Learner Affect Through the Looking Glass: Characterization and Detection of Confusion in Online Courses

Ziheng Zeng, Snigdha Chaturvedi, Suma Bhat
University of Illinois
Urbana-Champaign, USA
{zzeng13, snigdha, spbhat2}@illinois.edu

ABSTRACT

Characterizing the nature of students' affective and emotional states and detecting them is of fundamental importance in online course platforms. In this paper, we study this problem by using discussion forum posts derived from large open online courses. We find that posts identified as encoding confusion are actually manifestations of different learner affects pertaining to their informational needs–primarily seeking factual answers. We quantitatively demonstrate that the use of content-related linguistic features and community-related features derived from a post serve as reliable detectors of confusion while widely *outperforming* currently available algorithms of confusion detection. We also point out that several prediction tasks in this domain (e.g., confusion and urgency detection) can be correlated, and that a model trained for one task can effectively be used for making predictions on the other task without requiring labeled examples. Finally, we highlight a very significant problem of adapting the classifier to unseen courses.

Keywords

Confusion characterization, discussion forum analysis

1. INTRODUCTION

Discussion fora constitute a central feature of learner interaction in online course platforms, where learners post questions, opinions, and concerns, which are viewed, rated and answered by fellow-learners and/or teaching staff. In the particular instance of courses affording only virtual interactions, such as at-scale learning environments, forum posts constitute rich repositories of students' affective and emotional states captured in real time. The focus of this study is on *characterizing* the nature of students' affective and emotional states, manually identified as confusion in forum posts and developing automatic methods to *detect* them. Here, as in [25] and [2], we operationalize the definition of confusion as a state in which a student hits an impasse and is uncertain of how to move forward. As such, the reasons for confusion could be attributed to lack of clarity on the topic discussed or technical shortcomings of the learning interface, among others. Examples of such posts are shown in Table 1.

Table 1: Posts representing confusion and its absence.

I have also problems with the section "Pre course Survey" I have completed this section several times about 10, I have the final message "Thanks" but at each new connection appears in my courseware "pre course Survey (please complete)" Please help me, what I have to do ? (**Confusion**)
Interesting! How often we say those things to others without really understanding what we are saying. That must have been a powerful experience! Excellent! (**No confusion**)

The strong connection between learner affect, engagement, and learning outcomes has long been understood but studies on their effect on continued participation in internet-based learning environments such as MOOCs is only emerging (e.g.,[25, 2]). In addition to constituting supporting evidence to understand this association, mechanisms to automatically detect learner affect encoded via confusion in discussion fora serve the following ends. Firstly, they inform us about the aspects of a course that are frustrating for learners and hence need improvement [24, 21, 11]. Second, they can aid a timely and accurate intervention to struggling learners by providing critical insights into their emotional states[25], eventually leading to success of this critical course component.

For instance, when a student expresses confusion or misunderstanding about a concept, the immediacy with which the confusion is addressed impacts student satisfaction and course progress. Because of this, and the demands of an at-scale learning environment, efficient and automatic detection of confusion has become more important than ever before. With a steady increase in the number of courses on online course catalogs, and with limited means to control the instructor-to-student ratio in online platforms, the problem of detecting confusion as expressed in online fora is timely. Despite the critical need, relatively few studies analyze confusion in course discussion forum posts [25, 2].

While the explicit purpose of discussion fora is to engage the users in a way that develops a sense of community and communication within large-scale online courses, the posts themselves serve as proxy for learner affect and emotions expressed in various forms. Detecting this encoded affect from posts is an important challenge for natural language processing algorithms. This is because, at the outset, a post indicating confusion could be construed to be a question. Since question posts and confusion posts–forms of information seeking behavior–are remarkably similar, one would expect that approaches to detect questions (e.g., [7]) ought to be directly applicable. However, this is not always the case. Many times confusion posts do not have an explicit question making the two prob-

lems of question detection and confusion detection closely related but not the same. This makes the detection of confusion in a post a non-trivial problem partly because, for posts containing a question, the questions tend to occur with other declarative sentences. A second difficulty is the use of different question styles (informal, where standard features such as the question mark are likely to be absent or where the question is worded without a question mark). Hence, simple heuristics of using question mark or 5W1H words (who, what, which, where, why, how) are rendered inadequate.

Additionally, as observed in [18], finding patterns to identify non-questions is more challenging than finding patterns in questions (since they usually do not share common lexical and/or syntactic patterns). This is directly applicable to confusion posts where posts not indicative of confusion have diverse intent.

Prior studies in this direction (e.g., [6, 2, 25]) have led to the use of linguistic and structural features available from the discussion forum. While similar in spirit to these prior studies, this study sets itself apart from them in many ways. Firstly, we identify that confusion detection is different from simple/complex question detection. In order to solve this problem more effectively, we point out that the community needs a characterization of confusion instead of treating it as yet-another text-classification task. We present an in-depth analysis of types of 'confused posts' using high-quality and reliable manual annotations (Section 4). Motivated by this analysis, we then design features to detect confusion automatically in a supervised framework. We also point out that several prediction tasks in this domain (such as confusion and urgency detection) are correlated, and demonstrate that a model trained for one task can effectively be utilized for making predictions on the other task without requiring labeled examples. Finally, we highlight a very significant problem concerning the applicability of such classifiers to unseen courses. We summarize our contributions below:

Characterizing affective states and informational needs: We observe that nearly half of the posts encoding confusion and considered urgent pertain to users seeking answers to factual questions. Aside from indicating an information need, these posts are also used to report course-specific issues such as concerns with assignments or quizzes as well as to report course-related technical issues (e.g., unavailability of a lecture video or a peer-assessment grade).

Efficient confusion detection: We quantitatively demonstrate that our use of content-related linguistic features of a post and a set of community-related features associated with it serve as reliable detectors of confusion while widely *outperforming* currently available algorithms of confusion detection.

Combined confusion and urgency detection: We show that the trained confusion classifier also functions as an efficient urgency detector when tested on confusion posts also labeled as 'Urgent'.

Scaling the effort to other courses and domains: Based on the dataset, we make concrete suggestions to explore domain adaptation towards building course-generic classifiers. Rather than aiming for course-independent classifiers, our proposal is to harness the utility of available course-specific classifiers for an unseen course, based on suitably defined cross-domain similarities.

By means of a thorough quantitative evaluation of our proposed features in a supervised machine learning model, we demonstrate its effectiveness as a *scalable* and *efficient* model for automatic detection of confusion that generalizes well to unseen courses.

2. RELATED PRIOR WORK

Confusion and its impact on learners: Studies modeling confusion and exploring its relation to learner affect have found that even though students seem to struggle when confused, the situation leads them to attempt to resolve barriers to their understanding of complex concepts [16, 10, 8]. However, it has also been pointed out that remaining confused has a negative effect that leads to student disengagement and eventual dropout, thus making it imperative that confusion be resolved immediately [15, 25]. This necessity is more immediate in the context of learning at scale given the impersonal and the distant nature of the learning process[14, 19]. Thus, detecting learner affect, particularly with respect to understanding the material has the potential to contribute to the design of interventions as shown in prior studies (e.g.,[9, 22]) can lead to increased learning effectiveness in computer-based learning environments such as online courses.

Detecting confusion: Focusing on MOOCs, where the only venue for learner-instructor interaction is the discussion forum, studies are now beginning to explore automated mechanisms to provide timely learner support by analyzing forum content. These include, predicting when instructor intervention is needed [5, 6], monitoring student's opinion towards the course [20], recommending questions to users for assisting students seeking answers [23], identifying acceptable answers [13], organizing the forum content into aspects or topics along with their sentiments to help instructors in promptly addressing common issues [17], identifying posts that express confusion to predict points of eventual student dropout [25], and detecting posts that express confusion to then map confused posts to course video clips as a way to automate interventions [2]. A common feature of these approaches to detect confusion is their reliance on textual and structural features of the discussion forums to design effective algorithms.

While [25] uses a set of linguistic features to detect confusion, it disregards the structural features (e.g. the number of times a post has been read or the number of up-votes) that are found to be useful in detecting the informational need or urgency [6], [2] uses a set of structural features in combination with a linguistic feature in addition to also relying on the other dimensions of a post, such as expression of a sentiment and the sense of urgency. This latter reliance on the other dimensions is not realistic given the manual effort of assigning the labels for sentiment and urgency (needed to design corresponding classifiers). Our study shares similarities with these prior studies in that we rely on the discussion forum information, but differs from them by the use of a novel set of features that encode content-related aspects of forum posts to account for and structural aspects of the forum posts.

We compare the performance of our detection approach to that in [2] and show that our approach *outperforms* current state-of-the-art by a wide margin both in-domain and across course domains. In addition, differing from prior work, we show that our confusion classifier can simultaneously detect urgency, thereby addressing the need for immediacy for learning effectiveness.

3. DATA DESCRIPTION

The forum posts analyzed in this study are from the Stanford MOOC Posts dataset, a corpus composed of 29,604 anonymized learner forum posts from eleven Stanford University public online classes [1]. The posts are taken from three course domains: Humani-

Table 2: Summary of posts from the three discussion forums

Category	No. of Posts	Not Confused	Confused	Confused & Urgent (%)	No. of sentences per post (mean, sd)
Education	9878	6714	640	67.5	(3.6, 2.8)
Humanities	9723	1358	2257	86.4	(4.5, 4.7)
Medicine	10001	1581	1598	38.9	(4.3, 3.7)

ties/Sciences, Medicine, and Education, with about 10,000 posts in each set.

A salient feature of the dataset is that each post is available with manually assigned labels for six dimensions indicating *confusion*, *urgency*, *question*, *opinion*, *answer*, and *sentiment*. We encourage the readers to refer to [1] for more details. In our study, we only consider the dimensions of *Confusion* and *Urgency*:

Confusion - encodes the extent to which the post expresses confusion, on a scale of 1 (expert knowledge) to 7 (extreme confusion);

Urgency - denoting the extent to which the post is interpreted to be urgent and requires that an instructor respond to the post with 1 denoting 'not urgent at all' and 7 denoting 'extremely urgent';

We divide the posts into two groups–"confused" and "not confused" based on their gold *Confusion* scores. A score above 4 is considered a *Confused* post, whereas a score below 4 is regarded as a *Not confused* one (we disregard posts with score $= 4$ from the analyses). Likewise, an Urgency score above 4 is regarded as an Urgent post, whereas a score of 4 and below is regarded as a non-Urgent post. A summary of the data set is provided in Table 2.

4. CHARACTERIZING CONFUSION

To understand how confusion is expressed in forum posts, two of the authors independently coded a random sample of 200 posts from the entire data set for the following 6 types:

1. **Factual**, if the post seeks clarification of a factual aspect of the course material, as in the post, "Does this mean logistic regression always gives adjusted ratios and the manually computed ratios are unadjusted?"
2. **Course-specific**, if the user seeks a course-specific clarification, such as "Dear Staff, Can you give atleast 2 attempts for each quiz. Giving only one attempt is making us loose interest in the course. Kindly consider."
3. **Course-technical**, if the user seeks clarification on technical aspects of the course. For example, "I am trying to download 5.R.RData, but I cannot open it, can please let me know how I can open this file. With kind regards,"
4. **Recommendation**, if the user is seeking a recommendation. For instance, consider the following post. "another question would you use this form throughout the whole essay? or would you shorten it after using the full phrase?"
5. **Frustration**, where the user expresses frustration, as in, "I had the same issue. Am I bad at finding the check button and bad at math???"
6. **Other**, for posts that belong to none of the above 5 types.

The inter-rater reliability, κ, was 0.81. Based on the instances where both coders agreed, we characterize the type of posts. True to the fact that the discussion forum is an avenue for learners to seek learning support from fellow learners, the most popular post type is *Factual* (54% of the annotated posts), where learners seek to clarify their misunderstandings of concepts presented in the course. This post type is then followed by *Course specific* (27%) and *Course technical* (12%). The remaining posts were categorized as *Recommendation* (3%), *Frustration* (2%) and *Other* (2%).

Overall, these observations confirm that posts indicative of confusion need to be addressed in a timely manner; even though some of them may not be explicit questions, they echo the information seeking nature and the uncertainty encoded in posts that are explicit questions. Additionally, we hypothesize that the inherent difference in the nature of affective states encoded as confusion could be responsible for the inconclusive nature of the effect of confusion on learning outcomes (e.g., confusion positively impacting learning in [10] and negatively impacting outcomes in [25].

5. DETECTING CONFUSION

Our next focus is on building a confusion detector that will allow for automatic identification of confusing posts to facilitate immediate response thereby enhancing the learning experience and reducing learner frustration. Towards this end, the confusion-detection features can be grouped into two categories: content-related and community-related features.

Content-related features: These features analyze the textual content of the post:

1. Automated readability index (ARI): Readability indices are designed to measure how understandable a piece of text is. We hypothesize that the posts encoding confusion, owing to their information seeking nature as well as owing to the tendency of learners to post verbatim course content, have higher readability indices (i.e., are more difficult to read) than those posts that do not encode confusion.
2. Post length in words;
3. Unigrams: These binary features encode whether a word occurred in the post or not.
4. Topicality (LDA): These features use supervised Latent Dirichlet Allocation (LDA) [4] to generate the LDA labels as features. Towards this, we first perform a preprocessing step involving stop-word removal (including numbers and punctuation); stemming; and removing high-frequency (top 1%) and low-frequency words (occurring fewer than 5 times). Then a supervised LDA (sLDA) model is obtained with the confusion labels. Here we use the confusion labels for each post to obtain two sets of LDA words (associated with presence/absence of confusion). This model predicts a label (confusion or not) based on the words in the post that occur in the respective LDA set.
5. Question mark: Since confusion is often expressed via questions, this feature checks for presence of a question mark.

Community-related features: A second set of predictors of whether a post encodes confusion or not is obtained by observing how the community of learners reacts to a post. In particular, a post that is of general interest to learners (such as one that is seeking a factual clarification, or that seeks resolution for a course-related technical problem) would be read by several viewers, thus leading to a rela-

Table 3: Performance of our approach and the two baselines. 'NR' stands for results that were not reported in the respective paper.

Course	Model	Accuracy	Precision	Recall	F-measure	Cohen's Kappa
Humanities	Our Model	84.38	90.38	77.16	**83.14**	0.69
	Unigrams Model[3]	71.99	71.00	82.21	75.28	0.44
	YouEDU[2]	NR	77.80	64.20	70.00	0.62
Education	Our Model	80.04	79.44	81.02	80.00	0.60
	Unigrams Model[3]	82.03	78.76	87.81	**82.96**	0.64
	YouEDU[2]	NR	NR	NR	38.30	0.36
Medicine	Our Model	83.75	86.67	80.14	**83.16**	0.67
	Unigrams Model[3]	70.39	72.82	65.33	68.69	0.41
	YouEDU[2]	NR	69.90	58.90	62.70	0.56

tively higher number of reads. Likewise, posts encoding confusion are considered important resulting in higher up-votes. Accordingly, our set of features includes the number of (i) reads and (ii) up-votes of the post.

We cast the task of confusion detection as one of binary classification, where posts expressing confusion constitute the positive class. For the purpose of this study we do not use the confusion-types identified in the characterization. We trained an Elastic-net model, which is a regularization approach that uses a mixture of L_1 and L_2 penalties to perform variable selection [26].

6. EXPERIMENTS

Datasets: From Table 2 we can see that for majority of the courses, the data is biased towards the negative (not-confusion) class. This makes learning difficult, especially for the positive (confusion) class. In order to alleviate this problem, for each course, we down-sample the negative class (randomly) such that the two classes are balanced. Additionally, forum posts from 'Education', contains very few (640) confusion posts. This resulted in a very small resampled dataset for this course (compared to the posts in Humanities and Medicine) after down-sampling the negative class. Noting that this dataset was prone to over-fitting due to very few posts as compared to the number of features, we up-sampled the positive class to twice its original size before down-sampling the negative class as before.

We also tokenized the content of the posts; removed stopwords (175 unique words); stemmed [12]; and removed infrequent words (with count less than 5). The final vocabulary lists for these courses contained about 2400, 1400, and 1750 words respectively.

Evaluation Measure: From the perspective of helping students, the positive (confusion) class, indicative of learner affect, is more important than the negative class. An ideal classifier would, therefore, identify all confusion posts bringing them to the instructor's attention (high recall for the positive class). Additionally, a high precision for the positive class is also important so that the instructor's efforts are not wasted in analyzing false-positives. Therefore, it seems natural to evaluate models using the F-measure of the positive class (in-line with related prior work). For the sake of completeness, we also report accuracy and Cohen's Kappa.

6.1 Confusion Detection

Table 3 compares 10-fold CV results of our model with two prominent baselines: (i) Unlike our model, our first baseline [2] uses manual annotation for dimensions such as Opinion and Question (apart from ground truth confusion labels for training). We include their performance as reported in their paper. (ii) The second baseline [3] uses only Unigram features. We replicated this baseline in our experiments. Also, a random baseline would get a score of 50%. However, we do not include this result in the tables for clarity.

We can see that, for Humanities and Medicine, our model performs significantly better than the baselines. For instance, for the Humanities course, our model achieves 10.4% and 18.8% relative improvements in F-measure over the two baselines. Similarly, on the Medicine course, our model achieves 21.1% and 32.3% relative improvements in F-measure. Our model's Cohen's Kappa (and accuracy when reported) are also better than the baselines. *This indicates the utility of our features in not only learning the positive class, but also performing well on the overall classification task.*

For the Education course, our model outperforms the YouEDU[2] model significantly. Our model achieves an F-measure of 80.0% as opposed to only 38.3% by the YouEDU model. We would like to remind the reader that the data for the Education course was particularly skewed towards the negative class (not-confusion) with only 6.5% of the posts belonging to the positive class (confusion). *This stark difference in performances of the two models, emphasizes the need for models that can pay particular emphasis on the minority class, which in this case is more significant than the majority class.*

Interestingly, for this course, the performance of our model is comparable to the unigrams model [3], with the latter performing slightly better. Both the models use the same dataset and so neither suffers from the rare-class problem. The seemingly disadvantageous nature of our features for this course is not consistent with the results obtained for the other two courses, and requires further investigation. However, in general, the features proposed in our approach provide a considerable boost in performance.

6.2 Effect of Degree of Confusion

As mentioned in the data description, the dimension of Confusion was annotated on a scale of 1-7 (denoting the degree of confusion), which could be potentially construed to correspond to a scale of affective states. While we had conflated all the positive confusion levels (rep. negative levels) for the purpose of detection, here we evaluated the performance of our detector on its ability to detect the degree of confusion. We examined the performance (here, accuracy) at every Confusion degree and report the results in Table 5. We observe that the accuracy monotonically increases with confusion level, suggesting the classifiers suitability for real applications (e.g., potentially informative to instructional designers).

6.3 Feature ablation analysis

Table 4 compares the predictive importance of our various features by removing them one at a time. For convenience, the first row for each course depicts the performance with the full feature set (same as Table 3). From the table, 'Unigram' and 'Question-mark' seem to be the most valuable. For instance, the model for Education re-

Table 4: Feature ablation. For each course, the top row corresponds to the complete feature set. The subsequent rows represent performance with one of the features removed. Removing any feature (except 'LDA') decreases performance, indicating its utility.

Course	Feature-class	Removed Feature	Accuracy	Precision	Recall	F-measure	Cohen's Kappa
Humanities	–	None	84.38	90.38	77.16	**83.14**	0.69
	Community-related	Number of Reads	84.16	89.86	77.24	82.96	0.54
		Score	84.24	89.86	77.40	83.06	0.55
	Content-related	ARI	84.12	89.66	77.40	82.95	0.55
		Post Length	83.76	89.40	76.88	82.53	0.54
		Unigrams	80.73	88.44	70.82	78.53	0.24
		LDA	84.52	90.01	78.05	83.43	0.70
		Question Mark	70.91	72.64	73.03	72.00	0.53
Education	–	None	80.04	79.44	81.02	**80.00**	0.60
	Community-related	Number of Reads	76.99	75.67	79.30	77.26	0.54
		Score	77.46	75.98	80.16	77.88	0.55
	Content-related	ARI	77.62	76.04	80.47	78.04	0.55
		Post Length	76.88	75.42	79.53	77.25	0.54
		Unigrams	62.15	60.67	69.77	64.70	0.24
		LDA	85.16	83.33	87.97	85.44	0.70
		Question Mark	76.64	73.86	82.58	77.88	0.53
Medicine	–	None	83.75	86.67	80.14	**83.16**	0.67
	Community-related	Number of Reads	83.65	86.41	80.20	83.10	0.67
		Score	83.72	86.49	80.26	83.17	0.67
	Content-related	ARI	83.78	86.51	80.39	83.25	0.68
		Post Length	83.72	86.59	80.14	83.13	0.67
		Unigrams	80.43	86.65	72.93	78.91	0.61
		LDA	83.62	86.06	80.64	83.14	0.67
		Question Mark	70.04	73.90	62.49	67.55	0.40

Table 5: Accuracy of the model in detecting Confusion at different levels. Numbers in () show number of instances. Performance improves with increasing scores. Confusion at levels higher than 5.5 did not have sufficient instances.

Course	4.5	5	5.5
Education	0.76 (521)	0.80 (93)	0.87 (24)
Humanities	0.69 (463)	0.79 (553)	0.79 (190)
Medicine	0.71 (641)	0.86 (762)	0.90(154)

lies heavily on the Unigram features (removing which decreases the F-measure from 80% to 64.7%). Removing any of the other features like 'Number of reads', 'Post Length' also hurt model performance, albeit to a lower degree. Experiments reveal that the inclusion of LDA as a feature hurts more than helping the model's performance. Overall, we can conclude that removing most of our features reduces the performance of the model to various degrees, indicating their utility.

6.4 Testing on Unseen Courses

Our supervised model requires having labeled training data. However, considering the short duration of most online courses, manually annotations for an ongoing course is not only expensive but also infeasible due to time and privacy constraints. Hence, domain-independence of such classifiers is extremely desirable. In our next experiment, we test a given model on an unseen course in order to estimate the domain-independence of existing methods. Table 6 shows the results of this experiment. The last column of the table shows the change in model's performance when tested on a course not seen during training. We can see that the model performance always decreases when it is tested on a new course. However, the decrease can be expected to depend on the difference in the class-conditional distributions of the train and the test sets. *From this perspective, one could argue that the post from Humanities and Medicine are more similar to each other than to the posts from Education, as far as this task is concerned.* From instance, when a model trained on data from Humanities is tested on data from Medicine, and vice-versa, the decrease in F-measure is only about of 4 points. On the other hand, the model suffers a much greater decrease in performance when it is trained on data from Medicine (or Humanities) and is tested on data from Education, and vice-versa.

This result indicates that domain-adaptation methods, that aim to build course-independent classifiers, should not blindly aim for classifiers that perform well on all courses. Instead, a more opportunistic alternative would be based on assessing the similarity between the data from the source (training) and the target (testing) courses.

6.5 Urgency Prediction

In Table 2 we can see that there is a high correlation between the 'Confused' and 'Urgent' labelings. For instance, 86.4% of the posts from Humanities labeled as 'Confused' are also labeled as 'Urgent'. Therefore, it would be of interest to investigate how well a model trained for detecting confusion would perform on the task of detecting urgency. Table 7 shows the results of this experiment. For this table we train our model using ground-truth Confusion labeling, and use the trained model to make predictions on the test instances. We then judge model's performance by comparing predicted positive/negative class with the ground truth Urgent/not-urgent class. Note that we use urgent/not-urgent labelings only during evaluation and not training. Like before, we are primarily interested in the F-measure of the positive (urgent) class. From the table we can see that we achieve a reasonably high F-measure especially for Humanities (75.78%) and Medicine (80.68%). This suggests that for the two related tasks, classifiers trained for one task could be used for the other task with little modifications.

7. FUTURE DIRECTIONS

We have presented detailed analysis of posts indicative of confusion from a collection of discussion forum posts from learners on online courses spanning 3 domains. Our detailed manual analysis of the types of confusion posts suggests that subsequent explo-

Table 6: Model performance decreases when tested on unseen courses. Performance drops indicate a need for more aggressive domain-adaptation efforts on *diverse* pairs (like Education-Humanities), as compared to *similar* ones (Humanities-Medicine).

train-Course	test-Course	Acc.	Precision	Recall	F-measure	Kappa	Change in F-measure
Humanities	Humanities	84.38	90.38	77.16	83.14	0.69	–
Education		70.25	67.86	76.95	72.12	0.40	-11.02
Medicine		79.16	78.95	79.53	79.24	0.58	-4.10
Education	Education	80.04	79.44	81.02	80.00	0.60	–
Humanities		71.88	81.60	56.48	66.76	0.44	-13.24
Medicine		70.82	77.17	59.14	66.96	0.42	-13.04
Medicine	Medicine	83.75	86.67	80.14	83.16	0.67	–
Humanities		81.06	87.03	73.00	79.40	0.62	-3.76
Education		65.15	61.40	81.59	70.07	0.30	-13.09

Table 7: Model trained for detecting confusion performs well on the Urgency prediction task without using urgency labels.

Course	Accuracy	Precision	Recall	F	Kappa
Humanities	80.50	72.07	80.59	75.78	0.60
Medicine	83.02	76.57	85.54	80.68	0.66
Education	61.95	30.13	88.15	44.10	0.26

rations could consider more specific models involving dedicated components for each of the confusion types.

Future work could also focus on supplementing our results with qualitative analyses, e.g. via interviews of learners, to explore specific findings in greater depth. Another related direction for future exploration is the inclusion of clickstream information in the analysis to afford a broader view of learner-content interactions in the presence of confusion.

8. ACKNOWLEDGMENTS

This work is supported in part by the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) – a research collaboration as part of the IBM Cognitive Horizons Network.

9. REFERENCES

[1] A. Agrawal and A. Paepcke. The stanford mooc posts dataset, December 2014. Available from http://datastage.stanford.edu/StanfordMoocPosts/.

[2] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke. YouEDU: addressing confusion in mooc discussion forums by recommending instructional video clips. In *EDM 2015*, pages 297–304. ACM, 2015.

[3] A. Bakharia. Towards cross-domain mooc forum post classification. Learning @ Scale, pages 253–256. ACM, 2016.

[4] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, pages 121–128, 2007.

[5] M. K. Chandrasekaran, M. Kan, B. C. Y. Tan, and K. Ragupathi. Learning instructor intervention from MOOC forums: Early results and issues. In *EDM*, pages 218–225, 2015.

[6] S. Chaturvedi, D. Goldwasser, and H. Daumé III. Predicting instructor's intervention in MOOC forums. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1501–1511, 2014.

[7] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun. Finding question-answer pairs from online forums. In *ACM SIGIR conference on Research and development in information retrieval*, pages 467–474. ACM, 2008.

[8] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, 29(3):241–250, 2004.

[9] R. S. J. de Baker, S. M. Gowda, M. Wixon, J. Kalka, A. Z. Wagner, A. Salvi, V. Aleven, G. Kusbit, J. Ocumpaugh, and L. M. Rossi. Sensor-free automated detection of affect in a cognitive tutor for algebra. In *EDM*, pages 126–133, 2012.

[10] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170, 2014.

[11] C. Geigle and C. Zhai. Scaling up online question answering via similar question retrieval. In *L@S*, pages 257–260, 2016.

[12] K. Hornik. Snowball: Snowball stemmers, 2007. R package version 0.0-1.

[13] M. Jenders, R. Krestel, and F. Naumann. Which answer is best?: Predicting accepted answers in MOOC forums. In *WWW*, pages 679–684. ACM, 2016.

[14] R. W. Larson and M. H. Richards. Boredom in the middle school years: Blaming schools versus blaming students. *American journal of education*, pages 418–443, 1991.

[15] D. M. C. Lee, M. M. T. Rodrigo, R. S. J. de Baker, J. O. Sugay, and A. Coronel. Exploring the relationship between novice programmer confusion and achievement. In *International Conference on Affective Computing and Intelligent Interaction*, pages 175–184. Springer, 2011.

[16] B. Lehman, S. D'Mello, and A. Graesser. Interventions to regulate confusion during learning. In *Conference on Intelligent Tutoring Systems*, pages 576–578. Springer, 2012.

[17] A. Ramesh, S. H. Kumar, J. R. Foulds, and L. Getoor. Weakly supervised models of aspect-sentiment for online course discussion forums. In *ACL*, pages 74–83, 2015.

[18] K. Wang and T. Chua. Exploiting salient patterns for question detection and question retrieval in community-based question answering. In *COLING*, pages 1155–1163, 2010.

[19] Y.-C. Wang, R. Kraut, and J. M. Levine. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *ACM conference on Computer Supported Cooperative Work*, pages 833–842, 2012.

[20] M. Wen, D. Yang, and C. P. Rosé. Sentiment analysis in MOOC discussion forums: What does it tell us? In *EDM*, pages 130–137, 2014.

[21] A. F. Wise, Y. Cui, and J. Vytasek. Bringing order to chaos in MOOC discussion forums with content-related thread identification. In *Conference on Learning Analytics & Knowledge, LAK*, pages 188–197. ACM, 2016.

[22] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164, 2009.

[23] D. Yang, D. Adamson, and C. P. Rosé. Question recommendation with constraints for massive open online courses. In *Eighth ACM Conference on Recommender Systems, RecSys*, pages 49–56, 2014.

[24] D. Yang, M. Piergallini, I. K. Howley, and C. P. Rosé. Forum thread recommendation for massive open online courses. In *EDM*, pages 257–260, 2014.

[25] D. Yang, M. Wen, I. Howley, R. Kraut, and C. Rose. Exploring the effect of confusion in discussion forums of massive open online courses. In *Learning@ Scale*, pages 121–130. ACM, 2015.

[26] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.