Using Student Annotated Hashtags and Emojis to Collect Nuanced Affective States

Amy X. Zhang
MIT CSAIL
Cambridge, MA, USA
axz@mit.edu

Michele Igo
University of California, Davis
Davis, CA, USA
mmigo@ucdavis.edu

Marc Facciotti
University of California, Davis
Davis, CA, USA
mtfacciotti@ucdavis.edu

David Karger
MIT CSAIL
Cambridge, MA, USA
karger@mit.edu

ABSTRACT
Determining affective states such as confusion from students' participation in online discussion forums can be useful for instructors of a large classroom. However, manual annotation of forum posts by instructors or paid crowd workers is both time-consuming and expensive. In this work, we harness affordances prevalent in social media to allow students to self-annotate their discussion posts with a set of hashtags and emojis, a process that is fast and cheap. For students, self-annotation with hashtags and emojis provides another channel for self-expression, as well as a way to signal to instructors and other students on the lookout for certain types of messages. This method also provides an easy way to acquire a labeled dataset of affective states, allowing us distinguish between more nuanced emotions such as confusion and curiosity. From a dataset of over 25,000 discussion posts from two courses containing self-annotated posts by students, we demonstrate how we can identify linguistic differences between posts expressing confusion versus curiosity, achieving 83% accuracy at distinguishing between the two affective states.

Author Keywords
Massive Open Online Courses (MOOCs); Forums; Emotion; Hashtags; Emojis; Confusion; Curiosity; Online Discussion.

ACM Classification Keywords
H.5.3. Group and Organization Interfaces: Asynchronous interaction; Web-based interaction

INTRODUCTION
Many large courses today use online discussion forums in order to allow educators and students to help one another understand the material. However, in a large course, it can be difficult for instructors to find the students and the comments that need the most assistance, especially if they do not have the capacity to read or respond to every comment. As a result, researchers have become interested in predicting the emotions that student convey in online discussion threads in order to provide further attention to particular students and threads. To do this, prior studies have developed labeled datasets through manually tagging students' posts with their affect using crowd workers hired from systems such as Mechanical Turk [1, 5]. Particular categories that have been annotated include confusion, sentiment, and urgency, among others. However, it is time-consuming and costly to have instructors or paid crowd workers annotate posts.

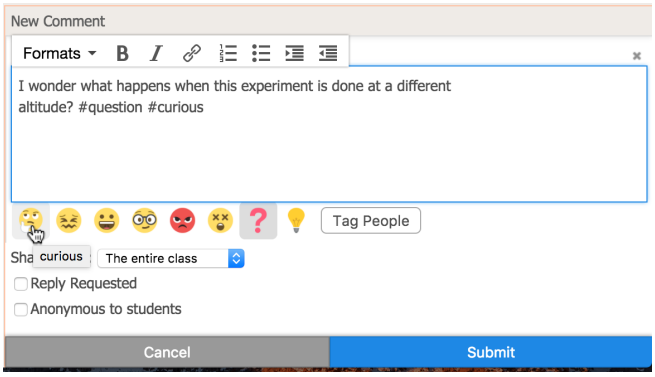

Figure 1. The comment box in Nota Bene that allows the author to add multiple hashtags, along with shortcut emoji buttons, to their post.

Additionally, annotations by paid crowd workers are made by people that are likely not acquainted with the course or the material in question. This may make it difficult to distinguish between more nuanced affective states, such as *confusion* and *curiosity*. In prior work predicting confusion, the appearance of a question mark was found to be the most important feature [5]. However, this does not adequately distinguish confusion from curiosity, as shown in the following examples that have been annotated by the author of the question:

L@S 2017, April 20–21, 2017, Cambridge, MA, USA.

http://dx.doi.org/10.1145/3051457.3054014

- **Curious**: *I wonder why bacteria multiply so rapidly as opposed to other organisms? Could it be that they are mainly unicellular organisms?*
- **Confused**: *Why would ATP be higher in a cancer cell than in a white blood cell? I don't understand how this information helps us understand the cycle of ATP pools.*

This information, while perhaps difficult for paid annotators to infer, can be easily determined by the original poster, since they are intimately aware of their own feelings. We note that the *emotional* state of confusion, which is what we focus on, is different from a confused *understanding* of the material, which may be something best annotated by instructors. Given the nature of what we wish to collect, authors of the post may be a better source than crowd workers for gold data annotations.

To address these issues, we present a novel strategy of collecting annotations using student-provided hashtags and emojis. Self-annotated data from hashtags has been previously explored in the context of Twitter [2] and used towards capturing subtle variations in sentiment. We choose this method as student authors may be one of the best sources regarding their own affective state. Additionally, students may be eager to self-annotate their posts as another form of self-expression or if they know that instructors or other students are using the information to go through posts. Finally, given the prevalence of social media, they are likely comfortable with using hashtags or emojis in online conversation already.

We develop an interface supporting self-coding (shown in Figure 1), and from our deployment on a large discussion system, we collect a dataset of over 25,000 annotated posts. From this dataset, we show how we can develop models to distinguish between the affective states of confusion and curiosity at 83% accuracy. From this analysis, we are also able to learn indicative terms for each category. This suggests that there are indeed some lexical cues that can help convey curiosity versus confusion.

The contribution of our work is in presenting a practical strategy for collecting nuanced affective state in educational forums at scale. We additionally present a model to distinguish posts expressing confusion from posts expressing curiosity. Determining these nuanced affective states has ramifications for instructors aiming to provide interventions for their students, as an instructor would likely have different responses for a student expressing curiosity versus a student expressing confusion.

DATA COLLECTION

We now describe the interface for collecting annotations as well as the annotated dataset that was collected.

Discussion Interface

In the discussion system, buttons containing emojis appear below the comment textbox. As shown in Figure 1, clicking on an emoji adds a hashtag to the textbox. Students can write in their own hashtags but the emoji buttons provide a shortcut to a set of 8 default tags we chose with the help of instructors, as shown in Table 1. As can be seen, users can add as many hashtags as they like to their post. We deployed the buttons on Nota Bene (NB)[1] [6], a textbook annotation system that classrooms can use to have threaded discussions "in the margins", or anchored to a particular place on a page. With a single click, students can create a private or public post anywhere in the margins of the course material. While reading, students can see each others' public posts and respond to them, and instructors can respond to all posts in the course. The system has generated over a million posts and has been used in thousands of courses by over a hundred thousand students.

Hashtag	Count	Emoji	Hashtag	Count	Emoji
#interested	5550		#confused	2612	
#question	5493		#idea	2464	
#curious	5122		#help	840	
#useful	3311		#frustrated	172	

Table 1. Hashtags and associated emoji in the discussion interface.

Hashtags	Count	Hashtags	Count
#curious #question	697	#confused #help	321
#curious #interested	655	#help #question	307
#confused #question	601	#confused #curious	239
#interested #useful	540	#interested #question	221

Table 2. The top 8 pairs of hashtags that appear together in the same post.

Data

The hashtag feature was rolled out to NB in the spring of 2016. While anyone using the system can make use of the buttons, we focus on the data provided by two courses using NB that explicitly encouraged students to use the hashtags. These included summer and fall quarter iterations of a course at University of California, Davis titled *Introductory Biology 2A*. Each iteration of the course enrolls over a thousand students. Reading assignments from textbooks were posted on NB, and course points were awarded for commenting in NB. The use of hashtags was not required for credit but was encouraged by the instructors as a means for communicating affect to others. From the two courses, 293,316 posts were made by 2,353 unique authors. 17.3% or 50,773 of these posts were replies to other posts. From the discussions, we extracted all the posts that contain a hashtag in the text of the post. From the two courses, we collected a total of 25,564 posts, 3,275 of which were replies, containing hashtags written by 1,356 unique authors. This constitutes 8.7% of all posts and 57.6% of all authors. In Table 1 and Table 2, we show counts for each of the hashtags as well as the most frequent pairs of hashtags that appear in the same post.

CLASSIFYING CONFUSION VERSUS CURIOSITY

While there are many use cases for our data, we now focus on the task of distinguishing confusion versus curiosity, an

[1] nb.mit.edu

Model	Accuracy	Precision	Recall	F1	AUC
Baseline	0.69	0.48	0.69	0.56	0.50
LR	**0.83**	**0.83**	**0.83**	**0.83**	**0.78**
SVM	0.81	0.81	0.81	0.81	**0.78**
ADT	0.77	0.77	0.77	0.76	0.71
RF	0.76	0.76	0.76	0.74	0.67

Table 3. Results for predicting confusion versus curiosity posts across the different models.

Feature Set	Accuracy	Precision	Recall	F1	AUC
All Features	0.83	0.83	0.83	0.83	0.78
Content	0.79	0.79	0.79	0.78	0.72
Author	0.71	0.70	0.72	0.69	0.61
Sentiment	0.68	0.64	0.68	0.58	0.52

Table 4. Results for predicting confusion versus curiosity posts using the LR model and using only one of the feature categories at a time.

issue of particular importance to instructors trying to target confused students or improve course materials. In the Introduction, we show an example of a post containing a `#curious` hashtag versus a post containing a `#confused` hashtag. As can be seen, both are posed as questions asking "why" a phenomenom occurs and might both be identified as "confused" by an outside annotator, though the author of the post labeled the first one as "curious".

As posts can have more than one hashtag, we gather all the posts that contain a `#curious` hashtag but no `#confused` hashtag and all posts that contain a `#confused` hashtag but no `#curious` hashtag. This led to 4,875 curious posts and 2,365 confused posts for a total of 7,240 posts, including replies. We also strip all hashtags from the posts.

Models

We experiment with four different classification algorithms and compare the performance. The algorithms we choose are Logistic Regression (LR), Support Vector Machines (SVM) with a linear kernel, Adaptive Boosted Decision Trees (ADT), and Random Forests (RF). We use 10-fold cross validation and average the results. We also have a baseline which is simply tagging all posts as curious.

Features

We make use of three feature categories.

Content: The first is unigrams from the text of the post. We use a word tokenizer that keeps punctuation as separate tokens. We also reduce words to their word stem using the Snowball stemmer. Finally, we use TF-IDF weighting, and set a minimum document frequency of 5 posts. Though we experimented with bigrams and trigrams, we found no improvements using these additional features.

Author: The second feature is the author of the post. This was chosen as some students may tend to post more curious or confused comments.

Curiosity		Confusion	
Feature	Importance	Feature	Importance
wonder	3.055	confus	4.202
curious	2.728	understand	2.189
remind	2.143	write	1.913
scientist	2.080	strip	1.902
g2	2.023	wouldn	1.838
interest	1.950	sentenc	1.789
telophas	1.850	thought	1.729
similar	1.756	don	1.668
cool	1.692	explain	1.619
fascin	1.677	moment	1.603

Table 5. Important unigram word stem features for curiosity and confusion based on coefficients learned from a linear SVM.

Sentiment: The sentiment of the post may be related to curiosity versus confusion as curiosity has positive connotations while confusion has negative connotations. We calculate the frequency of positive and negative terms used in the post using Linguistic Inquiry and Word Count (LIWC) dictionaries[2] [3].

We also experimented with other feature categories that have been used in the past towards predicting confusion [5] or engagement [4], such as certainty and tentativeness, use of negation or personal pronouns, and use of cognitive processes or insight words, all using dictionaries taken from LIWC. While several of these features had weak correlations with the two categories, they did not yield improvements in our best performing model so we omit them here.

Results

The LR model achieves the best accuracy score of 0.83 using all the features. However, as the dataset is unbalanced, area under the curve (AUC) may be more relevant, and here the LR and SVM models have the best score of 0.78. This constitutes a 28% absolute improvement in AUC over the baseline of 0.5.

Also, we note in Table 4 that the LR model with only unigram features has an AUC of 0.72, while the model with only author features achieves a 0.61 AUC. This demonstrates that the content features are the most important towards making the prediction. A model without author features is useful for cases when there is no prior information about the authors on which to train, for instance when needing to annotate posts from a new iteration of a course with new students.

In Table 5, we show the most important unigram features for confusion and curiosity using coefficients determined by a linear SVM model. For the curiosity class, important word stems include "wonder" and "interest". This echoes prior work showing terms signifying cognition or application can predict engagement [4], as curiosity is related to higher engagement. The terms "remind" and "similar" also suggest that the student is making comparisons to other topics. Finally, curiosity is positively correlated with positive emotion (Spearman's rank correlation, ρ=0.115, p<0.0001).

[2] https://liwc.wpengine.com

On the confusion side, terms such as “understand", “thought", and “explain" show a focus on comprehension. We also see the use of negation in word stems such as “wouldn" and “don", and we determine that confusion is positively correlated with negative emotion (Spearman’s rank correlation, ρ=0.155, p<0.0001). As can be seen, features that signal whether a post contains a question, such as presence of a question mark, are not useful for distinguishing between confusion and curiosity.

DISCUSSION AND FUTURE WORK

Using a new method of collecting student-annotated hashtags, we build a dataset of 25,000 posts annotated with a variety of tags with little cost or effort on the part of instructors. With these annotations, we can build novel features within discussion systems, such as allowing educators and students to filter posts by certain hashtags or even augment the reading material with highlights based on emotions expressed. These additions to a discussion interface could further motivate students to voluntarily annotate their posts with affect. In the case of the two courses used in our data collection, the instructors encouraged the use of hashtags by mentioning it during lecture. In other iterations of the course where they did not explicitly discuss hashtags, we saw usage drop considerably. Thus more work is necessary to consider how instructors and system design can encourage annotation.

The creation of this dataset also allows us to differentiate between nuanced emotions such as confusion and curiosity at 83% accuracy and 78% AUC. Given that the most important feature category for our best model is unigrams, this demonstrates that there are indeed some lexical differences between the two categories of posts. This suggests that contrary to our assumptions, it may be possible for crowd workers to distinguish between the two to some degree, even though student self-annotations are still cheaper and faster. Future work will need to determine how well student and crowd annotations align. There may also be some interesting differences between student and instructor annotations. In the case of instructor annotations, the instructor may additionally be able to state whether the student has a confused *understanding* of the material, even if the student may not *feel* confused.

These results suggest future improvements in our model that can account for cases where the student feels confused but there are no clear lexical signifiers. In those cases, underlying knowledge of the course could be taken into account. For instance, to be able to separate confusion and curiosity, one might be guided by what the course has already covered as well as what topics are outside the scope of understanding the material and require inference. A question about a fundamental topic already covered would signify a misunderstanding of something important, or confusion, while a question musing about an unexplained connection would signify curiosity. This would be an interesting future line of work to explore, and our student-annotated dataset could be used for evaluation.

Finally, the ability to detect nuanced signals of affect is an improvement over models that ignore emotions such as curiosity when attempting to identify confusion, as this gives educators the ability to better tailor their actions in response. The differentiation between curiosity and confusion is particularly important as it suggests opposite actions though both types of posts may be asking questions. Students expressing curiosity might be directed to different resources or prompt different responses than students expressing confusion. Instructors may also wish to focus more attention towards helping confused students rather than curious students. Additionally, these different signals can help instructors with improving their course material. Course material that evokes curiosity should be promoted while material that evokes confusion should be targeted for improvement. Since we developed our tagging feature on a textbook annotation system where students can select arbitrary portions of the page on which to comment, we can determine down to the sentence or paragraph level which portions of the material evoke confusion or curiosity.

CONCLUSION

In this work, we develop a method for collecting annotations of nuanced affective states at scale by allowing students to self-annotate their forum posts with hashtags and emojis. For students, this is a new way to express themselves as well as a way to signal to their peers and instructors what kind of attention they want in response. After collecting over 25,000 annotated posts using this strategy from two courses, we demonstrate the usefulness of this data towards understanding nuanced emotions by developing models to distinguish posts expressing confusion from posts expressing curiosity. Our best model achieves an accuracy of 83%, and we also show the most important terms from the text of the post that signify confusion or curiosity.

ACKNOWLEDGEMENTS

We would like to thank Ya’akov Gal and Eran Yogev for useful discussions and Roy Fairstein for developing the hashtag feature.

REFERENCES

1. Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. 2015. YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips. (2015).
2. Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING*. ACL, 241–249.
3. Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
4. Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2014. Linguistic Reflections of Student Engagement in Massive Open Online Courses.. In *ICWSM*.
5. Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. 2015. Exploring the effect of confusion in discussion forums of massive open online courses. In *Learning@Scale*. ACM, 121–130.
6. Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *CHI*. ACM, 1883–1892.