Sentiment Analysis to Improve the Usability of an Online Practice Space

By Natalie Mionis

B.S Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2018

Submitted to the
Department of Electrical Engineering and Computer Science
In Partial Fulfilment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

June 2019

Author: _____
Department of Electrical Engineering and Computer Science
May 24, 2019

Certified by: _____
Justin Reich, Director of the MIT Teaching Systems Lab, Thesis Supervisor
May 24, 2019

Accepted by: _____
Katrina LaCurts, Chair, Master of Engineering Thesis Committee

Sentiment Analysis to Improve the Usability of an Online Practice Space

By Natalie Mionis

Submitted to the Department of Electrical Engineering and Computer Science
On May 24, 2019, in partial fulfilment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

I develop a new confusion measurement feature for use in Teacher Moments, an online practice space for teachers in training. Teacher Moments, developed at the Teaching Systems Lab (TSL) at MIT, is an online platform that simulates difficult classroom experiences. Teacher candidates can use this platform to gain experience handling realistic, uncomfortable situations that they may eventually face in a real classroom. Teacher educators who instruct teacher candidates can then review teacher candidates' responses and progress in Teacher Moments. The sentiment analysis feature detects confusion in responses to Teacher Moments scenarios. This feature, named the confusion measurement tool, allows teacher educators to identify teacher candidates who may be exhibiting confusion. By identifying confused teacher candidates, teacher educators can then modify their class activities or provide additional attention to any teacher candidates who may need it. I find that the confusion measurement tool is effective at identifying responses that do not exhibit confusion, but is less effective at identifying confused responses. I also find that the confusion measurement tool helps teacher educators use their time more effectively when analyzing responses to Teacher Moments scenarios.

Thesis Supervisor: Justin Reich
Title: Executive Director, MIT Teaching Systems Lab

Acknowledgments

I have had a fantastic time working in the Teaching Systems Lab, and I am so grateful for the opportunity to do my thesis there. I'd like to give a huge thank you to Justin Reich and YJ Kim for helping me find the MEng position in the lab and supporting me throughout my thesis work.

My two direct advisors, Garron Hillaire and Meredith Moore, have been immensely helpful in answering all my questions and providing direction to my work. Thank you for teaching me so much about education and sentiment studies, and for being so welcoming and available to answer my many questions. It's been a lot of fun working together!

I'd like to thank Kevin Robinson for providing invaluable technical expertise about React and web development.

I'd like to thank Meredith Thompson for her insight into teacher educator needs, and for helping set up playtests and interviews for my studies.

I'd like to thank all the anonymous teacher educators for partaking in various surveys and interviews throughout my thesis work.

I'd like to thank everyone in the Teaching Systems Lab for being so welcoming and inclusive. It's been a great place to work.

I'd like to thank my friends for teaching me some of the most valuable lessons I've learned over my five years at MIT.

Lastly, I'd like to thank my family - Mom, Dad, Scott, Erika, and Julia – for their endless support and love. All these degrees wouldn't have been possible without you!

Contents

1 Introduction **11**

2 Background **14**
2.1 Existing Teacher Education Methods 14
2.1.1 Mixed Reality Simulation: Mursion 15
2.1.2 Clinical Simulation: Dotger 16
2.2 Confusion in Education 16
2.2.1 Confusion Theory 17
2.3 Sentiment Analysis in Education 18
2.4 Research Questions 20

3 Methodology **21**
3.1 Coding Scheme Development 21
3.2 Data Gathering and Coding 22
3.2.1 Instrument 1 – Expert Labels 22
3.2.2 Instrument 2 – User Labels 24
3.3 Developing the Confusion Measurement Tool 26
3.3.1 System Architecture 26
3.3.2 Pre-Processing of Data 26
3.3.3 Sentiment Analysis Models 30
3.3.4 Text Analysis Accuracy 31
3.4 User Testing 32

4 Results **34**
4.1 Performance Metrics 34
4.1.1 Training Data Cross Validation 34
4.1.2 Face Validity with User Labeled Data 37
4.1.3 False Negative Predictions 38
4.1.4 False Positive Predictions 42
4.2 Usability Measures 45
4.2.1 Confusion Measurement Tool Use Cases 45
4.2.2 Timing Implications of Confusion Measurement Tool 46

5 Discussion **50**
5.1 Summary of Findings 50
5.2 Algorithmic Bias and Usability Recommendations 51

6 Future Work
6.1 Coding Scheme Modifications 52
6.1.1 Gathering Data for New Coding Scheme 52
6.2 Usability Improvements 53

A Survey for A/B Playtest **55**
References **56**

List of Tables and Figures

Figure 1: Teacher Moments Workflow 12
Figure 2: Existing Teacher Education Methods 14
Figure 3: Kort *et al.* Model of Learning and Affect 17
Figure 4: Class Imbalance in the Training Dataset 24
Figure 5: Collecting User Labeled Data 25
Figure 6: Class Imbalance in the Testing Dataset 25
Figure 7: N-gram Example 27
Figure 8: Model Architecture with N-grams of Different Sizes 29
Figure 9: Model Architecture with N-grams of One Size 29
Figure 10: A/B Test Results for Time Taken to Rate Confusion of One Teacher Candidate 47
Figure 11: A/B Test Results for Time Taken to Choose Most Confused Teacher Candidate 48
Table 1: Confusion Coding Schemes in Teacher Education Journals Literature Review 19
Table 2: Coding Scheme for Confusion 21
Table 3: F-Score Values for SVM Models 34
Table 4: F-Score Values for LSTM Models 35
Table 5: F-Score Values for Logistic Regression Models 35
Table 6: Precision, Recall, and F-score for 3 Models and Random Baseline 36
Table 7: SVM-UBT Results with User Labeled Data 37
Table 8: Confusion Matrix for SVM-UBT on User Labeled Data 38
Table 9: Comparison of IBM Watson Transcription vs. Human Transcription 39
Table 10: False Negative Predictions Also Coded as Not Confused by a Human Rater 39
Table 11: Sample of False Positive Predictions 42

Chapter 1
Introduction

In their pre-service coursework and instruction, teacher candidates are trained to handle a variety of difficult situations that may arise in a classroom. During difficult conversations with students or parents, teacher candidates can experience cognitive dissonance as a result of realizing inconsistencies between their values and behavior [1]. Cognitive dissonance occurs when an individual encounters beliefs or behaviors that conflict with their own, resulting in mental discomfort [2]. Cognitive dissonance may manifest as confusion in these difficult classroom situations. A sentiment analysis model that goes beyond polarity measurements and identifies and measures confusion could therefore serve as a useful tool for teacher educators to evaluate the confusion levels of their teacher candidates

In this project, I develop a confusion measurement tool for use in Teacher Moments. Teacher Moments, created by TSL, is a software platform in which teacher candidates can practice facing realistic classroom challenges. With Teacher Moments, teacher candidates can select classroom scenarios from a wide range of options. Each scenario models a challenge that a teacher may face in a classroom, and sets the scene with videos or written explanations. Then, teacher candidates interact with the scenario either through textual responses or audio recordings. The response data is then stored in the Teacher Moments dashboard for teacher educators to review in order to understand how teacher candidates performed in the simulated scenario.

Teacher Moments poses three types of prompts to teacher candidates. These three types are anticipatory, in-simulation, and reflection prompts. Before beginning a simulated interaction, the teacher candidate reads about the simulated situation and answers anticipatory prompts to predict how the simulation might go. Then, the scenario is simulated via a series of videos and text. During the simulation, the teacher candidate responds to in-simulation prompts, which are in the form of audio responses or text responses. After the simulation, the teacher candidate answers reflection prompts to reflect on their performance. This flow is outlined in Figure 1.

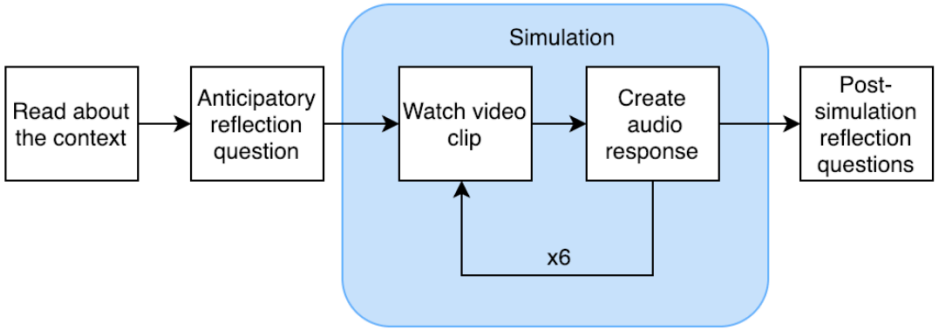

Figure 1: Teacher Moments Workflow

With the confusion measurement tool developed in this project, teacher educators can measure the confusion levels of their teacher candidates. Instead of manually reading responses in the dashboard to determine whether they think a teacher candidate is confused, teacher educators can use the confusion measurement tool to provide a measurement of confusion. In this way, the confusion measurement tool may allow teacher educators to spend their time more effectively when analyzing response data. Additionally, by knowing which of their teacher candidates are confused about how to handle difficult classroom simulations, teacher educators can adjust the pace of the course or provide additional attention to those who need it.

In order to detect confusion, confusion must first be defined. To sufficiently define confusion as it appears in Teacher Moments simulation responses, I develop a coding scheme for confusion. A set of Teacher Moments simulation responses are coded according to the coding scheme and used to train a confusion detection model. Responses from all three types of prompts – anticipatory, in-simulation, and reflection – are coded and used to train the model. The model is then tested against a second set of data. This data contains anticipatory, in-simulation, and reflection responses labeled as confused or not confused by outside users, instead of researchers using the coding scheme. I use a testing dataset of user labeled responses in order to assess how well the researcher construct of confusion compares to users' construct of confusion. Lastly, the model is incorporated into Teacher Moments, and I perform user testing to understand how teacher educators use the confusion measurement tool, whether they gain any additional insight from using the tool, and whether the tool allows them to more effectively use their time analyzing response data.

The rest of this paper is structured as follows. Section 2 provides background on existing teacher education methods and introduces the use of sentiment analysis in education. I also introduce my research questions in Section 2. Section 3 details the methodology I follow while developing the coding scheme and building the sentiment analysis tool. Section 4 details quantitative performance measures and qualitative usability findings for the confusion measurement tool. In Section 5, I summarize the results and discuss algorithmic bias concerns. In Section 6, I detail future work to further improve the confusion measurement tool.

Chapter 2

Background

2.1 Existing Teacher Education Methods

The common goal amongst all teacher education methods is to prepare teacher candidates for the classroom. Traditional teacher education approaches have focused on discussions of textbooks about best practices for teaching. Practice-based education takes a more hands-on approach to teacher education by providing teacher candidates with opportunities to practice their teaching. Dieker *et al.* point out that "there is a gap in teacher education instruction where teacher candidates and struggling teachers can rehearse their skills, improve their skills, and build confidence in their abilities" [3]. Practice-based education has arisen to help fill this gap. Practice-based education focuses on practicing classroom interactions in low-stakes environments. Within practice-based education, there are multiple different approaches. One is live rehearsals, which may include activities such as practicing presenting a lesson to a class. Clinical simulations often involve a teacher candidate practicing a one-on-one interaction with an actor who is portraying a student or parent. Mixed reality simulations also simulate classroom interactions, but through the use of virtual reality. Teacher Moments is an example of a practice space, which incorporates elements of clinical simulations onto a software platform. A diagram of teacher education approaches is shown in Figure 2.

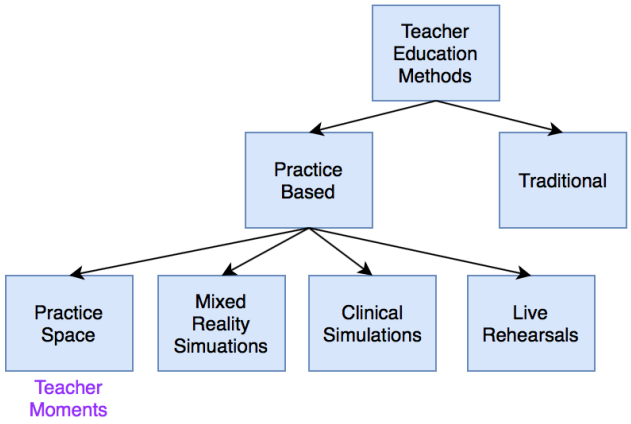

Figure 2: Existing Teacher Education Methods

As a website that provides clinical simulations through videos and audio, Teacher Moments incorporates aspects of mixed reality simulations and clinical simulations. To better understand how the functionality of Teacher Moments compares to other practiced-based approaches for teacher education, two examples of mixed reality simulations and clinical simulations are given below.

2.1.1 Mixed Reality Simulation: Mursion

Mursion is a virtual reality environment in which participants can practice communication skills in multiple fields, including corporate, education, healthcare, and defense [4]. In the educational arm of Mursion, pre-service teachers can practice engaging with students in contexts that pose interpersonal and communicative challenges. The pre-service teacher stands in front of a large screen where up to five virtual students, who are controlled by a separate human actor, sit in a classroom. The pre-service teacher stands in front of a monitor to communicate with the virtual students, while a Kinect tracks the pre-service teacher's movements. The teacher educator can control the type of challenges that the pre-service teacher will face in the virtual environment. For example, the teacher educator can set parameters such as the number of students in the classroom and their characteristics. After the simulation is over, a number of statistics such as the wait time between questions and answers are presented.

Although Mursion does not incorporate teacher educator feedback into the simulation, some teacher educators may watch their teacher candidates engage in the simulation and then provide feedback afterwards. Lisa Dieker, a professor at the University of Central Florida involved in the development of Mursion, emphasizes the importance of a "cyclical process" for teacher education, which is composed of simulations, observations, and discussions about performance [3]. One way in which Teacher Moments differs from Mursion is that Teacher Moments incorporates teacher educators into the feedback process. Through the Teacher Moments dashboard, teacher educators can review teacher candidate responses to anticipatory, in-simulation, and reflection questions. Additionally, the confusion measurement tool allows teacher educators to gain further insight into the response data and potentially adjust the activities in their classroom as a result.

2.1.2 Clinical Simulation: Dotger

A clear alternative to mixed reality simulations are clinical simulations. Clinical simulations are different from mixed reality simulations in a few ways. Most notably, clinical simulations use on another actor to play the role of a student or parent rather than simulating those characters with mixed reality. Additionally, clinical simulations often simulate one-on-one interactions with a single actor, whereas mixed reality simulations can more easily simulate multiple students or parents in one interaction.

Ben Dotger, a professor at the Syracuse School of Education, has developed a series of one-on-one clinical simulations that help pre-service teachers practice interpersonal skills they will likely rely on in the classroom. In these simulations, the teacher candidate meets one-on-one with an actor, who plays the role of a student, parent, or colleague. The actor does not have a specific script, but does have a list of important points to focus on during the conversation. These points include "verbal triggers", or statements which are meant to focus the conversation on a specific topic. These verbal triggers are often somewhat defensive or unjustified statements which capture the teacher candidate's attention and set the course of the conversation. Dotger says that the purpose of these simulations "is to bring to life some of the most frequent, most common situations that occur in a teacher's classroom or school environment—as well as some of the situations that don't occur all that often but that are really important" [5]. Simulations may be focused on particular school subjects or may be more broadly related to the student's learning and well-being. For example, the simulations range from a student asking questions about a math problem to a student showing signs of neglect and abuse at home. Dotger believes that simulating these difficult conversations produces cognitive dissonance and allows teacher candidates to self-regulate their speech and behavior in challenging scenarios. Teacher Moments also aims to induce cognitive dissonance in teacher candidates, so that they can practice facing situations which challenge their beliefs. Measuring confusion is one way to determine whether teacher candidates are experiencing cognitive dissonance.

2.2 Confusion in Education

Teacher Moments simulations aim to induce cognitive dissonance in teacher candidates, as these simulations present a number of challenging scenarios that teacher candidates may not

know how to approach [6]. Since teacher candidates experiencing cognitive dissonance may show signs of confusion, I adopt a specific focus on using sentiment analysis to detect confusion. By understanding when teacher candidates are confused, teacher educators can meet with teacher candidates accordingly or plan and conduct debrief discussions to mitigate any confusion.

2.2.1 Confusion Theory

The theory of confusion for the purposes of this project follows the model proposed by Kort *et al*. [7]. In this model, there are two axes in the learning process - affect and learning. The learning axis ranges from "constructive learning" to "un-learning", and the affect axis ranges from "positive affect" to "negative affect". A diagram of these axes is shown in Figure 3.

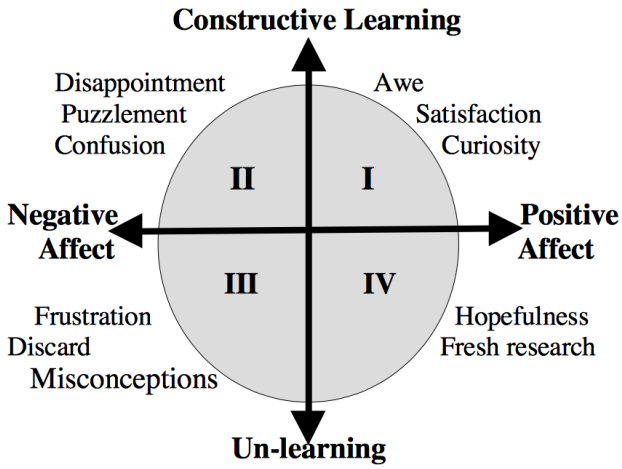

Figure 3: Kort et al. Model of Learning and Affect

The first quadrant is characterized by positive affect and constructive learning. In this scenario, the student is working through material and is actively engaged. In the second quadrant, the student is still somewhat engaged in the work, but is having troubles with some aspects. If the student becomes increasingly stuck in their work, she moves into quadrant three, where she is likely feeling frustrated and is no longer learning. Lastly, a student in quadrant four may still be facing some unknowns, but has a positive outlook on how the work can move forward.

This model was originally developed to characterize confusion in the context of an Intelligent Tutoring System (ITS). An ITS models the user's understanding of the material, contextualizes the user's responses, and responds in an appropriate manner. Identifying confusion in an ITS system allows the system to adjust accordingly when the user shows signs of confusion. Similarly, the confusion measurement tool in Teacher Moments aims to identify confusion so that the teacher educator can respond appropriately to teacher educators who show signs of confusion.

Identifying confusion in teacher candidates is important, since experiencing confusion may impact a teacher candidate's ability to work though cognitive dissonance and reconcile their beliefs or behavior with contradicting information. More generally, confusion can greatly impact a student's ability to learn, although the effect of confusion on learning is somewhat disputed. Craig *et al*. find that confusion is beneficial to learning, when students resolve their confusion and finish their work with a greater understanding of the material [8]. Rodrigo *et al*. find opposite results, and argue that confusion inhibits learning as students may become frustrated and stuck in their work [9]. The research by Craig *et al.* and Rodrigo *et al.* focuses on confusion in students answering technical and straightforward questions. Teacher Moments presents teacher candidates with anticipatory, in-simulation, and reflection prompts, which do not have standard correct answers. Therefore, these findings may not effectively characterize the effect of confusion on teacher candidates while using Teacher Moments.

2.3 Sentiment Analysis in Education

Motivation to build the confusion measurement tool came largely from recent work in the space of sentiment analysis for education. While sentiment analysis has proven applications in education, it hasn't yet found widespread use in teacher education technology. Sentiment analysis has often been used to analyze student sentiments as students progress through a course at a school or university. Sentiment analysis used in these educational contexts can tell us whether students are enjoying course material and keeping up with the pace of the course. Altrabsheh *et al.* ask students to post feedback to Twitter about courses they are taking, and then use sentiment analysis to categorize these tweets as displaying primarily positive or negative emotions [10]. This information can then be provided to the course instructor to adjust the course accordingly. Rosé *et al*. run a similar study in which they measure the sentiments of students

taking massive open online courses (MOOCs) via the student's posts to the class forums [11]. Rosé *et al*. find a correlation between sentiment and number of students who drop out of the MOOCs each day. Clearly, sentiment analysis can be a useful tool in student education.

In this thesis, I focus on the less-explored use case for sentiment analysis in teacher education. One of the main contributions of this paper is a coding scheme for confusion as it appears in Teacher Moments. I conducted a systematic review of existing coding schemes for teacher education. On Google Scholar, I used terms "coding", "scheme", "confusion" to search four teacher education journals for articles from 2009-2019. The searched journals were *Transactions on Learning Technologies*, *The Journal of Teacher Education*, *The Journal of Technology and Teacher Education*, and *The Journal of Digital Learning in Teacher Education*. A total of 40 search results appeared, and the most relevant are displayed in Table 1. In this search, I found no existing coding schemes for confusion.

Journal	Example	Sources
Transactions on Learning Technologies	Use coding schemes that do not include confusion	Gamification for Engaging Computer Science Students in Learning Activities: A Case Study (Ibáñez 2014)
		Language and Discourse Are Powerful Signals of Student Emotions during Tutoring (D'Mello 2012)
		Using the Tablet Gestures and Speech of Pairs of Students to Classify Their Collaboration (Viswanathan 2017)
Journal of Teacher Education	Use coding schemes that do not include confusion	Comparing the impact of online and face-to-face professional development in the context of curriculum implementation (Fishman 2013)
		Teacher questioning to elicit students' mathematical thinking in elementary school classrooms (Franke 2009)
		Effects of video club participation on teachers' professional vision (Sherin 2009)
		Preservice EAL Teaching as Emotional Experiences: Practicum Experience in an Australian Secondary School (Nguyen 2014)
		History teachers' knowledge of inquiry methods: An analysis of cognitive processes used during a historical inquiry (Voet 2017)
Journal of Technology and Teacher Education	Use coding schemes that do not include confusion	Evaluating Teacher's Support Requests When Just-In-Time Instructional Support Is Provided to Introduce a Primary Level Web-Based Reading Program (Wood 2011)
		Epic Fails: Reconceptualizing Failure as a Catalyst for Developing Creative Persistence within Teaching and Learning Experiences (Smith 2015)
		Missed Opportunities, Misunderstandings, and Misgivings: A Case Study Analysis of Three

		Beginning English Teachers' Attempts at Authentic Discussion With Adolescents in a Synchronous CMC Environment (Groenke 2011)

Table 1: Confusion Coding Schemes in Teacher Education Journals Literature Review

2.4 Research Questions

I focused on answering these research questions in my thesis work:

- Question 1: To what extent can we model and detect confusion using a sentiment analysis tool?
- Question 2: How do teacher educators use the confusion measurement tool, and do they gain any additional value from it?
- Question 3: Does the confusion measurement tool help teacher educators effectively use their time when analyzing Teacher Moments data?

Chapter 3
Methodology

3.1 Coding Scheme Development

In this project, the model proposed by Kort *et al*. provided a baseline definition for confusion. The definition of confusion was then refined to be specifically applicable to Teacher Moments scenario responses. However, as is discussed in Chapter 5, this coding scheme has potential to be used for other applications in the field of teacher education. The purpose of this coding scheme is to give examples of traits that may appear in a confused response. It was used for coding responses to anticipatory, in-simulation, and reflection prompts as confused or not confused to create a labeled dataset. The coding scheme is shown in Table 2.

Category	Example
Expressing doubt (explicitly)	"I'm not sure, but I think it could be ..." "I don't really know"
Asking questions to clarify; asking for explanation or justification to help understand (inferred confusion). NOT questions posed to students in the simulation.	"What did that mean?" "Why are they asking me this?"
Hesitation (non-verbal cue)	"I think %HESITATION that ..." "Maybe [pause] this could be..." Short communication

Table 2: Coding Scheme for Confusion

The characteristics of confusion shown in the coding scheme seem quite intuitive, and they also align with the model of confusion proposed by Kort *et al*. When a participant explicitly expresses doubt, or asks clarifying questions, they are indicating that they are actively engaged in the activity and are paying attention to whatever is confusing them. However, they are clearly experiencing some difficulty with the material, and are displaying negative affect. One important revision made to the coding scheme during the coding process was that clarifying questions refer to questions posed about the simulation, rather than questions asked as part of the simulation. For example, a teacher candidate may ask questions to simulated students during the simulation, and

these questions would not be classified as confusion. Only questions relating the the simulation as a whole may be classified as confusion.

The last category of confusion is non-verbal cues. Jokinen *et al*. show that many interpret hesitation as a sign of uncertainty, confusion, or doubt [12]. In Teacher Moments, hesitation is captured from audio responses and represented in text as "%HESITATION" or "[pause]".

3.2 Data Gathering and Coding

Two types of data are gathered. Researchers use the coding scheme to code one set of data as confused or not confused. This dataset serves as the training set. The second type of data, which serves as a testing dataset, is user labeled data. To gather the user labeled data, Teacher Moments users were asked to complete a Teacher Moments scenario exercise and then label their own responses as confused or not confused. To measure the accuracy of the confusion classifier and answer research question 1, both types of data are necessary. The researcher labeled data grounds the classifier in a theoretical basis and the user label data tests it against a practical perspective.

3.2.1 Instrument 1 – Expert Labels

The first type of data gathered is used to train the confusion model. This data is gathered from playtests, which are testing sessions run by TSL in which outside participants are invited to test various projects that TSL is working on, including Teacher Moments. Over the past two years, playtests have been run with Teacher Moments for various Teacher Moments scenarios. In these playtests, outside participants take the role of pre-service teachers and complete simulated scenarios. If the participant gives consent, their response data is recorded. This collected data serves as training data for the confusion measurement tool. Approximately 500 user responses consisting of 790 total sentences were used as training data. These responses are from anticipatory, in-simulation, and reflection prompts in Teacher Moments. This amount of data is sufficient for a binary classification problem. Cetintas *et al*. use a similar amount of data – 930 sentences – to train a sentiment analysis binary classifier [13]. All three types of prompt responses are used in order to have a sufficient amount of training data.

This training data included both textual responses, in which the user types a response with the keyboard, and transcript responses, which are transcribed text from users' audio responses. The transcription data is the output from the IBM Watson speech to text transcription service. When teacher candidates respond to via audio recordings, the recordings are first transcribed by IBM Watson to text, and then used as training data for the classifier.

After gathering the training data, three researchers coded each response in the training data as confused or not confused. Researchers coded each piece of data with binary labels of 0 (not confused) or 1 (confused). To code the training data, the coding scheme for confusion in Table 2 was used. The three human coders included the author of this paper and two researchers at TSL. Tie breaks were enforced when there were not unanimous decisions for whether a response represented confusion or not. During the coding processes, the coders decided to throw out four pieces of response data. These four responses had such bad audio transcriptions that no sensible interpretation was possible. An example of one of the four disregarded responses was "chasing after [PAUSE] as [PAUSE] eighteen what". Presumably, the teacher candidate did not actually say such a nonsensical phrase, so the data point was discounted. Additionally, all coding decisions were discussed and sometimes coders changed their opinions based on the discussion. Of the 497 pieces of data, coders did not have unanimous labels on 146 pieces of data (29%). For these 146 pieces of data where there was some disagreement, tie breaks and discussions were effectively able to resolve disagreements.

The labels from the three researchers who coded the training dataset were used to calculate inter-rater reliability. The parameter used to measure inter-rater reliability was Cohen's Kappa. Cohen's Kappa measures the agreement between raters who each classify N items into C mutually exclusive categories. In this case, the raters classified approximately 500 sentences into two categories - confused, or not confused. Cohen's Kappa was calculated as 0.30165. This kappa suggests there was a non-negligible amount of disagreement between raters. The primary cause we identified for this disagreement was the misunderstanding of how to interpret clarifying questions in the coding scheme. Originally, the coding scheme did not distinguish between questions posed to students as part of the simulation, and questions posed about the simulation. For example, one sentence that originally received disagreement from the coders was "how specifically would you build this tower?" This question is clearly posed to a student during the simulation, rather than a question posed about the simulation during the anticipatory or reflection

prompts. Distinguishing between responses to the three different types of prompts in Teacher Moments may help alleviate disagreement about questions in responses and improve Cohen's Kappa. Such ways to improve Cohen's Kappa in the future are discussed in Chapter 6.

The coded training dataset displays significant class imbalance, as only 18.5% of the training data is classified as confused. This class imbalance is shown in Figure 4.

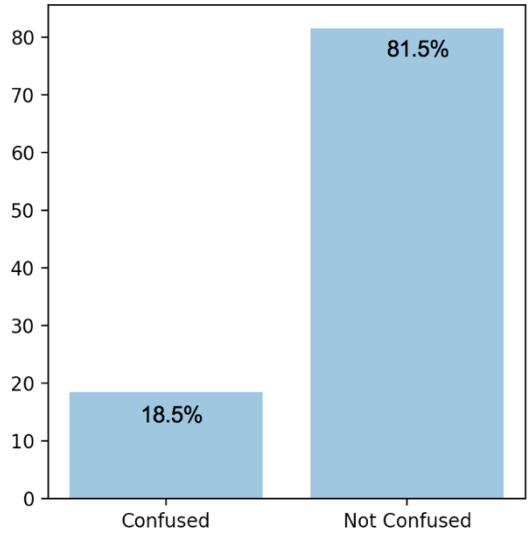

Figure 4: Class Imbalance in the Training Dataset

3.2.2 Instrument 2 – User Labels

The second type of data used in developing the sentiment analysis feature was user labeled data. This data served as a testing dataset for the classifier. The testing dataset is composed of user-labeled responses in order to test how the researcher construct of confusion aligns with users' construct of confusion. In two trials of the Teacher Moments software, users were asked to complete a certain scenario in Teacher Moments. These users were both playtest participants at TSL and real pre-service teachers. At the end of the simulation, a series of questions were posed that asked the user to assess their own level of confusion when they gave certain responses. A screenshot of this question is shown in Figure 5.

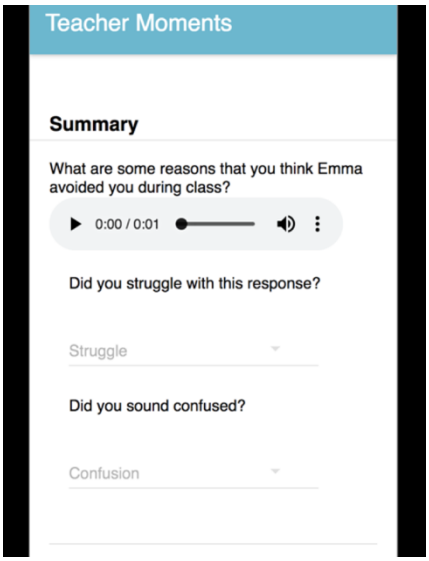

Figure 5: Collecting User Labeled Data

Users were asked to label their own responses to anticipatory, in-simulation, and reflection prompts as confused or not confused after they completed a Teacher Moments scenario. The user labels of confused or not confused are used as ground truth for the testing dataset. This dataset also showed class imbalance, but less so than the training dataset. In the testing dataset, 26% of responses are labeled as confused. Figure 6 shows class imbalance in the testing dataset.

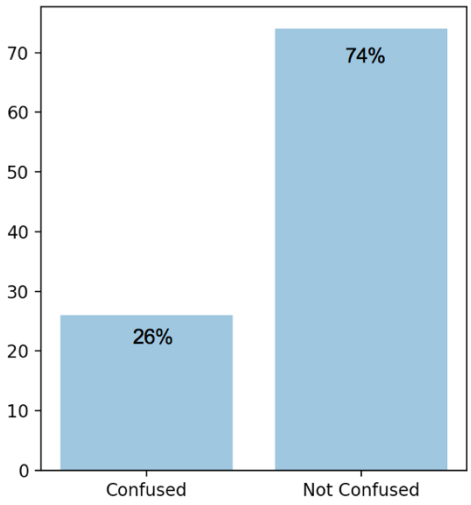

Figure 6: Class Imbalance in the Testing Dataset

3.3 Developing the Confusion Classifier

3.3.1 System Architecture

The code for this thesis built on the pre-existing Teacher Moments codebase, which is primarily in React, a javascript library for user interfaces. Teacher Moments uses a postgres database and Amazon Simple Storage System (S3) to store permissions, accessibility, and scenario response data. Teacher Moments also integrates the IBM Watson text-to-speech API to transcribe audio recording data to textual responses. Lastly, the live version of Teacher Moments is hosted on the cloud platform Heroku.

The confusion measurement tool was implemented through a mix of React, Python, and PSQL code. React was used to alter and add to the Teacher Moments code base to create functionality for the tool. Python was used to train and implement the neural network. Lastly, PSQL was used to interact with the database of Teacher Moments users and check permissions.

The Python code for sentiment analysis made use of a few different modules. The keras and scikit-learn modules were utilized to train various types of models. The natural language toolkit module (nltk) was utilized to perform preprocessing on textual data. Lastly, numpy was used for various matrix operations on the processed data.

3.3.2 Pre-Processing of Data for Classifier

Before the response data was used to train a model for detecting confusion, it had to be processed. The response data begins as strings of typed text or text from transcribed audio recordings. Each response may contain several sentences. First, the responses were tokenized by sentence, so each entry was only one sentence long. Each sentence is then tokenized, so that each punctuation mark and word was separated as an individual component of the sentence. Then, stopwords and punctuation were removed from the sentences. The stopwords used in this application are from the Python nltk module. These stopwords are words such as “a”, “it”, and “the” [14]. Removing stopwords has been shown to improve the accuracy of sentiment classifiers [15]. Punctuation was also removed from the training data. Removal of punctuation is enforced via the Python string module, which defines punctuation as the ASCII characters which are

considered punctuation in the C locale [16]. Punctuation was removed to help normalize all sentence data.

After stopwords and punctuation marks were removed from the sentences, the sentences were converted into N-grams. N-grams are collections of N consecutive words in a given sentence. An example of N-grams of different sizes is shown in Figure 7.

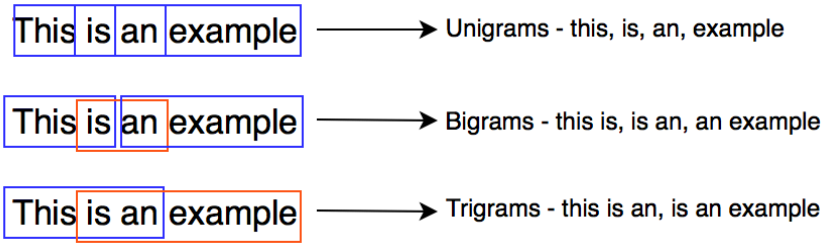

Figure 7: N-gram Example

N-grams are commonly used in sentiment analysis of textual data. The advantage of using N-grams with N > 1 rather than individual words for sentiment analysis is that N-grams can infer sentiments from collections of words. For example, N-grams are very useful for negations, where the negating word, such as “not”, may reverse the sentiment of the following words. N-grams can effectively capture collective sentiment whereas considering words on a singular basis cannot.

Choosing the size of the N-grams was largely an experimental decision. There exists scientific literature that shows both smaller and larger sizes of N have their advantages and may work better in certain contexts. Pang *et al*. report that unigrams outperform bigrams for sentiment classification of movie reviews [17]. However, Dave *et al.* show that bigrams and trigrams outperform unigrams for product review sentiment classification [18]. In order to determine the appropriate N-gram sizes for this application, accuracy measures were recorded for unigrams, bigrams, and trigrams. Using unigrams, bigrams, and trigrams together was also tested as an approach. This approach takes advantage of the benefits of both smaller and larger N-grams. Larger size N-grams can better capture patterns of sentiments that may relate to confusion, such as the phrase “I don’t know”, which is a trigram. However, smaller size N-grams may be better for capturing pauses and hesitations, since those appear as the single words “%HESITATION” or “[pause]” in the transcription.

Each unique unigram, bigram, or trigram appeared with a certain frequency amongst all the training data sentences. Many N-grams only appeared once. In fact, amongst the 8322 unique trigrams that appear in the training data, 8150 trigrams only appeared once. Experiments were performed to measure the precision, recall, and f-score values of models with different frequencies of N-grams allowed. These experiments aimed to determine whether including N-grams that only appeared once in the dataset would improve or weaken the model. The results of these tests are shown in Section 4.1.1.

The last step prior to training the models was to format the processed data into the right shape. The training data was formatted into matrix X of binary values. The number of rows of X equaled the number of sentences of training data, and the number of columns of X equaled the number of N-grams (unigrams, bigrams, and trigrams) which occur with a frequency greater than a parameter MIN_FREQ. For example, an N-gram frequency (MIN_FREQ) of one mandates that all N-grams that appear at least once in the training data can be used. An N-gram frequency of two mandates that only N-grams that appear two or more times in the can be included. The presence of an N-gram was treated as a binary feature. In other words, each sentence either contained a certain N-gram and had a 1 in the corresponding entry of X, or did not contain that N-gram and had a 0 in the corresponding entry of X. Y was a column vector of binary values, where a 0 represented no confusion in the sentence and a 1 represented confusion. A diagram of this process with N-grams of size 1, 2, and 3 is shown in Figure 8.

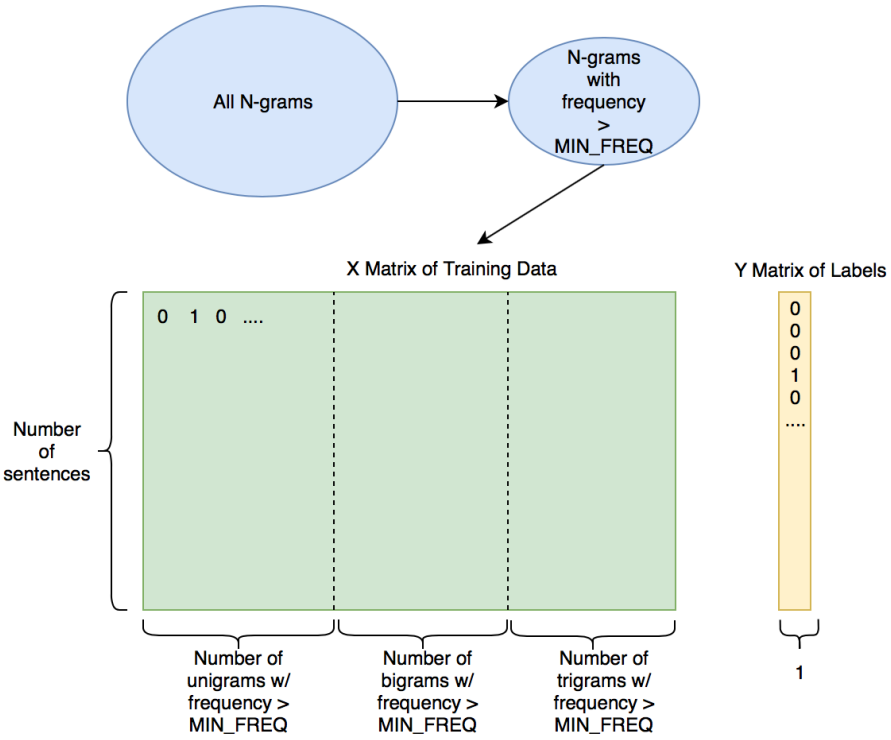

Figure 8: Model Architecture with N-grams of different sizes

A diagram of this process for N-grams only of one size N $\in \{1,2,3\}$ is shown in Figure 9.

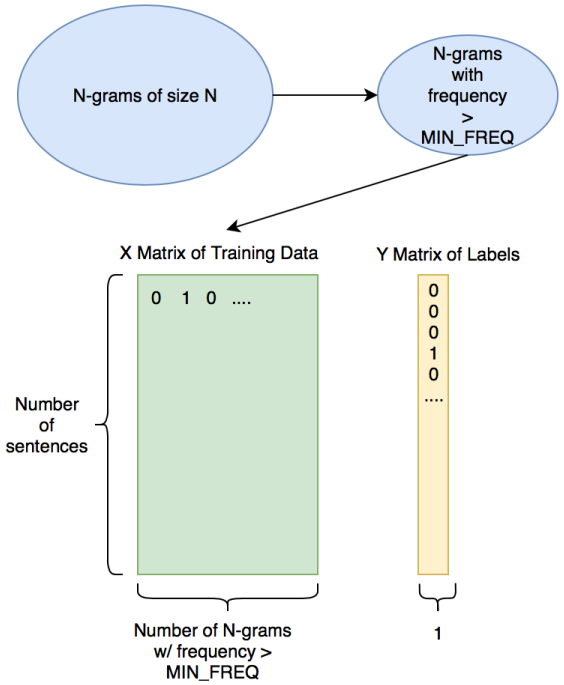

Figure 9: Model Architecture with N-grams of one size

3.3.3 Sentiment Analysis Models

Since to my knowledge there is no existing research on which type of classifier works best for detecting confusion in a teacher education application, I created and tested three different types of models for the confusion measurement tool. These three models were a long short term memory network (LSTM), a support vector machine (SVM), and a logistic regression model. These types of models have all been proven effective for sentiment analysis applications, particularly for binary classification problems.

LSTM networks, a type of recurrent neural network (RNN), have become an effective method of language modeling. Jozefowics *et al*. show that the use of N-grams as input features for a LSTM outperforms the use of N-grams alone for sentiment analysis [19]. RNNs, such as LSTMs, used in conjunction with N-grams, have become a popular method of creating language models for applications such as sentiment analysis [19]. In this project, LSTMs with N-gram input features were tested to detect confusion in Teacher Moment scenario responses.

The LSTM was trained with a mean squared error loss and Adam optimizer. Since this is a binary classification problem, the softmax activation function is applied. After approximately 20 epochs, the LSTM achieved its maximum accuracy and stopped improving.

Support vector machines (SVM) are conventional choices for binary classification problems. SVMs use a kernel function to map a set of data points that may not be linearly separable to a space in which they can be linearly separated. Support vector machines have found many applications for sentiment analysis. Mullen *et al*. use an SVM to classify texts as having positive or negative sentiment [20]. This project used SVMs for a similar binary classification problem. SVMs with different kernel functions were tested, and a linear kernel performed the best.

Logistic regressions are often the go-to models for binary classification problems. Logistic regression models fit binary classification data to a logistic function. Logistic regressions have natural applications to sentiment analysis systems which classify positive and negative sentiments, such as in the sentiment polarity research conducted by Hamdan *et al.* [21]. Since this project also aims to perform binary classification, a logistic regression seemed like a natural fit. The accuracy results of these three types of models are shown in Section 4.1.1.

3.3.4 Text Analysis Accuracy

Each of these three models was trained and tested via five-fold cross validation. In five-fold cross validation, the training data was split into five groups. One group was used as a temporary test set, while the remaining four groups of data were used to train the model. Accuracy was measured against the temporary test set, and this process repeated until each of the five sections of data had been used as a test set. From these accuracy measures, recall, precision, and F-scores were recorded for each model.

Precision is the ratio of correctly predicted labels of a certain class to all data predicted to be in that class. For example, a precision measurement for the confusion class represents what percentage of data that is labeled as confused by the classifier actually have a ground truth label of confused. Similarly, a precision measurement for the not confused class represents the ratio of data that gets labeled as not confused by the classifier to the data that have a ground truth label of not confused. Recall is the ratio of correctly predicted labels of a certain class to all data with a ground truth label of that class. For example, a recall measurement for the confusion class represents what percentage of responses with a ground truth label of confused will actually be predicted as confused by the model. The F-score is simply a weighted average of precision and recall. F-score was used for the purposes of comparing the LSTM, SVM, and logistic regression models. Micro and macro averaging are used to calculate recall, precision, and F-score. Micro-averaging considers each prediction individually, while macro-averaging considers each class independently. In situations when there is a class imbalance between the confused and not confused classes where the minority class underperforms as compared with the majority class, micro-averaging will produce higher F-scores and macro averaging will produce lower F-scores. This is because micro-averaging checks accuracy across the entire dataset while macro-averaging checks accuracy of each class first and then averages the accuracy of two classes. In this study, there is a significant amount of class imbalance, so micro and macro averaging produce different performance measures.

Although there are no existing studies that develop a coding scheme for confusion and use it for binary classification, other related studies can provide examples of what performance levels are achieved in sentiment analysis classification problems. Cetintas *et al.* use binary classification to predict whether student's questions are relevant or irrelevant to the activities in their classroom [13]. The group achieves F-scores ranging from 0.69 to 0.87. Calvo *et al.* build

classifiers to recognize four different emotions - anger/disgust, fear, joy and sadness – on three different datasets [22]. When considering the best performance of the classifiers across the three datasets, the reported F-scores range from 0.28 to 0.79.

Since there are no existing studies that use sentiment analysis methods to detect confusion in teacher education applications, there are no F-scores to directly compare the results of this study against. I instead use a random baseline model to serve as a benchmark for comparing F-scores achieved in this project. The random baseline model randomly labels responses as confused or not confused. When comparing my models to the random baseline, I focus on macro averaging F-score, since it better represents the model's performance given significant class imbalance. However, it is still worthwhile to analyze macro and micro averaging results to understand the model's performance on the minority and majority classes. Similarly, it's important to analyze the false negative and false positive errors, in order to understand why the classifier commits these errors and be able to prevent them in the future.

In summary, to answer research question 1 and determine the accuracy of a classifier that detects confusion in Teacher Moments responses, I look at a few different accuracy measures. First, I measure the precision, recall, and F-score of different types of classifiers on the training dataset coded by researchers using the developed coding scheme. I then use the best performing classifier on the testing dataset of user labeled responses and measure precision, recall, and F-measure for that dataset. In order to better understand the shortcomings of the classifier, I analyze the causes of false positive and false negative errors.

3.4 User Testing

The goal of user testing is to understand how teacher educators would use a confusion measurement tool in their classroom. Particularly, these tests are aimed at gathering the data necessary to answer research questions 2 and 3.

First, I measured how the confusion measurement tool impacted the time that teacher educators spend interpreting student data in the dashboard. To answer this question, quantitative measurements were taken via an A/B test. A group of 12 participants were invited to TSL to partake in a playtest, and divided into two subgroups. One group was given the newer version of the Teacher Moments dashboard which had the confusion measurement tool. The other group was given a version of Teacher Moments without the tool. Both subgroups were given a survey

that asked the participants to complete various tasks within the Teacher Moments dashboard. Timing measurements were taken for how long each participant took to complete these tasks. The participants in this A/B test were not all teacher educators, simply because it was not feasible to bring together a large enough sample size of teacher educators to conduct a meaningful A/B test. However, the participants were involved in education somehow and had ties to TSL. The participants were given specific instructions for the task to complete within the dashboard, so the fact that they are not teacher educators should not impact the results significantly. A copy of the survey of tasks used in the A/B test is included in Appendix C.

The second type of user test was aimed at understanding whether teacher educators could gain any new, valuable insights from using the new confusion measurement tool. This test sought out qualitative data relating to teacher educators' use of the new features. I aimed to understand whether and how teacher educators would incorporate the confusion measurement tool into their classroom, and whether they may change any of their classroom activities given the results of the tool. One teacher educator was provided with the new tool in the Teacher Moments dashboard. The teacher educator was given a few days to test out the tool on the scenario response data that she had previously acquired in her teacher candidate classroom. After the teacher educator had a chance to test out the tool, an interview was conducted to understand how she used the tool and what additional value the tool added to Teacher Moments, if any.

Chapter 4

Results

4.1 Performance Metrics

4.1.1 Training Data Cross Validation

Three types of models - SVMs, LSTMs, and logistic regressions - were trained with different sized N-grams and N-gram frequencies. Five-fold cross validation was used to measure precision, recall, and F-score. Unigrams, bigrams, and trigrams were used individually, along with a combination of all three sizes of N-grams. N-gram frequencies of one and two were tested. Table 3 shows F-score values for the SVM models.

	N-gram frequency = 1		N-gram frequency = 2	
	Micro	Macro	Micro	Macro
Unigrams	0.64	0.53	0.62	0.53
Bigrams	0.73	0.48	0.54	0.51
Trigrams	0.76	0.45	0.65	0.50
Unigrams, Bigrams, and Trigrams	0.69	0.56	0.65	0.56

Table 3: F-score Values for SVM Models

In Table 3, the best classifier seems to be Unigrams, Bigrams, and Trigrams with a N-gram frequency of 1. This model achieves the highest macro averaging F-score (0.56) while also achieving the highest micro averaging F-score (0.69). Macro averaging F-score is considered first since it better represents the model's performance given the significant class imbalance. Table 4 shows F-score values for LSTM models.

	N-gram frequency = 1		N-gram frequency = 2	
	Micro	Macro	Micro	Macro
Unigrams	0.24	0.19	0.24	0.19
Bigrams	0.24	0.19	0.24	0.19
Trigrams	0.24	0.19	0.24	0.19
Unigrams, Bigrams, and Trigrams	0.24	0.19	0.24	0.19

Table 4: F-score Values for LSTM Models

The LSTM models all perform equally poorly and achieve low F-scores, so I do not consider any of these models to move forward with. Table 5 shows F-score values for logistic regression models.

	N-gram frequency = 1		N-gram frequency = 2	
	Micro	Macro	Micro	Macro
Unigrams	0.64	0.54	0.63	0.55
Bigrams	0.76	0.43	0.76	0.43
Trigrams	0.76	0.43	0.76	0.43
Unigrams, Bigrams, and Trigrams	0.68	0.56	0.65	0.56

Table 5: F-score Values for Logistic Regression Models

From this data, it's clear that all models achieve a higher micro averaging F-score than macro averaging F-score due to the significant class imbalance. Many models with higher micro averaging F-scores have lower macro averaging F-score, meaning that these models perform better at predicting the majority class of no confusion than they do at predicting the minority class of confusion. Two such models are the SVM with bigrams and a frequency of 1, which will be labeled Model SVM-B, and the SVM with trigrams and a frequency of 1, which will be labeled Model SVM-T. It appears that the SVM with unigrams, bigrams, and trigrams with a frequency of 1 has the highest micro averaging F-score amongst models with the highest macro averaging F-score of 0.56. This will be labeled Model SVM-UBT.

The precision, recall, and F-measure scores for Models SVM-B, SVM-T, and SVM-UBT are shown in Table 6. These measures achieved via cross validation on the training dataset of researcher coded responses serve as a prediction of how the model will fare on user labeled data. In order to better contextualize these results, a random baseline model is also included in the comparison.

	Confused			Not Confused			Micro			Macro		
	P	R	F	P	R	F	P	R	F	P	R	F
SVM-B	0.27	0.07	0.11	0.76	0.94	0.84	0.73	0.73	0.73	0.51	0.50	0.48
SVM-T	0.40	0.02	0.04	0.76	0.99	0.86	0.76	0.76	0.76	0.58	0.51	0.45
SVM-UBT	0.34	0.31	0.32	0.79	0.80	0.80	0.69	0.69	0.69	0.56	0.56	0.56
Random Baseline	0.49	0.48	0.43	0.49	0.48	0.43	0.49	0.49	0.49	0.49	0.48	0.43

Table 6: Precision, Recall, and F-score for 3 Models and Random Baseline

All the precision, recall, and F-score values for the random baseline model are close to 0.5, since the model has a 0.5 chance of guessing that a response is confused or not confused. This baseline model can be compared to Models SVM-B, SVM-T, and SVM-UBT to see where the other models excel and where they have room for improvement.

From these recall, precision, and F-score results, it is clear that SVM-B, SVM-T, and SVM-UBT have different amounts of bias towards predicting confusion. SVM-T is least the likely of the three models to predict confusion, while SVM-UBT is the most likely to predict confusion. SVM-UBT generally has lower precision, recall, and F-score values for the not confused class than SVM-B and SVM-T. However, SVM-UBT also has much higher precision, recall, and F-score values for the confusion class than SVM-B and SVM-T. Therefore, SVM-UBT has a lower micro averaging F- score and a higher macro averaging F-score than the other models. Based on the macro averaging F-score, SVM-UBT performs the best of the three and is a sensible choice for the Teacher Moments confusion measurement tool.

However, it's clear that SVM-UBT still has room for improvement. Precision, recall, and F-score for SVM-UBT in the confused class are all lower than the random baseline model. This means that SVM-UBT is worse at predicting confusion than a model that randomly guesses.

Although SVM-UBT has a higher F-score than SVM-B and SVM-T for the confusion class, it is still worse than the random baseline model. SVM-UBT produces a significant amount of false negatives. It labels responses that have a ground truth of confusion as being not confused. Methods to reduce the number of false negatives produced by a confusion classifier for Teacher Moments are further discussed in Chapter 6.

4.1.2 Face Validity with User Labeled Data

A set of user labeled data served as a test set to measure how well the models performed on another dataset. The purpose of using this test set was to assess how well the researcher definition of confusion aligned with users' definition of confusion. In this test, the user labels of confused or not confused are considered ground truth, instead of researcher labels. High F-scores in this test indicate not only indicate that the model can effectively predict confusion. High F-scores also indicate that the researchers had a perception of confusion that largely aligned with a 3rd party's idea of confusion. SVM-UBT was selected to use with the test dataset, since it displayed the highest micro averaging F-score amongst models with the highest macro averaging F-score. The results for the face validity test with SVM-UBT are shown in Table 7.

	Confused			Not Confused			Micro			Macro		
	P	R	F	P	R	F	P	R	F	P	R	F
SVM-UBT	0.30	0.29	0.30	0.75	0.76	0.76	0.64	0.64	0.64	0.53	0.53	0.53

Table 7: SVM-UBT Results with User Labeled Data

SVM-UBT achieves results in the user labeled testing dataset that are fairly similar to its results in the researcher labeled training dataset. Cross validation with the researcher labeled training dataset predicted a macro averaging F-score of 0.56; testing with the student labels gave a macro averaging F-score of 0.53. While these F-scores are similar, it's important to compare predicted accuracy from cross validation with tested accuracy from the student labeled dataset by examining the precision and recall for the confused and not confused classes. For the confused class, cross validation predicts a F-score of 0.32, precision of 0.34, and recall of 0.31, and the test with student labels resulted in a F-score of 0.30, precision of 0.30, and recall of 0.29. These

testing results align with the predictions from cross validation. When comparing the not confused class, the predicted F-score from cross validation was 0.80, precision was 0.79, and recall was 0.80. On the testing dataset for the not confused class, the F-score was 0.76, precision was 0.75, and recall was 0.76, which are comparable measures to the predicted measures via cross validation. In terms of research question 1, these results indicate potential to model confusion using researcher coding, as the tested results are in line with the predicted results. However, given that the recall, precision, and F-score for the confused class are below the random baseline in both cross validation and user labeled testing, more work is needed to adequately model confusion to match student perception.

4.1.3 False Negative Predictions

One contribution of this work is a coding scheme for identifying confusion in teacher education applications. However, given the non-negligible false negative predictions by the classifier, it's necessary to discuss whether and how the coding scheme may play a role in creating the false negative predictions.

A confusion matrix displays the true positive, true negative, false positive, and false negative predictions for a classifier. The better illustrate how SVM-UBT performs on the test dataset of user labeled data, the confusion matrix for SVM-UBT is shown in Table 8.

	Predicted: Not Confused	Predicted: Confused
Ground Truth: Not Confused	103	32
Ground Truth: Confused	34	14

Table 8: Confusion Matrix for SVM-UBT On User Labeled Data

Based on this confusion matrix, it's clear that SVM-UBT makes 34 false negative predictions and 32 false positive predictions. In this section, I analyze 20 random data points of those 34 false negatives (59%). For the inspection I listened to the audio recordings and examined the transcripts to determine the extent to which transcription error might be a contributing factor. After comparing the transcriptions to the audio I inspected the text to see if any of the items would be identified as confusion based on the coding scheme to determine if the

reason for false negatives was rooted in the algorithm's limitation to predict the target construct. Finally, I inspected the false negatives to see if there was evidence of confusion in text that would merit revising the coding scheme.

I confirmed that the translations appeared to be a reasonable representation of the communication. I listened to audio recordings of the 20 random responses and determined that all 20 transcriptions captured the audio with reasonable accuracy. To illustrate my standards for reasonable accuracy, I present three examples of what IBM Watson provided and a transcription written based on listening to the audio. I felt that these transcriptions were reasonable. These transcriptions are shown in Table 9.

IBM Watson	Human Transcriber
I think college community helps you stalk December on your goals and maybe even help you find other passion	I think college communities help you focus and refine your goals and maybe even help you find other passions
thats her points but even if you do runners oaks on art and once that and want to do things in our you have to be pretty proficient and know you are [PAUSE] so even then [PAUSE] going on in our schools are important [PAUSE] yeah	That's her point but even if you do wanna focus on art and [pause] and want to do things in art you have to be proficient and know your art. So even then going to art school is very important. [pause] and yeah
I hope to convince her to register or the ABCS	I hope to convince her to register for the AP CS test.

Table 9: Comparison of IBM Watson Transcription vs. Human Transcription

After manually coding the 20 items, none appeared to be expressions indicating confusion. To illustrate this point, I present all 20 from the sample in Table 10.

I think college community helps you stalk December on your goals and maybe even help you find other passion
no I think thats [PAUSE] I mean Ill see you know you do your work and your but quiet thats thats okay and I think you know you are definitely more than welcome we encourage that passions outside of school what what are some things
Emma you have the choice to not yeah Im fine [PAUSE] youre really a good student I believe [PAUSE] I would like to [PAUSE] yeah you should consider signing up
I think I heard the [PAUSE] why are you sure eighty six Ammon maybe at reconsider her choice but %HESITATION theres definitely a lot more work to do on does she call me is that this actually provide value to our

there are a lot of reasons [PAUSE] yeah [PAUSE] moreover if you start studying the entire month a week before yes its true there is not just your best I can help you study a lot more in the van and help you prepare or D. S.
I anticipate that she %HESITATION asked me why she would do that is she she would not the wild
thats her points but even if you do runners oaks on art and once that and want to do things in our you have to be pretty proficient and know you are [PAUSE] so even then [PAUSE] going on in our schools are important [PAUSE] yeah
I hope to convince her to register or the ABCS
MMA beyond receptive to teachers back and [PAUSE] one willing to talk to me about the X.
I think thats a good thing for you to realize [PAUSE] but you shouldnt have to follow in your familys footsteps [PAUSE] I think youre really bright student and if you feel as though college is right for you [PAUSE] you should do it but you shouldnt follow in the footsteps of anyone else because they dont determine how you live your life
if this situation happen another soon I think I would [PAUSE] be a little bit more assertive instead of beating around the bush
so my original goal on this conversation was trying to get to the bottom of why am I felt so distant and why issue was starting to slack off in the class overall [PAUSE] at the end of the conversation I feel like I made some headway even though MFL and seem just a little bit hesitant to continue talking and even though she didnt really reach a realization [PAUSE] she [PAUSE] she said shed seen in class tomorrow and I think Im a little bit hopeful that she wants to keep thinking about this and she is still mulling it over [PAUSE] in that respect I think Ive made some steps towards my goal just making sure her mind at ease
I think that first and foremost because of the fact that she is awaiting me in conversation itll be hard to get her to open up to talk about these things and I think itll [PAUSE] difficult to get at the core reasons of why I think its really great to actually understand this but I do think that her general shyness might [PAUSE] make it difficult conversation piece [PAUSE] and thats something that we all need to work with
yeah
a lot
I think she is going to normal our system my questioning and %HESITATION [PAUSE] may be [PAUSE] for ten like she is interested in finding out where when in reality shes just going to [PAUSE] but I hope I am wrong
the exam doesnt really count for [PAUSE] all that much I mean [PAUSE] who could get [PAUSE] born in college or not [PAUSE] depending on the score you dont do well it is not a very high chance [PAUSE] so [PAUSE] you know you can just try it if thats the reason why youre doing really well in the class so I dont think you have anything to be worried about if you use them all [PAUSE] I dont think it will be difficult for you is [PAUSE] get it when youre doing really well so Im not worried about you
clear mom [PAUSE] that generation it might be easier [PAUSE] to find a job [PAUSE] %HESITATION calls and it is now it is for people going to college so called hard for a lot more things it was [PAUSE] new college can also be a good experience for [PAUSE] learning about [PAUSE] you want to do [PAUSE] or not %HESITATION [PAUSE] learning how to be on your own without being totally alone its kind of like an ice cream store until late [PAUSE] regular adult wise because you still

kind of have a safety net here a lot more independent than the high school so sure the social reason and for some of the opportunities that or do you call it might still be worth it okay
%HESITATION really %HESITATION well what exam what are you thinking about doing after school [PAUSE] does it online at all with the AP exam that youre gonna take like computer science
what do you what kind of job do you want to get your high school if youre not

Table 10: False Negative Predictions Also Coded as Not Confused by a Human Rater

The sample of 20 items inspected shows agreement between the coding scheme and the classifier. A researcher using the coding scheme determined all responses in Table 10 to be not confused, therefore achieving the same results as the classifier. Since the classifier seems to give the same labels as the coding scheme would, I instead focus on the user construct of confusion as a possible source of the false negative errors.

There are a few possible reasons that users may have labeled these responses as confused, even though the coding scheme presented in this paper would identify them as not confused. First, the users who labeled this data may have used more than just the text transcriptions to code the responses as confused or not confused. Since the users completed a Teacher Moments simulation and then coded their own responses as confused or not confused, they had additional insight into how they felt during the simulation. When coding their responses, the users may have remembered how they felt while answering that question, even if none of those feelings manifested in the textual responses. Users may also have listened to the audio clips of their responses, and judged their intonation to determine whether or not their response sounded confused. The coding scheme only uses text transcriptions of the audio clips, so it cannot rely on intonation to inform its labels. However, pauses and hesitations are represented in the text transcripts. Secondly, the users may have been unreliable judges of confusion. Users may have identified aspects of their responses that they believed indicated confusion, but actually were excluded from the construct of confusion for the purposes of this paper. Since no data was collected about how users decided on confusion labels for their responses, it's impossible to understand their thought process while coding their responses. Lastly, the coding scheme may under-represent the construct of confusion. Perhaps the users had a broader definition of confusion that the coding scheme did not completely capture. Data that asks users why they labeled each response as confused or not confused would help clarify what elements of confusion may be missing from the researcher definition.

In conclusion, further data is required to fully understand the cause of the false negative predictions. A researcher using the coding scheme identified these responses as not confused, while users identified these responses as indicating confusion. Given this evidence, I argue that the definition of confusion set forth by the coding scheme has construct underrepresentation. This means that the coding scheme does not fully capture the construct of confusion. In Chapter 6, I discuss what data can be collected in order to refine the coding scheme.

4.1.4 False Positive Predictions

The other errors committed by the classifier are false positive predictions. These occur when a user has labeled data as not confused, but the classifier labels that data as showing confusion. As can be seen in the confusion matrix in Table 8, the classifier makes 32 false positives, meaning that the classifier incorrectly predicts 32 responses to be confused when they are actually labeled as not confused by the user. Using the coding scheme, I coded the 32 false positive responses myself to see whether the cause of the false positives originates with the algorithm's limitation to predict confusion. A sample of 20 false positives is shown in Table 11.

I didnt want to hold you back [PAUSE] just
I think maybe [PAUSE] listen to the student more and Im sure they would have a more dynamic response [PAUSE] important
Id imagine I I like to go about this like really respectful you know they have a lot of empathy for towards her hopefully we can connect [PAUSE] Amanda she might be unwilling it firstly Im Im sure will be extra couple talking to teachers out there I think we got the same page especially if like the setting isnt from the people
first of all to be recognized sake [PAUSE] whatever your I. me as soon maybe the problem may not be the problem itself son opening creating a space where she can feel comfortable to express what the true issue is %HESITATION is the first step [PAUSE] just creating a space theres a reason why she doesnt feel comfortable talking to me in the first place so [PAUSE] creating a space that she can feel [PAUSE] capable of expressing ourselves
Emma doesnt really have any confidence [PAUSE] she [PAUSE] shes also really intelligent but [PAUSE] she she kind of your worries and group even %HESITATION you know shes a hard working student %HESITATION and she doesnt like the extra attention on her from the teacher
college can definitely be a scary thing you havent seen in years [PAUSE] but theres definitely no reason you should [PAUSE] is there something
okay well what do you think that you would like to do [PAUSE] once you get out of
is there a reason why you dont want to take the AP exam [PAUSE] I know that youve been doing really well in my class and I think if the exam is a great fit for you
with this conversation I can understand why shes [PAUSE] money might be has done [PAUSE] taking [PAUSE] and my [PAUSE] theres anything you
so I wanted to understand like why [PAUSE] no I understand a lot more about a month you know she was really more interested in is like no [PAUSE] like why would I dont know I mean [PAUSE] I have

a desire to keep [PAUSE] I dont know if thats like so much Michael understanding where shes coming from
whats going on [PAUSE] I dont really understand why you would feel this way the mind tell me why
lacking confidence and shes been getting these are only some eighties action right thing which is going to fail the exam or not do as well she wants to [PAUSE] she may be worried about the implications of that [PAUSE] or just stress about the exam it doesnt really understand why she should be doing if she doesnt [PAUSE] your mission of warning me because you feel [PAUSE] she doesnt want to take
the new shows to be really [PAUSE] try to be
schools [PAUSE] you should
youre probably have different reasons for not wanting to take [PAUSE] listening
I see that such as soon my so I was off really quiet [PAUSE] top political is coming [PAUSE] I honestly care promises and on more than just the academic [PAUSE] it also has [PAUSE] global grow and learn [PAUSE] theres a lot of [PAUSE] these %HESITATION let me in and more about you know [PAUSE] if you are what you do
only only
even though [PAUSE] you shall [PAUSE] your [PAUSE] your [PAUSE] I have a really
really hands on this
right

Table 11: Sample of False Positive Predictions

Of the 32 false positives that a I coded with the coding scheme, I coded 14 as showing confusion (44%). Since these responses were labeled as not confused by the users, and using the coding scheme, I labeled 44% of them as confused, there seems to be a discrepancy between the definition of confusion set forth in the coding scheme and the users' construct of confusion. As mentioned in section 4.1.4, the users had access to audio clips and personal insight when classifying their responses. To better understand why the coding scheme did not produce the same results as the user labels, more information about the users' thought process while coding data is needed. Gathering this information is discussed in Chapter 6.

In addition to a possible discrepancy between users' construct of confusion and the researcher construct of confusion, there is also an apparent difference between human coder predictions and classifier predictions. A human coder using the coding scheme coded 44% of these responses as being confused, while the classifier considered all of them to be confused. One possible reason for this difference is how the classifier handles questions in responses. As can be seen in the bolded responses in Table 11, some of the responses ask questions posed at the student in the simulation. The coding scheme only considers clarifying questions posed about the simulation as confusion. Questions posed to simulated students are not included in the coding scheme for confusion, and ideally should not be classified as confused by the classifier.

However, I noticed that a number of false positive responses that were classified as not confused by the human coder were questions posed to simulated students. As indicated by the existence of questions posed to students that get classified as confused, the classifier seems to have difficulty differentiating between questions posed about a simulation and questions posed to simulated students. This is perhaps because all three types of Teacher Moments prompts – anticipatory, in-simulation, and reflection – are used together in the datasets. Considering responses to these three types of prompts separately may help the classifier differentiate between questions posed to students and questions posed about the simulation. Questions posed to students would only occur during the simulation, and not during the anticipatory and refection phases. Training classifiers on the three types of prompts separately could therefore distinguish between different types of questions and reduce the number of false positives. This is discussed more in Chapter 6.

In addition to the 32 false positives, the classifier correctly classifies 14 responses as being confused (true positives). In total, the classifier predicts 32 + 14 = 46 responses as being confused. The number of responses labeled as confused by users equals 14 (true positives) plus 34 (false negatives), for a total of 48 user labeled confused responses. Of these 48 responses that have been labeled by users as confused, the classifier only predicts 14 correctly. However, the frequency at which the classifier predicts confusion (46 responses) is similar to the frequency that confusion occurs in the dataset (48 responses). This similarity in frequency implies that although the classifier may not predict the correct responses as showing confusion, it predicts a fairly accurate percentage of responses as showing confusion. Therefore, the classifier may be more useful in certain applications and uses than others.

In terms of use cases for the confusion measurement tool, the existence of false positive errors has some implications for how a teacher educator may use the tool. In a scenario where a teacher educator uses the current classifier to predict confusion for a single piece of response data, the effect of false positive predictions would be more impactful. As is shown in the confusion matrix in Table 8, there is a non-negligible chance that the classifier predicts false positives. In this scenario, since the teacher educator is considering only one response on an individual basis, the incorrect prediction would give the teacher educator the wrong interpretation of the data.

In a scenario where a teacher educator uses the classifier to assess the confusion level an entire classroom, with many pieces of response data, the effect of false positive predictions is

less noticeable. In this scenario, the teacher educator is only wanting to measure the confusion of a large set of responses collectively, rather than individual responses. Since the classifier predicts confusion with the correct frequency, the classifier is likely to predict the correct level of confusion for a large collection of responses, even though it may not correctly identify which of those individual responses display confusion.

4.2 Usability Measures

4.2.1 Confusion Measurement Tool Use Cases

To answer research question 2, qualitative data was gathered from an interview with a teacher educator and from a playtest. Prior to the interview, the teacher educator was given a few days to test out the confusion measurement tool on Teacher Moments response data from her own classroom of students. Then, the teacher educator was asked how useful the tool was for reviewing Teacher Moments response data.

The teacher educator had reservations about relying on the confusion measurement tool to predict confusion in her classroom. She mentioned that she felt unfamiliar with how the tool worked, and therefore did not want to rely on the confusion prediction values. She mentioned that if she had used the tool previously, she would feel more comfortable making sense of the predictions and incorporating them into her teaching plans. Overall, she did not find the results of the confusion measurement tool to be very trustworthy. Had she been more comfortable trusting the confusion prediction values, she said she might have used them to group more confused students with less confused students in future classroom activities. She also may have met with more confused students separately to discuss why they may be feeling confused.

The teacher educator had a few suggestions for additions and modifications to the tool, which would help her find it more trustworthy. These modifications will be discussed in Chapter 6.

Information from the A/B playtest also conveys information about how teacher educators would use the confusion measurement tool and what additional value can be gained from it. During the A/B playtest, participants were asked to answer questions based on response data in the Teacher Moments dashboard. Seven participants were given access to the tool, and five

participants were not given access. Of the seven participants with access to the tool, not all of the participants chose to use it. However, the responses of participants who did use the tool illuminate how a teacher educator might use it.

Participants in the A/B test were asked to assess how confused some student response data seemed. For these tasks, some participants with access to the confusion measurement tool chose to make use of it. One participant used the confusion measurement tool to gain an initial sense of how confused a student seemed. Then, the participant read through the student's responses, and decided to modify the original confusion rating from the confusion measurement tool. Another participant also used the confusion measurement tool to get an initial confusion rating for a student. Then the participant listened to some audio clips and agreed with the rating from the confusion measurement tool. These examples demonstrate use of the tool combined with personal opinion to reach a conclusion about student confusion. A third participant relied solely on the confusion measurement tool to assess student confusion. This participant did not read any response data or listen to any audio clips. This participant noted that she felt there was an overwhelming amount of response data. By using only the confusion measurement tool, she could more quickly understand confusion levels in the response data.

4.2.2 Timing Implications of Confusion Measurement Tool

Timing measurements recorded during the A/B playtest help answer research question 3. Each task in the survey had a corresponding hidden timer which recorded how long the participant took to answer the question. However, some participants answered questions more thoroughly and correctly than others, so the answers to the questions must also be considered when looking at timing results.

The participants were asked to rate a given student's confusion level between 0 (not confused at all) to 100 (very confused). Participants with access to the features had the option to use the confusion measurement tool to complete this task, but not all of the participants chose to use it. The confusion measurement tool returned a confusion rating of approximately 7% for the given student. The average of the confusion ratings given by the group without the tools was 27.6%, and the average was 23.4% for the group with tools. Although these confusion ratings are fairly similar, the group with access to the tools seemed to be more confident in their responses. When asked to rank their confidence in their confusion rating between 1 (not confident at all) to

7 (very confident), the group with tools had an average of 5.3/7 confidence while the group without tools had an average of 3.2/7 confidence. The group with access to the tools also took a larger median amount of time answer the confusion rating question. A box and whisker plot is shown in Figure 10.

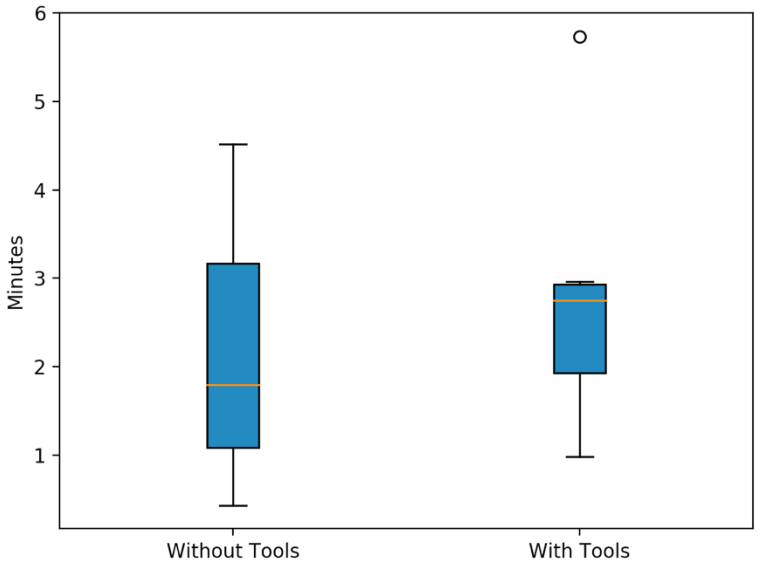

Figure 10: A/B Test Results for Time Taken to Rate Confusion of One Teacher Candidate

However, only two of the seven participants with access to the confusion measurement tool indicated that they used the tool for this question. Those two participants answered the question in 1 and 3 minutes. The participant who use the confusion measurement tool and answered the question in 3 minutes noted that she also read through the responses in addition to using the tool.

Then, participants were asked to choose which of the nine students displayed in the dashboard seemed most confused. The results to this question are especially interesting. While only two out of seven participants with access to the tools chose to use the tools for rating one student's confusion, four out of seven participants used the tool for this question which involved looking at all students' responses. Of the four participants who made use of the confusion measurement tool, three also read some responses or listened to audio clips to support their answer. Perhaps when tasked with looking at so much data, more participants felt the need to rely on the tool rather than reading through all responses. Additionally, participants with access

to the tool took a significantly longer time to answer the question. A box and whisker plot of these timing records is shown in Figure 11.

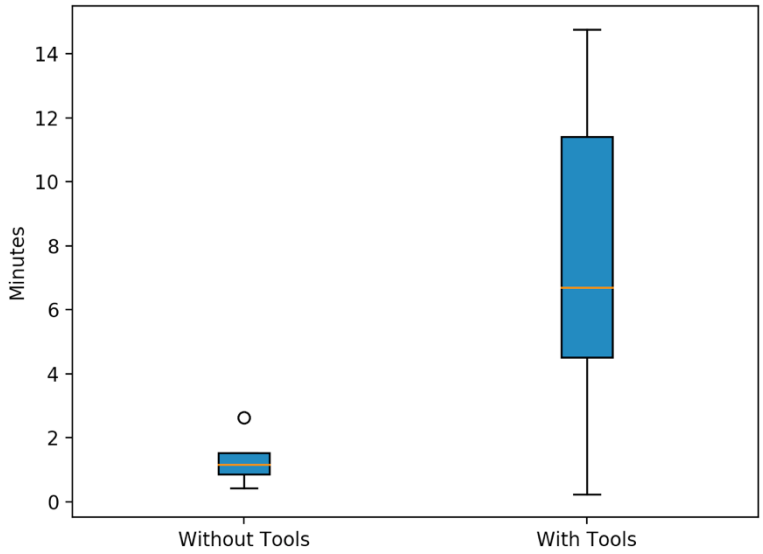

Figure 11: A/B Test Results for Time Taken Choose Most Confused Teacher Candidate

These results seem contradictory to the expectation that having access to the confusion measurement tool would help participants more quickly measure confusion. The explanations of participants about how they arrived at their answers to this question helps explain this trend. Three out of five participants without access to the tools either did not give an explanation for their answer, or noted that the data seemed to large to work with. Considering that this was the last question in the survey, I think the group of participants without the tools may have been mentally fatigued. They may have felt daunted by a task asking them to analyze such a large set of data. Facing a difficult question without any tools to assist them, these participants may have put less effort into answering it. Meanwhile, the participants with access to the confusion measurement tool took longer on average because they felt less daunted by the task. With access to the confusion measurement tool, these participants may have felt that the task was doable, and may have put in a greater amount of effort since they saw a clear approach to arrive at an answer. With the confusion measurement tool, participants made more effective use of their time. Rather than being daunted by a large amount of data, they used the tool to dissect and analyze the data. Participants with the tool gave more elaborate explanations of why they chose a certain student as most confused, indicating that they were able to use the tool to aid their decision process as

they analyzed the data. Unlike the group without the confusion measurement tool, the group with the tool seemed to spend their time analyzing the responses instead of giving up. Although the users with the tool took longer, they seemed to use their time more effectively to identify confused students.

Chapter 5
Discussion

5.1 Summary of Findings

In terms of research question 1, I found that a confusion measurement tool for Teacher Moments can detect confusion with at least more overall accuracy than a random baseline model. For detecting no confusion, the developed classifier achieves a much higher F-score than the random baseline, but for predicting confusion, the classifier performs significantly worse than the random baseline model. False negative errors – when the classifier predicts a response to be not confused even though it has a ground truth of confused – primarily explain the low accuracy when detecting responses as confused. In order to reduce the number of false negative errors, the coding scheme may need to be modified. Ways to reduce the number of false negative errors, and consequences of doing so, are discussed in Chapter 6.

In terms of research question 2, interviews with teacher educators and the A/B test showed examples of how the confusion measurement tool may be used in a classroom. Participants in the A/B test used the classifier in addition to reading response data to get an overall sense of how confused students were. Some participants relied solely on the confusion classifier to identify confused students, while some participants had difficulty trusting the classifier. The interviewed teacher educator also had difficulty trusting the classifier, and was reluctant to incorporate the output of the classifier into her classroom activities as a result. In the future, modifications can be made to the classifier as it appears in the Teacher Moments dashboard to make it more trustworthy. These alterations are discussed in Chapter 6.

To answer research question 3, participants in the A/B test were timed as they answered questions involving the use of the confusion measurement tool. Based on the amount of time taken to answer the questions, and the explanations given by participants, it seems that the confusion measurement tool allows teacher educators to more effectively manage their time analyzing large amounts of data to understand which teacher candidates show signs of confusion. Those who used the tool often used it in conjunction with reading some responses manually in order to determine which teacher candidates were confused. In short answer questions, these participants detailed their thought process of using the tool and reading responses. Those without

the tool seemed daunted by the amount of data, did not take much time to submit an answer, and gave little or no explanation to their thought process. Overall, it seems that having access to the tool allows users of the Teacher Moments dashboard to feel equipped to search for confusion amongst a large amount of data. With the confusion measurement tool, users were more likely to spend their time effectively, by using the tool and reading responses to detect confused students.

5.2 Algorithmic Bias and Usability Recommendations

One important factor in developing a sentiment analysis model that I did not incorporate into this project is possible algorithmic bias. Since the confusion model has been trained on audio transcripts, a users' accents may impact the transcription results and therefore the confusion rating. Additionally, hesitations and pauses, which are incorporated into the audio transcriptions, may be more prevalent in some cultures than others. A model such as those developed in this paper may be prone to cultural, racial, or socioeconomic bias. Any further development of the confusion measurement tool should include measures to prevent any such biases.

While errors and biases in the model should be minimized, teacher educators can also use the confusion measurement tool in a way that further minimizes any errors or biases. Every model has a chance for error, either by leaving confused students unidentified or by misidentifying confusion in students who are not confused. To counteract these errors in the confusion measurement tool, the teacher educator should use the tool in conjunction with her own opinions. In most teacher education scenarios, a teacher educator will personally know her teacher candidates, and will be able to form her own opinions on which teacher candidates seem confused by coursework. By reconciling the results of the confusion measurement tool with her own opinions on her students' performance, a teacher educator can alleviate any errors in the confusion measurement tool and more effectively identify confusion in her classroom.

Chapter 6
Future Work

6.1 Coding Scheme Modifications

One way to reduce the number of false negative and false positive errors committed by the classifier is to consider responses to anticipatory, in-simulation, and reflection prompts separately. In the current classifier, there is only one coding scheme used for all responses, and no distinction between responses to the three types of prompts. In future work, three new coding schemes could be made that represent confusion in each of the three types of prompts. Then, three classifiers can be trained based on each of the three coding schemes.

As described in Chapter 4, false negatives likely arise due to construct underrepresentation in the current coding scheme for confusion. This means that users' construct of confusion seems to include some aspects that do not appear in the construct of confusion represented in the coding scheme. By making coding schemes for anticipatory, in-simulation, and reflection responses separately, the construct of confusion can be narrowed and specialized for each type of response. This way, the researcher constructs of confusion may be more likely to match the users' constructs of confusion, since the activity in which confusion may occur is more constrained and defined to one of three particular tasks.

One current source of false positive errors is that the classifier has trouble distinguishing between questions posed to students in simulation, and questions posed about the simulation. Creating three coding schemes for the anticipatory, in-simulation, and reflection prompts would help alleviate this issue. Questions posed to students will only occur during the simulation, so these types of questions can be disregarded as signs of confusion, whereas questions during the anticipatory and reflection phase may be considered signs of confusion. Correcting the misinterpretation of different types of questions may reduce the number of false positives.

6.1.1 Gathering Data for New Coding Scheme

In order to create coding schemes that better match the users' construct of confusion, their construct of confusion must be better understood. To better understand their construct of

confusion, another user labeled data set could be gathered. While labeling their responses as confused or not confused, users should not be given access to audio clips. Rather, they should only be able to see the text transcripts of their responses. This way, users will not be able to use intonation as a marker for confusion. Users should label responses to anticipatory, in-simulation, and reflection responses. After users label each response, the survey should also ask questions to understand their reasoning. Two such questions could be "What about this response made you label it as confused?" and "What information did you use to decide on a confusion label?". The answers to these questions will help researchers identify what elements of confusion the user has identified in their response, and how those elements may differ between anticipatory, in-simulation, and reflection prompts. These elements of confusion can then be considered and added to the three new coding schemes.

In this project, approximately 500 responses from all three types of prompts were coded. With the three new coding schemes, more data will be required so that each coding scheme can be used on about 500 responses. Therefore, a total of approximately 1500 responses, with about 500 responses of each prompt type, will be needed. Hopefully, higher inter-rater reliability measures will be achieved as a result of using three coding schemes. Three classifiers can then be trained and tested for each of the anticipatory, in-simulation, and reflection datasets. The goal of developing three coding schemes is to achieve higher F-scores for the confusion class, since that is the current classifier's accuracy measure most in need of improvement. Higher F-scores for the confusion class would make the classifiers more effective for use in Teacher Moments. Once the F-scores for the classifiers are all above F-scores for the random baseline, the coding schemes could become applicable to a wider variety of teacher education applications.

6.2 Usability Improvements

Currently, the biggest usability issue with the confusion measurement tool is the lack of context for the confusion prediction. The interviewed teacher educator and A/B test participants had difficulty trusting the confusion prediction, since they had little context for what the prediction meant. They felt that a confusion rating for a particular student may be hard to make sense of without any other confusion ratings to put that number in context. To solve this reliability issue in the future, a few different approaches could be taken. The confusion prediction value for a given student could be presented in the context of the confusion levels for

the entire class. For example, if the classifier rates a given student as less confused than the average rating for the entire class, the classifier could simply report that the student is less confused than average. This approach puts an individual's confusion rating in the context of a larger group's confusion rating, which may help make the individual's rating more meaningful. Additionally, confusion measurements could be calculated over time, throughout different response data sets. For example, if a teacher candidate completes five different Teacher Moments scenarios throughout a semester-long class, her confusion could be displayed for each of those five scenarios. The teacher educator could see how the teacher candidate has progressed throughout the course. The teacher educator could also compare whether a teacher candidate struggles more in certain simulations than others. By comparing confusion ratings from different scenarios, the individual confusion ratings are put in the context of other ratings, and therefore become more meaningful.

Appendix A

Survey for A/B Test

Question	Response Type
How many students participated in the Teacher Moments simulation?	Short answer
How many response(s) indicate "surprise"? If you don't think any responses indicate surprise, please put 0.	Short answer
Copy and paste the response text of an example response that you think indicates surprise. If you don't think any responses indicate surprise, please just put N/A.	Short answer
How did you determine that this response indicates surprise? If you don't think any responses indicate surprise, please just put N/A.	Short answer
Did “student_email” seem confused in this simulation? Please indicate how confused you think “student_email” was during the simulation on a scale from 0 (not at all confused by the simulation) to 100 (very confused by the simulation).	Sliding scale from 0 to 100
You provided a number between 0 and 100 to rate “student_email” s confusion. How confident are you in the rating? 1 is not confident at all, 7 is very confident.	Sliding scale from 1 to 7
How did you determine how confused “student_email” was during the simulation?	Short answer
Which of the students appears to be the most confused? Please choose the email of the student you think was most confused.	Dropdown list of all student emails who participated in the simulation
Why did you think that student was most confused? How did you use the interface to find evidence of confusion? When considering the evidence of confusion, how did you evaluate it?	Short answer

References

[1] Cook, Jennifer. "'Coming Into My Own as a Teacher': Identity, Disequilibrium, and the First Year of Teaching." *The New Educator*, 2009, doi:10.1080/1547688X.2009.10399580.

[2] McLeod, Saul. "Cognitive Dissonance." *Simply Psychology*, 2008, www.simplypsychology.org/cognitive-dissonance.html.

[3] Dieker, Lisa, et al. "The Potential of Simulated Environments in Teacher Education: Current and Future Possibilities." *The Journal of the Teacher Education Division of the Council for Exceptional Children*, Feb. 2014, doi:10.1177/0888406413512683.

[4] https://www.mursion.com

[5] Boll, Carol. "Research Profile: Clinical Simulations Put Future Teachers to the Test." *Syracuse University News*, Apr. 2018, news.syr.edu/blog/2018/04/25/research-profile-clinical-simulations-put-future-teachers-to-the-test/.

[6] Thompson, Meredith, et al. "Teacher Moments: An Online Platform for Preservice Teachers to Practice Parent-teacher Conversations." SocArXiv, 19 June 2018.

[7] Kort, Barry, et al. "External Representation of Learning Process and Domain Knowledge: Affective State as a Determinate of Its Structure and Function." *Workshop on Artificial Intelligence in Education*, May 2001, affect.media.mit.edu/projectpages/lc/AI-ED.PDF.

[8] Craig, Scotty, et al. "Affect and Learning: an Exploratory Look into the Role of Affect in Learning with AutoTutor." Journal of Educational Media, vol. 29, Oct. 2004, doi:10.1080/1358165042000283101.

[9] Rodrigo, Mercedes, et al. "Affective and Behavioral Predictors of Novice Programmer Achievement." *ACM SIGCSE Bulletin*, Aug. 2009, doi:10.1145/1595496.1562929.

[10] Altrabsheh, Nabeela, et al. "SA-E: Sentiment Analysis for Education." *Frontiers in Artificial Intelligence and Applications*, June 2013, doi:10.3233/978-1-61499-264-6-353.

[11] Rosé, Carolyn, et al. "Sentiment Analysis in MOOC Discussion Forums: What Does It Tell Us?" *ResearchGate*, Jan. 2014, www.researchgate.net/publication/264080975_Sentiment_analysis_in_MOOC_discussion_forums_What_does_it_tell_us.

[12] Jokinen, Kristiina, and Jens Allwood. "Hesitation in Intercultural Communication: Some Observations on Interpreting Shoulder Shrugging." 2010, sskkii.gu.se/jens/publications/bfiles/B91.pdf.

[13] Cetintas, Suleyman, et al. "Microblogging in a Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging-Supported Classroom." *Transactions on Learning Technologies* , Dec. 2011, doi:10.1109/TLT.2011.14.

[14] https://www.nltk.org

[15] Ghag, Kranti Vithal, and Ketan Shah. "Comparative Analysis of Effect of Stopwords Removal on Sentiment Classification." *IEEE International Conference on Computer, Communication and Control*, Sept. 2015, doi:10.1109/IC4.2015.7375527.

[16] https://docs.python.org/2/library/string.html

[17] Pang, Bo, et al. “Thumbs up? Sentiment Classification Using Machine Learning Techniques.” *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, July 2002, doi:10.3115/1118693.1118704.
[18] Dave, Kushal, et al. “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews.” *Proceedings of the 12th International Conference on World Wide Web*, May 2003, doi:10.1145/775152.775226.
[19] Jozefowicz, Rafal, et al. “Exploring the Limits of Language Modeling.” *Computing Research Repository*, 2016, arxiv.org/pdf/1602.02410.pdf.
[20] Mullen, Tony, and Nigel Collier. “Sentiment Analysis Using Support Vector Machines with Diverse Information Sources.” *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Jan. 2004, www.aclweb.org/anthology/W04-3253.
[21] Hamdan, Hussam, et al. “Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis.” *SemEval@NAACL-HLT*, 2015, doi:10.18653/v1/S15-2128.
[22] Calvo, Rafael, and Sunghwan Mac Kim. “Emotions in Text: Dimensional and Categorical Models .” *Computational Intelligence* , vol. 29, Jan. 2012, doi:10.1111/j.1467-8640.2012.00456.