YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips

Akshay Agrawal
Stanford University
akshayka@cs.stanford.edu

Jagadish Venkatraman
Stanford University
vjagadish@cs.stanford.edu

Andreas Paepcke
Stanford University
paepcke@cs.stanford.edu

ABSTRACT

In Massive Open Online Courses (MOOCs), struggling learners often seek help by posting questions in discussion forums. Unfortunately, given the large volume of discussion in MOOCs, instructors may overlook these learners' posts, detrimentally impacting the learning process and exacerbating attrition. In this paper, we present YouEDU, an instructional aid that automatically detects and addresses confusion in forum posts. Leveraging our publicly-available Stanford MOOCPosts corpus, we train a heterogeneous set of classifiers to classify forum posts across multiple dimensions. In particular, classifiers that target sentiment, urgency, and other descriptive variables inform a single classifier that detects confusion. We then employ information retrieval techniques to map confused posts to minute-resolution clips from course videos; the ranking over these clips accounts for both video-clickstream data and textual similarity between posts and closed captions. We measure the performance of our classification model in multiple educational contexts, exploring the nature of confusion within each; we also evaluate the relevancy of materials returned by our ranking algorithm.

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

* Proliferation of MOOCs
* Volume of posts high
* Difficult to get a birds-eye view of the course, difficult to address it.

* Work looking into sentiment thus far is limited by datasets
* Work has been done on confusion, but not so much on MOOCs (save RosÃl')
* Work into intelligently intervening + aiding the instructor
* Previous work has found forum to perhaps not be the most useful, even

* We suspect that the forum's perceived lack of usefulness is not instrinsic but rather lack of attention lack of instructor tools + isolation from other parts of the classroom.
* We accordingly set out to adress both of these problems – mining for affect gives instructors a pulse on the state of the course, and linking to videos marries forum and other course resources.
* Why video snippets as opposed to videos? [6] – in a retrospective study of four edX courses, the maximum median engagement, regardless of video length, was six minutes. * open sourced our entire implementation

The remainder of this paper is organized as follows. We examine related work in section two, present the Stanford MOOCPosts corpus in section three, sketch the architecture of YouEDU in section four, detail YouEDU's constituent classification and recommendation phases, evaluating both and interpreting results in sections five and six, and propose future work in section seven.

2. RELATED WORK

3. THE STANFORD MOOCPOSTS CORPUS

A precondition to automatically detecting affect in MOOC discussion forums was manually identifying it; given that no publicly-available corpus of tagged MOOC discussion forum posts existed prior to our research, we set out to create our own. The outcome of our data compilation and curation was the Stanford MOOCPosts dataset: a corpus composed of 29,604 anonymized learner forum posts from eleven Stanford University public online classes. Freely available to academic researchers, the MOOCPosts dataset was designed to enable computational inquiries into the nature of both affect and content in MOOC discussion forums.

Each post in the MOOCPosts dataset was scored across six dimensions – confusion, sentiment, urgency, question, answer, and opinion – and subsequently augmented with additional metadata. In this section, we detail the data collection methodology, define each of the six dimensions, and briefly present some insights gleaned by mining the set.

3.1 Methodology: Compiling the Dataset

Nine judges from oDesk were hired to ...

3.2 Insights and Discussion

We report insights gleaned into the nature of affect, etc. across these courses.

3.2.1 Relationship between Variables

In this section, we report the pairwise correlations between variables to 1) shed some light into the nature of each and also 2) to motivate a YouEDU design choice.

4. YOUEDU: DETECT AND RECOMMEND

YouEDU is a personalized intervention system that recommends educational video clips to learners. Figure 1 illustrates the key steps that comprise YouEDU. YouEDU takes as input a set P of forum posts, processing them in two distinct phases: (I) detection and (II) recommendation. In the first phase, we apply a classifier to each post in P, outputting a subset P_c consisting of posts in which the classifier detected confusion. The confusion classifier functions as a *combination* classifier in that it combines the predictions from classifiers trained to predict other post-related qualities.

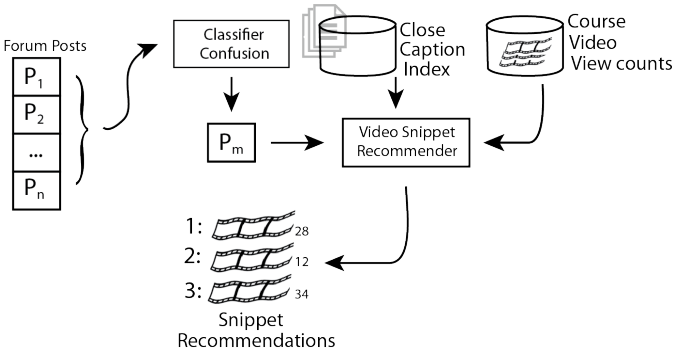

Figure 1: YouEDU Architecture. The YouEDU pipeline consists of two phases: post classification and video snippet recommendation.

The second phase takes P_c as input and, for each confused post in $p \in P_c$, outputs a ranked list of educational video snippets that address the object of confusion expressed in p. In particular, for a given post, the recommender produces an initial ranking across a number of one-minute video clips by computing a similarity metric between the post and closed caption sections. The ranking of videos in the retrieved set is then further informed by video-clickstream data.

While YouEDU outputs minute-resolution video clips, it does not necessarily guarantee that these clips fully address the exhibited confusion – indeed, several minutes of instructional content are often required to explain a single concept. Rather, the video snippets collectively form an ad-hoc index. For example, say that for a given post, YouEDU outputs three video snippets with start times s_1, s_2, s_3, in order of decreasing relevance, and say that these snippets were contained in videos v_1, v_2, v_3, respectively, v_1, v_2, v_3 not necessarily unique. In order to clarify his or her confusion, the author of the post should begin watching video v_1 at s_1 – the learner can autonomously set the end time of the snippet, and can move on to the next video, start time pair if any confusion still lingers.

In the following two sections, we delve further into both phases of YouEDU, describing them in detail and relating the results of empirical evaluations.

5. PHASE I: DETECTING CONFUSION

We frame the problem of detecting confusion as a binary one: Given a discussion forum post p with a true label L in {not confused, confused}, apply some hypothesis h that correctly divines L. Posts with a confusion rating greater than four in the MOOCPosts dataset fall into the "confused" class, while all other posts fall into the "not confused" class.

We craft a rich feature space that fully utilizes the data available in our MOOCPosts dataset, choosing logistic regression with l_2 regularization as our statistical model. Results from empirical evaluations demonstrate that our classifier performs reasonably well, while simultaneously providing insight into the nature of confusion across multiple courses.

5.1 Feature Space and Model Design

Our feature space is composed of three types of inputs, those derived from: the post body; post metadata; and other classifiers. The confusion classifer we train functions as a combining layer that folds in the predictions of other classifiers; these classifiers are trained to predict variables correlated with confusion. We expand upon each type of input here.

5.1.1 Bag-of-Words

We take the bag-of-words approach in representing documents, or forum posts. Each document is represented in part as a vector of indicator variables, one for each word that appears in the training data – the i-th indicator is one if the i-th word in the vocabulary is present in the document, zero otherwise. A word is defined as either a sequence of one or more alphanumeric characters or a single punctuation character (one of {. , ; ! ?}).

Documents are pre-processed before they are mapped to vectors. We prune out stop words, using a subset of the stop word list published by the Information Retrieval Group at the University of Glasgow [1]. Removed words inlcude, but are not limited to, interrogatives ("who", "what", "where", "when", "why", "how"), words that identify the self ("I", "my"), verbs indicating ability or the lack thereof, negative words ("cant", "cannot", "couldnt"), and certain conjunctions ("yet", "but"). We ignore alphabetic case[1] and lemmatize numbers, LaTeX equations, and URLs. Intuitively, the presence of numbers and equations in a forum post might indirectly convey confusion or the lack thereof, in that the learner may be asking a question about some quantity or perhaps providing an answer to a quantitative question; similarly, a knowledgeable learner might answer a question by citing a URL.

The unigram document representation, while simple, pervades text classification and often achieves high performance [3]. We employ l_2 regularization in order to prevent overfitting, a risk that is aggravated when the dimension of the feature space exceeds the training set size [9].

[1]All-caps discussion certainly does communicate affect in some Internet forums – it is typically associated with aggression and is considered a breach of "nettiquette" [7]; however, we assume that MOOC forum-goers are somewhat civil, and so accounting for case would needlessly inflate our feature space.

5.1.2 Post Metadata

The feature vector derived from unigrams is augmented with post metadata, including:

- The number of up-votes accumulated by the post. We rationalized that learners might express interest in posts that voiced confusion that they shared.
- The number of reads garnered by the post's containing thread.
- Whether the poster elected to appear anonymous to his or her peers or to the entire population. It has been shown that anonymity in educational discussion forums enables learners to ask questions without fear of judgement [5], and our dataset demonstrates a strong correlation between questions and confusion.
- The post author's grade in the class at the time of post submission, where "grade" is defined as the numer of points earned by the learner (e.g., by correctly answering quiz questions) divided by the number of points possible. The lower the grade, we hypothesized, the more likely the learner might be confused about a topic.
- The post position – that is, whether or not the post was the first message in a thread. In order to seek help on a forum, a learner must first a post; most likely, we hypothesized, the learner will create a new thread for that post.

5.1.3 Classifier Combination

In section 3.2, we demonstrated that, at least in the humanities and medicine courses, confusion is significantly correlated with questions, answers, urgency, sentiment and opinion. As such, in predicting confusion, we take into account the predictions of five distinct classifiers, one for each of the aforementioned variables. We use the fine-grained method of combining classifiers in which the outputs of several classifiers are fed as input to a *combination function* [2]. In our case, the combination function is itself a classifier.

For a given train-test partition, let D_{train} be the training set and D_{test} be the test set. In both sets, each example is tagged along the six variables. Let H_q, H_a, H_o, H_s, and H_u be classifiers for the question, answer, opinion, sentiment, and urgency variables, respectively. We call these classifiers *constituent* classifiers. Each constituent is trained on D_{train}, taking as input bag-of-words and post metadata features, as described in the previous two subsections.

Let H_c, a classifier for confusion, be our combination function. Like the constituent classifiers, H_c is trained on D_{train} and takes as input bag-of-words and metadata features. Unlike the constituents, H_c also treats the ground-truth labels for the question, answer, opinion, sentiment, and urgency variables as features. When testing H_c on an example $d \in D_{test}$, the constituent classifiers each output a prediction for d. These five predictions are appended to the bag-of-words and metadata features derived from d. The resulting vector is given as input to H_c, which then predicts d's confusion class.

A few subtleties: H_s uses an additional metadata feature that the other classifiers do not – the number of negative words (e.g., "not", "cannot", "never", etc.). H_q, H_a, H_u, and H_c treat the number of question marks as an additional feature, given the previously presented correlations; [11] also used question marks in predicting confusion. And while H_q, H_a, and H_o are by nature binary classifiers, H_s and H_u are multi-class. The latter three predict values in $\{1 + 0.5x \mid x \leq 12\}$. By preserving the full range, we provide H_c with granular information about the non-confusion variables.

5.2 Evaluation and Discussion

For clarity, we refer to the confusion classifier that uses all the features described in the section 5.1 as the *complete* classifier. In this section, we evaluate and interpret the performance of both the complete classifier and confusion classifiers with pared-down feature sets, reporting insights and intuitions gleaned about the nature of confusion in MOOCs along the way.

We quantify performance primarily using two metrics: F_1 and Cohen's Kappa. We favor the Kappa over accuracy because the former accounts for chance agreement [4]. Unless stated otherwise, reported metrics represent an average over 10 folds of stratified cross-validation.

5.2.1 Confusion at the Course-Set Granularity

Table 1 presents the performance of the complete classifier on the humanities and medicine course sets. Both sets are somewhat heterogenous collections of courses. The former contains two economics courses, two statistics courses, a global health course, and an environmental physiology course, totaling to 9,709 posts, while the latter contains two iterations of the same medical statistics course, a science writing course, and an emergency medicine course, totaling to 9,998 posts. In our dataset, not-confused posts outnumber confused ones – in the humanities course set, but 23% of posts exhibit confusion, and in the medicine course set, 16% do. As such, our classifier was naturally better at identifying non-confused posts than confused posts.

5.2.2 The Language of Confusion Across Courses

Table 2 presents the performance of the complete classifier on select courses, sorted in descending order by Kappa. Our classifier performed best on courses that traded in highly technical language. Take, for example, the following two posts from *Managing Emergencies*:

> What could have caused the hemoptysis in this patient with pnuemonia?

> At what doses is it therapuetic for such a patient because at high doses it causes vasoconstrition through alpha1 interactions, while at low doses it causes dilation of renal veins and splachinic vessels.

Both of these posts were tagged as exhibiting confusion, and both of them are fairly saturated with medical terms. Incidentally, our classifier achieved its highest performance when cross-validating on this course (Kappa = 0.741). This makes intuitive sense – a vocabulary so technical and esoteric is likely only used when a learner is discussing or asking a question about a specific course topic. Indeed, inspecting

Course Set	Not Confused			Confused			Kappa
	Precision	Recall	F_1	Precision	Recall	F_1	
Humanities	0.898	0.943	0.919	0.778	0.642	0.700	0.621
Medicine	0.924	0.946	0.935	0.699	0.589	0.627	0.564

Table 1: Complete Confusion Classifier Performance, Course Sets. For each course set listed, this table reports the average performance across 10 folds of stratified cross-validation.

Course	# Posts (% Confused)	F_1: Not Confused	F_1: Confused	Kappa
Managing Emergencies	279 (18%)	0.963	0.771	0.741
Statistical Learning	3,030 (30%)	0.909	0.767	0.677
Economics 1	1,583 (23%)	0.933	0.741	0.675
Statistics in Medicine (2014)	1,218 (28%)	0.908	0.748	0.658
Statistics in Medicine (2013)	3,320 (21%)	0.916	0.671	0.589
Science Writing	5,181 (10%)	0.961	0.527	0.491
Women's Health	2,141 (15%)	0.933	0.506	0.445
How to Learn Math	9,878 (6%)	0.970	0.383	0.359

Table 2: Complete Confusion Classifier Performance, Individual Courses. Our classifier performed best on courses whose discourse was characterized by technical diction, like statistics or economics. In courses like *How to Learn Math* that facilitated open-ended and somewhat roaming discussions, our model found it more difficult to implicitly define confusion.

our model's weights revealed that "systematic" was the 11^{th} most indicative feature for confusion (odds ratio = 1.23) and "defibrillation" was the 15^{th} (odds ratio = 1.22). Similarly, in *Statistical Learning*, "solutions" was the sixth most indicative feature for confusion (odds ratio = 1.75), and"predict" was the ninth (odds ratio = 1.65).

A glance at Table 2 suggests that our classifier's performance degrades as the discourse becomes less technical. Posts like the following were typical in *How to Learn Math*, an education course about the pedagogy of mathematics:

> I am not sure if I agree with tracking or not. I like teaching children at all levels. It seems that if you teach kids at the same level then it becomes the Sam ethnic *[sic]* over and over. ... In a normal class setting the lower level learners can learn from the higher learners and vice versa. Although I do find it very hard to find a middle ground. There has to be an easier way.

The above post was tagged as exhibiting confusion, but the language is much more subtle than that seen in the posts from *Managing Emergencies*. In technical courses, there was a sharp shift in diction between confused posts and not-confused posts. The same could not necessarily be said for non-technical courses, and it is not surprising that we saw our lowest Kappa (0.359) when classifying *How to Learn Math*. Indeed, in the education course, learners tend to voice more confusion about the structure of the class than the content itself – "link", "videos", and "responses" are the fourth, fifth, and seventh most indicative features for the confusion class, respectively.

Examining the feature weights learned when cross-validating on the humanities and medicine course sets provides us with a more holistic view onto the language of confusion. Domain-specific words take the backseat to words that convey the learning process. For example, in both course sets, "confused" was the word with the highest feature weight (odds ratios equal to 3.19 and 2.97 for humanities and medicine, respectively). In the humanities course set, "?", "couldn't", "question", "haven't", and "wondering" came next, in that order. The importance of question-related features in particular is consistent with [13] and with the correlations in the MOOCPosts dataset. In medicine, the next highest ranked words were "explain", "role", "understand", "stuck", and "struggling".

Training Course	Test Course	Kappa
Stats. in Medicine (2013)	Stats. in Medicine (2014)	0.629
Stat. Learning	Stats. 216	0.590
Economics 1	Stats. in Medicine (2013)	0.267
Stats. in Medicine (2013)	Women's Health	0.175

Table 3: Nature of Confusion Across Domains. Training and testing on similar courses typically resulted in high performance, especially in the case of technical courses.

Table 4 displays the most informative features for humanities and medicine course sets, as well as *How to Learn Math* and *Managing Emergencies*.

5.2.3 Training and Testing on Distinct Courses

We ran a series of experiments in which we trained the complete classifier on posts from one course and then tested it on posts from another one, without cross-validation. The results of these experiments are tabulated in Table 3.

Our highest Kappa (0.629) was achieved when training on *Statistics in Medicine 2013* and testing on *Statistics in Medicine 2014*; this makes sense, since they comprise two runs of the same course. Many instructors plan to offer the same MOOC multiple times [8]. If an instructor could tag but one of those runs, then a classifier like ours deployed in an online setting might be met with success. Even if that tagging process is too expensive, our results when training on *Statistical Learning* and testing on *Statistics 216* suggest that an online classifier might perform well so long as its training data derives from the same domain as the test data. Performance might suffer, however, if the domains of the training

Humanities	Medicine	How to Learn Math	Managing Emergencies
constituent:urgency (6.59)	constituent:question (4.05)	constituent:question (6.64)	constituent:urgency (2.47)
constituent:question (3.47)	confused (2.98)	constituent:urgency (2.13)	constituent:question (2.34)
confused (3.20)	explain (2.71)	hoping (1.94)	? (1.73)
? (3.14)	role (2.41)	link (1.76)	metadata:#? (1.54)
couldn't (2.40)	understand (2.36)	available (1.63)	hope (1.40)
report (2.23)	stuck (2.27)	responses (1.62)	what (1.31)
manual (1.94)	struggling (2.25)	middle (1.62)	understand (1.29)
question (1.91)	constituent:urgency (2.25)	support (1.60)	dr (1.24)
haven't (1.84)	sentence (2.20)	discussing (1.60)	how (1.24)
wondering (1.83)	little (2.09)	instruction (1.60)	meant (1.23)

Table 4: Most Informative Features, Odds Ratios. Features prefixed with "constituent:" correspond to the predictions of constituent classifiers, while those prefixed with "metadata" correspond to post metadata features. All other features are unigram words.

and test data are non-overlapping, as is the case in the last two experiments in Table 3.

5.2.4 *Constituent Classifiers and Post Metadata*

Figure 2 illustrates the performance of each constituent classifier when cross-validating on the humanities and medicine course sets, as well as on the education course. The constituent question classifier outperforms all the others by a large margin, likely because the structure of questions is fairly consistent. (Note that the constituent classifiers are not themselves fed by a lower level of classifiers; if we were attempting to predict, say, sentiment instead of confusion, we could try to improve over the performance shown here by creating a sentiment combination function that was informed by its own set of constituent classifiers.)

The combining fuction of our complete classifier consistently determined that the constituent classifiers for the question and urgency variables were particularly indicative of confusion (see Table 4). Table 5 shows the results of an ablative analysis in which one constituent classifier was removed from the complete classifier at a time, until we were left with a classifier with no constituent classifiers (call it a *flat* classifier). The flat classifier performed worse than the complete classifier in the two course sets and the education course, with the drop in performance most pronounced in the medicine course set. For both course sets, the urgency constituent seemed to be the most helpful of the five constituents – this makes intuitive sense, since we would expect that instructors would prioritize posts in which learners were struggling to understand the course material. However, the same was not true for *How to Learn Math*, which is consistent with the fact that no significant correlation between confusion and urgency was found (see section 3.2).

The post position metadata feature also contributed positively to the classifier's performance, as shown in Table 5 – removing it from the flat classifier for medicine dropped the Kappa by 0.03. The other metadata features, however, did not appear to consistently or appreciably affect classifier performance, and so we chose to omit them from our ablative analysis.

Table 4 shows that the number of question marks was also an informative feature, at least in the *Managing Emergencies* course.

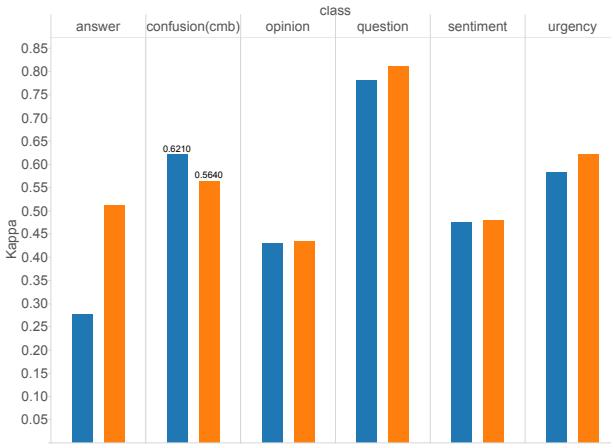

Figure 2: Constituent Classifier Performance. This figure visualizes the performance of the constituent classifiers, as well as the performance of the complete classifier (confusion(cmb), where "cmb" is short for combined). The constituents for answer, opinion, and question are binary classifiers, while those for sentiment and urgency are multiclass, as described in section 5.1.

	Humanities	Medicine	Learn Math
Complete Classifier	0.621	0.564	0.359
Minus Question	0.621	0.559	0.350
Minus Answer	0.620	0.555	0.345
Minus Opinion	0.619	0.553	0.310
Minus Sentiment	0.621	0.555	0.292
Minus Urgency	0.618	0.512	0.337
Minus Post Position	0.614	0.482	0.337

Table 5: Ablative Analysis. This table shows how the removal of the constituent classifiers and the post position feature from the complete classifier affects performance, as measured by the Kappa. Minus question is the complete classifier without the constituent question classifier; minus answer is the minus question classifier without the constituent answer classifier; and so on.

6. PHASE II: RECOMMENDING CLIPS

6.1 The Recommendation Algorithm

6.1.1 Retrieval

6.1.2 Ranking

6.2 Evaluation

Two experts were hired ...

7. FUTURE WORK

Future work might focus on strengthening the link between the classifiers and the reccomendation system; in particular, it would behoove us to devise a way to filter our set of confused posts to a subset for which recommendation makes sense. Additionally, we might want to make our classifiers better and index back into the previous course to retrieve answers for courses. Deploying this system live is another thing that we might do.

YouEDU's two phases need not be packaged together; in an online setting, they could operate as independent, complemetary services. The output of Phase I could be presented directly to instructors, many of whom express interest in understanding activity in discussion forums [10]. As for Phase II, the recommendation system might live as a search-box of sorts: learner would type natural language queries in which they voiced their confusion, and our system would serve them relevant resources.

8. CONCLUSION

YouEDU takes an initial step towards building automated confusion intervention ...

9. ACKNOWLEDGMENTS

This section is optional[12]; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

10. REFERENCES

[1] I. R. G. at University of Glasgow. Stop word list.

[2] P. N. Bennett, S. T. Dumais, and E. Horvitz. The combination of text classifiers using reliability indicators. *Information Retrieval*, 8(1):67–100, 2005.

[3] C. Boulis and M. Ostendorf. Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. Citeseer, 2005.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960.

[5] M. Freeman and A. Bamford. Student choice of anonymity for learner identity in online learning discussion forums. *International Journal on E-learning*, 3(3):45–53, 2004.

[6] P. J. Guo, J. Kim, and R. Rubin. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 41–50, New York, NY, USA, 2014. ACM.

[7] S. Hambridge. Netiquette guidelines. 1995.

[8] F. M. Hollands. MOOCs: Expectations and reality. 2014.

[9] A. Y. Ng. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 78–, New York, NY, USA, 2004. ACM.

[10] K. Stephens-Martinez, M. A. Hearst, and A. Fox. Monitoring MOOCs: Which information sources do instructors value? In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 79–88, New York, NY, USA, 2014. ACM.

[11] M. Wen, I. Howley, R. Kraut, and C. Rosé. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second ACM Conference on Learning @ Scale Conference*, L@S '15, New York, NY, USA, 2015. ACM.

[12] M. Wen, D. Yang, and C. P. Rosé. Sentiment analysis in MOOC discussion forums: What does it tell us? *Proceedings of Educational Data Mining*, 2014.

[13] N. Wilson. Learning from confusion: Questions and change in reading logs. *English Journal*, pages 62–69, 1989.