Sentiment Analysis of a Student Q&A Board for Computer Science

Saul Wyner, Erin Shaw, Taehwan Kim, Jia Li, and Jihie Kim
Information Sciences Institute
University of Southern California
swyner@ucla.edu, shaw@isi.edu, taehwan@isi.edu, jiali@isi.edu, jihie@isi.edu

Abstract

Students' emotions and attitudes are discernible in messages posted to online question and answer boards, and are vital aspects of natural argumentation. Understanding student sentiment may help analyze student and instructor argumentation, with such potential benefits as helping instructors identify students with potential course issues, optimize help-seeking, and potentially improve performance, as well as identify both positive and negative actions by instructors and provide them with valuable feedback, within the framework of a computer supported pedagogical collaborative system. We present a set of context-independent emotion speech acts used by students in a university-level computer science course to express certainty, frustration, tension, and politeness. We develop viable, functional, automatic classification approaches to emotion acts. To explore potential in profiling by emotion act in argumentation analysis, we present a heuristic-driven analysis of thread success and emotional profiles, and detail future avenues of research.

1 Introduction

Online discussion boards are widely used in higher education, extending the availability of instructors, assistants, and materials to students beyond the traditional classroom. Students use discussion forums to collaborate, exchange information, and seek answers to problems from their instructors and classmates- which are all activities that are open for argumentation analysis. Discussion board use is associated with improved academic performance and greater student satisfaction [Kumrow, 2005; Newman and Schwager, 1995].

Our previous work on analyzing student discussions and argumentation has been based on rhetorical speech acts, course topics, and problem tasks [Kim *et al.*, 2007; The PedDiscourse Project, 2009], and classification systems for these features enable us to automatically identify student problems. Similarly, understanding student sentiment may help instructors identify students with potential course issues, optimize help-seeking, and potentially improve performance. In the same vein, by examining different instructor interactions through student sentiment, instructors could receive valuable feedback on their actions and impact.

In this paper we present a set of dialogue features, or emotion acts, that characterize student sentiment with respect to 1) tension and frustration, 2) certainty/confidence and 3) politeness. These were exhibited by students in a Q&A board in an undergraduate computer science course. A discussion corpus consisting of almost 1,030 student posts was manually labeled with the emotion acts using XML tags. We then show the first stages of the development of automatic emotion act classification.

In subsequent analysis of the corpus, to explore the potential and validity of emotion act analysis, we analyze and explore profiles of emotion acts generated by an automated heuristic-driven analysis of thread success and examine the implications of the results. We then detail potential successive methods of both student and discussion profiling for fully-automated, direct functional course instructor benefit.

2 Emotion Annotation

It is extremely difficult to devise a category of emotion labels given the gradations and subtlety of the way emotion is expressed in language. It is not surprising then that there is no general agreement on how to label emotional content and that instead there exist a number of different labeling schemes for different domains [Ordelman and Heylen, 2005]. However, preliminary work does suggest that at least some emotional content in English, such as certainty, can be identified and selected for, independent of context [Rubin *et al.*, 2006]. As such, it stands to reason that exploring context-independent emotion annotation has great potential.

Emotion coding is an iterative process, beginning with the desire to identify students' self-efficacy and attitudes. Confidence, interest and mastery begot urgency, understanding and technicality as an approach to measuring academically pertinent characteristics of messages. We were also interested in the orthogonal dimensions of humor and politeness. Our current work focuses on poster certainty/confidence, frustration, and interpersonal tension.

The methodology of emotion act annotation involved identifying repeatable speech fragments that indicate identified emotion acts throughout the corpus of student posts.

Tension	**Examples**
Instructor Judgments: Possible student issues with class attendance, judgment or choices	If you really want to do this; I stated in class on at least 2 occasions
Student Judgments: Possible student issues with questioner or target	Result of this sucks; Wow… That was..
Frustration	**Examples**
Repetitious Actions, Continual Actions: Descriptions of continuous actions without real progress	A lot (15+ times); Never seems to end; High rate of redundancy; Another can of worms
Large Quantities: Descriptions of overwhelming amounts of work and other material	Zillions of references; Super-huge; Simply gargantuan; Monstrous, super-verbose
Difficulty/Impassability, Material Denigration: Statements of explicit difficulty in either solution or understanding of issues, as well as frustration about the material itself	Serious disk quota problems; Severe annoyances; A pain to fix; Makes it really hard
Self-Denigration/Lack of Confidence: Declarations of a personal belief in a lack of ability on the part of the poster	I have spent FAR too long; …I'm stumped; Longer than they should have
High Certainty	**Examples**
Specificity of Question/Answer: Specific phrasing that concisely explains through examples and pre-conditions	The only way; I found the answer; It only appears
Ease of Understanding/Completeness: Emphasis of the simplicity or completeness of a solution or question	The trick is; Just wait till; Will be simple; All you need to do is
Necessity: Specifically stating that the presented solution is required, or in the case of a question, its importance	Must be able to; Vitally important task; Must have something; You will
Logical Presentation: A method of presenting a proposition, solution, or question that makes it a logical proposal	I assume that; Granted,; Likewise,; On the other hand,
Low Certainty	**Examples**
Vagueness in Question/Answer: Statements that imply only general or surface understanding of the material at hand by stating personal understandings over factual presentation	What is wrong?; If I understand; Seems to me; Read it somewhere
Lack of Understanding: Statements that clearly state a lack of understanding; differs from other Speech Acts as it implies a continuing lack, rather than an individual issue	I am still confused; Not sure if I understand; I follow most; I'm not sure
Optional Nature: Statements indicating a not strongly recommended or vital issue, solution, or question	Should be compiled from the network directory; You might try; …maybe I'll try making; What is wrong?
Weakened Presentation: Phrases that weaken or justify logical proposal statements	Correct me if I am wrong; Apparently; I am guessing that is the way; As far I know
Politeness	**Examples**
Positive: Language strategies used according to formal cultural rules to avoid losing face. Commonly identified as typical polite speech	Thanks; Okay thanks; Good luck with your project
Negative: Dealing with a face-threatening act, by lightening the request or response into a less pressing, informal status.	I was wondering if; Thought I'd throw this out there; Get this cleared up early; Just a head's up,
Bald on record: Dealing with a face-threatening situation by ignoring or emphasizing the consequences of the threat	I question the; don't bzero anything; Change it to this:; Do we?
Off record: Attempting to change the request or response into a non-face-threatening statement, i.e., by generalizing a query to a rather than asking for direct help	Has anyone else had this problem; What would do; Asking for answers directly is way easier

Table 1. Types and examples of emotion acts (EAs).

This was complicated by the greatly irregular nature of the post content, in terms of frequent misspellings and grammar/syntactical errors, stemming from common parlance, simple carelessness, and Computer Science student subculture language use. This necessitated a high level of selectivity and repeatability in all annotations, as well as reliance on specific patterns of distinct phrases and grammar from within the corpus rather than whole statements. In addition, all annotations were classified by a select number of sub-characteristics of each emotion act, detailed in Table 1.

2.1 Emotion Acts (EAs)

Two annotators worked together with four other project members for over two months to define the final emotion acts, and to label a dataset of 1,030 messages in 210 threads. Inter-annotator agreement was unacceptably low, the determined cause being difficulty identifying textual emotion content by foreign-born annotators, an interesting subject currently being researched. Because of this, all annotations were reviewed and corrected when necessary by the lead annotator, preserving consistency and validity.

Labeling consisted of a set of XML tags, containing the emotion act classification as well as contextual information on the post content, such as the type of response (question/query or answer/statement), the type of poster (student, instructor, or TA), and the type of poster to which the post was in response. Table 1 describes the set of emotion labels for coding student messages and their specified sub-characteristics, with examples.

2.2 Automatic Classification Approach of EAs

For developing automatic classification of emotion acts, we used a similar approach that was previously applied to identify speech acts [Kim *et al.*, 2009]. We classify each post into its constituent emotion acts. Each post is considered as one data point in this approach and hence need to be converted to some representation which could then be easily treated. Then, we classify each as positive or negative for the individual emotion acts.

We chose to focus on certainty and frustration for our initial classification work, as they are the most plentiful emotion acts and the most directly relevant, as they pertain directly to student performance. First, we use data pre-processing for discussion threads with human annotation. It replaces some typical words and contents with fixed keywords- for instance, programming code fragments are replaced with a "code" keyword, and URL links are also converted to a reserved keyword. Also, contractions such as "I'm" and "You're", are replaced to "I am" and "You are".

After necessary pre-processing, we need to specify feature space to use machine learning algorithms. We consider the same feature space as suggested in [Kim *et al.*, 2009], which is detailed as follows.

F1: Cue phrase and their position in the post
All possible forms of cue phrase as an n-gram, and the position in the post as in the first part, last part or elsewhere.

F2: Message position in the thread.
Indicates if the post is the first post, last post or one of the other posts.

F3: The emotion acts of the previous message.

F4: Poster class.
Identifies the poster as a Student or Instructor.

F5: Poster change.
Checks if the current poster is the same as the previous poster.

F6: Post length
Categorizes the post as Short (1-5 words), Medium (6-30 words), or Long (>30 words).

First, we generate all possible features as candidate features by taking in all cases in the training data. For instance, for the F1 feature, we consider all combinations of cues, such as unigram, bigram, unigram and multi-unigram from the annotated cue phrases. From F2 to F6, the features, as categorical features, already contain all possible values.

After generating all candidate features, we specify which features to use by considering their classification powers. Here, we use information gain [Yang and Pedersen, 1997]-we apply each feature to training data and record information gain. Then we sort potential features by their information gain and pick the top two hundred features.

Using this fixed feature space, we convert all messages in training data to vectors by taking into account each feature. After transformation, we apply Support Vector Machine [Chang and Lin, 2001] because of its effective power to broad application areas, such as natural language processing. To use these algorithms, we divide the whole annotated data into training data and test data. We then evaluate the classification result by considering its F-score.

We divide annotated discussion threads to training data of 159 threads and test data of 52 threads. Then we train the model by using training data. After training, we apply the model to test data. The result is shown in Table 2.

	Test Data Results		
Emotion Act	Precision	Recall	F-Score
High Cert.	0.712	0.755	0.733
Low Cert.	0.627	0.595	0.61
Frustration	0.6	0.323	0.42

Table 2. Automatic annotation test results for certainty and frustration.

As one can see, these initial results show that our approach might be able to be feasible for the automatic classification of emotion acts, both in this specific case and in further argumentation work. Due to the relatively small set size of available manually-annotated training data, espe-

		Emotion Act Percentage From Each Group									
Subset for Analysis	Total # of Posts	High Cert. Level	Low Cert. Level	Med. Cert. Level	N/A Cert. Level	Frust.	Tension	Bald-On-Record Polite.	Positive Polite.	Negative Polite.	Off-Record Polite.
All Posts	1030	49.03%	13.50%	3.79%	33.69%	24.27%	2.72%	19.71%	17.38%	24.66%	6.89%
All Success-ful	916	50.11%	12.99%	3.17%	33.73%	23.47%	2.84%	19.43%	17.36%	23.91%	6.33%
All Unsuc-cessful	114	40.35%	17.54%	8.77%	33.33%	30.70%	1.75%	21.93%	17.54%	30.70%	11.40%
Successful Answers	645	56.12%	10.23%	2.95%	30.70%	18.60%	3.72%	22.79%	11.16%	20.31%	2.02%
Unsuccessful Answers	79	43.04%	10.13%	8.86%	37.97%	25.32%	1.27%	29.11%	13.92%	32.91%	6.33%
Successful Questions	271	35.79%	19.56%	3.69%	40.96%	35.06%	0.74%	11.44%	32.10%	32.47%	16.61%
Unsuccessful Questions	35	34.29%	34.29%	8.57%	22.86%	42.86%	2.86%	5.71%	25.71%	25.71%	22.86%
Successful Instructor Answers	233	62.66%	2.58%	0.43%	34.33%	5.58%	8.58%	35.19%	3.00%	7.30%	0.43%
Unsuccessful Instructor Answers	25	48.00%	0.00%	8.00%	44.00%	0.00%	4.00%	44.00%	8.00%	16.00%	0.00%
Successful Original Posts	180	33.89%	17.78%	2.22%	46.11%	33.33%	0.56%	13.89%	37.22%	32.78%	20.00%
Unsuccessful Original Posts	30	43.33%	16.67%	3.33%	36.67%	50.00%	0.00%	23.33%	36.67%	40.00%	23.33%
Successful Final Posts	180	45.56%	16.11%	4.44%	33.89%	17.22%	2.22%	17.22%	26.11%	24.44%	5.00%
Unsuccessful Final Posts	30	26.67%	13.33%	3.33%	56.67%	20.00%	0.00%	30.00%	6.67%	13.33%	6.67%

Table 3. The distribution of relevant emotion acts in successful vs. non-successful threads.

cially for frustration, the result is not yet at a level where it can be applied in a functional setting. However, we strongly expect these results to greatly improve as more training data become available.

3 Successfulness Analysis with EAs

While the implementation of emotion acts is a significant development, emotion acts only represent the lowest level of potential emotional analysis of student argumentation. With consistent and functional emotion acts, posters, posts, and entire threads can be analyzed in terms of repeatable emotion act profiles, revealing important data on argument content. As a proof of concept, we wished to develop an independent heuristic to classify threads with a hypothetically robust emotional distinction, and examine the resulting emotion act profile for such a distinction.

We chose the concept of successful vs. unsuccessful discussion threads, defined by containing a provided final solution or ratification of issues, and beneficial discussion and instruction on the pathway to that end. As such, we experimented with several classification measures of successful threads, based upon observed trends in annotated threads. To fulfill the need for a conclusion, we focused on threads that concluded with an answer, or an acknowledgment of thanks for a provided solution. For the need for discussion, we only included the subset of threads that also contained equal numbers of or more answer/statement posts, to ensure a basic level of back-and-forth pedagogic discourse. The generated results by these criteria were examined by the annotators and found to closely conform to their intrinsic impressions of the success of the individual threads.

Those threads that did not fit the qualifications of a successful thread were classified as an unsuccessful thread. This categorization revealed 180 successful threads, and 30 unsuccessful threads, which was deemed to be a reasonable expectation for a high-level Q&A board.

After this classification, both successful and unsuccessful threads were treaded to a further break-down into relevant subsets for emotion act analysis. The analysis was based upon a simple presence test for specific emotion acts, and the percentage of posts within the subset that contained that emotion act. Certainty, however, as the most common emotion act, was instead calculated as a level, defined by containing over 75% of a specific type of either high or low certainty emotion acts. If the ratio was less than 75%, it was designated as medium certainty, and not applicable, or N/A, if certainty emotion acts were not present. While rudimentary, this examines the potential for more rigorous profiling, by revealing any obvious difference among successful and unsuccessful threads.

3.1 Results of Successfulness Analysis

The results show a clear distinction between successful and unsuccessful threads. Distinctions were noted when there existed at 10% or above difference from successful vs. unsuccessful versions of the chosen subset.

Within the certainty measures, high certainty is shown to strongly influence the successfulness of a thread in answers, while having little effect in questions. However, in the initial posts, high certainty seems to counter-indicate success. In contrast, low certainty seems to have minimal effect, except in the case of questions, in which it is strongly represented in unsuccessful questions. A lack of certainty emotion acts also strongly indicates successful vs. unsuccessful questions and original posts, while it shows the inverse in final posts.

In terms of frustration and politeness, frustration is unsurprisingly well-represented in unsuccessful posts, though most notably in original posts. Bald-on-record politeness shows strongly in unsuccessful instructor answers, original posts, and final posts. Positive politeness is seen greatly in successful questions and final posts, while negative politeness is greater in successful final posts. Off-record politeness shows little effect overall.

While these results show many interesting possible relationships between expressed emotion acts and topic success, the clear and immediate indication shows that emotion acts can show distinctions between different types of posts and threads, which prove their potential use as a profiling mechanism for argumentation.

4 Conclusion

As the distinctions between successful and unsuccessful threads show that profiling and automatic identification by emotion act is fully possible in this specific type of argumentation analysis, it is important to look forward toward methods and directions of higher-level interpretation, in both pedagogical discourse and other types of argumentation. The procedure used in investigating successfulness is only for broad proof-of-concept, rather than developing specific profile criteria for automatic categorization. As such, future development in profiling arguments will require specific categories, defined by interactions within posts between different emotion acts in a repeatable manner. This will be able to operate as a potentially automatic process that can present information regarding important qualities of posts, threads, and students.

We have described an important first step towards the semi-automatic identification of emotion speech acts: We have identified common emotion acts used by students in a computer course who interact within a question and answer board, developed the first stages of a promising automatic classification approach, and have shown that these acts are significant within the corpus through an investigation of successfulness of threads. Even with a dataset of only 1,030 labeled posts, there are many research avenues to explore.

In combination with existing metrics based on rhetorical speech acts, contribution quantity and technical depth, the new measures will assist instructors and researchers in understanding how students learn and how to predict their course performance, creating a functional computer-supported collaborative argumentation environment for pedagogy. This study complements previous work on analyzing student discussions using rhetorical speech acts, course topics, and problem tasks.

Acknowledgments

This work was supported by a National Science Foundation, CCLI-Phase 2 (Expansion) grant (Award No. 0618859).

References

[Kim *et al.*, 2009] Kim, J., Kim, T. and Li, J. (2009) Identifying Unresolved Issues in Online Student Discussions: A Multi-Phase Dialogue Classification Approach, Proc. of AIED 2009.

[Kim *et al.*, 2007] Kim, J., Shaw, E. Chern, G. and Herbert, R. (2007) Novel tools for assessing student discussions: Modeling threads and participant roles using speech act and course topic analysis, Proc. of AIED 2007.

[Kumrow, 2005] Kumrow, D. (2005). Student Self-Regulatory Resource Management Strategies and Academic Achievement in a Web-based Hybrid Graduate Nursing Course. In Education and Technology, Editor(s): T.C. Montgomerie, J.R. Parker, ICET 2005.

[Newman and Schwager, 1995] Newman, R. and Schwager, M. (1995). Students' Help Seeking During Problem Solving: Effects of Grade, Goal, and Prior Achievement. American Educational Research Journal, v32 n2 p352-76.

[Ordelman and Heylen, 2005] Ordelman, R. and Heylen, D. (2005). Annotation of Emotions in meetings in the AMI project.

[The PedDiscourse Project, 2009] The PedDiscourse Project 2009, http://ai.isi.edu/discourse

[Rubin *et al.*, 2006] Rubin, V., Liddy E., and Kando, N. (2006). Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series), Editor(s): J. Shanahan, Y. Qu, J. Wiebe, Springer-Verlag New York, Inc.

[Yang and Pedersen, 1997] Yang, Y. and Pedersen, J., (1997). A Comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning, pp.412-420.

[Chang and Lin] Chang, C. and Lin, C. (2001). LIBSVM: a library for support vector machines.