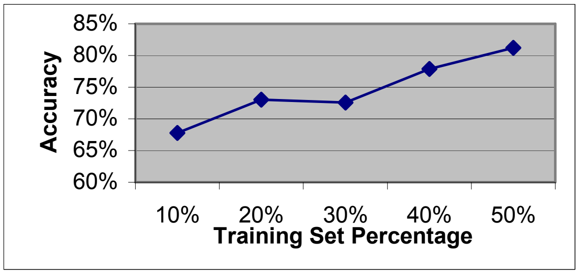

Figure 1: Performance of Categorization

Table 2 shows the precision and recall rate of querying G and A messages. Precision and recall are the standard evaluation indicators in information retrieval.

Precision is the proportion of correctly classified messages against all messages classified as that type. Recall is the proportion of correctly classified messages against all messages belonging to that type.

The performance of querying *Academic* messages is sensitive to the percentage of training set used. No such trend is observed in the *General* messages. A reason of this trend is because the Academic messages are made up of messages from 5 related but distinctive topics. When the training set gets larger, every topic is more likely to be included through the random selection process. We manually craft a 10% training set that made up of messages from all 5 topics, and the resulting precision and recall rate for Academic messages are 62% and 56% respectively.

Table 2: Performance in Precision and Recall

Training Set	*G*		*A*	
Percentage	*Precision*	*Recall*	*Precision*	*Recall*
10%	71%	85%	57%	37%
20%	77%	81%	58%	53%
30%	78%	82%	60%	54%
40%	80%	89%	72%	57%
50%	83%	90%	77%	64%

4.2 Experiment 2: Seek and Contribute

The performance of the text categorizer against the percentage of messages used as the training set is shown in Figure 2. The accuracy wanders between 60% and 80%. Compared to the categorization of *Academic* and *General* messages, this categorizer is less sensitive to the training set percentage.

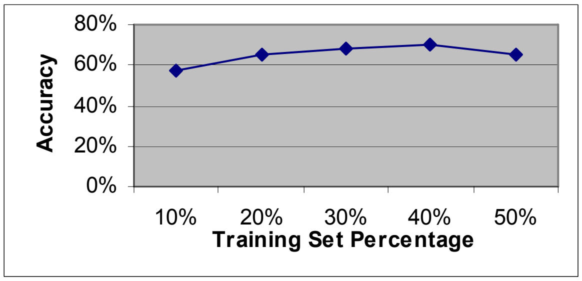

Figure 2: Performance of Categorization

Table 3 shows the precision and recall rate of querying S and C messages. The results reveal that the query of S performs very poorly, with the recall rate around 30 percent.

Table 3: Performance in Precision and Recall

Training Set	*S*		*C*	
Percentage	*Precision*	*Recall*	*Precision*	*Recall*
10%	50%	20%	59%	85%
20%	67%	32%	65%	89%
30%	85%	31%	65%	96%
40%	79%	38%	68%	93%
50%	75%	35%	62%	90%

4.3 Experiment 3: Topic Categorization

In this experiment the messages are categorized into 5 topics. The performance of the text categorizer against the percentage of messages used as the training set is shown in Table 4.

Table 4: Topic Categorization Performance

Training Percentage	*N Correct*	*J Correct*	*I Correct*	*V Correct*	*A Correct*
10%	63%	92%	47%	0%	59%
20%	65%	100%	54%	17%	85%
30%	75%	82%	46%	33%	89%
40%	73%	100%	55%	40%	92%
50%	70%	100%	63%	40%	100%

The performance of most categories is satisfactory, and the poor performance of *Video* messages is clearly hampered by the low number of samples available in the experiment. The overall recall rate at 50% training percentage is 77%.

5. Discussion and Conclusion

Table 5 summarizes the performance of text categorization in each of the three experiments.

Table 5: Summary of Performance

Experiment	*Remarks*	*Accuracy*
1	Academic (A), General (G)	65% - 85%
2	Seek (S), Contribute (S)	30% - 70%
3	Topic Categorization	70% - 100%

To understand whether the accuracy is sufficient, table 6 is developed as a reference of the desired accuracy of various content analysis purposes. Student assessment tasks general require a very high accuracy to ensure fairness and consistency. On the other hand, the information obtained from text categorization is sufficient for an indicative purpose in the monitoring of learning progress. For example text categorization can reveal that the current topics of online discussion do not match the prescribed schedule. There is little harm that teacher notified of this trend by an automatic mechanism actually found the analysis inaccurate. The figure of the desired accuracy of content analysis research is based on the commonly adopted inter-coder reliability standard.

Table 6: Desired Performance of Various Content Analysis Purposes

Experiment	*Relevant Purposes*	*Desired Accuracy*
1	Assessment	90% - 100%
	Learning Monitoring	60% - 100%
2	Content Analysis Research	70% - 80%
3	Learning Monitoring	60% - 100%

For the purpose of assessment, a near perfect accuracy would be desired but yet attainable. However, a semi-automated mechanism that includes the teacher in-a-loop looks a promising approach. In such a mechanism the teacher would review the preliminary assessment and make corrections if required. The high precision and recall rate of the *General* category could be exploited so that messages considered marginal in the categorization requires review.

For the purpose of content analysis research, text categorization again seems fallen short of the desired accuracy. In the experiment carried out, the categorization of *Seeking* and *Contributing* messages would require the building-in of sentence structural information into document representation, which is beyond the capability of word-based vector-space classification algorithms. This categorization experiment would be benefited from natural language processing approaches such as part-of-speech tagger [10], which has been employed in document genre classification with promising performance [11].

For the purpose of learning progress monitoring, text categorization is found to provide sufficient accuracy. The technology could be integrated into online discussion forum software to provide teachers with tools to analyse the learning progress of online learners. Lecture notes or textbook content could replace discussion messages as the training set and potentially improve the accuracy further because of their more comprehensive content coverage.

This paper describes a research work that evaluated the performance of automatic text categorization in several content analysis tasks. It contributed to a better understanding of the strength and limitation of text categorization. The potential and the application of text categorization in assessment, content analysis research, and monitoring of learning progress have been discussed. We envisage that this paper would provide useful information to education technology practitioners interested in applying text categorization in teaching and learning situations.

6. References

[1] Rourke, L., Anderson, T. Garrison, D.R., & Archer, W., "Methodological issues in the content analysis of computer conference transcripts", *International Journal of Artificial Intelligence in Education*, 12, 2001.

[2] Rourke, L., Anderson, T. Garrison, D.R., & Archer, W., "Assessing social presence in asynchronous, text-based computer conferencing". *Journal of Distance Education*, 14(2), 1999.

[3] A.K. Lui et al., "An Evaluation Framework of Expertise Presence in Computer Conferences", *British Journal of Educational Technology*, accepted for publication.

[4] J.E. Ingvaldsen et al., "Financial News Mining: Monitoring Continuous Streams of Text", in *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, 321-324, 2006.

[5] K. Swan, J. Shen, R. Hiltz, "Assessment and collaboration in online learning", *Journal of Asynchronous Learning Networks*, 10 (1), 45-62, 2006.

[6] A.K. Lui et al., “A Learning Thermometer: Improving Visibility of Learning Activities in Online Discussion Forums”, *Distance Education and Technology: Issues and Practice*, D. Murphy et al. eds., Open University of Hong Kong Press, 293-307, 2004.

[7] M.F. Porter, "An algorithm for suffix stripping", *Program*, 14(3), 130-137, 1980.

[8] T.K. Landauer, et al., "An Introduction to Latent Semantic Analysis", *Discourse Processes*, 25(2-3), 337-354.

[9] N. Glance et al., “Deriving Marketing Intelligence from Online Discussion”, in *Proceeding of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining KDD '05*, 419-428, 2005.

[10] M. Banko and R.C. Moore “Part of Speech Tagging in Context”. In *Proceedings of 20th International Conference on Computational Linguistics*, Geneva, 556-561, 2004.

[11] A. Finn and N. Kushmerick, “Learning to Classify Documents According to Genre”. *Journal of American Society for Information Science and Technology*, 57(11), 1506-1518, 2006.

[12] R.J., Mooney, P.N. Bennett, and L. Roy(1998). "Book recommending using text categorization with extracted information." In Papers of the AAAI-98/ICML-98 Workshop on Learning for Text Categorization and Papers of the AAAI-98 Workshop on Recommender Systems., Madison,WJ.