Incorporating student opinion into opinion mining: A student-sourced sentiment analysis classifier

Book Section

Chapter 13

Incorporating student opinion into opinion mining

A student-sourced sentiment analysis classifier

Garron Hillaire, Bart Rienties, Mark Fenton-O'Creevy, Zdenek Zdrahal and Dirk Tempelaar

13.1 Introduction

> Yeah, well, you know, that's just, like, your opinion, man.
> Jeffrey "the Dude"
>
> Lebowski talking to Jesus Quintana in *The Big Lebowski*

In Open World Learning we focus on free online learning resources and explore how to support more students to benefit from these resources. To better understand the student experience there is a need to focus on emotional measures as emotions are considered integral to the learning process (Immordino-Yang & Damasio, 2007). As one of the ubiquitous modes of communication in online learning is text, we focus on sentiment analysis (SA), which is an affective computing measure that can interpret emotions in text by classifying if text is positive, negative, neutral, or mixed (both positive and negative). In Chapter 13, we focus on how student perceptions relate to and are affected by predictions about their emotional expression in text. By exploring how student opinions relate to and are potentially influenced by SA we explore the validity and utility of SA.

When SA classifiers are built the process starts with establishing the correct labels for text, referred to as ground truth. Establishing ground truth relies on human judgements. We ironically reference the quote from *The Big Lebowski* "that's just, like, your opinion, man." to light-heartedly call into question how truth is established. In our opinion, there is justifiable reason to anchor truth for SA to the opinions of students. While there is inherent subjectivity when anchoring truth to student opinions, SA commonly purports to measure how the opinion of the author of the text elicits a reaction from the intended reader of the text (Balahur & Steinberger, 2009).

Early SA work used text from product reviews and as well as star ratings (e.g., 1-star reviews considered negative; 5-star reviews considered positive) (Liu, 2010). Effectively, the labels for text were inferred by a star rating that came from the author of the text. While it is commonly held that SA technologies work best when used on text for similar contexts as to the context where data used to train the classifier originated, it is not commonly held that the labels for the text should also come from people from the context. For example, a very common practice in

DOI: 10.4324/9781003177098-15

SA research is to have researchers rate text using trained raters on established coding schemes (Thelwall, 2013), or use anonymous raters from crowd-sourcing platforms such as Amazon's Mechanical Turk (MTurk) (Mohammad & Turney, 2013) where the wisdom of the crowd typically replaces the training of raters.

The choice of who is best situated to rate the valence of text is directly related to the definition of emotion – which is still a highly debated concept. If emotion is universal, then there are attributes we can identify as characteristics of emotions. For example, when someone is happy they might say "I am going to Disneyland!". This phrase comes from an advertising campaign in the late 1980s where the most valuable player from the super bowl would shout this phrase after winning the game. While the Disney corporation likely wants this phrase to be universal there are people in the world who may have never heard the phrase or even know about Disneyland (a popular theme park). In contrast to the universal perspective, the Constructed Theory of Emotion (CTE) would suggest that only those with familiarity of the social context would understand the emotional expression (Feldman Barrett, 2018). In Chapter 13, we test CTE by considering the perspectives of the social group of students from the classroom and contrast this with a social group of anonymous raters.

In conjunction with the debate on the definition of emotion there is a further multi-level debate on how emotion is best measured. The first emotional measurement debate is between discrete measurement of emotions such as happiness and anger in contrast with the perspective that emotion is best measured in dimensional terms such as the dimension of valence from positive to negative (Feldman Barrett & Russell, 1998). In Chapter 13, we focus on the dimensional measurement of valence. We adopt four possible categories of valence: positive, negative, neutral, and mixed. Specifically, we explore if the social consensus used rate text should be from a contextual group (the students) or an anonymous out-of-context group (Mechanical Turk). Finally, we examine the accuracy of our proposed classifier by showing the predictions of the classifier to the students during interviews. We shared predictions with students to see if students viewed the predictions as accurate and useful. To situate this work in the broader context of SA research, we first review related work.

13.2 Related work

It is important to note that not all emotional measures share a common aim and not all measurement adoption explicitly states the assumptions of the measures (Weidman, Steckler, & Tracy, 2016). This makes comparison between work difficult as SA studies consider accuracy of those measures based on completely different definitions of truth (e.g., universal vs. social). Two key assumptions of measurement adoption are related to debates both on what emotion is and how it should be measured. To illustrate these debates, we review three theories on emotion, three approaches to measure emotion in text, and finally classify 15 existing studies in the context of learning within this taxonomy of emotional theory and measurement based on how they evaluate accuracy of the measures.

13.2.1 Three perspectives on emotion

Basic Emotion Theory (BET) considers some emotional experiences to be so fundamental that they are described as universal. For example, people may have a common experience of emotion when it comes to some specific emotional responses, such as anger and happiness. Typically, researchers who adopt the BET position on emotion focus on five to thirteen emotions that are considered fundamental to the human experience: Happiness, Enjoyment, Sadness, Fear, Anger, Disgust, Interest, Contempt, Rage, Love, Lust, Care, and Surprise (Tracy & Randles, 2011). One limitation for BET is that there is minimal relevance for basic emotions in learning activities that span 30 minutes to 2 hours (Calvo & D'Mello, 2010).

CTE is a perspective that suggests that the manner by which emotion is interpreted is through the influence of social factors. An example of how social theorists interpret emotion is illustrated in the book *How Emotions Are Made* by Lisa Feldman Barrett when she used a picture of Serena Williams. The photo was taken immediately after Serena beat her sister, Venus Williams, in the 2008 U.S. Open. The picture Barrett presents is a cropped image of Serena's facial expression and Barrett suggests that looking at the facial expression in isolation of context might be categorised as an expression of terror when using a basic perspective on emotion. However, by taking context into consideration we should instead interpret the image to mean something closer to exultation (Feldman Barrett, 2018, p. 42). Barrett argued that emotion consists of making meaning, prescribing action, regulating the body, emotion communication, and social influence. Two of the components, emotion communication and social influence, are considered social as they are aspects of emotion that cannot be done in isolation.

Situated Affectivity Theory (SAT) considers the goal as the focal point for interpreting all of the components of emotion (Wilutzky, 2015). With this goal orientation, a manipulation between an individual and their environment is the basis for stimulation for emotion. The physiological response represents a physical experience that resonates with the interaction with the environment. Emotional communication is thought to be used by people to achieve goals.

13.2.2 Three perspectives on valence

Valence is a dimensional perspective on organising emotions commonly considering positive and negative. There are three competing perspectives on how valence should be organised. The bipolar model considers positive and negative to be the opposite ends of the same spectrum (Russell & Carroll, 1999). For example, the emotion happy can be placed on the positive end of the spectrum and the emotion sad can be placed on the negative end of the spectrum. The bi-variate model suggests a co-activation where emotions can be categorised as simultaneously activating positive and negative (Watson, Wiese, Vaidya, & Tellegen, 1999). In the bi-variate model, there are two variables (one for positive and one for negative). The evaluative space model (ESM) suggests that emotions are both bipolar and bi-variate (Cacioppo, Gardner, & Berntson, 1999). Effectively, ESM argues that valence should

be thought of as a plane. We can consider the *Y*-axis of the plane to range from neutral to negative and the *X*-axis of the plane to range from neutral to positive. Points on the *X*- and *Y*-axes represent bipolar categories of emotion. Chapter 13 adopts ESM by considering the four valence categories of positive, negative, neutral, and mixed.

13.2.3 Reviewing sentiment analysis in education

SA research shows promise regarding investigations into the complex role of emotion in learning. Given the potential for SA in educational research, it is essential to consider the validity and reliability of SA. To begin considering validity and reliability it is essential to precisely clarify what SA purports to measure. As it is common for researchers to use emotional measures without explicitly stating their theoretical perspective on emotion (Weidman et al., 2016), first we reviewed the 15 identified SA in studies in the context of learning and classified how accuracy was evaluated in relation to the three emotion theories reviewed. The results are reported in Table 13.1.

We classified five studies that used methods that are best described as BET. In these studies, the researchers believed that they could identify what was accurate as this indicated that emotion expression was identifiable by someone other than students in the context of learning. For example, BET studies included an examination of teacher evaluations where researchers read the teacher evaluations, and coded the "actual" sentiments based on the perspective of the researcher reporting an overall accuracy of 86.28% (Rajput, Haider, & Ghani, 2016).

Table 13.1 Interpretation of emotion theory of sentiment analysis studies in education

Studies	*#*	*BET*	*SAT*	*CTE*	*None*
Ortigosa et al. (2014); Troussas, Virvou, Espinosa, Llaguno, & Caro (2013)	2	☑	–	–	–
Chaplot et al. (2015); Crossley, Paquette, et al. (2016); Wen et al. (2014); Wyner, Shaw, Kim, Li, & Kim (2008)	4	–	☑	–	–
Calvo & Kim (2010)	1	–	–	☑	–
Munezero et al. (2013); Jagtap & Dhotre (2014); Shapiro et al. (2017); Chang, Maheswaran, Kim, & Zhu (2013); Kagklis, Karatrantou, Tantoula, Panagiotakopoulos, & Verykios (2015)	5	–	–	–	☑
Rajput et al. (2016); Santos et al. (2013)	2	☑	–	☑	–
Hillaire, Rienties, et al. (2018)	1	☑	☑	–	–
Total	15	5/15	5/15	3/15	5/15

Note: BET = Basic Emotion Theory; SAT = Situated Affect Theory; CTE = Constructed Theory of Emotion; None = No Evaluation of Accuracy.

Three studies used evaluation of accuracy methods best aligned with CTE. For example, one study compared course ratings on a Likert scale to determine which ratings were positive and inferred comments in the review were positive (Calvo & Kim, 2010). None of these studies directly asked participants their opinion about the text analysed by SA (sometimes referred to as opinion mining) which is a clear gap in educational research.

We classified five studies all using discussion forums as reflecting SAT when the focus was on correlations between SA and outcomes (e.g., student retention), because this placed an emphasis on the relationship between emotion expression and goal orientation. For example, when predicting student attrition in an online course SA was used in conjunction with other measures to generate two predictive algorithms which reported a Kappa statistic of 0.403 and 0.432 when predicting attrition (Chaplot, Rhim, & Kim, 2015). Next, we evaluated the same 15 studies to examine which valence categories were measured considering the four valence categories identified in our review on valence theory (see Table 13.2).

When considering valence categories measured when applying SA to the context of learning, there appears to be an emphasis in the existing literature on measuring positive and negative valence. Of the 15 studies reviewed, all of the studies measured both positive and negative valence as indicated in Table 13.2. About half of the studies, seven out of 15, measured the category of neutral, and only two out

Table 13.2 Valence categories of sentiment analysis studies in education

Studies	#	*Positive*	*Negative*	*Neutral*	*Mixed*
Munezero, Mozgovoy, Montero, & Sutinen (2013); Jagtap & Dhotre (2014); Troussas, Virvou, Espinosa, Llaguno, & Caro (2013); Crossley, Paquette, Dascalu, McNamara, & Baker (2016); Wen, Yang, & Rosé (2014); Wyner, Shaw, Kim, Li, & Kim (2008); Chang, Maheswaran, Kim, & Zhu (2013)	7	☑	☑	–	–
Calvo & Kim (2010); Ortigosa et al. (2014); Chaplot et al. (2015); Hillaire, Rienties, et al. (2018); Shapiro et al. (2017); Kagklis, Karatrantou, Tantoula, Panagiotakopoulos, & Verykios (2015)	6	☑	☑	☑	–
Rajput et al. (2016)	1	☑	☑	–	☑
Santos et al. (2013)	1	☑	☑	☑	☑
Total	15	15/15	15/15	7/15	2/15

of 15 studies measured a category of mixed emotion. One study (Santos et al., 2013) measured all four categories of positive, negative, neutral, and mixed. However, they referred to mixed as ambivalence - which they defined as both positive and negative. In the second study considered mixed expression (Rajput et al., 2016), the authors used neutral and mixed interchangeably when describing the results but reported statistics for the category of mixed expression.

13.3 Student-sourcing, crowd-sourcing ground truth for a classifier with students

We explore centring students with *student sourcing*, using crowd-sourcing methods with students evaluating their own group discussions. In doing so we flip the assumption from the perspective that crowd ratings are by default noise to the default assumption they are accurate. When establishing ground truth more single labels are better in the condition where raters are considered reliable. Based on the shifting the assumption that student ratings are by default accurate single ratings is considered useful. A common approach is using the Expectation Maximisation (EM) algorithm (Dempster, Laird, & Rubin, 1977), which selects the best label using crowd-sourcing label data by considering both the prevalence of each valence category and the categorical accuracy of each rater. Effectively the uniqueness of student opinions is favoured by this approach because the EM algorithm adopts single ratings as ground truth. Where multiple ratings occur, the EM algorithm selects a best fit as a proxy for what social consensus might evolve between students. As the approach is novel we evaluate the work using both standard approaches to reliability, and benchmark this specialised classifier with general crowd-sourcing approaches.

Typically, with crowd sourcing a large number of people are recruited to categorise text by providing labels frequently generating five labels for each item being categorised. Providing both Fleiss' Kappa and Krippendorff's alpha are suggested for crowd-sourced labels in social computing (Salminen, Al-Merekhi, Dey, & Jansen, 2018) because the expectation is that agreement is usually low with crowd-sourcing methods. For example, Krippendorf's alpha scores around 0.10 were frequently found when evaluating crowd-sourcing methods (Alonso, Marshall, & Najork, 2013). We use crowd sourcing as one of the benchmarks for student's sourcing where students provide labels instead of anonymous MTurk raters disconnected from the classroom context and then validated the outcome of training on MTurk ratings by predicting student labels that close the loop by validating with student labels. This approach is contrasted with centring students where using artificial intelligence approaches we instead train a classifier based on student labels and then use the student-sourced classifier to predict student labels (see Figure 13.1)

Finally, we conducted interviews with respective students involved in the experiments to further lean into student perspectives. Therefore, to investigate the assumption that we can accept student opinions as correct for opinion mining two research questions need to be assessed:

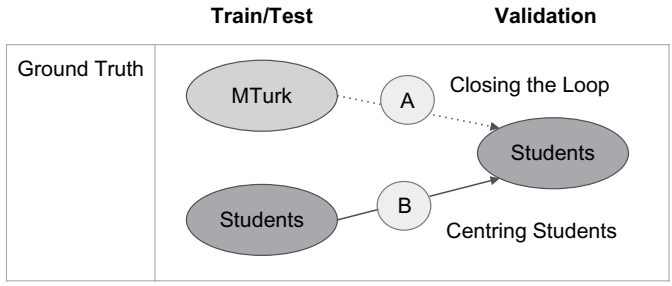

Figure 13.1 Comparing closing the loop with centring students with ground truth.

1 To what extent are student-sourced examples reliable?
2 To what extent can we use student-sourced examples to train an SA classifier?

13.4 Methods

Chapter 13 is based on two extensive studies undertaken with two separate cohorts of students at a university in the Netherlands as part of the thesis of the first author (Hillaire, 2021). While the university recruits international students, courses are taught in English. Cohort 1 included 767 freshmen in a statistics course in Fall 2016 who (1) worked on an online group assignment where students chat with one another and (2) reviewed their discussions and provided examples of messages for valence categories. There were 304 females and 463 males. The population was international, including 191 domestic (Dutch), 529 European Students, and 47 non-European students. Mechanical Turk was used to generate five labels for messages selected by Cohort 1 Students. Cohort 2 included 484 freshmen in the same statistics course in Fall 2017 who (1) completed an online group assignment, (2) provided examples of messages for valence categories (see Figure 13.2).

We generated Data Set 1 with the EM algorithm which selected the ground truth label for each text message based on the example text and labels provided by Cohort 1 Students. We generated Data Set 2 with the EM algorithm which selected the ground truth label for text message from text examples provided by Cohort 1 Students and labels for the text provided by Mechanical Turk workers. We generated Data Set 3 with the EM algorithm which selected the ground truth label for each text messages based on the example text and labels provided by Cohort 2 Students (see Figure 13.2).

Finally, we used Data Set 1 to train Classifier 1 (a logistic regression classifier). We used Data Set 2 to train Classifier 2 (a logistic regression classifier). Both Classifier 1 and Classifier 2 categorised text messages as positive, negative, neutral, or mixed. Finally, we used Classifier 1 and Classifier 2 to predict labels for Data Set 3. To ground the comparison between Classifier 1 and Classifier 2 we compared them to general SA classifiers used on Data Set 3 (see Figure 13.2).

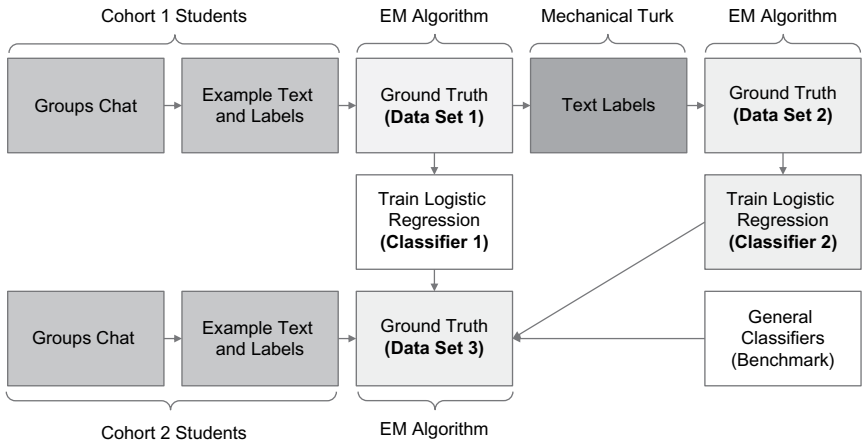

Figure 13.2 Research design generating three data sets and two classifiers.

Finally, we interviewed six students from Cohort 2 to evaluate the accuracy of Classifier 1 as well as the utility of having access to SA predictions on their own text messages.

13.5 Procedure

In Cohort 1, students (n = 767) were assigned randomly to groups of five (M = 4.73 SD = 0.84) in a laboratory setting, whereby each student had a desktop computer, and all written communication was online as part of a regularly occurring lab session for their course. Previous research on this task reported that overall students enjoyed working together in groups (Mittelmeier, Rienties, Tempelaar, Hillaire, & Whitelock, 2018). The group work activity for Cohort 2 was the same as for Cohort 1 with a small change to the warmup exercise. The post-activity was changed in that participants no longer provided examples of ambiguous messages and the final modification was a series of interviews conducted with six students to examine the trustworthiness of the algorithm's predictions.

In the post-activity Cohort 1 participants were first given a set of instructions to provide 1–3 examples of positive, negative, neutral, mixed, and ambiguous messages (for Cohort 2 removed the Ambiguous valence category). For Cohort 2, the interview consisted of three parts. Part 1 asked students to review a subset of messages from their group chat and identify if the message was positive, negative, neutral, or mixed. Part 2 asked participants to compare their rating with the prediction from the student-sourced classifier in conjunction with the text features the algorithm used to predict the valence. If the prediction was different than the student label provided in Part 1 the student was asked if the algorithm prediction changed their mind. Finally, at the end of the interview participants were asked if the predictions were useful.

13.5.1 Analysis

To answer RQ1 we computed inter-rater agreement for Data Sets 1, 2, and 3 and compared the results to benchmarks of agreement for crowd sourcing in social science. Low agreement in crowd ratings does not mean the opinions of labels are incorrect; it may simply indicate they have different opinions (Salminen et al., 2018).

To answer RQ2 we generated Classifiers 1 and 2 (logistic regression classifiers) based on Data Sets 1 and 2, respectively, and compared the accuracy of Classifiers 1 and 2, with General Benchmarks when predicting valence labels for Data Set 3. We also benchmarked the accuracy of Classifiers 1 and 2 with general measures. Finally, we interviewed students from Cohort 2 to evaluate the accuracy and utility of predictions from Classifier 1 used to interpret their text data.

13.5.2 Results

To answer RQ1, we first established three datasets and then computed agreement statistics. Data Set 1 was generated by 767 students providing examples for positive, negative, neutral, mixed and ambiguous, resulting in 2512 records with 1979 distinct messages. Data Set 2 was generated by using the EM algorithm to select the ground truth label for Data Set 1 which resulted in 1778 messages categorised as positive, negative, neutral, and mixed (we excluded the 201 messages categorised as ambiguous). We next used Mechanical Turk where five raters classified the 1778 messages as positive, negative, neutral, and mixed. Data Set 3 was generated by 484 students providing examples for positive, negative, neutral, mixed. This resulted in 986 records with 755 distinct messages. After generating the three datasets we computed agreement statistics resulting in Krippendorff's alpha scores of 0.44, 0.25, and 0.42 for Data Sets 1, 2, and 3, respectively. Datasets 1 and 2 generated a range of between one and five ratings per unique message so we further computed and report Fleiss' Kappa scores for agreement statistics based on the number of ratings. For Data set 2 we had five raters for every unique message and report Fleiss' Kappa for completeness (see Table 13.3).

In Table 13.3, we observe that both Data Set 1 (alpha=0.44) and Data Set 3 (alpha=0.42) had similar Krippendorff's alpha scores indicating (1) students had moderate agreement with one another on the valence labels from their own chat data; and (2) student agreement was above Mechanical Turk raters (alpha=0.25) as well as below the average Krippendorff's alpha score of 0.60 found in crowdsourcing studies in social science. These results show promise that crowd sourcing with students has the potential to do better than using services such as Mechanical Turk, but also indicates that agreement is below the average indicating room for improvement.

To answer RQ2, we first established a series of benchmarks using general SA technologies making predictions about Data Set 3 with labels from Cohort 2 students. The f-measures for the best benchmarks was VADER with an f-score of 0.43. Next, we trained Classifier 1 using Data Set 1 with labels from Cohort 1

Table 13.3 Agreement statistics for three data sets

	Raters	*1 rater*	*2 raters*	*3 raters*	*4 raters*	*5 raters*	*Krippendorff's alpha*
Data Set 1 (Fleiss' Kappa)	Cohort 1 Students	1586 (–)	330 (0.42)	56 (0.52)	6 (0.30)	1 (–)	0.44
Data Set 2 (Fleiss' Kappa)	Mechanical Turk					1778 (0.25)	0.25
Data Set 3 (Fleiss' Kappa)	Cohort 2 Students	577 (–)	139 (0.41)	30 (0.50)	4 (0.36)	5 (−0.15)	0.42

students and trained Classifier 2 using Data Set 2. The best cross-validation f-scores for Classifier 1 was 0.475 and the subsequent validation F-score was 0.462. The best cross-validation f-scores for Classifier 2 was 0.550 and the subsequent validation F-score is 0.456. When comparing these results Classifier 2 had a higher cross-validation score and Classifier 1 had a higher validation score (see Table 13.4). This means that when we tested the two classifiers on Data Set 3 (with labels from Cohort 2 students) that Classifier 1 trained on Cohort 1 student labels was more accurate than Classifier 2 trained on Mechanical Turk labels.

Finally, we interviewed six students from Cohort 2 about the predictions from classifier 1. Across the six students interviewed they reviewed 113 messages of which they agreed with the algorithm 36 times, and disagreed 77 times. For the 77 disagreements, they changed their mind to agree 21 times (27% or 21/77) after seeing the algorithm's predictions (see Table 13.5). When considering the initial agreement (36 times) and when they changed their mind (21 times) the students considered the prediction accurate 50% of the time (57/113).

Participants changed their mind to agree with the algorithm one to three times with the exception of one student who changed their mind eleven times. Students who found the algorithm to be useful had final agreement that ranged from 42% to 67% (initial agreements 5-9 messages; final agreements 8–20 messages) with a). The one student who did not find the algorithm to be useful, Student-6, only initially agreed with the algorithm once and changed their mind to agree with it two times for a total of three agreements out of 12 messages (25%). While sample

Table 13.4 Agreement statistics for three data sets

	Train/test data	*Validation data*	*Cross validation*	*Validation*
Classifier 1	Data Set 1 (Cohort 1 Students)	Data Set 3 (Cohort 2 Students)	0.475	0.462
Classifier 2	Data Set 2 (Mechanical Turk)	Data Set 3 (Cohort 2 Students)	0.550	0.456

Table 13.5 Agreement, disagreement, final agreement, and usefulness of SSAC

Participant	*Agree*	*Disagree (change)*	*Final agreement %*	*Useful*
Student 1	7	10 (1)	47%	Yes
Student 2	9	21 (11)	67%	Yes
Student 3	8	9 (2)	59%	Yes
Student 4	5	13 (3)	44%	Yes
Student 5	6	13 (2)	42%	Yes
Student 6	1	11 (2)	25%	No

size from interviews is small, it is noteworthy that Student-6 who did not find it useful had a final agreement of 25%, which was the same as the unweighted chance levels of accuracy for predicting four categories, while all of the students who found it useful had above chance levels of agreement. This result suggests above chance levels of accuracy is necessary for students to find the classifier useful. Five out of six students interviewed said that the algorithm was useful. When describing the usefulness of the algorithm, participants described benefits including: (1) better understanding their own communication (e.g., "I started thinking more about what I said"), (2) better understanding communication of other students (e.g. "I started analysing the way others said it"), and (3) seeing an alternate interpretation that changed their mind which they described as learning from the algorithm.

13.6 Discussion and moving forwards

Chapter 13 illustrated how a student-sourced SA could build a better understanding of the online student experience and emotions in particular. What is novel about our findings is that we demonstrated that (1) student labels had a higher level of inter-rater agreement than Mechanical Turk labels, (2) Mechanical Turk labels generated a higher cross-validation score than student labels, and (3) student labels trained a classifier with higher accuracy than the classifier trained using Mechanical Turk labels. A potential explanation for this result is that the consensus established by Mechanical Turk workers was simply divergent from the consensus of students. We could reframe this to say what Mechanical Turk workers consider to be the true labels for text has higher consistency, but their idea of truth is different from students. From the perspective of the CTE the consensus established by members of the social context is the very definition of emotion. Interpreting these results from a CTE perspective suggests there is potential benefit in having raters that come from the context where the text was originally generated. This finding builds on the existing belief that SA classifiers are context sensitive and perform best when used in contexts similar to the context where training data for the classifier was collected by contributing evidence that context sensitivity may also include the relationship between the raters of text and the context where the text was collected.

13.6.1 Implications for practice

Practitioners that use educational technology should be cautious when they incorporate SA classifiers trained on data dissimilar to classroom data as general technologies had low performance. Practitioners should consider how to centre the lived experience of students when integrating classifiers that seek to model highly subjective topics such as SA. Not only is there reason to share SA predictions with students to anchor accuracy with student opinion, but students reflecting on SA predictions demonstrated the benefit of thinking about both what they say to their peers and what their peers say to them in terms of emotional expression. Future

work should explore supporting and evaluating student awareness of emotion expression in text.

References

Alonso, O., Marshall, C. C., & Najork, M. (2013). A human-centered framework for ensuring reliability on crowdsourced labeling tasks. In *AAAI Workshop – Technical Report, WS–13–18*, 2–3.

Balahur, A., & Steinberger, R. (2009). Rethinking sentiment analysis in the news: From theory to practice and back. In *Proceedings of the '1st workshop on opinion mining and sentiment analysis*, 1–12. Retrieved from http://langtech.jrc.it/Documents/09_WOMSA-WS-Sevilla_Sentiment-Def_printed.pdf

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, *76*(5), 839–855. doi:10.1037/0022-3514.76.5.839

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37. doi:10.1109/T-AFFC.2010.1

Calvo, R. A., & Kim, S. M. (2010). Sentiment analysis in student experiences of learning. In *Third International Conference on Educational Data Mining (EDM2010)*, 111–120.

Chang, Y. H., Maheswaran, R., Kim, J., & Zhu, L. (2013). Analysis of emotion and engagement in a STEM alternate reality game. *CEUR Workshop Proceedings*, *1009*, 29–32. doi:10.1007/978-3-642-39112-5-82

Chaplot, D. S., Rhim, E., & Kim, J. (2015). Predicting student attrition in MOOCs using sentiment analysis and neural networks. In *Workshops at the 17th International Conference on Artificial Intelligence in Education, AIED-WS 2015*, *1432*, 7–12. doi:10.1016/j.evalprogplan.2016.04.006

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge – LAK '16*, 6–14. doi:10.1145/2883851.2883931

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Feldman Barrett, L. (2018). *How emotions are made*. UK: Pan Macmillan.

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*(4), 967–984.

Hillaire, G. (2021). *Understanding emotions in online learning: Using emotional design and emotional measurement to unpack complex emotions during collaborative learning*. Milton Keynes, UK: The Open University.

Hillaire, G., Rienties, B., & Goldowsky, B. (2018). Struggling readers smiling on the inside and getting correct answers. In *AERA 2018 Annual Meeting, Symposium Session: Perspectives on Emotion and Engagement Among Struggling Adolescent Readers: Findings from an Online Literacy Intervention*. New York.

Immordino-Yang, M. H., & Damasio, A. (2007). We feel, therefore we learn: The relevance of affective and social neuroscience to education. *Mind, Brain, and Education*, *1*(1), 3–10. doi:10.1111/j.1751-228X.2007.00004.x

Jagtap, B., & Dhotre, V. (2014). SVM and HMM based hybrid approach of sentiment analysis for teacher feedback assessment. *Ijettcs. Org*, *3*(3), 229–232. Retrieved from http://ijettcs.org/Volume3Issue3/IJETTCS-2014-06-25-132.pdf

Kagklis, V., Karatrantou, A., Tantoula, M., Panagiotakopoulos, C. T., & Verykios, V. S. (2015). A learning analytics methodology for detecting sentiment in student fora: a case study in distance education. *European Journal of Open, Distance and E-Learning*, *18*(2). https://doi.org/10.1515/eurodl-2015-0014

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 1, 1–38. doi:10.1145/1772690.1772756

Mittelmeier, J., Rienties, B., Tempelaar, D., Hillaire, G., & Whitelock, D. (2018). The influence of internationalised versus local content on online intercultural collaboration in groups: A randomised control trial study in a statistics course. *Computers and Education*, *118*, 82–95. doi:10.1016/j.compedu.2017.11.003

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*. doi:10.1111/j.1467-8640.2012.00460.x

Munezero, M., Mozgovoy, M., Montero, C. S., & Sutinen, E. (2013). Exploiting sentiment analysis to track emotions in students' learning diaries. In *Koli Calling*. doi:10.1145/2526984

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, *31*, 527–541. https://doi.org/10.1016/j.chb.2013.05.024

Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-based sentiment analysis of teachers' evaluation. *Applied Computational Intelligence and Soft Computing*, *2016*, 1–12. doi:10.1155/2016/2385429

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin*, *125*(1), 3–30. doi:10.1037/0033-2909.125.1.3

Salminen, J. O., Al-Merekhi, H. A., Dey, P., & Jansen, B. J. (2018). Inter-rater agreement for social computing studies. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (Vol. 49, pp. 80–87). IEEE. doi:10.1109/SNAMS.2018.8554744

Santos, O. C., Salmeron-Majadas, S., & Boticario, J. G. (2013). Emotions detection from math exercises by combining several data sources. In *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 7926 LNAI, pp. 742–745). https://doi.org/10.1007/978-3-642-39112-5-102

Thelwall, M. (2013). *Sentiment strength detection for the social web*. Retrieved from http://sentistrength.wlv.ac.uk/documentation/course.html

Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, *3*(4), 397–405. doi:10.1177/1754073911410747

Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013). Sentiment analysis of facebook statuses using naive bayes classifier for language learning. In *IISA 2013 - 4th International Conference on Information, Intelligence, Systems and Applications*, 198–205. doi.10.1109/IISA.2013.6623713

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two activation systems of affect: Structural findings, evolutionary considerations, and psychbiological evidence. *Journal of Personality and Social Psychology*, *76*(5), 820–838. https://doi.org/0022-3514/99/$3.00

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2016). The Jingle and Jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. doi:10.1037/emo0000226

Wen, M., Yang, D., & Rosé, C. P. (2014). Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the 8th International Conference on Weblogs and Social Media ICWSM 2014*, *525534*, http://www.scopus.com/inward/record.url?eid=2-s2.0-84909951147&partnerID=40&md5=80f121bcfc587505feaee3a3d6675c59, 525–534. https://doi.org/doi:10.1016/j.bspc.2014.01.007

Wilutzky, W. (2015). Emotions as pragmatic and epistemic actions. *Frontiers in Psychology*, *6*, 1–10. doi:10.3389/fpsyg.2015.01593

Wyner, S., Shaw, E., Kim, T., Li, J., & Kim, J. (2008). Sentiment analysis of a student Q & A board for computer science. *The 9th KOCSEA Technical Symposium*. Vienna, VA.