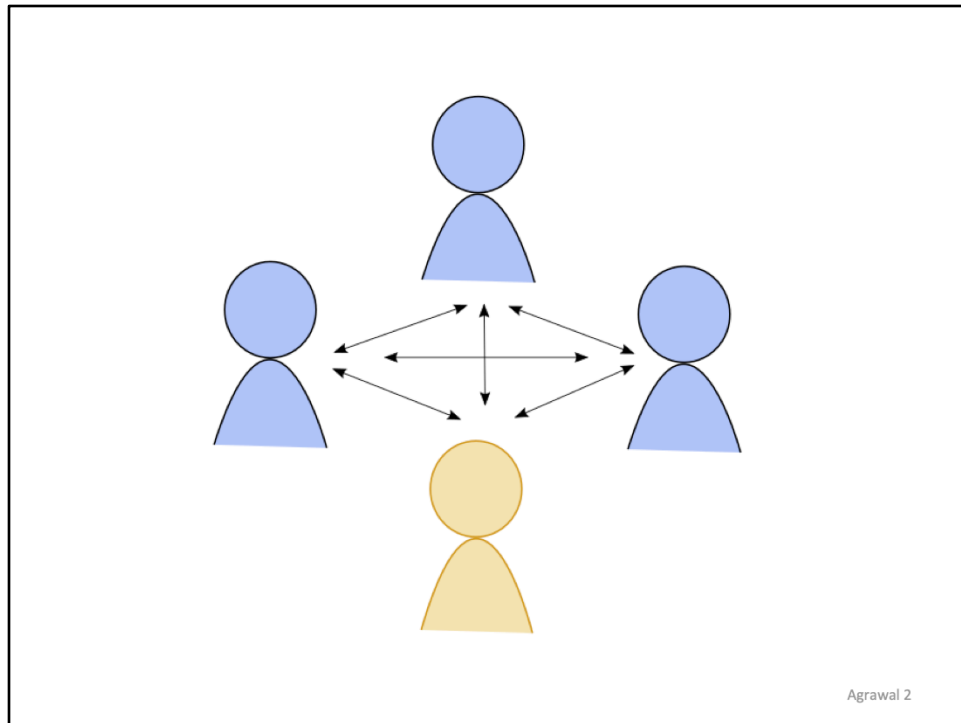


YouEDU

Staging Intelligent Interventions in MOOC Discussion Forums

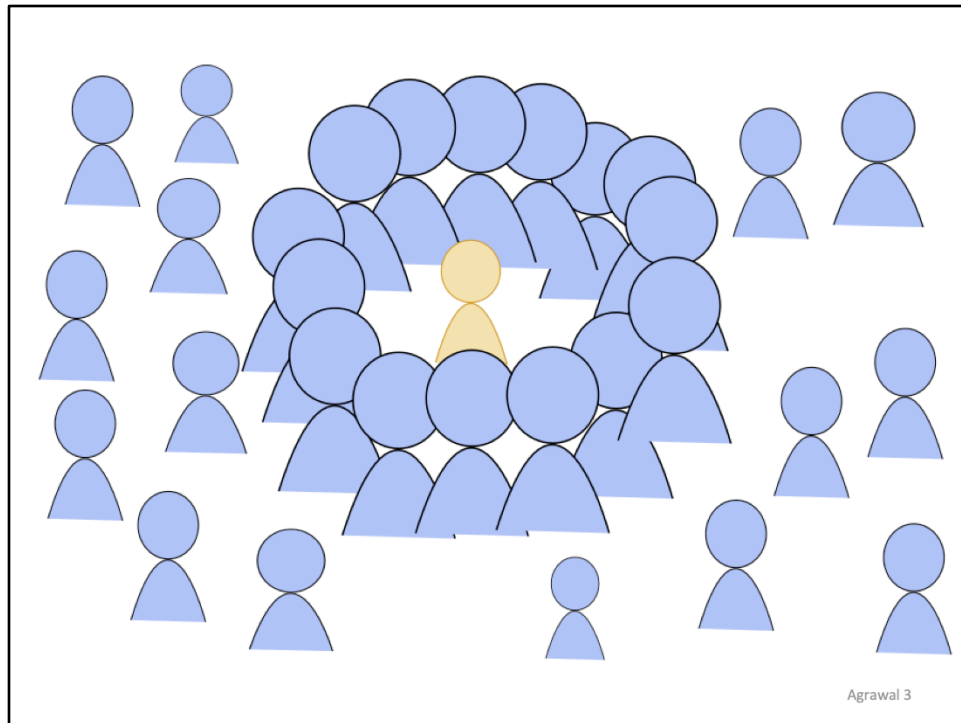
Akshay Agrawal
Jagadish Venkatraman
Shane Leonard
Andreas Paepcke

Hi — my name is Akshay, and today I'm going to talk about YouEDU, a prototype that my colleagues and I built that stages intelligent interventions in MOOC discussion forums.

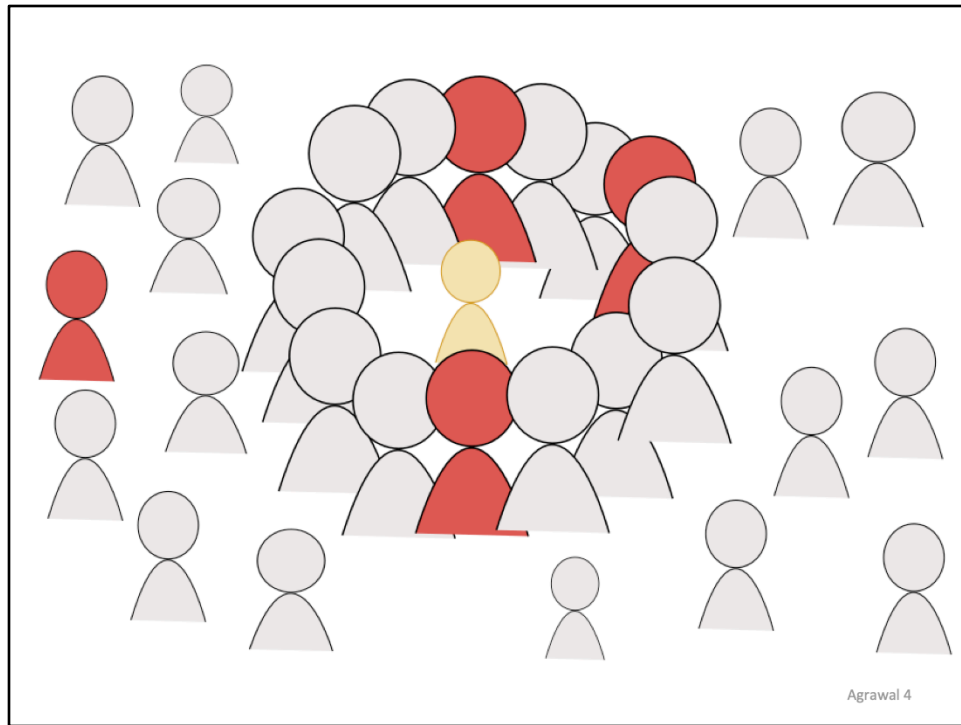


When I think of discussion in an educational context, this is the classical picture that pops into my mind: A few students (the blue figures here) engaging in conversation both with each other and with an instructor, the gold figure here. The number of participants should be small enough so as to allow both the instructor and students to be fully engaged in the discussion, and to really derive something meaningful from it.

Unfortunately, in Massive Open Online Courses, or MOOCs, the reality ends up being more like ...



this: A mob thousands of students vying for the attention of a single instructor, rendering authentic “discussion” impractical; it becomes more of a Q&A. I know this from experience, because I worked as a TA for a Stanford MOOC on Computer Networking last year; my job was to sift through the forum and help students who were struggling with the material.



And, you know, I thought – Wouldn't it be great if the discussion forum could filter out the noise and highlight the learners who were confused about the material and the posts in which they asked for help?

An Intelligent Forum

Agrawal 5

This motivated the idea of a discussion forum that was intelligent in two ways or phases ...

An Intelligent Forum

- Phase 1: Detect Confusion

Agrawal 6

In the first phase, the forum would detect confusion in forum posts and ...

An Intelligent Forum

- Phase 1: Detect Confusion
- Phase 2: Intervene Automatically

Agrawal 7

in the second phase, it would stage some sort of automatic intervention designed to mitigate the confusion that hung over these students.

Challenges

Agrawal 8

We soon found out that there are some challenges, however, to building an intelligent forum.


Challenges

- Scale:  $\times 10^4$

Agrawal 9

The first is scale – a given MOOC might have 10s of thousands of learners in it – increasing the complexity of the problem.


Challenges

- Scale:  $\times 10^4$
- Vocabulary of confusion is domain-dependent

Agrawal 10

Another challenge is that the way confusion is expressed – in other words, the vocabulary of confusion – is largely dependent upon the particular course in which it arises. For example, a learner expressing confusion in a mathematics class will likely use different linguistic structures than one expressing confusion in a humanities class.

Challenges


- Scale:  $\times 10^4$
- Vocabulary of confusion is domain-dependent
- TA overloaded



Agrawal 11

And a third challenge is related to interventions. Since TAs often have their hands full, we'd like our interventions to be independent of them

Challenges

- Scale:  $\times 10^4$
- Vocabulary of confusion is domain-dependent
- TA overloaded
 - lack of forum structure



Agrawal 12

But mitigating confusion automatically seems difficult, particularly because forum posts and the LMS aren't very structured.

Why are forums still dumb?

- No tagged forum datasets
- Availability of large-scale forum data

Agrawal 13

OK, but surely we could surmount these challenges somehow. So why are forums still *dumb*? It mainly boils down to data.

Given domain-specificity, want to take a machine learning approach. Most ML approaches need tagged data, and these datasets are expensive to generate. No such dataset for forums existed prior to our work. What's more, large-scale forum data was also not easily available; this is changing, because Stanford is making much of the data generated by its MOOCs open to researchers.

YouEDU

- Pre-work: Dataset
 - Stanford MOOCPosts
- Phase 1: Confusion Detection
 - Classify individual posts w.r.t. confusion
- Phase 2: Intervention
 - Map confused posts to video snippets

Agrawal 14

So it's against this backdrop that we present YouEDU, our proposed solution to the intelligent forum problem. This is an outline of what remains of the talk:

- + I'll begin by describing a human tagged dataset of forum posts that we compiled that enabled the rest of our work.
- + I'll then talk about the first phase of our system, in which we use machine learning to detect confusion in forum posts.
- + After that, I'll talk about the second phase of our system, in which we stage interventions to automatically mitigate the confusion found in posts. In this phase, we use information retrieval techniques to recommend a list of snippets from instructional videos that we feel might address the confusion voiced in a particular post.

MOOCPosts Dataset

- 30,000 posts, 11 courses
 - Humanities/Sciences, Medicine, Education
- Human-coded, 6 dimensions
 - confusion, sentiment, urgency [1-7]
 - opinion, question, answer [0/1]

<http://datastage.stanford.edu/StanfordMoocPosts/>

Agrawal 15

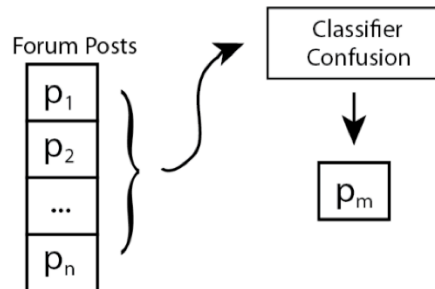
The dataset we compiled, called the MOOCPosts dataset, contains 30,000 forum posts collected from 11 Stanford MOOCs. These 11 courses were partitioned into three categories – Humanities/Sciences, Medicine, and Education. Each partition contains 10,000 posts. The sciences and medicine partitions contained fairly technical courses, and the education set consisted of a single course, *How to Learn Math*, in which teachers discussed pedagogical best practices when it came to teaching math.

Each course partition was coded by 3 distinct human raters, for a total of 9 raters. Each post was scored along 6 dimensions. Three were rated on a scale from 1-7: to what degree does this post express confusion, with 1 being not at all and 7 being a lot, what is the sentiment of this post, 1 being very negative and 7 being very positive, and how urgent is it that an instructor respond to this post, 1 being not at all and 7 being very much so.

The other three dimensions were binary variables: Is this post an opinion, does it contain a question, and does it offer an answer?

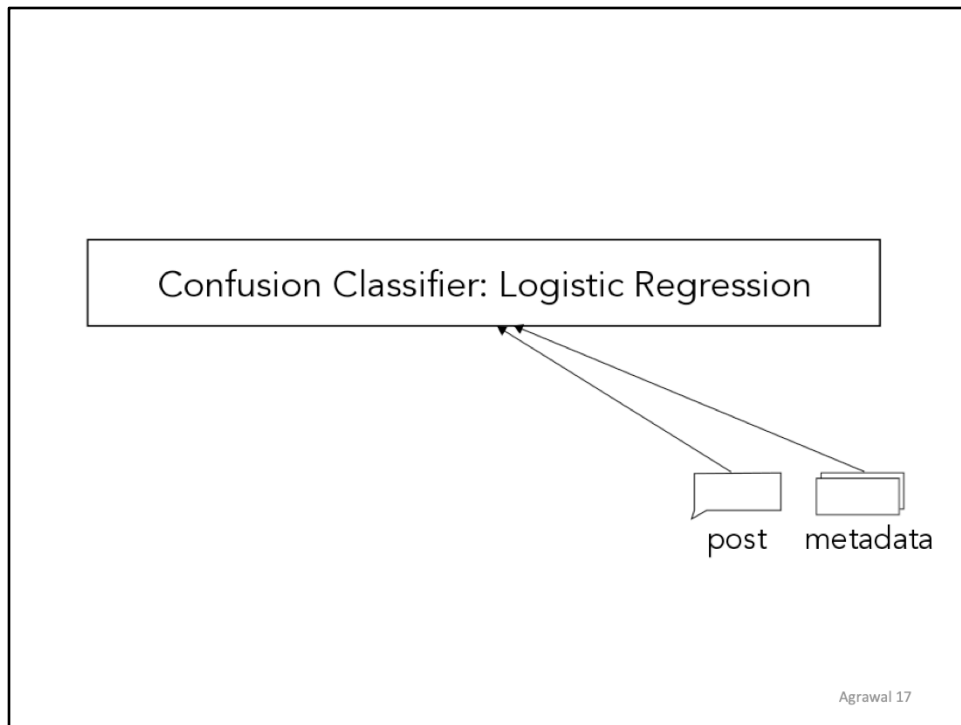
The dataset is available for researchers, and you can read more about it in our paper and at datastage.stanford.edu.

Phase 1: Detecting Confusion

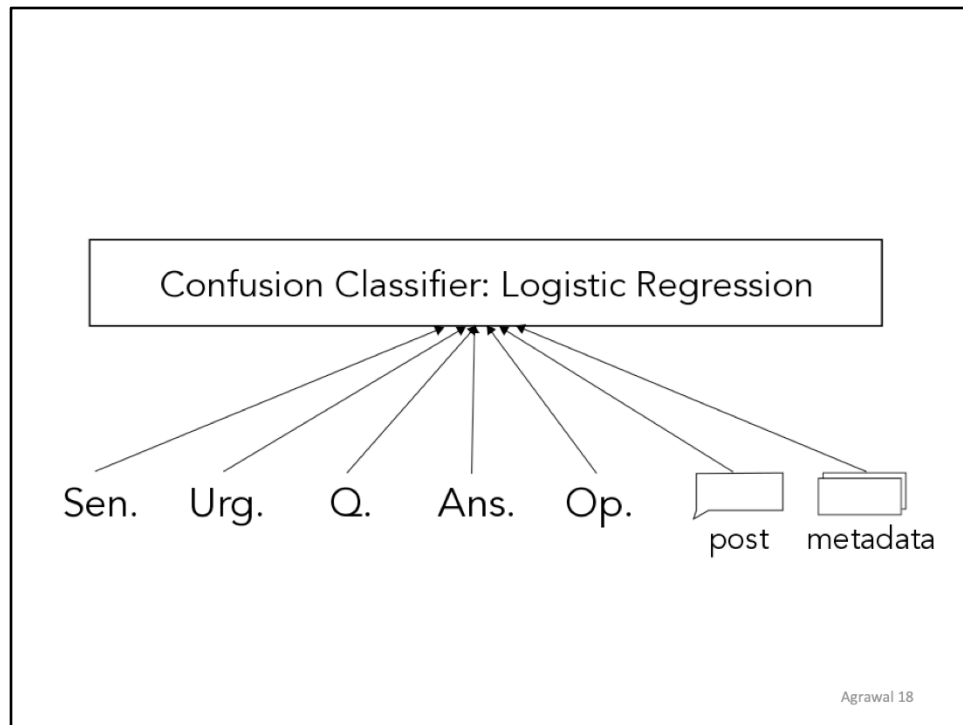


Agrawal 16

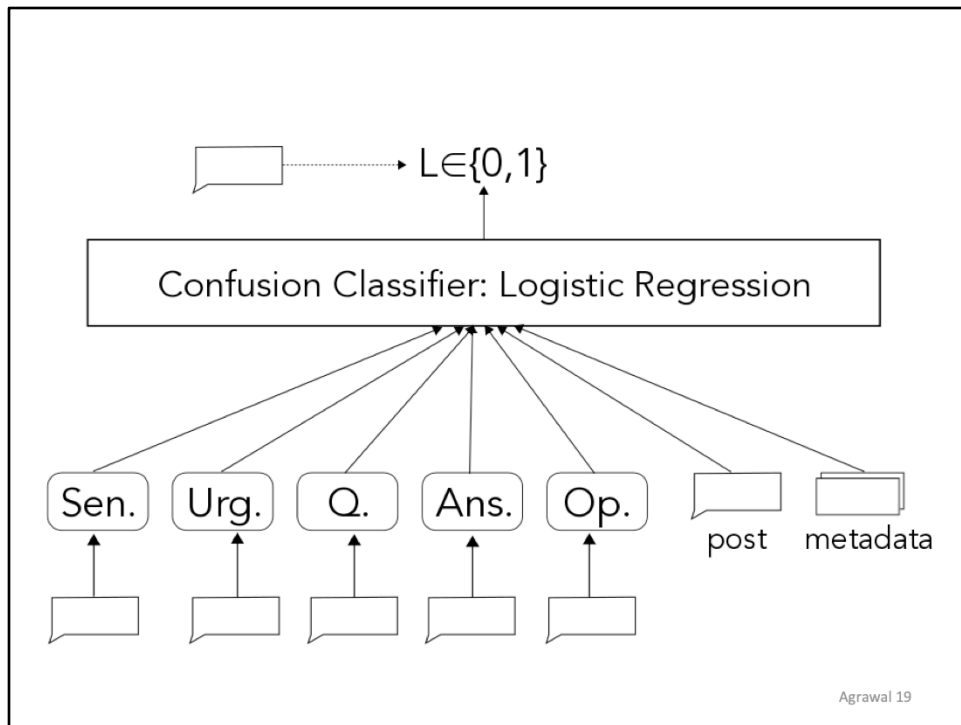
The MOOCPosts dataset is what enabled phase 1 of YouEDU, in which we detect confusion. In particular, in this phase, we take as input a series of forum posts, one-by-one, and feed them into a classifier. In screening these posts for confusion, we frame the classification problem as a binary one: is the forum poster confused?



We used a logistic regression layer as our classifier. The feature vector for our classifier includes a bag-of-words representation of the body of the forum post, as well as some additional metadata about it, including the position of the post within the thread – i.e., did the post start the thread or was it a reply – whether the poster chose to be anonymous, and so on. The intuition here was that people who start threads might be more likely to be seeking help, a student might choose to be anonymous because they were embarrassed about expressing confusion.

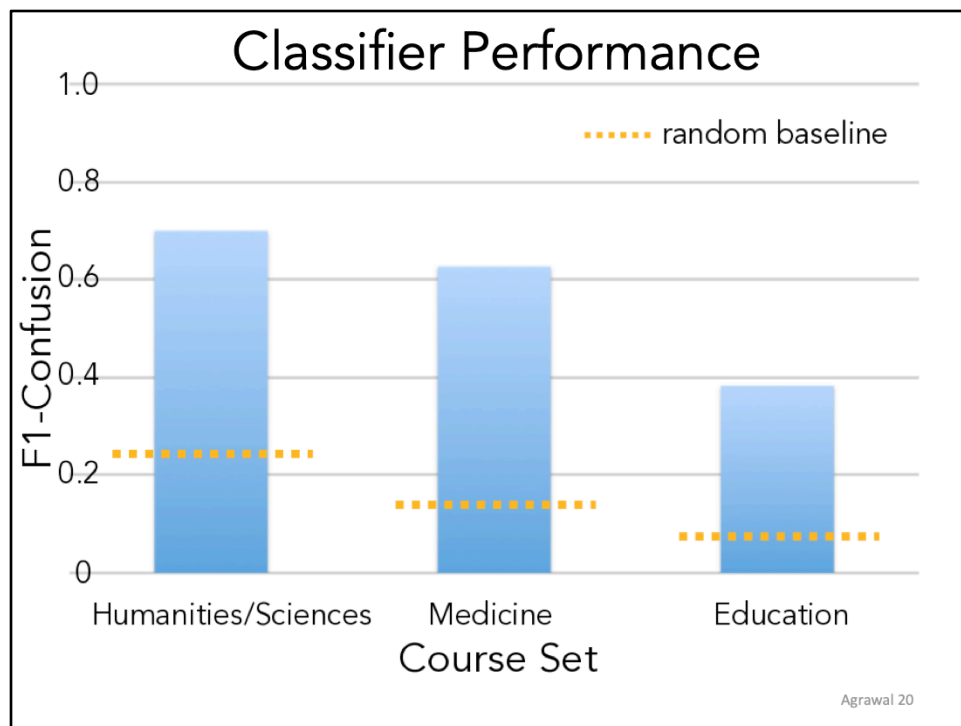


When we train our classifier, the feature vector also includes the ground truth labels for the five other variables from our MOOCPosts dataset – sentiment, urgency, question, answer, and opinion. An analysis of the dataset found that these variables were correlated with confusion. In the training phase, we also build classifiers for the five non-confusion variables – these sub-classifiers are not nested in that they only include the post and metadata as their feature vectors.



When testing, unlike before, instead of using ground-truth values for the five non-confusion variables, our vector includes *guesses* for these values generated by the sub-classifiers we created when training. Our logistic regression classifier folds in all these guesses along with the other features and outputs a binary label indicating whether or not it believes the post voices confusion. We experimented with using guesses as opposed to ground-truth in training as well but found no significant difference in performance.

If you're curious about the relative importance of each of these different types of features, I'd encourage you to look at our paper.

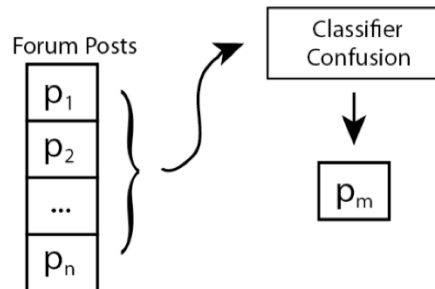


Here, we've got a graph of how well our classifier performed when trained/tested on the three course partitions. The x-axis displays the partitions – hum/science, medicine, and education – and the y-axis is the F1 for the confusion class. The dashed orange lines indicate the expected performance of a random baseline classifier that assigns a post in a given course set as confused with probability equal to the percentage of posts that are actually confused in said course set.

In absolute terms, you can see here that we perform comparably on the sciences and medicine courses, but we perform significantly worse on the *How to Learn Math* course. This result is intuitive, because the science and medicine course sets contained technical courses. And in technical courses, the language of confusion is fairly straightforward and constrained – You know, for example, -- Can someone please explain logistic regression for me? Or "I don't understand such-and-such concept". But, in the *How to Learn Math* course, the language of confusion is complex and wide-ranging, and only six percent of posts expressed confusion.

The upshot of all of this is that, as is often the case when it comes to MOOCs, we are better at solving our problem for math-y courses and not so great at doing so for courses that consist of more authentic discussion or complex thought. The underlying reason for this, we suspect, is that our concept of confusion is not well-defined for these latter courses.

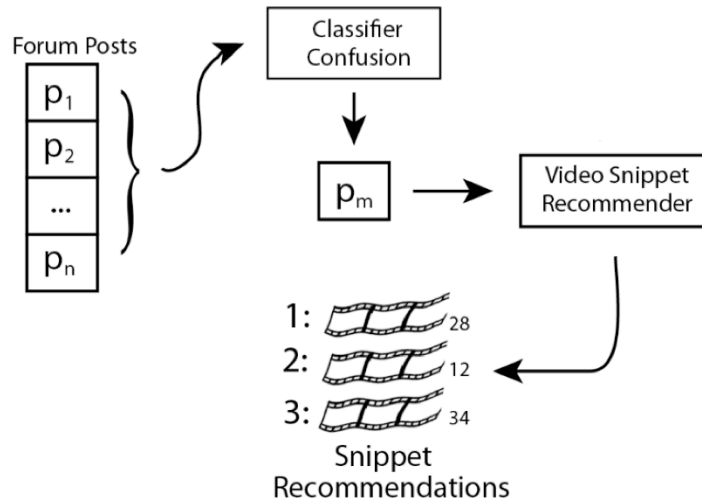
Recap: Phase 1



Agrawal 21

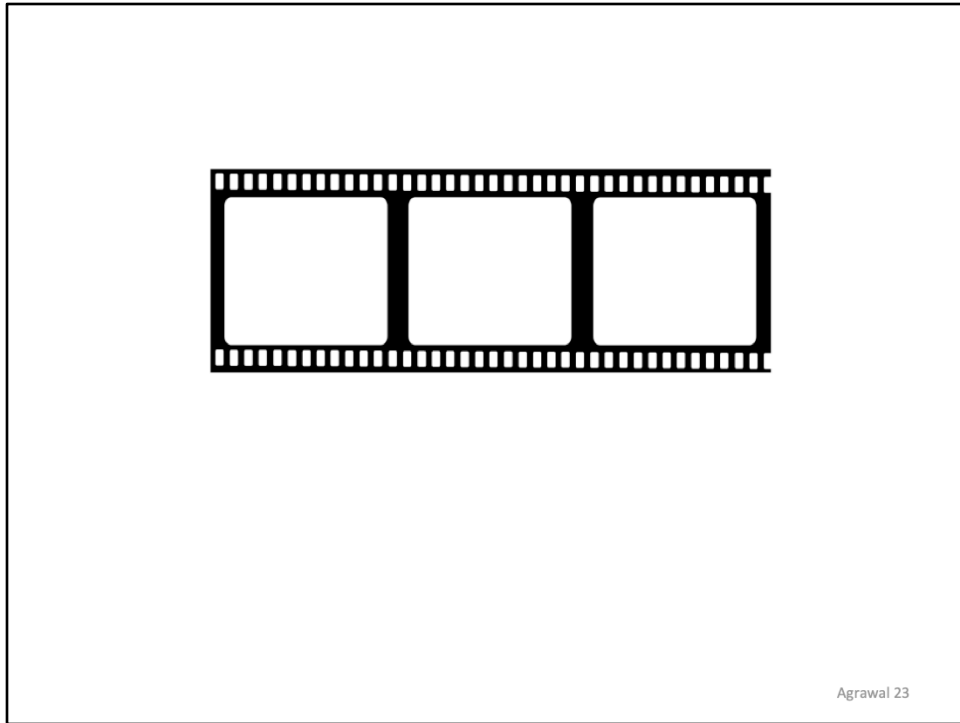
So, to recap the story so-far, the MOOCPosts dataset enabled us to engineer phase 1 of our system, in which we screen posts for confusion.

Phase 2: Staging Interventions

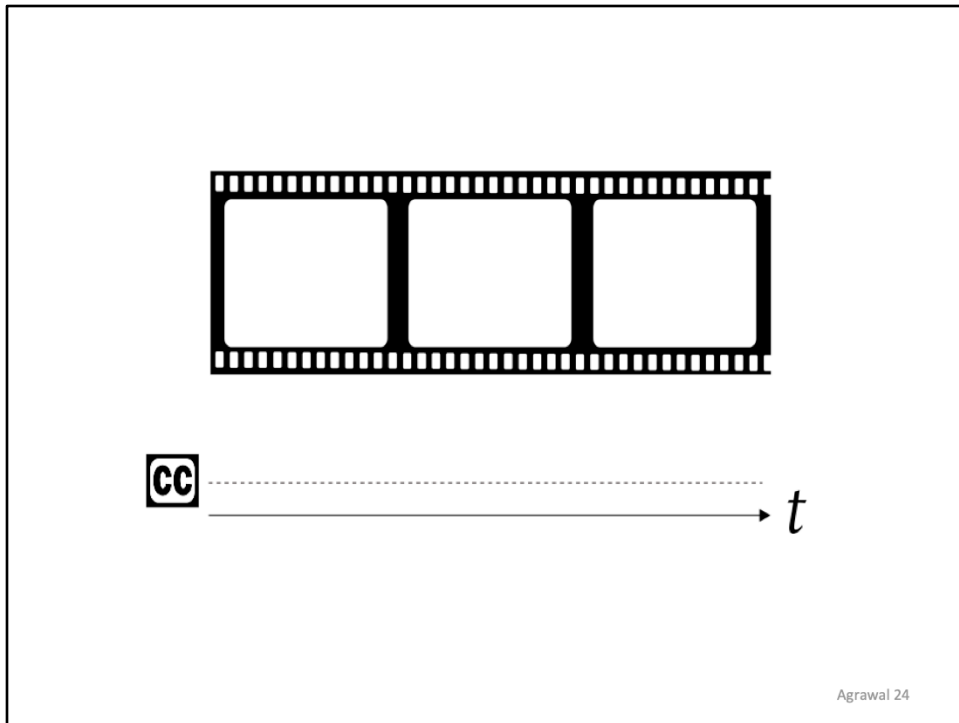


Agrawal 22

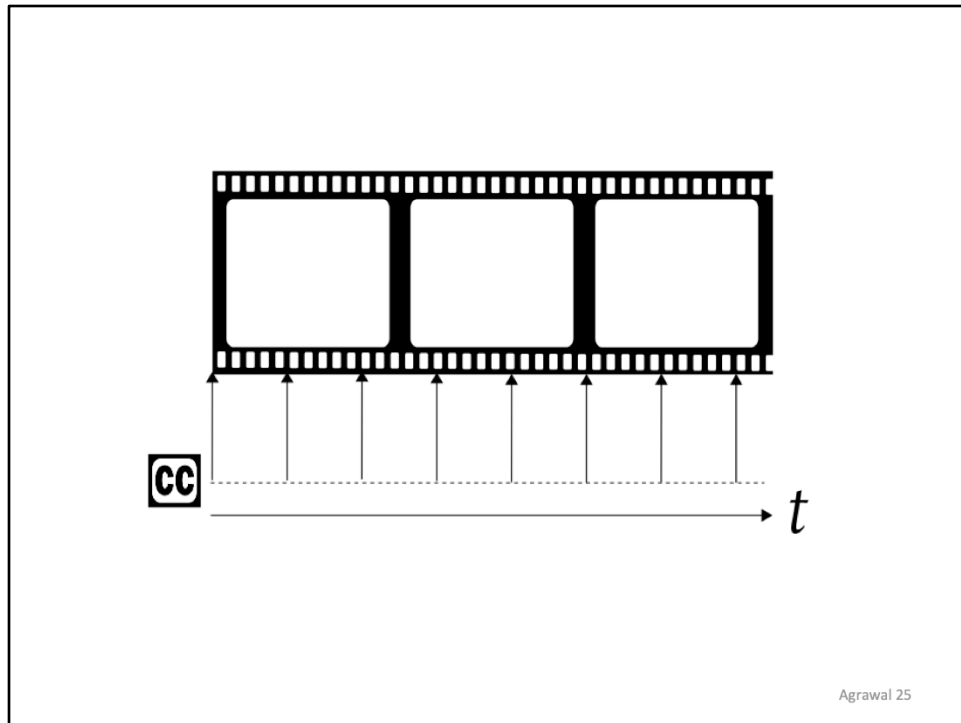
We pick up from there in phase 2, in which we take a confused post and recommend a few video snippets (so video start times) that might address the confusion in that post.



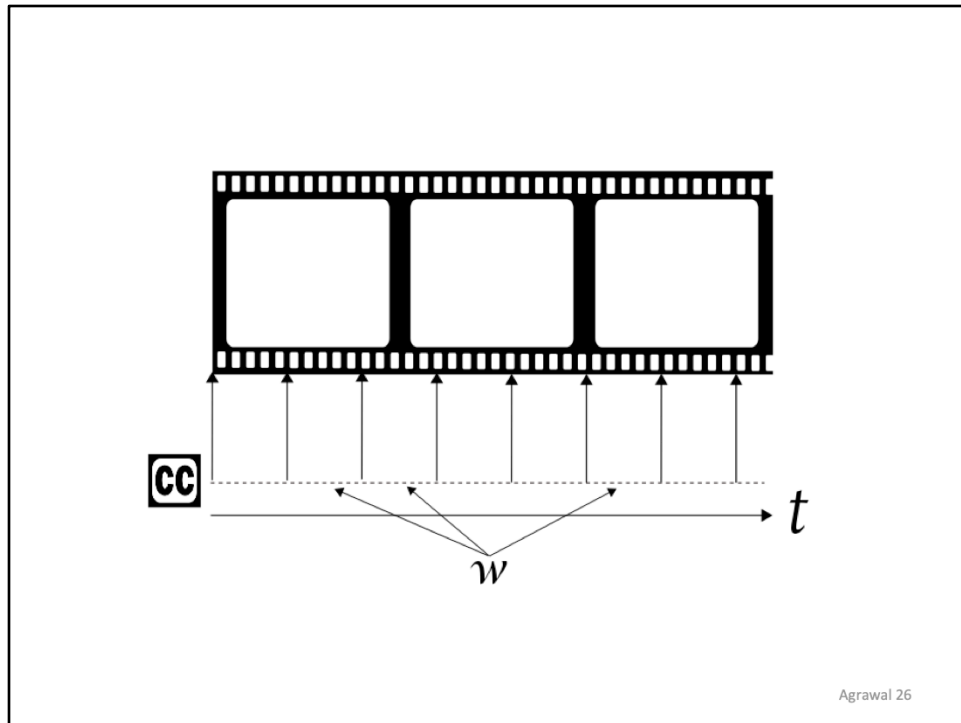
In order to recommend video snippets, we need to have a way of indexing into all of the instructional videos in a course. But, it's difficult to reason about video – it's not clear how to relate posts to videos – so we decided to add a level of indirection.



Luckily for us, our law mandates that these instructional videos be subtitled. So for each video, we have a time-stamped, textual caption file.



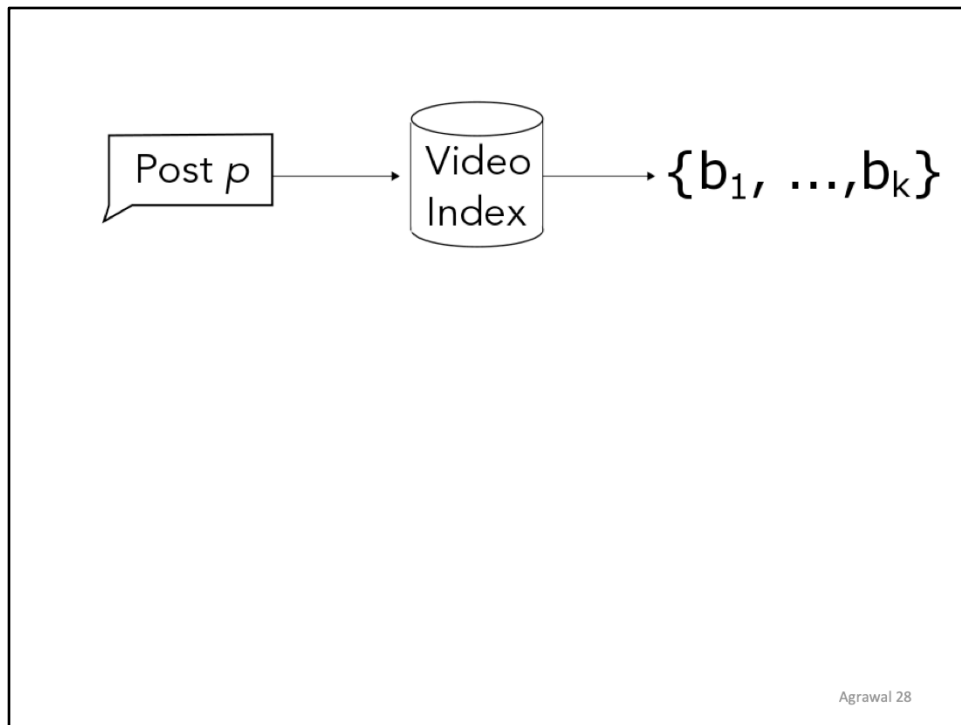
We use that caption file in dividing the video into one-minute chunks, or *bins* – we treat these bins as the fundamental items to be retrieved in phase 2 of YouEDU, as they map directly to video snippets. Each bin is a triplet consisting of the `video_id`, `start_minute`, and the list of noun phrases that occurred in it.



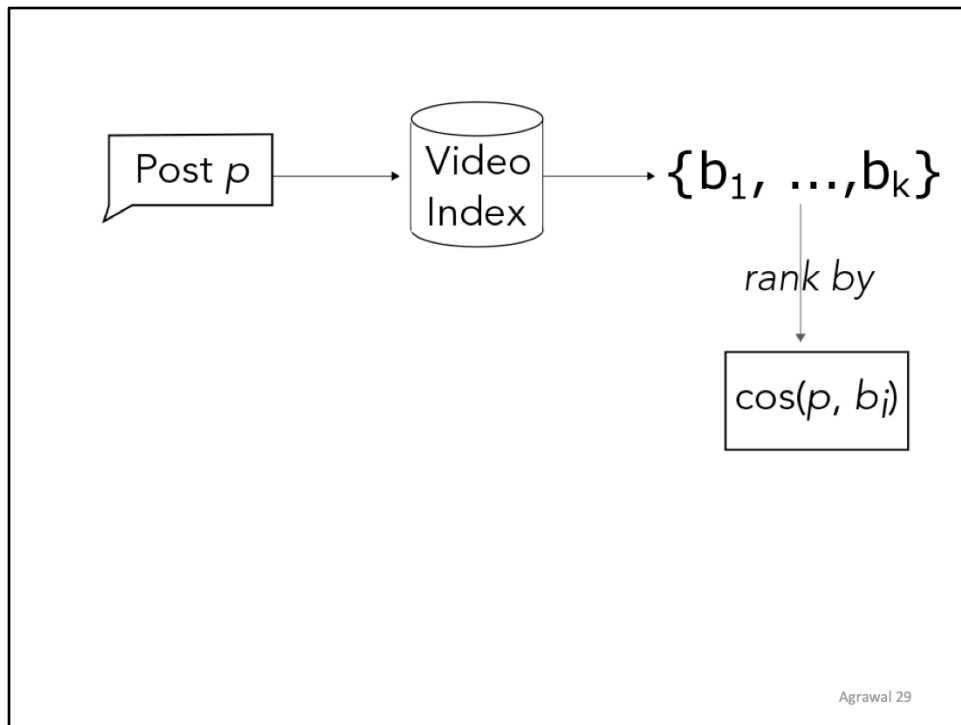
We then scan through *all* of the bins, over all videos, and build a single index mapping each word in our vocabulary to the bins in which the word appeared. This index – from words to bins -- will be used to retrieve video snippets.



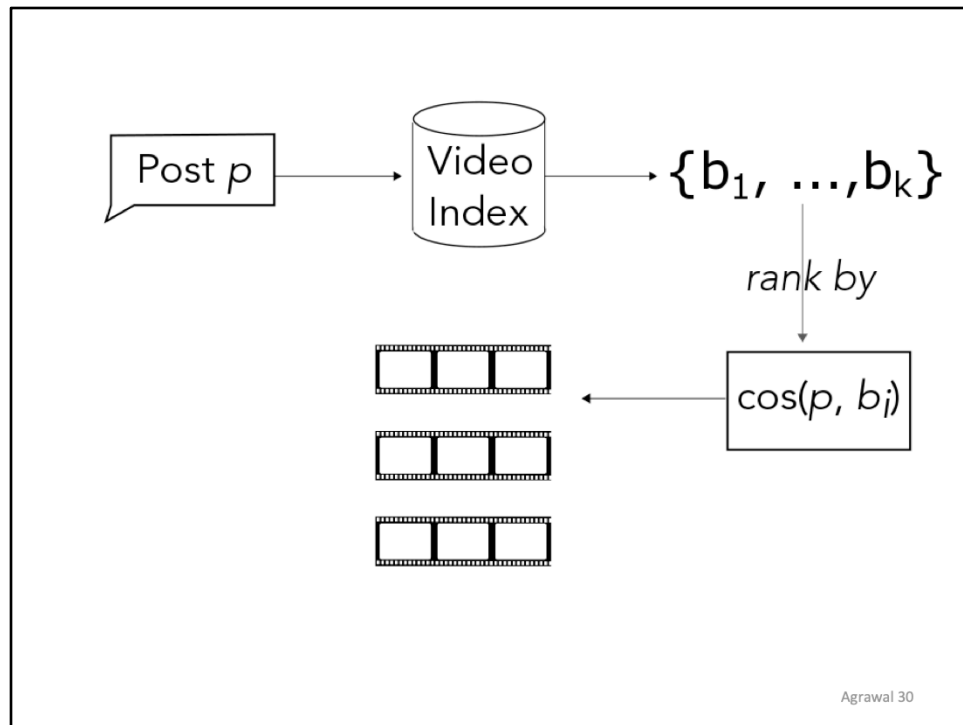
Finally, we frame the recommendation problem as a classical IR problem. Our post is our query, and we want to retrieve the most relevant bins for it.



We begin by pre-processing our post and querying our video index to retrieve all the bins that include at least one word that appeared in our post, narrowing our search space.



Bins and posts are represented as term-frequency vectors over the vocabulary of all the caption files in a given course, so we proceed to *rank* the bins with respect to their cosine similarity with the post.



So, finally, we output a ranked list of video snippets that we hope are related to the content of the post. So, if a learner is confused about, say, the Normal distribution, then these clips should be instructional segments that explain that particular distribution.

Right – so how well did we actually do in making these recommendations?

Interventions: Evaluation

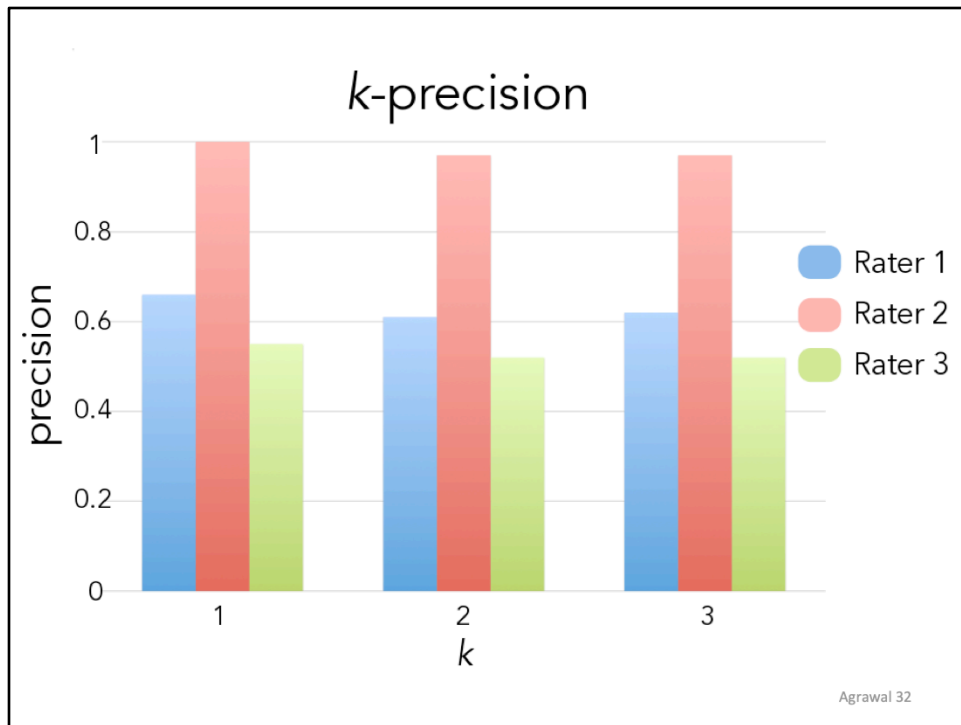
- 20 confused posts, course in statistics
- Six recommendations for each
 - randomize order
- 3 human raters
 - Label each snippet as relevant or irrelevant
- k-precision
 - precision of retrieved documents, limited to the first k recommendations

Agrawal 31

We evaluated our recommender by taking a random sample of 20 confused posts from a course in statistics; we hand-pruned our sample of posts that expressed confusion about, say, how to operate the video-player, as such posts are not in the domain of our recommender system.

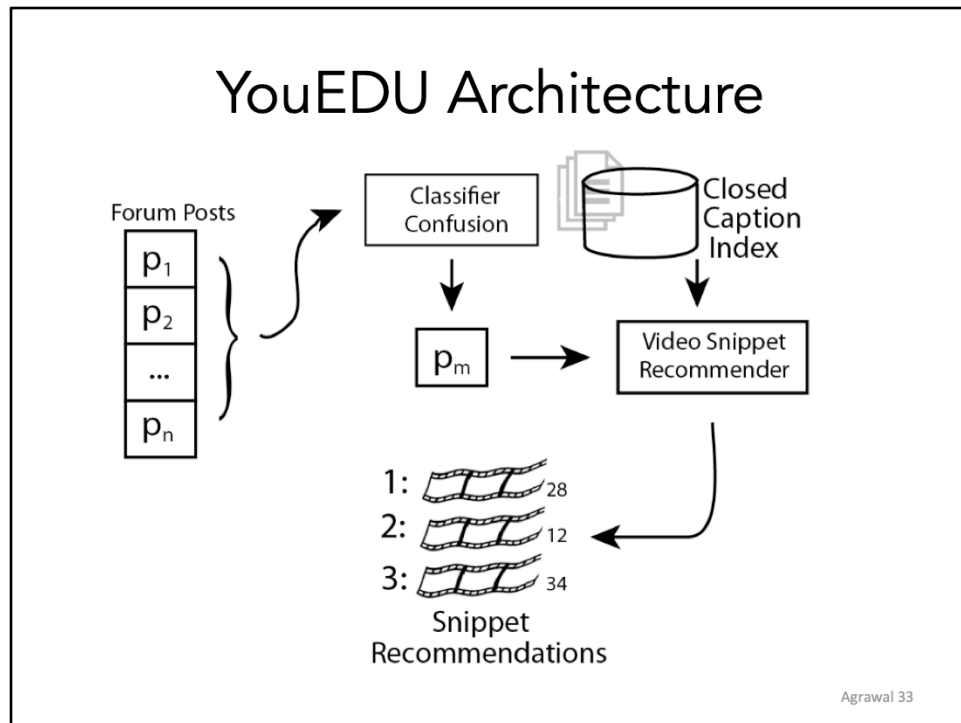
We then used our recommender to generate a ranked list of six recommendations for each of these 20 posts, and we presented them to 3 human raters in a randomized order, obscuring our recommender's ranking. For each post, the raters were asked to label each of its recommendations as either relevant or irrelevant.

One of the metrics we used to quantify our performance was the k-precision, defined as the precision of our video snippets limited to the first k recommendations.



This graph charts our k-precision for $k=1$ to 3. In interpreting this chart, there are a couple of things to note

- though the actual values for each raters were pretty different, the trends from $k=1$ through 3 were consistent across all these raters
- Say that a MOOC consists of 50 10-minute videos. That's 500 bins; for any given post, only a small fraction are likely to be relevant, so a precision of 50% is likely significantly better than random chance.
- That said, there is, of course, still room for improvement.



To summarize – here’s our architecture in full. We begin by screening forum posts for confusion, then use our recommender and our closed caption index to retrieve relevant video snippets.

We demonstrated that our classifier was robust, and, though this is but a prototype, our experiments suggest that something like YouEDU might actually work well in a live setting.

And our work here – we’ve only just scratched the surface. We defined intelligent forums in a narrow way. We could imagine a much more robust forum that did all this but also monitored course sentiment, automatically paired together learners, and self-organized in a way that encouraged authentic discussion. (And all this work is applicable to self-paced MOOCs, too.) We hope that the MOOCPosts dataset will prove useful in enabling researchers and engineers to continue improving the online learning experience.

Classification Suite

<http://github.com/akshayka/edxclassify>

MOOCPosts Dataset

<http://datastage.stanford.edu/StanfordMoocPosts/>

Contact

akshayka@cs.stanford.edu

Agrawal 34

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.

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.

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.

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.