Automated Cognitive Presence Detection in Online Discussion Transcripts

Vitomir Kovanovic
Simon Fraser University
Vancouver, BC, Canada
vitomir_kovanovic@sfu.ca

Srecko Joksimovic
Simon Fraser University
Vancouver, BC, Canada
sjoksimo@sfu.ca

Dragan Gasevic
Athabasca University
Edmonton, AB, Canada
dgasevic@acm.org

Marek Hatala
Simon Fraser University
Vancouver, BC, Canada
mhatala@sfu.ca

ABSTRACT

In this paper we present the results of an exploratory study that examined the use of text mining and text classification for the automation of the content analysis of discussion transcripts within the context of distance education. We used Community of Inquiry (CoI) framework and focused on the content analysis of the cognitive presence construct given its central position within the CoI model. Our results demonstrate the potentials of proposed approach; The developed classifier achieved 58.4% accuracy and Cohen's Kappa of 0.41 for the 5-category classification task. In this paper we analyze different classification features and describe the main problems and lessons learned from the development of such a system. Furthermore, we analyzed the use of several novel classification features that are based on the specifics of cognitive presence construct and our results indicate that some of them significantly improve classification accuracy.

1. INTRODUCTION

One of the important aspects of modern distance education is the focus on the social construction of the knowledge by the means of asynchronous discussion groups [2]. Their increased usage in distance education has produced an abundant amount of records on the learning processes [7]. Educational researchers recognized the importance of this "gold-mine of information" [14] about the learning process, and used it mainly for research, usually long after the courses are over. Nowadays, there is a need to analyze this learners generated data in automatic and continuous fashion in order to inform instructors, and student about the current student performance and possible learning outcomes. *Learning Analytics*, an emerging research field that aims to make a sense of the large volume of educational data in order to understand and improve learning [21], is a promising area of research that could be successfully used to analyze and understand the discussion transcript logs in their full complexity. However, at the moment the majority of the approaches for analysis of discussion transcripts are not based on the established theories of educational research, and focus mostly on the quantitative aspects of the trace and log data. Given the need to assess the qualitative aspects of the learning products this is not enough. To address this issue, we base our transcript analysis approach on the well established Community of Inquiry (CoI) model of distance education [10, 11] which is used for more than a decade to answer this type of questions.

In this paper we present the results of a study that focused on the automation of the content analysis of discussion transcripts using Community of Inquiry coding technique. We developed an SVM-based classifier for automatic classification of the discussion transcripts in accordance with the CoI framework, and we discuss in detail the challenges and issues with this type of text classification, most notably the creation of the relevant classification features.

2. BACKGROUND WORK

We based our work on the theoretical foundations of the Community of Inquiry framework and previous work done in the field of text classification. In this section we will present an overview of the Community of Inquiry framework and the relevant findings in text classification field that informed our approach.

2.1 Community of Inquiry (CoI) Framework

Among the different techniques for assessment of quality of distance education environments, one of the best-researched models that comprehensively explain different dimensions of social learning is Community of Inquiry (CoI) model [10, 11]. The model consists of the three interdependent constructs that together provide comprehensive coverage of distance learning phenomena [10, 11]: i) *Social presence* describes relationships and social climate in a learning community [10], ii) *Cognitive presence* describes the different phases of students' cognitive engagement and knowledge construction [11], and iii) *Teaching presence* explains the instructor's role in the course planning and execution [10].

For our study the most important is the *Cognitive Presence* construct which is defined as *"an extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication."* [10, p .89]. The model defines four different phases of cognitive presence:

1. *Triggering event*: In this phase some issues, dilemma or problem is identified. Often, in the formal educational context, they are explicitly defined by the instructors, but also can be created by any student that participates in the discussions [11].

2. *Exploration*: In this phase students move between their private reflective world and shared world where social construction of knowledge happens [11].

3. *Integration*: This phase is characterized by the synthesis of the ideas that are generated in the exploration phase and ultimately construction of meaning.

4. *Resolution*: In this phase students analyze practical applicability of the generated knowledge, test different hypotheses, and ultimately start a new cycle of knowledge construction by generating a new triggering event.

The framework comes with its own content analysis scheme and it attracted a lot of attention in the research community resulting in the fairly large number of replication studies and empirical testing of the framework [12]. However, even though Community of Inquiry proved to be a viable model for assessing learning quality in an online educational contexts, the practical issues of applying CoI analysis and its coding scheme remain; It is still a manual, time consuming process which makes the coding of the messages very expensive. For example, for the study presented here, it took approximately one month for the two coders to manually code the 1747 discussion messages. This need for manual coding has been pointed as one of the main reasons why many transcript analysis techniques had almost no impact on educational practice and never moved out of the domain of educational research [7]. In order to support broader adoption of CoI framework there is a need for an automation of the coding process, and that is the *exact purpose of this study*. We focus on the coding of the cognitive presence, however the overall goal is to automate content analysis for all three CoI presences in order to provide a comprehensive picture of the learning process. This would allow instructors to adopt CoI framework for guiding instructional interventions, and to provide learners with feedback making them more aware of their own learning and learning of their peers.

2.2 Text Classification and Automatic Content Analysis Approaches

In order to automate the content analysis of discussion transcripts, we adopted text mining classification techniques [1]. As the cognitive presence is a latent construct and not clearly observable, we based our work on the previous work that focused also on mining latent constructs. The work done on opinion mining of online product reviews [3, 15, 23], gender style differences [13] and sentiment analysis [4] are some of the main areas of research that informed our classification approach.

The text classification tasks have been studied in the context of several different areas. In general, the majority of the studies extensively used lexical features such as N-grams, Part-of-Speech (PoS) tags and word dependency triplets, or some mixture of them as their main type of features. For example, for the problem of classifying online product reviews as either based of qualified or unqualified claims, Arora et al. [3] used the combination of N-grams, PoS bigrams and dependency triples with the approximation of syntactic scope [3]. Authors achieved Cohen's Kappa of .353 and classification accuracy of 72.6% for their binary classification task. For the similar problem, Joshi and Penstein-Rosé [15] used word dependency triplets `<Rel, HeadWord,ModifierWord>` as features where `Rel` is a grammatical relation between the words (e.g., Adjective), while `HeadWord` and `ModifierWord` are either a concrete words (e.g., Camera, Great) or PoS classes (e.g., Noun, Adverb). Their study showed that in the context of opinion mining use of the PoS class as a `HeadWord` and the concrete word as a `ModifierWord` provides a small, but statistically significant improvement over baseline unigram model [15].

ID	Phase	Messages	(%)
0	Other	140	8.01%
1	Triggering Event	308	17.63%
2	Exploration	684	39.17%
3	Integration	508	29.08%
4	Resolution	107	6.12%
	All phases	1747	100%

Table 1: Number of Messages in Different Phases of Cognitive Presence

Another type of features that are also utilized are word pattern features. For example, for sarcasm detection in online product reviews Tsur and Davidov [23] used K-nearest neighbors (KNN) classifier with the patterns of 1-6 *content words* and 2-6 *high frequency words* (i.e., words that occur frequently in many reviews) as classification features. In the context of stylistic differences among genders, the idea of pattern features was further expanded by Gianfortoni et al. [13] with more complex notion of word patterns, however, reported results showed very modest improvement in classification accuracy achieving Cohen's Kappa of only 0.18 in the best case.

Finally, there are other approaches as well, most notably the ones which are based on the use of Latent Semantic Analysis (LSA) in the context of automate assessment of student essays [8] or the use of more complex features which make a use of genetic programming [4, 18].

In terms of the classification methods used, the majority of approaches use K Nearest Neighbors (KNN) or Support Vector Machines (SVM) algorithms. SVM is particularly popular algorithm for text classification and according to Aggarwal and Zhai [1], *"text data is ideally suited for SVM classification because of the sparse high-dimensional nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories"*[pg. 195]. SVM classifiers also work well with a large number of weak predictors, which is the case of text classification where typically the majority of features are very weakly predicting class label [18].

3. METHODS

3.1 Data set

For the purpose of our study, we used the data set obtained from a graduate level course in Software Engineering from a Canadian fully distance learning university. The data set consists of 1747 messages which were coded by two human coders for the levels of cognitive presence. Coders achieved excellent interrater agreement (*percent agreement*=98.1%, *Cohen's Kappa=0.*974) indicating the quality of the content analysis scheme. The most frequent type of messages were exploration messages occurring on average 39% of the time (Table 1) while the least frequent were resolution messages occurring on average in 6% of the cases. These large differences in the category distributions are not surprising as they are shown by the previous work in CoI research field. The reason for this is that the majority of students are not progressing to the later stages of cognitive presence [11] which in turn limits the potential for development of their critical thinking skills. Thus, even though we have 5 categories, the baseline accuracy using the simplest majority vote classification is 39%.

Feature Type	Feature Names
N-grams	`unigrams`, `bigrams`, `trigrams`
Part-of-Speech N-grams	`pos-bigrams`, `pos-trigrams`
Back-Off N-grams	`bo-bigrams`, `bo-trigrams`
Dependency Triplets	`dep-triplets`
Back-Off Dependency Triplets	`h-bo-triplets`, `m-bo-triplets`, `hm-bo-triplets`
Named Entities	`entity-count`
Thread Position Features	`is-first`, `is-reply-first`

Table 2: Extracted Features

Feature Set	Additional Features	Classification Accuracy	Cohen's Kappa	P-val
majority vote baseline	0	0.392	0.000	
`unigrams` baseline	2241	0.547	0.364	
+ `bigrams`	3155	0.556	0.376	0.427
+ `trigrams`	911	0.554	0.374	0.571
+ `pos-bigrams`	737	0.561	0.385	0.249
+ `pos-trigrams`	2810	0.560	0.382	0.304
+ `bo-bigrams`	6953	0.560	0.381	0.323
+ **`bo-trigrams`**	17986	**0.584**	**0.410**	0.006
+ **`dep-triplets`**	1435	**0.564**	**0.386**	0.062
+ **`h-bo-triplets`**	1931	**0.571**	**0.396**	0.031
+ **`m-bo-triplets`**	2771	**0.579**	**0.406**	0.003
+ `hm-bo-triplets`	1375	0.558	0.379	0.359
+ **`entity-count`**	1	**0.559**	**0.381**	0.030
+ **`is-first`**	1	**0.555**	**0.375**	0.037
+ `is-reply-first`	1	0.550	0.367	0.665

Table 3: Classification Results. Bold typeface indicates statistically significant features

3.2 Feature Extraction

Based on our literature review described in Section 2, we extracted a wide variety of features that were frequently used in the similar studies (Table 2). We extracted the commonly used N-gram features (i.e., unigrams, bigrams and trigrams) and Part-of-Speech (PoS) bigrams and trigrams. In addition, similarly to the works of Joshi and Penstein-Rosé [15] we extracted: i) back-off versions of bigrams and trigrams by replacing one or more words in a N-gram by the corresponding PoS tag, and ii) word dependency triplets and their back-off versions. Finally, in addition to the features found in the research literature, we extracted an additional set of features which we thought might be useful given the specifics of the cognitive presence construct.

Given the difference among the phases of Cognitive Presence, we extracted the `entity-count` feature, which shows the number of named entities that were mentioned in the message using DBPedia Spotlight [19] web service. The rationale behind this feature is that different phases of cognitive presence could be potentially characterized by the different number of constructs that were discussed in the message. For example, it might be the case that exploration messages contain on average a larger number of concepts, as one of the key characteristics of exploration is brainstorming of different problem solutions and ideas [11].

Another important aspect of cognitive presence is that it develops over time through the communication with other students [11]. In practice this means that triggering and exploration messages are more likely to be observed in the early stages of discussions, while integration and resolution messages are more likely in the later stages of the discussions. To test this hypothesis, as the first step we extracted two simple features: i) `is-first`, which indicates whether a message is the first in the discussion topic, and ii) `is-reply-first` which indicates whether a message is the reply to the original discussion opening message.

3.3 Classifier Implementation

For the purpose of this study we decided to use SVM classification as it is a well known and popular technique especially well suited for the purpose of text classification as we described in Section 2. In order to maximize classification quality and assess the usefulness of different types of features, we experimented with the several different sets of features and evaluated them using 10-fold cross validation, which is considered a good compromise between sizes of training and test data [20]. We used only features that had support threshold of 10 or more (i.e., occurring 10 or more times in the data) in order to keep the number of features reasonable and to protect from overfitting the classifier to the noise in the data which is captured by the low supported features. We used linear kernel and default values of parameters ($C = 1, gamma = 1/k$). In order to compare different set of features we used McNemar's test [9] as it is shown to have low Type I error rate [6].

To implement the classifier and feature extraction we used several popular open source tools and libraries. In the feature extraction step we used Stanford CoreNLP suite[1] of tools for tokenization, Part-of-Speech tagging [22] and dependency parsing [17]. We used the popular Weka [24] data mining toolkit and LibSVM library [5] for developing the classifier, and to implement the McNemar's test we used Java Statistical Classes (JSC) library[2].

4. RESULTS

Table 3 shows the results of our classification experiment. The baseline unigram model achieved 54.72% accuracy which is slightly less than in the case of the more complex models with larger number of features. The biggest improvement was observed by adding the back-off version of trigrams which improved classification accuracy to 58.38% and Cohen's Kappa to 0.41 which is accompanied with the largest increase in the feature space.

Our results are similar to the ones of Arora et al. [3] and Joshi and Penstein-Rosé [15] with our classifier having somewhat lower absolute levels of accuracy and a slightly bigger values of Cohen's Kappa metric. Our results also show that adding both head-backoff and modifier-backoff versions of dependency triplets improves the classification accuracy, as well as the ordinary dependency triplets. With respect to the three features that we proposed, the use of the indicators for the number of named entities (i..e, `entity-count`) and discussion starters (i.e., `is-first`) also showed statistically significant improvement over the baseline unigram model. In addition, the use of those features has an almost nonexistent impact on the classifier feature space making the building of classification model faster and more interpretable.

5. CONCLUSIONS AND FUTURE WORK

As our results show, the proposed approach for automating content analysis seems promising. The current level of Cohen's Kappa

[1] http://nlp.stanford.edu/software/corenlp.shtml
[2] http://www.jsc.nildram.co.uk/index.htm

is at the lower end of 0.4-0.7 range which is considered to be a fair to good agreement [16]. However, in order to replace manual message coding, the Cohen's Kappa should be above 0.7 level which is still out of reach.

One important aspect of coding discussion transcripts that we observed and which does not affect the work that we reviewed is *message quoting*. We observed many instances in which student puts direct quotation of others' message into his own which makes a problem for classification based on lexical features such as N-grams, PoS tags or Dependency triplets. In our future works we will look for a ways to address this issue and to estimate the impact of quoting on classification accuracy.

We also showed the potential of novel features which are based on the deeper theoretical understanding of the latent construct under interest and its coding instrument. They could provide a significant improvement of the classification accuracy without a big impact on the feature space complexity.

References

[1] C. C. Aggarwal and C. Zhai. *Mining Text Data*. Springer, Feb. 2012.

[2] T. Anderson and J. Dron. Three generations of distance education pedagogy. *The International Review of Research in Open and Distance Learning*, 12(3):80–97, Nov. 2010.

[3] S. Arora, M. Joshi, and C. P. Rosé. Identifying types of claims in online customer reviews. In *Proceedings of the HLT-NAACL 2009*, page 37–40, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[4] S. Arora, E. Mayfield, C. Penstein-Rosé, and E. Nyberg. Sentiment classification using automatically extracted subgraph features. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, page 131–139, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[6] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, Oct. 1998.

[7] R. Donnelly and J. Gardner. Content analysis of computer conferencing transcripts. *Interactive Learning Environments*, 19(4):303–315, 2011.

[8] R. M. Duwairi. A framework for the computerized assessment of university student essays. *Computers in Human Behavior*, 22(3):381–388, May 2006.

[9] B. Everitt. *The analysis of contingency tables*. Chapman and Hall, 1977.

[10] D. R. Garrison, T. Anderson, and W. Archer. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education*, 2 (2–3):87–105, 1999.

[11] D. R. Garrison, T. Anderson, and W. Archer. Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, 15(1):7–23, 2001.

[12] D. R. Garrison, T. Anderson, and W. Archer. The first decade of the community of inquiry framework: A retrospective. *The Internet and Higher Education*, 13(1–2):5–9, Jan. 2010.

[13] P. Gianfortoni, D. Adamson, and C. P. Rosé. Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, page 49–59, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] F. Henri. Computer conferencing and content analysis. In A. R. Kaye, editor, *Collaborative Learning Through Computer Conferencing*, pages 117–136. Springer Berlin Heidelberg, Jan. 1992.

[15] M. Joshi and C. Penstein-Rosé. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference*, page 313–316, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[16] K. H. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 0 edition, Dec. 2003.

[17] M.-C. d. Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, page 449–454, 2006.

[18] E. Mayfield and C. Penstein-Rosé. Using feature construction to avoid large feature spaces in text classification. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, page 1299–1306, New York, NY, USA, 2010. ACM.

[19] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, page 1–8, New York, NY, USA, 2011. ACM.

[20] P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. In L. LIU and M. T. ÖZSU, editors, *Encyclopedia of Database Systems*, pages 532–538. Springer US, Jan. 2009.

[21] G. Siemens, D. Gasevic, C. Haythornthwaite, S. Dawson, S. B. Shum, R. Ferguson, E. Duval, K. Verbert, and R. S. d Baker. Open learning analytics: an integrated & modularized platform. *Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques*, 2011.

[22] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the HLT-NAACL 2003*, page 252–259, 2003.

[23] O. Tsur and D. Davidov. Icwsm – a great catchy name: Semi-supervised recognition of sarcastic sentences in product reviews. In *Proceeding of the International AAAI Conference on Weblogs and Social Media*, 2010.

[24] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. Morgan Kaufmann, 3 edition, Jan. 2011.