Measuring Learner Tone and Sentiment at Scale Via Text Analysis of Forum Posts

Michael Schubert
Georgia Institute of Technology
Atlanta, GA
mike.schubert@gatech.edu

Damian Durruty
Georgia Institute of Technology
Atlanta, GA
ddurruty@gmail.com

David A. Joyner
Georgia Institute of Technology
Atlanta, GA
djoyner3@gatech.edu

ABSTRACT

Instructors of online courses do not face their students directly and thus must rely on a different set of tools for gauging how a class is doing. Furthermore, the measurement of "how's it going" is vastly different between a group of 25 students and a group of 300. Noisy outliers may lead an instructor to believe things are going well when they are not, or conversely think the class is not understanding things and progressing when in fact they are. In this work, we present two approaches for assessing the sentiment of a large population of learners without the benefit of face to face interaction. The first is a process for analyzing a large body of learner generated posts and determining the overall sentiment. The second is an approach to analyzing individual learners in order to target specific interventions to maximize their success. Leveraging the combination of these approaches will enable instructors to know how a very large body of students are perceiving the work to be performed as well as personalize intervention techniques based on the situation an individual is facing.

Author Keywords

Learning at scale; sentiment analysis; tone analysis; MOOC

INTRODUCTION

Instructors of online courses do not face their students directly and thus must rely on a different set of tools for gauging how a class is doing. There are few visual or audial cues to distinguish a frustrated student from an enthusiastic one, and textual content can often be difficult to interpret, forcing both parties to "read between the lines." This in turn frequently results in inaccurate judgements or, in some cases, hidden grievances. This can be further complicated in online classes with students of varying degrees of comfort with online communication.

Even aside from the differences in medium, the measurement of "how's it going" is vastly different between a group of 25 students than it is in a group of 300 or 3000. Noisy outliers may lead an instructor to believe things are going well when they are not, or conversely think the class is not understanding things and is not progressing when in fact they are.

These problems are exacerbated by the relative newness of interactive online teaching techniques. The instructor may be new to the format and not have the tools to gauge effectiveness in place. New courses may come online that have not been taught in the format before and understanding the effectiveness of its delivery may not be an easily measurable metric. Additionally, existing courses experience content changes or changes in the instructor of record, and the need to gauge how a class is performing becomes pertinent yet again. Even absent these disruptions, online courses lend themselves to more potential complacency than typical courses as significant portions of the work are completed in advance.

This paper explores two systems we have built for evaluating sentiment in our classes in a large, online Master's of Science in Computer Science program [8, 9]. These methods entail the systematic retrieval of online forum posts, scoring individual and group sentiment through available text analytics tools, and returning sentiment measures that can be used to determine overall trends in class sentiment or individual student needs based on perceived negative posting patterns.

EXISTING SOLUTIONS

There are many solutions in the educational technology space that aim to the address the problem of regulating large classes. Tools such as DyKnow [1] and Alma [2] employ data-driven behavioral models to capture disparate student interactions to provide a 360-degree view and predict learning outcomes. There is also research supporting administering the Myers-Briggs Type Indicator test to identify learners with the personality types that may require additional help and enable instructors to accommodate their learning styles [3].

Other solutions use AI to augment human instructors. Cognitive Tutor, developed by Carnegie Mellon University, is an Intelligent Tutor that can be programmed for specific domains or problems [4], while the Andes Physics Tutor by Arizona State University targets beginning Physics students

in particular[5]. While these AI tutors can supplement human instruction, they target a different problem set by providing feedback on work rather than providing responses to open ended questions and answers.

Intelligent tutors are limited in scope, and typically not capable of learning or radically adapting to new circumstances. Goel and Polepeddi succeeded in building an intelligent agent that could respond to open ended questions via the creation of Jill Watson [6]. This agent automated the answering of general questions that repeated within and across semesters in addition to other capabilities. This led to freeing the human teaching assistants to focus on providing individualized and higher quality interactions with their students, especially given the higher volume of forum interactions in these online classes [9]. An open question of their work that our paper seeks to advance is whether such forum-based interventions—whether they are actively assisting students or instructors—are effective in enhancing student performance and improving student retention.

Other collaborations have sought to predict student attrition based on sentiment and included approaches with domain specific positive/negative words and phrases, as well as a general approach for setting up the analysis criteria for a particular course. Wen, et al. present visualizations that show more of a correlation rather than a prediction of future student drop out [7]. The overall sentiment trends layered on top of weekly course markers indicate promise of correlating events with sentiment. This is exhibited in their "Teaching Course" example, where there are several peaks and valleys throughout the course, and the valleys correlate with dropouts. The data is not present in this paper, but a hypothesis would be that these low points were either examinations or major assignments that learners may have performed poorly on.

SOLUTION AND IMPLEMENTATION

We developed two solution approaches in the context of an online Master of Science in Computer Science program that heavily leverages forum discussions as part of the course experience [8]. These tools are specifically built for the needs we observed as instructors, teaching assistants, and students in the program: we noted the challenge of identifying which students in a large online class are in need of individual intervention, as well as the ability of a class to seemingly turn very negative very quickly. Toward the first, we wanted to develop a way for individual students to be surfaced to instructors, and toward the second, we wanted to identify the latent trends in classroom sentiment that may predict an upcoming issue.

We constructed tools to address both of these problems. Each of these approaches consists of three different key components that are responsible for retrieving posts, scoring the body text of the post, and then visualizing the scores for sentiment of the population and tone of the individual.

Text Retrieval

The aforementioned Computer Science program leverages the Piazza platform for section-specific discussion forums. Piazza is a popular web-based forum tool used to facility Q&A and discussions in college classes. In this online graduate program, Piazza plays a critical role: some instructors estimate that students spend more time interacting with their classmates and instructors via Piazza than any other component of the course, including watching lecture videos, completing homeworks, and studying for tests. Most classes use Piazza for instructor-driven discussions, student-initiated questions, regular announcements, grade return, and social interactions. As such, Piazza is a common target for negative and positive posting patterns in response to trends like low assignment averages or high instructor responsiveness.

The retrieval harness was built in Python 2.7.6, and leverages the work performed by Hamza Faran in exposing the private API used by Piazza to serve mobile and desktop clients. The user is first authenticated against Piazza and a list of course sections the user has access to is presented. The current user only needs read access to the public portions of the course section.

The test cases in this release were only performed against the public posts. Should a user designated as an instructor within Piazza run the same analyses, the results may differ as private student posts would also be considered as part of the text retrieval. We noted as instructors, however, that identifying struggling students based on private posting problems is a smaller issue as these students are effectively already surfacing themselves to instructors; the greater issue is identifying students who are not deliberately seeking out instructor feedback.

The system prompts the user to select a course and then begins the text retrieval process. Each post is iterated and scraped for the body content of the text. All formatting is removed such that only the post's content is sent into the pipeline for analysis. This text is then sent to one of two different tools based on the goal: one for sentiment, one for tone.

Sentiment

Sentiment analysis is used for evaluating the overall classroom attitude. This pipeline starts with scoring, then provides the instructor a visualization.

Scoring of Sentiment

The Microsoft Text Analytics API was utilized for the scoring of sentiment in this solution. This API provides a sentiment service that scores the sentiment of an English language post on a scale from 0 to 1. Scores close to 1 represent a very positive body of text, while scores close to 0 represent a very negative body of text.

This API currently supports text of up to 5000 characters. Prior to calling the sentiment service, the body of a post is evaluated for length and truncated to 5000 characters if the

character count exceeds that threshold. No attempts are made to infer the sentiment of the lost text, nor to remove words or phrases that are articles (e.g. 'the'). Less than 1% of posts surpassed this length limit.

Scores are placed in a data frame on a row corresponding to the student that generated the post, and the date that the post was created. These are the key data points required to carry out the visualization aspects of the system.

Visualization of Population Sentiment

Pandas boxplot generation is utilized to create the visualization of sentiment on a week by week basis. Each student's individual scores are averaged each week, then visualized as data points in a boxplot. The median value is denoted, along with the spread of values. An instructor can then use this visualization to compare weekly sentiment to the course syllabus and correlate the learners' attitudes with the materials and objectives in that timeframe. This approach was derived after experimentation and investigation into approaches in similar domains.

These measures can be plotted in aggregate in order to view the changes in sentiment over time. Twitter's use in tracking US elections has shown correlation between events and rallies taking place leading up to election driving sentiment for a candidate either up or down ultimately being reflected in the election [10].

This solution's visualization applies a similar trend analysis to change in sentiment for each period and promotes viewing those trends in conjunction with what was taking place in the class on a weekly basis. A student with decreasing sentiment values may indicate a learner that is struggling as the material progresses. A classroom full of learners exhibiting the same trend may indicate material that has opportunities for improvement.

Tone

Tone analysis is used for identifying individual students who may need instructor intervention. Like sentiment, tone is evaluated first by scoring, then by providing a visualization to the instructor.

Scoring of Tone

The IBM Watson Tone Analyzer service was used to score the tone of language in this solution. The service is a suite of closed-source, proprietary algorithms capable of automatically processing mass textual content. Its API accepts up to 128,000 characters and returns both a sentiment and personality profile for the content passed to it. The scoring returned is then persisted for later visualization, as well as trend analysis over time.

Visualization of Tone

Tone is represented graphically using color coded bar charts rendered by the D3 JavaScript framework. The bars are filled for each of the categories represented based on the scoring returned from the Tone Analyzer service. Each component of analysis (style, social, etc.) is grouped together to assist the viewer in understanding the breadth of emotion conveyed in text the student provided.

APPLICATION AND DISCUSSION

The two solution approaches outlined above work together to assess the sentiment of a large population of learners.

The usefulness of these approaches was measured based on personal experiences in two different classes offered in the aforementioned online Master of Science in CS program. This proof sets the system up for more rigorous assessment with other classes, especially those that see a greater number of instances of negative sentiment or tone.

The following sections explore the results of applying these approaches in live class forums, providing both a top-down population view of sentiment as well as a discrete view of an individual's emotional expression.

Population Sentiment

The first class investigated is visualized in Figure 1. The posts were evaluated as of December 10, 2016, plotted on a weekly basis, and resulted in the above graph. The median sentiment of each week is represented by a red line in the boxplot and is demonstrably high over the entire course.

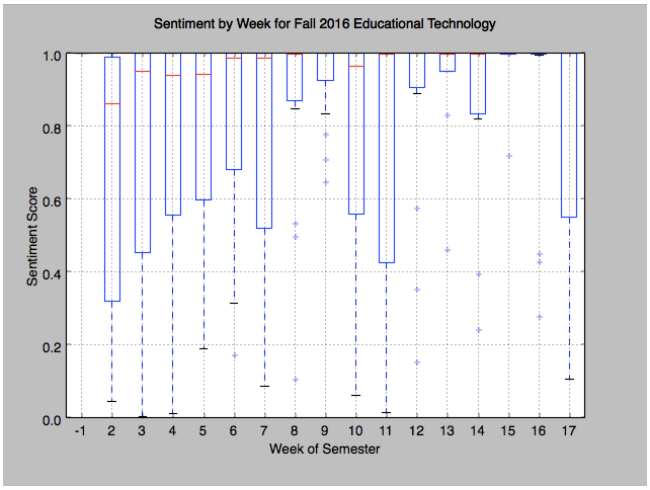

Figure 1. Sentiment analysis by week of the Fall 2016 online section of Educational Technology.

The spread of values varies widely in the first few weeks, but a notable trend of tightening to the generally positive range is notable. This trend is recognizable to students and instructors alike as relatable to the natural ambiguity of assignment expectations and general orientation to classroom procedures that normally take place in the first several weeks of a semester. The tightening of the range represents a trend toward positive sentiment, and generally stays that way until weeks 10-11 when particular milestones (and thus ambiguity) occurred in the class.

Although evaluation of the accuracy of this analysis is difficult due to the absence of another objective measurement of sentiment over time, the visualization does confirm the prediction of the course teaching team: they

anticipated greater instances of negative sentiment immediately following assignment deadlines (and grade return deadlines, which in this class typically follow within the same calendar week). The emergence of this trend in the visualization suggests both that the sentiment analysis is capturing some real trends, and thus that it may be useful in capturing trends we would not otherwise see, especially if applied at a smaller time scale.

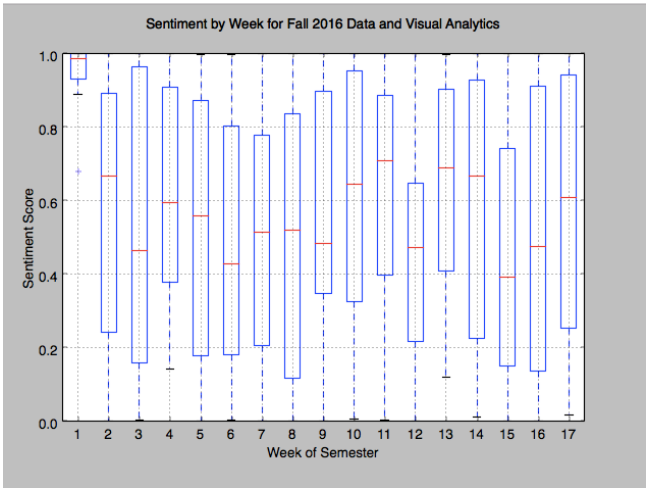

Figure 2. Sentiment analysis by week of the Fall 2016 online section of Data and Visual Analytics.

The second class investigated is visualized in Figure 2. This course was first offered in this format during the Fall 2016 semester. This particular scenario—that is, a class offered for the first time by an instructor without prior experience in teaching online—is one of the key drivers for this work, and this tool would assist the instructional team in measuring the efficacy of their week by week methods.

In this case, very positive sentiment is exhibited in week 1. This was correlated to the initial onboarding of the class, where students introduced themselves and discussed how excited they were about the material that was about to be consumed. From that point forward, the median sentiment of the class dropped, with swings up and down across all of the subsequent weeks.

Of particular interest is the wide spread of sentiment represented by the size of the boxes. Further investigation is required to determine the attribution of this spread, as well as whether it would be possible to account for this in the model. One possible reason is based on attitudes toward learning. Some individuals represent displeasure in having to figure out very tough problems, while others enjoy (or have less distaste for) the discovery process. These observations confirm the anecdotal observations of the professors, who note a significant difference between students framing the class as a learning endeavor and those framing the class as a credit-obtaining endeavor.

Another point of interest is the correlation between assignment due dates and decreases in median sentiment. Challenging assignments in this class were due in weeks 6, 9, 12 and 15. These were the low points in sentiment in the class.

It is not determinable from this visualization alone as to whether interventions were necessary in this class. A significant number of posts had tone that fell well below the 0.5 sentiment mark, indicating a negative sentiment. Grade data as well as feedback from the teaching staff would need to be incorporated to determine what, if any interventions are necessary.

One of the challenges experienced in accurately scoring this class was the number of 'anonymous' posts – those where a particular student posted without personal attribution. These posts were attributed to a single 'anonymous' student – whereas there were an unknown number of students that posted anonymously. Although this does not greatly interfere with aggregated data, it does mask whether dissatisfaction is widespread or attributed to a small number of disgruntled students. As noted above, if this tool were run from an instructor account, these identities would be discerned.

Individual Tone

In order to understand the emotions at an individual level, the generated Tone Analyzer results are processed and displayed in easy-to-understand bar graphs (see Figure 3 below). The emotional tone segments break down into anger, disgust, fear, joy, and sadness. There is also a language style summary consisting of analytical, confident, and tentative categories, as well as a social summary category consisting of five personality markers: openness, consciousness, extraversion, agreeableness, and emotional range.

Figure 3. Individual learner analytics

This personalized view will allow the instructor to determine the proper intervention for a learner that is expressing a negative sentiment in their virtual interactions. Based on the categorization of tones present in a forum posting, the instructor may choose to bolster confidence in a learner expressing fear, redirect a student expressing anger, or bring hope to someone expressing sadness.

The persisted individual results can also be presented in a dashboard view for the instructor. Whereas the individual profile presented in Figure 3 presents a detailed profile of an individual's tone, the view in Figure 4 aggregates the data and can be leveraged to highlight individuals for potential interventions that cross thresholds of tone or style.

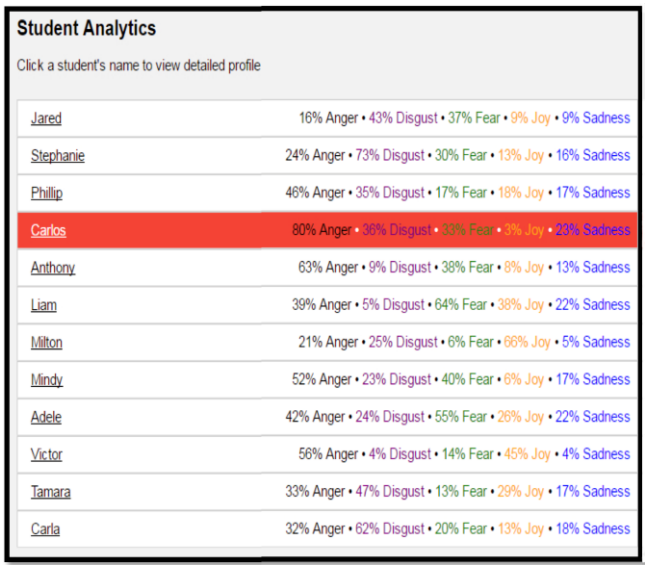

Figure 4. Aggregate student analytics highlighting a potential candidate for intervention

It is worth noting that the accuracy of these metrics in assessing a student's mood is a subject of debate. As such, interventions should be designed with caution. In our classes, we note that there is low risk associated with a misplaced personal positive message from an instructor to a student perceived to be struggling.

CONCLUSION AND FUTURE WORK

Gauging the level of understanding of a large body of learners has become more challenging in the increasingly popular massive online learning communities. Identifying and delivering the appropriate individual intervention is exacerbated due to the text-only nature of instructor-student interactions. This paper has presented approaches and findings in the area of sentiment and tone analysis for online learning forums that will help instructors gauge their efficacy and target interventions at both the population and individual levels. These approaches were applied in live online sections of collegiate courses, with results matching the anecdotal perceptions of the instructors.

This work presents multiple extension points for further refinement. Opportunities exist to tune the scoring of text based on several factors including the subject domain (e.g. English Literature has a different nomenclature than Computer Science), introduction of proportional weighting of responses vs original posts, and either weighting or removing instructor posts. Additionally, linking the outputs of these analyses to case management systems where event triggers and interventions applied could be tracked would provide further data to both refine the thresholds triggering an intervention, as well as enable review and improvement of the intervention approach applied. Finally, the accuracy of scoring provided by the Microsoft and IBM solutions were not in scope for exploration in this paper. This is an area we feel there are additional opportunities for improving the quality of our analysis at scale.

ACKNOWLEDGEMENTS

We thank IBM and Microsoft for providing educational access to their commercial services. We also thank Hamza Faran for his work exposing the Piazza API in Python.

REFERENCES

1. DyKnow. (n.d.). Retrieved February 21, 2016, from http://www.dyknow.com/
2. Meet Alma: The world's first truly integrated SIS LMS. (n.d.). Retrieved February 22, 2016, from http://www.getalma.com/
3. Offir, B., Bezalel, R., & Barth, I. (2007). Introverts, extroverts, and achievement in a distance learning environment. *The American Journal of Distance Education, 21*(1), 3-19.
4. Anderson, John R., et al. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence* 42.1: 7-49.
5. Vanlehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., ... & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. International Journal of Artificial Intelligence in Education, 15(3), 147-204.
6. Goel, Ashok K., and Lalith Polepeddi. (2016). *Jill Watson: A Virtual Teaching Assistant for Online Education*. Georgia Institute of Technology.
7. Wen, M., Yang, D., & Rose, C. (2014, July). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In Educational data mining 2014.
8. Joyner, D.A. (2017, April). Scaling Expert Feedback: Two Case Studies. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. 71-80. ACM.
9. Joyner, D. A., Goel, A., & Isbell, C. (2016). The Unexpected Pedagogical Benefits of Making Higher Education Accessible. In *Proceedings of the Third Annual ACM Conference on Learning at Scale.* Edinburgh, Scotland.
10. O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 122-129.