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

Akshay Agrawal
Stanford University
akshayka@cs.stanford.edu

Jagadish Venkatraman
Stanford University
jagadish@cs.stanford.edu

Shane Leonard
Stanford University
shanel@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 Stanford MOOCPosts corpus, we train a 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 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. Experimental results demonstrate that YouEDU achieves both its goals, paving the way for intelligent intervention systems in MOOC discussion forums.

1. INTRODUCTION

During recent years, many universities have experimented with online delivery of their courses to the public. Hundreds of thousands of learners across the world have taken advantage of these Massive Open Online Courses (MOOCs). While MOOCs are certainly more accessible than physical classes, the virtual domain brings with it its own challenges.

Lacking physical access to teachers and peer groups, learners resort to discussion forums in order to both build a sense of belonging and to better understand the subject matter at hand. Indeed, these forums could in theory be rich reflections of learner affect and academic progress. But, with MOOC enrollments so high, forums can seem unstructured and might even inhibit, rather than promote, community [17]. It becomes intractable for instructors to effectively monitor and moderate the forums. Learners seeking to clarify concepts might not get the attention that they need, as the greater sea of discussion drowns out their posts. The lack of responsiveness in forums may push learners to drop out of courses altogether [27].

The unattended, confused learner might revisit instructional videos in order to solidify his or her understanding. Yet video, a staple of MOOCs, is tyrannically linear. No table of contents or hyperlinks are available to access material in an organized fashion. Often presented with more than one hundred ten-to-fifteen-minute videos, learners might become discouraged when they realize that they will have to re-view footage to patch holes in their knowledge.

We concerned ourselves with solving the problems related to discussion forums and videos that arise when confusion goes unaddressed. In this paper, we present YouEDU, a unified pipeline that automatically classifies forum posts across multiple dimensions, staging intelligent interventions when appropriate. In particular, for those posts in which our classifier detects confusion, our pipeline recommends a ranked list of one-minute-resolution video snippets that are likely to help address the confusion. These recommendations are computed by using subsets of post contents as queries into closed caption files. That the snippets be short is important; [10] found that, regardless of video length, learners' median engagement time with videos did not exceed six minutes. Individual learners may watch beyond the minute we recommend, should they wish.

In order to enable YouEDU's classification phase, we hired consultants to tag 30,000 posts from three categories of Stanford MOOCs: Humanities and Sciences, Medicine, and Education. The set, dubbed the Stanford MOOCPosts Dataset, is available to researchers on request [2]. Besides describing the extent of confusion, each entry in the MOOCPosts set indicates whether a particular post was a *question*, an *answer*, or an *opinion*, and gauges the post's *sentiment* and *urgency* for an instructor to respond. In detecting confusion, our classifier takes into account the predictions of five other constituent classifiers, one for each of the variables (save confusion itself) encoded in our dataset.

The online teaching platforms that Stanford uses to distribute its public courses gather tracking log data comprising hundreds of millions of learner actions. We use a subset of these data as features for our confusion classification. Some of these data are also available in anonymized form to researchers upon request [1]. Until very recently, the data requisite for our classification approach—the MOOCPosts corpus and this additional metadata—simply did not exist.

The remainder of this paper is organized as follows. We examine related work in Section 2, present the MOOCPosts corpus in Section 3, and sketch the architecture of YouEDU in Section 4. In Sections 5 and 6 we detail, evaluate, and discuss YouEDU's classification and recommendation phases. We close with a section on future work and a conclusion.

2. RELATED WORK

Stephens-Martinez, et al. [21] find that MOOC instructors highly value understanding the activity in their discussion forums. The role of instructors in discussion forums is investigated in [22], which finds that learners' experiences are not appreciably affected by the presence or absence of (sparse) instructor intervention. The study did not, however, allow for instructors to regularly provide individual feedback to learners. Instructors interviewed in [12] stress the need for better ways to navigate MOOC forums, and one instructor emphasizes in particular the benefits to be reaped by using natural language processing to reorganize forums.

Wen, et al. [24] explore the relationship between attrition and sentiment, using a sentiment lexicon derived from movie reviews. Yang, et al. [27] conduct an investigation into the relationship between attrition and confusion. While [27] also presents a classifier for confusion, our classification approach differs from theirs in that it operates on a larger dataset and uses a different set of features, including those generated by other classifiers. Chaturvedi, et al. [7] predict instructor intervention patterns in forums. Our work is subtly different in that we predict posts that coders—who carefully read every post in a set of courses—deemed to be urgent, rather than learning from posts that the instructors themselves had responded to. The classification of documents by opinion and sentiment is treated in [20] and [4].

Yang, et al. [26] propose a recommendation system that matches learners to threads of interest, while Shani, et al. [19] devise an algorithm to personalize the questions presented to learners. The need for intervention systems to address confusion in particular is highlighted in [27]. Closed caption files were used in the Informedia project [23] to index into television news shows. To the best of our knowledge, the same has not been done in the context of MOOCs.

3. THE STANFORD MOOCPOSTS CORPUS

Given that no requestable 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. Available on request to academic researchers, the MOOCPosts dataset was designed to enable computational inquiries into 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.

3.1 Methodology: Compiling the Dataset

We organized the posts by course type into three groups: Humanities/Sciences, Medicine, and Education, with 10,000, 10,002, and 10,000 entries, respectively. Humanities/Sciences contains two economics courses, two statistics courses, a global health course, and an environmental physiology course; Medicine contains two runs of a medical statistics course, a science writing course, and an emergency medicine course; Education contains a single course, *How to Learn Math*.

Each course set was coded by three independent, paid oDesk coders. That is, three triplets of coders each worked on one set of 10,000 posts. No coder worked on more than one course set. Each coder attempted to code every post for his or her particular set. All posts with malformed or missing scores in at least one coder's spreadsheet were discarded. This elision accounts for the difference between the 29,604 posts in the final set, and the original 30,002 posts.

Coders were asked to score their posts across six dimensions:

- Question: Does this post include a question?
- Opinion: Does this post include an opinion, or is its subject matter wholly factual?
- Answer: Is this post an answer to a learner's question?
- Sentiment: What sentiment does this post convey, on a scale of 1 (extremely negative) to 7 (extremely positive)? A score of 4 indicates neutrality.
- Urgency: How urgent is it that an instructor respond to this post, on a scale of 1 (not urgent at all) to 7 (extremely urgent)? A score of 4 indicates that instructors should respond only if they have spare time.
- Confusion: To what extent does this post express confusion, or the lack thereof, on a scale of 1 (expert knowledge) to 7 (extreme confusion)? A score of 4 indicates neither knowledge nor confusion.

Coders were given examples of posts in each category. The following was an example of an extremely urgent post:

> *The website is down at the moment https://class.stanford.edu/courses/Engineering/Networking/Winter2014/courseware seems down and I'm not able to submit the Midterm. Still have the "Final Submit" button on the page, but it doesn't work. Are the servers congested? thanks anyway*

And

> *Double colons "::" expand to longest possible 0's If the longest is 0, will the address be considered valid ? (even if it doesn't make sense and there is no room for adding 0's) Can someone please answer ? Thanks in advance*

was given as an example of a post that was both confused (6.0) and urgent (5.0).

We created three gold sets from the coders' scores, one for each course type. We computed inter-rater reliability using Krippendorff's Alpha [11]. For a given post and Likert variable, the post's gold score was computed as an unweighted average of the scores assigned to it by the subset of two coders who expressed the most agreement on that particular variable. Gold scores for binary variables were chosen

	Humanities	Medicine	Education
Urgency	0.657	0.485	0.000*
Sentiment	-0.171	-0.098	-0.134
Opinion	-0.193	-0.097	-0.297
Answer	-0.257	-0.394	-0.106
Question	0.623	0.459	0.347

Table 1: Correlations with Confusion. The urgency and question variables are strongly correlated with confusion. All correlations, save the one denoted by *, were significant, with p-values < 0.01.

by majority votes across all three coders. We refer readers to our write-up in [2] for a more detailed treatment of our procedure and the complete inter-rater reliability results.

3.2 Discussion

We found significant correlations between confusion and the other five variables. In the humanities and medicine course sets, confusion and urgency were correlated with a Pearson's correlation coefficient of 0.657 and 0.485, respectively. In all three subdivisions of the dataset, confusion and the question variable were positively correlated (0.623, 0.459, and 0.347), while the sentiment, opinion, and answer variables were negatively correlated with confusion. Table 1 reports the entire set of correlations.

That questions and confusion were positively correlated supports the finding in [25] that confusion is often communicated through questions. The negative correlations can be understood intuitively. Confusion might turn into frustration and negative sentiment; as discussed in [16], confusion and frustration sometimes go hand-in-hand. If a learner is opining on something, then it seems less likely that he or she is discussing course content. And we would hope that learners providing answers are not themselves confused.

4. YOUEDU: DETECT AND RECOMMEND

YouEDU[1] is an 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 (Section 5).

The second phase takes P_c as input and, for each confused post $p_m \in P_c$, outputs a ranked list of educational video snippets that address the object of confusion expressed in p_m. In particular, for a given post, the recommender produces a ranking across a number of one-minute video clips by computing a similarity metric between the post and closed caption sections. In an online system, of course, learners may choose to watch beyond the end of the one-minute snippet—the snippets effectively function as a video index.

5. PHASE I: DETECTING CONFUSION

We frame the problem of detecting confusion as a binary one. 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 model.

[1]Our entire implementation is open-source.

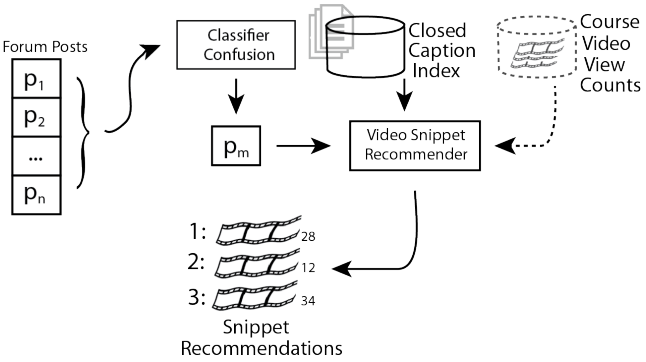

Figure 1: YouEDU Architecture. YouEDU consists of two phases: post classification and video snippet recommendation. The dotted-line module is under construction (see Section 7).

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 classifier 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. The unigram representation, while simple, pervades text classification and often achieves high performance [6]; we employ l_2 regularization to prevent overfitting [18]. Each document is represented in part as a vector of indicator variables, one for each word that appears in the training data. A word is a sequence of one or more alphanumeric characters or a single punctuation mark (one of {. , ; ! ?}).

Documents are pre-processed before they are mapped to vectors. We use a subset of the stop words published by the Information Retrieval Group at the University of Glasgow [14]. Words omitted from the stop word list include, but are not limited to, interrogatives, words that identify the self ("I", "my"), verbs indicating ability or the lack thereof, negative words ("never", "not"), and certain conjunctions ("yet", "but"). We ignore alphabetic case and collapse numbers, LaTeX equations, and URLs into three unique words.

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 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 [9], and our dataset demonstrates a strong correlation between questions and confusion.

- The poster's grade in the class at the time of post submission, where "grade" is defined as the number 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.
- The post's position within its thread—we hypothesized that learners seeking help would create new threads.

5.1.3 Classifier Combination

In Section 3, we demonstrated that 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 correlates. The outputs of these five classifiers are fed as input to a *combination function* [3]—that is, a classifier for confusion—that determines the confusion class for posts.

For a given train-test partition, let D_{train} be the training set and D_{test} be the test set. 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.

Let H_c, a binary 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, when training, 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—and not the ground-truth values—are appended to the vector v of bag-of-words and metadata features derived from d. In particular, if v_h is a vector of length five encoding the predictions of the constituent classifiers, then the concatenation of v and v_h is the final feature vector for H_c.

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; [27] 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. They predict values corresponding to negative (score < 4), neutral (score $= 4$), and positive (score > 4), providing H_c with somewhat granular information. Going forward, we refer to the confusion classifier that uses all the features described in this section as the *combined* classifier.

5.2 Evaluation and Discussion

In this section, we evaluate and interpret the performance of the combined classifier in contrast to confusion classifiers with pared-down feature sets, reporting insights 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 [8]. Unless stated otherwise, reported metrics represent an average over 10 folds of stratified cross-validation.

Table 2 presents the performance of the combined classifier on the humanities and medicine course sets. As mentioned in Section 3, both sets are somewhat heterogeneous collections of courses, with a total of nearly 10,000 posts in each set. In our dataset, not-confused posts (that is, posts with a confusion score of at most 4) outnumber confused ones—only 23% of posts exhibit confusion in the humanities course set, while 16% exhibit confusion in the medicine course set.

5.2.1 The Language of Confusion Across Courses

Table 3 presents the performance of the combined 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 post that was tagged as confused from *Managing Emergencies*, the course on which our classifier achieved its highest performance (Kappa = 0.741):

> *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.*

The post is saturated with medical terms. 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 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 (odds ratio = 1.75), and "predict" was the ninth (odds ratio = 1.65).

A glance at Table 3 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 ... 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 conveying confusion. The language is more subtle than that seen in the posts from *Managing Emergencies*, and it is not surprising that we saw our lowest Kappa (0.359) when classifying *How to Learn Math*. In this course, learners tended to voice more confusion about the structure of the class than the content itself—"link", "videos", and "responses" were the fourth, fifth, and seventh most indicative features, respectively.

Examining the feature weights learned from 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

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 2: Combined Confusion Classifier Performance, Course Sets.

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 (2013)	3,320 (21%)	0.916	0.671	0.589
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 3: Combined 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.

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", "report", "question", "haven't", and "wondering" came next, in that order. The importance of question-related features in particular is consistent with [25] and with the correlations in the MOOCPosts dataset. In medicine, the next highest ranked words were "explain", "role", "understand", "stuck", and "struggling". Table 4 displays the most informative features for the humanities and medicine course sets, as well as *How to Learn Math* and *Managing Emergencies*.

5.2.2 Training and Testing on Distinct Courses

We ran a series of experiments in which we trained the combined 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 5.

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 [12]. Ideally, an instructor would tag but one of those runs, allowing an online classifier to truly shine. Yet even if such tagging were infeasible, our experience learning and testing on similar courses, such as two different statistics courses, suggests that an online classifier might well exhibit good performance. Performance might suffer, however, if the domains of the training and test data are non-overlapping, as is the case in the last two experiments in Table 5.

5.2.3 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 outperformed all the others by a large margin, likely because the structure of questions is fairly consistent. Note that the constituent classifiers were 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.

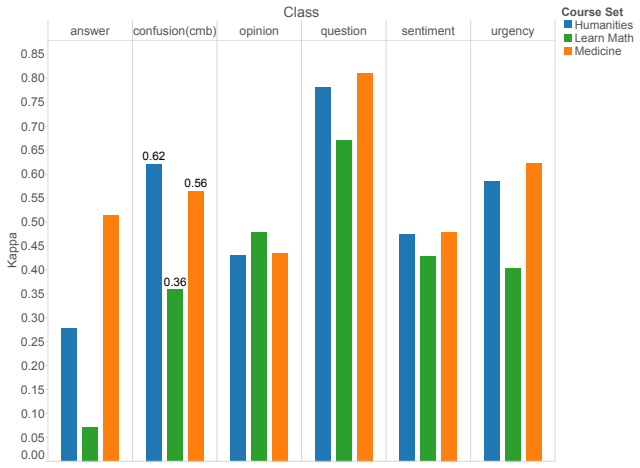

Figure 2: Constituent Classifier Performance. Confusion(cmb) is the combined classifier.

The combining function of our combined classifier consistently determined that the constituent classifiers for the question and urgency variables were particularly indicative of confusion (see Table 4). Figure 3 shows the results of an ablative analysis in which one constituent classifier was removed from the combined 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 combined classifier in the two course sets and the education course. For both course sets, the urgency constituent seemed to be the most helpful of the five constituents—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).

The post position metadata feature also contributed positively to the classifier's performance—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. (Though Table 4 shows that the number of question marks was an

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)

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

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

Table 5: Nature of Confusion Across Domains. Training and testing on similar courses typically resulted in high performance.

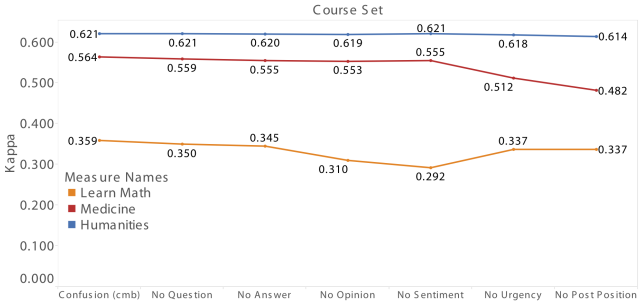

Figure 3: Ablative Analysis, Kappas. No Question is the combined classifier without the question constituent; No Answer is No Question without the answer constituent; and so on.

informative feature in the *Managing Emergencies* course.)

6. PHASE II: RECOMMENDING CLIPS

6.1 The Recommendation Algorithm

In this section, we describe how YouEDU recommends instructional material for a forum post that has been labelled as *confused* by Phase I. Every course can be thought of as a collection of several video lectures. Each video lecture on average is about 12-14 minutes long. We focus on the problem of identifying a ranked list of snippets, S, for each *confused* post. Each snippet s_i in S is a tuple *(video_id, seek_minute)* where *video_id* is an identifier for the recommended video and *seek_minute* is the time in the video to which the learner must seek and start playing the video. We would not necessarily need to recommend an *end_minute* in a deployed setting (learners could choose when to stop watching).

Phase II of YouEDU is divided into an offline indexing phase and an online retrieval phase. We define a *bin* as a time-indexed section of a video. Each bin b_i contains the transcribed text content of the video at a minute-long time interval i. We define *binscore(w, b)* of a word w and bin b as the number of times word w appears in bin b. We formulate video recommendation to learners as a classical information retrieval problem. In classical IR, the goal is to retrieve the top documents that match a user's query. In our case, the query corresponds to a confused post, and the document corresponds to a bin. We want to retrieve a ranked list of bins that addresses the content of the confused post.

6.1.1 *Offline—Indexing Pipeline*

In the indexing pipeline, we first divide each video into bins. We then use a part-of-speech tagger [5] to pre-process each bin. Nouns and noun-phrases tend to produce keywords that typically express what the content is about [13]. Hence, we represent a bin as a triplet (*video_id*, *start_min*, *noun_phrase_list)* where *noun_phrase_list* is a collection of only the nouns and noun-phrases in the bin.

We scan through each of the pre-processed bins and build an index from each word to the corresponding bin that the word appears in. This index would enable us to retrieve the list of bins B_w that corresponds to time epochs in the entire course when the word w was discussed. We also maintain a data structure that keeps track of *binscore(w,b)* for every word and bin. The constructed index and data structures are serialized to disk and are used by the retrieval phase.

6.1.2 *Online—Retrieval and Ranking:*

In the online phase, we take as input confused posts, processing each with a part-of-speech tagger. Similar to the technique we used for bins, we represent each post as a list of its constituent nouns and noun-phrases. Scanning through each of the words in the pre-processed post, we add bin b to the candidate set of retrieved bins if at least one term in the pre-processed post was mentioned in b. Since we have the index constructed offline, we can use it to prune candidates from a large number of available videos (and hence, bins) in the corpus.

We convert each post and bin into a V dimensional vector, where V is the size of the vocabulary computed over all words used in all lectures of the course. In this vector, the value on the dimension corresponding to word w_i is *binscore(w_i,bin)*. We define *simscore(P,B)* as the cosine similarity of the post and the bin.

$$simscore(P, B) = \frac{P \cdot B}{\sqrt{\sum_{i=1}^{V} P_i^2}\sqrt{\sum_{i=1}^{V} B_i^2}} \quad (1)$$

For each candidate bin C_i in the list of candidates C, we compute *simscore(C_i, post)*. We rank all bins in C by their *simscore* values and return the ranking.

6.2 Evaluation

We evaluated our ranking system on the 2013 run of the *Statistics in Medicine* MOOC, offered at Stanford University, which had 24,943 learners. We chose a random sample

of queries from our MOOCPosts dataset for that course. We ran each of those posts through Phase I of YouEDU and chose 20 random posts from the posts that were labeled as confused. For each of those confused posts our algorithm produced a list of six ranked video recommendations (that is, six bins, or one-minute snippets). We then randomized the order within each group of six, obscuring the algorithm's ranking decisions. Four domain experts in statistics at Stanford independently evaluated the relevance of each snippet to its respective post; the ratings of one expert were unfortunately lost due to technical difficulties. This process induced a human-generated ranking, which we then compared to the algorithm's rank order. The rating scale given to the raters is described below:

2: **Relevant**. The recommended snippet precisely address the learner's confusion.
1: **Somewhat relevant**. The recommended snippet is somewhat useful in addressing the learner's confusion.
0: **Not Relevant**: The recommended snippet does not address the learner's confusion.

6.2.1 Metrics

We used two metrics to evaluate the relevancy of our recommendations: NDCG and k-precision.

Normalized Discounted Cumulative Gain (NDCG): NDCG measures ranking quality as the sum of the relevance scores (gains) of each recommendation. However, the gain is discounted proportional to how far down the document is in the ranking. The underlying intuition is that the gain due to a relevant document (say, relevance score of 2) that appears as the last result should be penalized more than it would be if it appeared as the first result. Hence, the DCG metric applies a logarithmic discounting function that progressively reduces a document's gain as its position in the ranked list increases [15]. The base b of the logarithm determines how sharp the applied discount is.

If rel_i is the gain associated with the document at position i, the DCG at a position i is defined recursively as

$$DCG(i) = \begin{cases} rel_{\mathrm{i}} & i < b \\ DCG(i-1) + \frac{rel_i}{\log_b i} & otherwise \end{cases} \quad (2)$$

Since we want a smooth discounting function, we set b to 2. We use a graded relevance scale of 0, 1 and 2, corresponding to the types listed above, and computed the DCG for the ranked recommendations we obtained for each confused post. The ideal value of DCG (IDCG) is defined as the DCG based on the ideal ranking as judged by the raters. To obtain the IDCG, we sort the rankings given by the raters in decreasing order of relevance scores and compute the DCG of the sorted ranking. This corresponds to the maximum theoretically possible DCG in any ranking of the recommendations for that post. We normalize the DCG for our ranking by the IDCG to get the Normalized DCG (NDCG):

$$NDCG(i) = \frac{DCG(i)}{IDCG(i)} \quad (3)$$

If there are n recommended documents, then we report NDCG(n) as *NDCG*, the overall rating for the ranking.

Rater	NDCG	k-precision k=1	k=2	k=3
Rater1	0.66	0.66	0.61	0.62
Rater2	0.90	1.0	0.97	0.97
Rater3	0.82	0.55	0.52	0.52
Avg	0.79	0.74	0.70	0.70

Table 6: NDCG and k-Precision for recommendations

Precision at top k: We define the precision of a ranking R with n recommendations as the fraction of the recommendations that are relevant. The precision at k of a ranking R is defined as the precision of R restricted to its first k recommendations.

6.2.2 Results

Our results across the raters are summarized in Table 6. Our average precision at k=1 is 0.74. This intuitively means that on 74% of cases, the first video that we suggest to a learner (as a recommendation for his or her confused post) is a relevant video. The values at k=2 and k=3, at 0.70, are encouraging as well. Our NDCG numbers are high, indicating that we perform relatively well compared to the IDCG.

7. FUTURE WORK

The work we presented here is a first step; many opportunities for future work remain. We are actively investigating whether we can strengthen our snippet ranking further by considering which video portions learners re-visited several times. This analysis catalogs the number of views that occurred for each second of each instructional video in a course.

Another thrust of future work will use the question and answer classifiers to connect learners to each other. The challenge to meet in this work is to identify learner expertise by their answer posts, and to encourage their participation in answering questions related to their expertise. As in YouEDU, auxiliary data, such as successful homework completion, will support this line of investigation.

A third ongoing project in our group is the development of user interfaces for both instructors and learners. Using our classifiers, we have been experimenting with interactive visualizations of our classifiers' results. The hope is, for example, to have instructors see major forum-borne evidence of confusion in a single view, and to act in response through that same interface.

Video recommendations are not the only source of help for confused learners. Many online courses are repeated during multiple quarters. It should therefore be possible for our system to search forum posts of past course runs for answers to questions in current posts. Also, not all confusion is resolvable through videos. For example, difficulty in operating the video player is unlikely to have been covered in the course videos. Identifying such posts is an additional challenge.

8. CONCLUSION

We presented our two phase workflow that in its first phase identifies confusion-expressing forum posts in very large online classes. In a second phase, the workflow recommends excerpts from instructional course videos to the confused authors of these posts. Our approach utilizes new datasets of human tagged forum posts, data from learner interactions

with online learning platforms, and video closed caption files that are produced in concert with the videos for hearing-impaired learners. Evaluations of our classifiers and recommendations show that both phases of YouEDU perform well, and provide insight into the manifestations of confusion.

As novel online teaching methods are developed, the same underlying challenges will need to be met: keeping learners engaged, allowing them to feel like members of a community, and maximizing instructor effectiveness in the difficult environment of large public classes. Teaching online to very large numbers of learners from diverse backgrounds is formidable. But the potential benefits to underserved populations should encourage the investigative effort required for further research efforts.

9. ACKNOWLEDGMENTS

We sincerely thank Alex Kindl, Petr Johanes, MJ Cho, and Kesler Tannen for slogging through the snippet evaluations.

10. REFERENCES

[1] How to access the Stanford online learning data. `http://vpol.stanford.edu/research`, 2012+.

[2] A. Agrawal and A. Paepcke. The Stanford MOOCPosts Dataset. `http://datastage.stanford.edu/StanfordMoocPosts/`, December 2014.

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

[4] H. H. Binali, C. Wu, and V. Potdar. A new significant area: Emotion detection in e-learning using opinion mining techniques. In *Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on*, pages 259–264. IEEE, 2009.

[5] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[6] C. Boulis and M. Ostendorf. Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. Technical report, University of Washington, 2005.

[7] S. Chaturvedi, D. Goldwasser, and H. Daumé III. Predicting instructor's intervention in MOOC forums.

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

[9] 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.

[10] 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.

[11] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[12] F. M. Hollands and D. Tirthali. MOOCs: Expectations and reality, May 2014.

[13] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223. Association for Computational Linguistics, 2003.

[14] Information Retrieval Group at University of Glasgow. Stop word list. `http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words`. Accessed: 2015-02-05.

[15] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[16] Z. Liu, J. Ocumpaugh, and R. S. Baker. Sequences of frustration and confusion, and learning. In *Proc. Int. Conf. Ed. Data Mining*, pages 114–120, 2013.

[17] A. McGuire. Building a sense of community in MOOCs. `http://campustechnology.com/articles/2013/09/03/building-a-sense-of-community-in-moocs.aspx`, 2013. Accessed: 2015-02-01.

[18] 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.

[19] G. Shani and B. Shapira. Edurank: A collaborative filtering approach to personalization in e-learning.

[20] D. Song, H. Lin, and Z. Yang. Opinion mining in e-learning system. In *Network and Parallel Computing Workshops, 2007. NPC Workshops. IFIP International Conference on*, pages 788–792. IEEE, 2007.

[21] 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.

[22] J. H. Tomkin and D. Charlevoix. Do professors matter?: Using an a/b test to evaluate the impact of instructor involvement on MOOC student outcomes. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 71–78, New York, NY, USA, 2014. ACM.

[23] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens. Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52, May 1996.

[24] 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.

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

[26] D. Yang, M. Piergallini, I. Howley, and C. Rose. Forum thread recommendation for massive open online courses. In *Proceedings of 7th International Conference on Educational Data Mining*, 2014.

[27] D. Yang, 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.