ANALYSIS OF COLLABORATIVE ARGUMENTATION IN TEXT-BASED CLASSROOM DISCUSSIONS

by

Luca Lugini

B.S. in Computer Engineering, University of L'Aquila, 2012

M.S. in Computer Science, West Virginia University, 2014

Submitted to the Graduate Faculty of

the Computer Science Department in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

COMPUTER SCIENCE DEPARTMENT

This dissertation was presented

by

Luca Lugini

It was defended on

July 16th 2020

and approved by

Dr. Diane Litman, Department of Computer Science

Dr. Adriana Kovashka, Department of Computer Science

Dr. Erin Walker, Department of Computer Science

Dr. Kevin Ashley, School of Law

Dissertation Director: Dr. Diane Litman, Department of Computer Science

ANALYSIS OF COLLABORATIVE ARGUMENTATION IN TEXT-BASED CLASSROOM DISCUSSIONS

Luca Lugini, PhD

University of Pittsburgh, 2021

Collaborative argumentation can be defined as the process of building evidence-based, reasoned knowledge through dialogue and it is the foundation for text-based, student-centered classroom discussions. Previous studies for analyzing classroom discussions, however, have not focused on the actual content of student talk. In this thesis, we develop a framework for analyzing student talk in multi-party, text-based classroom discussions to understand how students interact and collaboratively build arguments. The proposed framework will simultaneously consider multiple features, namely argumentation, specificity and collaboration. We additionally propose computational models to investigate three aspects: 1) automatically predicting specificity; 2) automatically predicting argument components, and investigating the importance of speaker-dependent context; 3) using multi-task learning to jointly predict all aspects of student talk and improve reliability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

When asking other students for advise on whether to apply to a Ph.D. program the most common feedback I received was about it being like a roller coaster. The past few years have indeed been like a roller coaster, but the excitement of the high moments made it possible to overcome the low ones. I will always treasure my time at PITT, as it gave me the opportunity to explore several fields and eventually decide to specialize in Natural Language Processing.

I would first like to thank my wife, Kayenat Hamid. Her support, encouragement and understanding have been a significant part of me being able to complete this program. I am greatful for the unconditional support I received from my parents, Domenico Lugini and Immacolata Del Busso, and my brother, Franco Lugini. I'm also greatful for my family in the United States, and my Pittsburgh family. The times we spent together are some of my most precious memories.

Special thanks to my advisor, Diane Litman. Her guidance over the years was fundamental in developing research and analytical thinking skills needed for the Ph.D. program as well as future jobs. Among the main things she taught me was how to focus my curiosity towards a problem and formulate research questions so that I could then carry out the right experiments to get answers.

I would also like to thank my committee members: Adriana Kovashka, Erin Walker and Kevin Ashley. Their feedback was invaluable for better contextualizing my work and improving the quality of this thesis.

I was lucky to work on a main project over several years and I want to express my appreciation for the Discussion Tracker team: Diane Litman, Amanda Godley, Christopher Olshefski, Ravneet Singh and undergraduate students/mentees. Over time our work together taught me useful collaboration and interpersonal skills that I will carry with me.

Lastly, I would like to recognize friends, labmates and fellow graduate students at PITT. I have many fond memories of working late nights on projects or having discussions on new NLP approaches to try on our respective problems.

1.0 INTRODUCTION

Classroom discussions are regarded as one of the most effective pedagogical approaches for enhancing student skills. Extensive research in the educational community has shown that reasoning, reading, and writing skills can be positively affected by high-quality student-centered classroom discussions in English Language Arts (ELA) classrooms [Reznitskaya and Gregory, 2013, Graham and Perin, 2007, Applebee et al., 2003]. High quality discussions encourage student-to-student talk, negotiation of claims, supporting claims with evidence, and reasoning about those claims. Although the effectiveness of particular kinds of claims, evidence and reasoning can vary across disciplines, Chisholm and Godley [Chisholm and Godley, 2011] and Lee [Lee, 2006] showed that the specificity of these argument moves is related to discussion quality. Student-centered discussions and elaborated student talk during collaborative argumentation are also recognized as indicators of learning opportunities in different academic disciplines [Grossman et al., 2014, NGA & CSSO, 2010]. During classroom discussions in which questions are open-ended (i.e. with multiple possible right answers) working on collaborative argumentation gives students the opportunity to build disciplinary knowledge and learn how to critically think in a disciplinary way [Engle and Conant, 2002, Reznitskaya and Gregory, 2013]. There are several advantages of engaging in disciplinary argumentation collaboratively in a group rather than individually: students obtain a better understanding of the process of building and evolving disciplinary knowledge by considering multiple viewpoints, supporting and challenging each other's ideas [Engle and Conant, 2002]; they have the ability to return to previous ideas at any point in time to re-examine and improve them [Hong and Scardamalia, 2014]; they gain more complex disciplinary expertise by formulating, challenging and supporting different interpretations of an idea with evidence and reasoning.

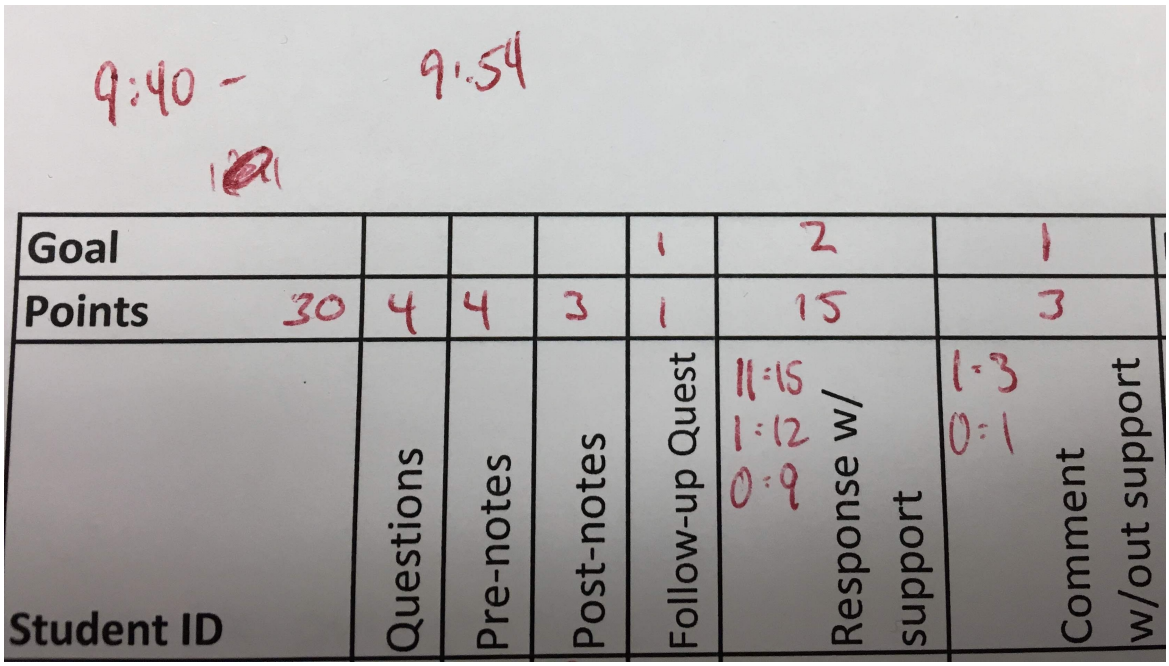


Figure 1: Rubric used by a teacher in grading a classroom discussion.

From the instructor's point of view, however, many teachers struggle to develop the skill necessary for effectively teaching collaborative argumentation [Lampert et al., 2010]. During a discussion a teacher must perform multiple tasks: manage the discussion, make sure all instructional goals are accomplished, keeping track of which students speak and how much they speak, all the while paying attention to the content and important features of students' talk and possibly reacting with planned interventions. As if this was not enough, usually no records exist after the discussion ends, so it is up to each teacher to recall what happened during the discussion in order to grade individual students and plan for the next discussion. Figure 1 shows the rubric used by a high school teacher when grading students in a classroom discussion. The figure shows that grades have a component based on homework assignment (questions and pre-notes) and one for students' reflections on the classroom discussion (post-notes). The rest of the grade comes from the students' contributions during the discussion: follow-up questions to other students' statements, and responses to other students' statements with particular attention on whether they provide support or not. The

teacher, therefore, needs to process each utterance to understand whether it poses a follow up question to a previous utterance, if it is related to a previous utterance at all, and if support for the student's statement is given. Since the teacher is also in charge of managing the discussion, at any point in time they may need to intervene with probes (e.g. asking for support in case it is not provided by a student or facilitating collaboration between students) and to make sure the discussion stays on topic. Effectively understanding and orchestrating collaborative argumentation requires advanced skills which are difficult to develop [Lampert et al., 2010].

Natural Language Processing (NLP) techniques have the potential of alleviating this problem by automatically analyzing classroom discussions and providing teachers with feedback on how to improve collaborative argumentation overall in the classroom. At the same time, students could also benefit from individual feedback on their contributions during discussions. While there is an increase in deploying automated tools in the classroom, prior research has not focused extensively on student-centered classroom discussions. Chen et al. [Chen et al., 2014] developed a tool for teacher self-assessment of classroom discussion through analyzing the frequency of participation of students in the discussion and teacher-student turn patterns. Blanchard et al. [Blanchard et al., 2016] proposed a system for detecting teacher questions from classroom discussion recordings. Gerritsen et al. [Gerritsen et al., 2015] developed a system with the goal of providing teaching assistants with feedback on instructor/student talk ratio and wait time. Researchers have also developed systems for identifying different instructional segments from recordings of classroom discussions [Kelly et al., 2018], and detecting dialogic properties of teacher questions [Samei et al., 2014]. Unfortunately, none of these works actually focuses on the content of student talk during discussion, making it unsuitable for analyzing collaborative argumentation. Similarly, prior systems developed for analyzing the content of discourse have not actually focused on data from spoken discussions. The limitations of prior work on collaborative argumentation also stem from the lack of publicly available corpora for multi-party, student-centered classroom discussions with related annotations.

The NLP community has also investigated problems related to the specific aspects of collaborative argumentation which are the focus of this thesis, namely argumentation and

specificity. With respect to former, several works analyzed aspects of argumentation in educationally-oriented domains such as student essays [Persing and Ng, 2015, Stab and Gurevych, 2014, Nguyen and Litman, 2015, Nguyen and Litman, 2016b, Nguyen and Litman, 2016a, Nguyen and Litman, 2018] while others focused on analyzing argumentation in multi-party online dialogues [Swanson et al., 2015, Misra et al., 2015, Habernal and Gurevych, 2017, Niculae et al., 2017]. These works, however, present some limitations when considering collaborative argumentation in classroom discussions: they either adapt a different argumentation scheme or disregard important labels (e.g. warrants) from their analyses, or were developed for performing tasks other than predicting argument components (e.g. analyzing quality of arguments, argument facets, argument strength, argument relations), and none of them focuses on multi-party educational discussions. As for specificity, previous research has mainly addressed the domains of newspaper articles [Louis and Nenkova, 2011, Louis and Nenkova, 2012, Li and Nenkova, 2015, Li et al., 2016, Carlile et al., 2018], and online/social media posts [Gao et al., 2019, Ke et al., 2018, Swanson et al., 2015]. Besides the difference in domain with our work on classroom discussions some of the limitations of the prior works are: they either annotate/predict specificity using a different unit of analysis that does not hold particular meaning in our corpora which derives from spoken discussions (e.g. sentences), or use a continuous numerical value instead of discrete categories; most importantly, none of them considered external resources when annotating/predicting specificity, which is a big limitation since our work is aimed at text-based discussions in which the definition of specificity is related to information about the text.

Overall, perhaps in part due to lack of publicly available corpora, there is a need in the NLP community for computational models specifically focused on analyzing collaborative argumentation and its components with respect to spoken, multi-party, text-based classroom discussions. We try to bridge this gap by presenting an end-to-end framework (from data collection, to manual annotation, to predictive models) that is targeted at helping teachers understand and improve collaborative argumentation in their classroom discussions.

1.1 THESIS STATEMENT

Motivated by the need for developing tools to support teachers and students in classroom discussions, the goal of this research is to develop a framework and computational models for analyzing collaborative argumentation in multi-party, text-based classroom discussions. The research hypotheses to be tested are as follows:

- **H1:** The first hypothesis relates to the effectiveness and reliability of the proposed framework, and is divided into two sub-hypotheses:
  - **H1.1:** The proposed annotation scheme can be used by humans to reliably annotate important aspects of student talk in classroom discussions.
  - **H1.2:** The proposed annotation scheme generates useful information for teachers.
- **H2:** The annotation of features of student talk can be reliably automated. This hypothesis is divided into four sub-hypotheses:
  - **H2.1:** The proposed model for specificity prediction delivers better performance than systems developed for other domains.
  - **H2.2:** Performance of an existing argument mining model can be improved by using features developed for argument mining in online dialogues.
  - **H2.3:** Modeling student-dependent contextual information will improve argument component classification performance.
  - **H2.4:** The proposed approach for modeling contextual information in argument component classification will be effective for multiple neural network models.
- **H3:** A model for jointly predicting all features of student talk using multi-task learning will outperform the individual models trained separately on each of the features.

The models we propose for testing hypotheses H2 and H3 consist of a hybrid combination of a neural network and handcrafted features: we can therefore augment the features extracted automatically by a neural network with the robustness of handcrafted features in a setting with a limited amount of training data.

1.2 CONTRIBUTIONS

For the educational community, we develop an annotation scheme for important aspects of classroom discussions grounded in theory of learning that can reliably be annotated by humans (i.e. argumentation, specificity and collaboration). The annotation scheme provides teachers with actionable information for planning classroom discussions. Additionally, we develop tools using NLP techniques to automatically generate annotations for student talk in classroom discussions. Such techniques can then be incorporated into practical tools to be used in the classroom. Lastly, we develop an analytics system capable of supporting teachers in understanding and fostering collaborative argumentation.

For the argument mining community, we develop new argument component classification models specifically designed for classroom discussions, and analyze the impact of student-dependent contextual information. We also carry out experiments to understand how argument component classification performance is affected by using different types of neural networks and different input granularity.

For the NLP community, we develop a hybrid approach to predict specificity which combines a neural network and handcrafted features in order to train a robust model given the limited amount of training data. We additionally evaluate how our proposed models for argument component classification can take advantage of contextual information. We show that the proposed models for contextual information generalize to other neural network types. Lastly, we perform experiments on simultaneously training models on multiple tasks including argument component classification, specificity, and collaboration. We show that multi-task learning can be used to train robust models, in particular for argument components and collaboration, that outperform the respective individual models.

As additional contribution, we released a dataset of classroom discussion transcripts annotated with the proposed annotation scheme for free use to the research community. An example excerpt of such discussion is given in Table 7.

1.3 OUTLINE

This chapter introduced the motivation and some of the challenges of this thesis. The rest of this thesis is organized as follows: Chapter 2 briefly describes the different datasets used in this work; Chapter 3 outlines the proposed framework for annotating different aspects of classroom discussions; Chapter 4 introduces computational models for automatically predicting specificity of student talk; Chapters 5 and 6 presents work on argument component classification in classroom discussions; Chapter 7 describes joint models for simultaneously predicting multiple important features of classroom discussions; Chapter 8 reports on findings from classroom deployment of an analytics tool, and Chapter 9 summarizes the thesis.

2.0 DATA AND RESOURCES

Since we have been working on developing an annotation scheme and computational models incrementally over the course of multiple years, we have used different datasets for testing research hypotheses. This chapter provides a description of all datasets used in our work. Table 1 shows the annotations available for the datasets and the tasks/chapters for which each is used.

The first dataset, D1, consists of manually transcribed text-based classroom discussions from English Language Arts high school classes. Text-based discussions are about a "text" (e.g., literature such as *Macbeth* and *Memoir of a Geisha*, a news article, a speech, etc.) and can either be mediated by a teacher or conducted exclusively among students. The number of students per discussion ranges from 5 to 13. The dataset was annotated for specificity according to the annotation scheme proposed in Chapter 3. Specificity labels distribution is shown in Table 2.

The dataset spans 23 classroom discussions and over 2000 argument moves. Two pairs of annotators coded specificity for 5 and 9 transcripts respectively, while the remaining 9 transcripts were single-coded. Inter-rater reliability on specificity labels for the two annotator pairs as measured by quadratic-weighted Cohen's Kappa is 0.714 and 0.9, indicating substantial agreement and almost perfect agreement, respectively. A gold standard set of labels for each double-coded discussion was obtained by resolving the disagreements between the two annotators. This dataset was used to develop our specificity prediction models in Chapter 4.

Dataset D2 consists of 73 transcripts of text-based classroom discussions, 5 of which are also present in D1. Some of the transcripts were gathered from published articles and dissertations, while the rest originated from videos which were manually transcribed by one

Table 1: Classroom discussion datasets annotated with number of discussions, number of argument moves, which annotations they contain, inter-rater agreement measures (unweighted Kappa for argumentation and collaboration, quadratic-weighted Kappa for specificity), and a description of their use in this work.

Dataset	**Disc**	**Argument Moves**	**Annotations**	**Cohen Kappa**	**Description**
D1	23	2057	Specificity	0.714, 0.900	Used for predicting specificity in Chapter 4
D2	73	2047	Argument component Specificity	0.629 0.641	Used for argument component classification in Section 5.2 and for preliminary analysis in Chapter 7
D3	29	3135 (2125 turns)	Argument component Specificity Collaboration	0.890 0.700 0.740	Used for argument component classification in Chapter 6 and multi-task learning in Chapter 7
D4	18	1942 (1467 turns)	Argument component Specificity Collaboration	0.971 0.813 0.578	Used for predicting argument components in Section 6.4.5 and in Chapter 8.

Table 2: Dataset D1 statistics.

	Annotation	Total Count	Percentage
Specificity	Low	730	35.49%
	Medium	974	47.35%
	High	353	17.16%
	Total	**2057**	**100.00%**

of our annotators. While detailed demographic information for students participating in each discussion was not available, our dataset consists of a mix of small group (16 out of 73)

Table 3: Distribution of class labels for argument component type and specificity in dataset D2.

	Annotation	Total Count	Percentage
Argumentation	Claims	1034	50.51%
	Evidence	655	32.00%
	Warrants	358	17.49%
	Total	**2047**	**100.00%**
Specificity	Low	710	34.69%
	Medium	996	47.66%
	High	341	16.65%
	Total	**2047**	**100.00%**

versus whole-class (57/73) discussions, both teacher-mediated (64/73) versus student only (9/73). Additionally, the discussions originated in urban schools (28/73), suburban schools (42/73), and schools located in small towns (3/73). After pre-processing (see Section 3.3 for details), we used the framework from Chapter 3 to annotate argument components and specificity. Only student turns were considered for annotations; teacher turns at talk were filtered out and do not appear in the final dataset. The distribution of argument component and specificity labels for D2 are shown in Table 3.

The average number of argument moves among the discussions is 27.3 while the standard deviation is 25.6, which shows a high variability in discussion length. The average number of words per argument move and standard deviation are 22.6 and 22.1, respectively, which also shows large variability in how much students speak. Dataset D2 will be used in Chapter 5 for testing the performance of argument component classification models.

Dataset D3 consists of 29 discussions: three rounds of discussions were recorded for each of the 10 teachers participating in a research study (one discussion was not text-based and therefore omitted), from three different high schools in the Pittsburgh area. Each discussion was 30 to 40 minutes long, with a mix of whole-class and small-group discussions: the number

of students ranged between 6 and 28. In terms of grade/level the classes were composed of three ninth grade, three tenth, two eleventh, and two twelfth grade; six of the classes were honors or AP level and four were "regular" level. A researcher observed each discussion while being recorded, created a map to link numerical student IDs to the location of each student, and took handwritten notes that were later used to align speaker IDs with the discussion transcript.

Each transcript was manually transcribed (by a professional transcription service) and pre-processed by an expert annotator in the same way as transcripts in dataset D2. Dataset D3 was also annotated following the annotation scheme from Chapter 3. The annotated transcript is publicly available for research use at http://discussiontracker.cs.pitt.edu and additional details on data collection and annotation can be found in [Olshefski et al., 2020].

The final D3 dataset consists of 3135 argument moves. The average number of argument moves per discussion is 108.1 and the standard deviation is 30.3, which suggests that discussions in D3 are much longer and with less variability in length compared to the dataset D2. The average number of words per argument move and standard deviation are 34.96 and 25.59 respectively, which shows that argument moves in D3 are generally longer than those in D2, but show a comparable variability. The annotators who provided labels for dataset D3 are the same as the ones used for dataset D2. Inter-rater reliability analyses were carried out by double-coding three discussions (the remaining 26 discussions were single-coded). Annotators achieved 92% agreement when deciding on argumentative vs. non-argumentative turns at talk. Unweighted kappa for argumentation and collaboration annotations was, respectively, 0.89 and 0.74. Quadratic-weighted kappa for specificity annotations was 0.70. Table 4 shows the class distributions for the D3 dataset. Dataset D3 will be used in Chapters 6 and 7.

Dataset D4 consists of 18 discussions collected in the Spring 2020 semester. Each discussion was transcribed using the same professional service used for D3 in order to maintain consistency of transcriptions. The data collection and annotation procedures (including annotators) were the same as in D3. As for D3, a researcher observed each discussion, created a map to link numerical student IDs to the location of each student, and took handwritten notes that were later used to align speaker IDs with the discussion transcript. The main

Table 4: Distribution of class labels for argument component type, specificity and collaboration for dataset D3.

	Annotation	Total Count	Percentage
Collaboration	New	802	37.69%
	Agree	37	1.74%
	Extensions	1015	47.70%
	Challenge	274	12.88%
	Total	**2128**	**100.00%**
Argumentation	Claims	2047	65.30%
	Evidence	762	24.30%
	Warrants	326	10.40%
	Total	**3135**	**100.00%**
Specificity	Low	1189	37.93%
	Medium	1071	34.16%
	High	875	27.91%
	Total	**3135**	**100.00%**

differences between D4 and D3 consist in the size of the dataset, and in the fact that D4 contains one discussion per teacher. Therefore, D4 consists of discussions from 18 teachers (10 of which also contributed to D3) from 4 high schools in the Pittsburgh area (3 of which appear also in D3). The average number of words per argument move is 48.84 and standard deviation is 34.4, which indicates that in general D4 contains longer argument moves than D3 - though not consistently. Inter-annotator reliability was computed on 3 transcripts for argumentation (kappa = 0.971) and specificity (quadratic-weighted kappa = 0.813) and collaboration (kappa = 0.578). D4 will be used in Section 6.4.5 to analyze the cross-dataset performance of argument mining models as well as in Chapter 8. Class label distributions for D4 are shown in Table 5.

Table 5: Distribution of class labels for argument component type, specificity and collaboration for dataset D4.

	Annotation	Total Count	Percentage
Collaboration	New	325	22.15%
	Agree	38	2.59%
	Extensions	790	53.85%
	Challenge	314	21.41%
	Total	**1467**	**100.00%**
Argumentation	Claims	1402	72.19%
	Evidence	345	17.77%
	Warrants	195	10.04%
	Total	**1942**	**100.00%**
Specificity	Low	554	28.53%
	Medium	697	35.89%
	High	691	35.58%
	Total	**1942**	**100.00%**

3.0 ANNOTATION OF CLASSROOM DISCUSSIONS

In this chapter we discuss the design of a framework for annotating student talk in text-based classroom discussions. We developed the annotation scheme with the intent of capturing three important aspects of classroom talk that are theorized to be important with respect to discussion quality and learning opportunities: argumentation (the process of systematically reasoning in support of an idea), specificity (the quality of belonging or relating uniquely to a particular subject), collaboration (relations between different turns in the discussion). This work is illustrated in [Lugini et al., 2018] and [Olshefski et al., 2020] and the contribution is shared with Christopher Olshefski.

3.1 RELATED WORK

Several studies in the educational domain have used argument moves, i.e. students' claims about the text, sharing textual evidence for claims, and reasoning to support claims, as measure of discussion quality [Reznitskaya et al., 2009, Chisholm and Godley, 2011]. Student reasoning in particular is believed to be of primary importance, especially when it is elaborated, highly inferential and based in evidence [Kim, 2014a, McLaren et al., 2010]. In the NLP field the main focus of educationally-oriented argumentation research has been on corpora of student persuasive essays [Ghosh et al., 2016, Klebanov et al., 2016, Persing and Ng, 2016, Wachsmuth et al., 2016, Stab and Gurevych, 2017, Nguyen and Litman, 2018]. In contrast with these previous works, our focus lies in multi-party spoken discussion transcripts from classrooms. Argumentation in online multi-party settings was studied by Swanson et al. [Swanson et al., 2015] and Misra et al.[Misra et al., 2015], however their focus was on

analyzing argument quality or argument facets unlike our work which is aimed at argument component classification. Furthermore, unlike studies in argument mining which only consider claims and premises [Stab and Gurevych, 2014, Stab and Gurevych, 2017, Nguyen and Litman, 2015, Nguyen and Litman, 2018], we include the warrant label as it is important to understand how students explicitly link evidence to claims through their reasoning.

In the educational community, previous studies found that specificity is another factor which impacts the quality of discussions [Chisholm and Godley, 2011, Sohmer et al., 2009] (where discussion quality was manually evaluated by experts based on argumentative structure and sociolinguistic content). Chisholm and Godley found a relation between the increase in specificity and the increase in quality of claims and reasoning. Within the NLP community, previous research analyzed specificity of sentences in the context of professionally-written newspaper articles and its role in their summarization [Li and Nenkova, 2015, Louis and Nenkova, 2011, Louis and Nenkova, 2012]. Li et al. [Li et al., 2016] subsequently improved the annotation scheme used in [Louis and Nenkova, 2011, Li and Nenkova, 2015] by considering contextual information, and by using a scale from 0 to 6 rather than binary specificity annotations. The annotation guidelines used in these studies work well for general purpose corpora [Gao et al., 2019] where the content is not directly related to a single source of information (e.g. a text), however in text-based discussions specificity must capture particular relationships between a discussion and the text it is based on (e.g. does it mention specific characters, does it provide a quote from the text). In order to capture such relationships we define specificity for classroom discussions in terms of a set of characteristics that simultaneously consider an utterance and the text currently discussed. Another difference with previous work is in the unit of analysis: we annotate argumentative discourse units since the concept of sentence is not clearly defined in speech. Carlile et al. [Carlile et al., 2018] also use argumentative discourse unit as unit of analysis for annotating specificity. Their definition of specificity, however, differs between (major) claims and premise, and since they do not include warrants in their annotation guidelines, their annotation scheme cannot be directly applied to our problem.

Prior work in the educational community focused on classroom discourse research analyzed collaboration in terms of accountable student talk and dialogue structure. While

developing guidelines to facilitate productive disciplinary engagement, Engle and Conant [Engle and Conant, 2002] found that accountability for building on others' ideas (integrating other people's contributions into the student's own ideas) and accountability for developing group expertise (sharing individual, personal expertise with other members of the discussion group) are important aspects for the collaborative learning process in the classroom. Keefer et al. [Keefer et al., 2000] analyzed student-centered classroom discussions by identifying different dialogue types within a discussion (e.g. critical discussion, consensus dialogue). They also used a graphical representations to show how students' utterances are connected throughout the discussion. However, these works represent small-scale expert analyses and do not provide actionable ways of annotating classroom discussions to understand if and how students collaborate. In Computer-Supported Collaborative Learning (CSCL), Samei et al. [Samei et al., 2014] evaluated how well humans and machine learning models can identify two dialogic properties of questions in classroom discourse, uptake (i.e. asking a question which is related to a prior statement) and authenticity (i.e. open-ended questions). Zhang et al. [Zhang et al., 2013] developed a classroom analytics system with the goal of improving knowledge co-construction through graphical visualizations of metadiscourse (how the discourse is organized) to identify threads of related ideas. Another construct in CSCL often used to analyze collaborative knowledge construction is transactivity [Gweon et al., 2013], in order to understand if participants in a conversation build on prior exchanges. Recent efforts have also been focused on developing automated models for predicting transactivity that are aimed at achieving state of the art results and improving generalization performance to out-of-domain data [Fiacco and Rosé, 2018]. A transactive exchange has two characteristics: *(i)* it contains reasoning; *(ii)* it references an idea voiced earlier in the discussion. It is therefore intrinsically linked with argumentation. While these related works share our research interest in understanding interactions between students, they are limited in the type of discourse annotated (e.g. only questions, only reasoning) and in how student utterances are connected (e.g. only identifying "build-on" relations, or transactive/non transactive exchanges). In order to fully capture nuanced interactions between students during a classroom discussion, we annotate all collaborative utterances with several distinct labels in addition to the utterance to which they refer. Richey et al. [Richey et al., 2016] collected and annotated the SRI

corpus of middle school, small group discussions of mathematical solutions with the goal of understanding how student talk patterns correlate with collaborative learning. The corpus includes multi-level annotations on collaboration: collaboration indicators (e.g. planning, acknowledging, explaining) and overall collaboration quality ratings (e.g. good collaboration, not collaborating). Though the SRI corpus is similar to the ones used in this thesis, there are some fundamental differences: SRI consists small group discussions (3 students on average) while our datasets consist of a mix of small group and whole-class discussions; SRI consists of audio recordings along with time-stamped annotations, while our datasets consist of written transcriptions with related synchronous annotations; SRI focuses solely on collaboration, whereas all our datasets except D1 include annotations on multiple dimensions of collaborative argumentation.

3.2 ANNOTATION SCHEME

This work uses two unit of analysis: for collaboration we use turns, which consist of a complete utterance; for argumentation and specificity we use argument moves, i.e. an utterance, or part of an utterance, containing a single argumentative discourse unit (ADU) [Peldszus and Stede, 2013]. In this document we use the terms argument move and ADU interchangeably. Table 6 shows the three dimensions of student talk considered in the definition of our annotation scheme: argumentation, specificity and collaboration. The complete annotation manual is provided in Appendix A.

The argumentation scheme is based upon Lee's work [Lee, 2006]. It consists of a reduced set of labels originating from Toulmin's argumentation model [Toulmin, 1958]. Our scheme explicitly specifies that for an argument move to be labeled as warrant it must come after claim and evidence, since by definition warrants cannot exist without these two components. A coded excerpt of a discussion is shown in Table 7. In the first row we can see student St1's turn being segmented into three argument moves. It show a natural expression of an argument: St1 first voices a claim, then provides historical events as evidence, and finally explains how the evidence links to the initial claim.

Table 6: Description of the labels sets in the proposed annotation scheme

Label	Definition
Argumentation	
Claim	An arguable statement that presents a particular interpretation of a text or topic
Evidence	Facts, documentation, text reference, or testimony used to support or justify a claim
Warrant	Reasons explaining how a specific evidence instance supports a specific claim
Specificity	
	Specificity elements: particulars, details, content language, chain of reasoning
Low	Statement that does not contain any of the specificity elements
Medium	Statement that accomplishes one of the specificity elements
High	Statement that accomplishes two or more specificity elements
Collaboration	
New	Expressing a new idea (or concept, or perspective) in the discussion
Extension	Building off a prior idea
Challenge	Challenging or questioning a prior idea
Agree	Expressing almost the exact idea of a prior turn

For specificity, the annotation scheme is based on [Chisholm and Godley, 2011]. The main goal of this scheme is to capture specificity as it relates to text-related characteristics mentioned by students. The three specificity labels in Table 6 stem directly from four elements of an argument move:

1. particulars: argument moves containing at least two terms that are particular (e.g., a person and a setting or a setting and an action) rather than a general group or situation

Table 7: Sample discussion excerpt from a discussion on the play "The Crucible".

Turn	Speaker	Talk	Col	Ref	Arg	Spec
1	St 1	My interpretation of it is that, without a middle ground, you are left with two very extreme points. Whether or not the middle ground directly centered, we have a range. We have a spectrum.[...]	New		Claim	Medium
		Throughout history, whether you go back to ancient Europe, and you look at tyrannies and dictatorships, not even ancient Europe. If you go back to the Holocaust and what Hitler was doing over in Germany [...] if you go back to Communism, as well [...]			Evidence	Medium
		Those are two extremes, and neither of them ended well, and just anarchy there. There is no order there, there is no civilized kind of society to base anything around. I think the middle ground is necessary just to create some kind of spectrum that we can go off of.			Warrant	Medium
2	St 9	I acknowledge your point, but there wasn't nobody going against anything until this happened, until this event occurred.	Challenge	1	Claim	Low
3	St 1	Does that make the way they were living right, thought?	Challenge	2	Claim	Low
4	St 9	If they were happy, I believe they were perfectly fine.	New		Claim	Low
5	St 17	My assessment of the topic at hand is, there needs to be a balance between state rights and user rights. [xx] slide, and to what extent was it off balance.	Extension	1	Claim	Medium

such as "you", "everyone", "books". Clichés or overgeneralizations are not particulars;

2. details: descriptions, explanations or elaborations that make the idea more understandable, contextualized, qualified, substantiated or vivid. ("Detailed" argument moves should avoid or at least explain general terms like good, bad, stupid, i.e., terms with multiple definitions);
3. content language: use of vocabulary or phrases that are specific to English Language Arts (such as "irony", "simile", "tragedy", etc.) or the text being discussed (such as quotes or expressions);
4. chain of reasoning: phrases or clauses that attempt to rationalize, justify or explain an idea(s). They link or synthesize at least two pieces of information or ideas;

Although these four elements were not directly coded by annotators, their definition was helpful in the training stages for achieving higher reliability. The first three argument moves in Table 7 all contain the second specificity element, as they provide definitions or elaborations. However, no content-specific vocabulary, clear chain of reasoning, or particulars are provided; therefore all three moves are labeled as medium specificity. Argument moves 2, 3 and 4 do not provide any specificity element and are labeled as low specificity. In a later argument move (not shown in the excerpt in Table 7) Student 4 says the following: *Back, in their society, they had a Puritan society, so being self serving wasn't accepted, so if you didn't follow God, then it was frowned upon, I have a quote, hold on. On page 198, Hale said, "In the book, a record that Mr. Parris keeps, I note that you are rarely in church on Sabbath day"*. This argument move was labeled as high specificity because it contains content language, a direct quote from the text, as well as details, explaining what they meant by "Puritan society".

The main goal of collaboration is to understand how a particular turn in a discussion relates to prior turns. The specific set of codes was developed through multiple iterations, starting with a set of 9 different collaboration codes based on a synthesis of related work in classroom discourse research [Keefer et al., 2000, Engle and Conant, 2002] and CSCL [Zhang et al., 2013, Gweon et al., 2013, Fiacco and Rosé, 2018, Richey et al., 2016], with particular focus to theory of accountable talk [Michaels et al., 2008, Michaels et al., 2010]. After a first round of annotation we decided to merge/remove ones that either never actually

happened in discussions (e.g. synthesis), or ones that created confusion for the annotators (e.g. rebuttal). Beside the collaboration label, annotators also marked a reference to the specific prior turn in the discussion that the current turn relates to. After analyzing part of collaboration annotations on dataset D3 we found that 95% of turns had a reference within the prior four turns. We therefore decided to include this as a guideline in the codebook, requiring annotators to only code collaboration with respect to the four prior turns (unless the speaker explicitly references an earlier turn). Detailed examples of each collaboration label can be found in Appendix A. In the overall scope of this thesis, collaboration plays a less central role compared to specificity and argumentation. While we extensively explored features and models to automatically predict argument components and specificity for ADUs, less emphasis was placed on collaboration: the main goal for collaboration with respect to the proposed NLP models, is to be incorporated as one of multiple tasks for jointly predicting two or three collaborative argumentation aspects.

The complete coding manual in Appendix A provides a more extensive set of example discussion excerpts with related labels that annotators can use when coding discussions.

3.3 DATA ANNOTATION

The data collection procedure was different for each dataset and is detailed in Chapter 2. Prior to applying the annotation scheme described in the previous section, each transcript was preprocessed using a procedure that was generally the same for all datasets in this study:

1. starting from a transcription of a classroom discussion, an expert annotator augmented each turn with a turn number and the ID of the student who voiced it (using the hand-written notes described in Chapter 2);
2. the annotator marked down non-argumentative turns, where a student utterance did not contain substantive argumentation (e.g. procedural talk, off-topic talk, meta-discourse talk); non-argumentative turns were filtered out when building NLP models in this thesis;
3. the same annotator further segmented argumentative turns from (2) into argument discourse units.

To ensure consistency the same expert annotator performed the preprocessing step for all datasets. After preprocessing, each transcript contained all the elements needed to applying our proposed annotation scheme. For different datasets, then, annotators used the coding manual we developed to label each student turn for collaboration (D3 and D4) and each ADU for argumentation (D2, D3 and D4) and specificity (all datasets).

3.4 RELIABILITY OF THE ANNOTATION SCHEME

In order to assess whether the proposed scheme can be reliably used to annotate classroom discussions, we conducted multiple reliability analyses. We evaluated reliability of distinguishing between argumentative and non-argumentative turns by double coding one transcript from dataset D3. Although this analysis yielded low Cohen kappa (0), raw percentage agreement was high at 92%. The difference between the two results is based on high class imbalance: annotator A only labeled 2 turns as non-argumentative (which B noted as argumentative), while B only marked 1 turn as non-argumentative (which A labeled as argumentative). We then conducted a reliability study on turn segmentation with two annotators on a subset of dataset D2 consisting of 53 transcripts. Our analysis is based on the same metric used by Habernal and Gurevych [Habernal and Gurevych, 2017], where they used Krippendorff unitized alpha (α_U) [Krippendorff, 2004] to evaluate the reliability of identifying argumentative discourse units in user-generated web discourse. Whereas Habernal and Gurevych used Krippendorff α_U to evaluate the task of identifying ADU boundaries and at the same time assigning labels to ADUs, our segmentation task is simpler since it only involves identifying ADU boundaries. We obtained a Krippendorff α_U of 0.952, which shows that turns at talk can be reliably segmented.

In the following sections we will outline the results of reliability analyses on annotations for each dataset used in this thesis.

3.4.1 Reliability Analysis on Dataset D1

The annotators were initially trained by using the annotation scheme to code transcripts one at a time and discussing disagreements at end of each transcript. Then, we used five text-based discussions for testing inter-rater reliability after training. Annotator pair P1 annotated discussions of *The Bluest Eye, Death of a Salesman*, and *Macbeth*. Pair P2 annotated two discussions based on the speech *Ain't I a Woman*. Overall, more than 40 students participated in the discussions, which generated 250 argument moves (consisting of more than 8200 words). Given the ordinal type of specificity labels, quadratic-weighted kappa (qwkappa) was used as measure of inter-rater reliability.

As previously noted in Table 1, the two annotator pairs achieved qwkappa of 0.714 and 0.9, which indicates substantial agreement and almost perfect agreement, respectively [McHugh, 2012]. Upon inspecting the confusion matrix for annotators P2 (similar trends observed for P1) we noticed relatively few low-high label disagreements as compared to low-med and med-high (4 vs. 18, respectively). This is also reflected in the quadratic-weighted kappa as low-high disagreements will carry a larger penalty (unweighted kappa of 0.797 compared to qwkappa of 0.900). The main reasons for disagreements over specificity labels come from two of the four specificity elements discussed in Section 3.2: whether an argument move is related to one character or scene, and whether it provides a chain of reasons. With respect to the first of these two elements we observed disagreements in argument moves containing pronouns with an ambiguous reference. Of particular note is the pronoun *it*. If we consider the argument move *"I mean even if you know you have a hatred towards a standard or whatever, you still don't kill it"*, the pronoun *it* clearly refers to something within the move (i.e. the standard) that the student themselves mentioned. In contrast, for argument moves such as *"It did happen"* it might not be clear to what previous move the pronoun refers, therefore creating confusion on whether this specificity element is accomplished. Regarding specificity element (4) we found that it was easier to determine the presence of a chain of reasons when discourse connectives (e.g. because, therefore) were present in the argument move. The absence of explicit discourse connectives in an argument move might drive annotators to disagree on the presence/absence of a chain of reasons, which

is likely to result in a different specificity label. Additionally, annotators found that shorter turns at talk proved harder to annotate for specificity.

3.4.2 Reliability Analysis on Dataset D2

Reliability analysis on dataset D2 was performed by double-coding 50 (out of 73) discussions, for a total of 1049 argument moves. While we used quadratic-weighted Cohen kappa for specificity since it is of ordinal type, argument component is a categorical variable therefore unweighted kappa is used in this case. The two annotators achieved a kappa of 0.629 for argumentation and qwkappa of 0.641 for specificity, indicating substantial agreement for both annotations [McHugh, 2012].

3.4.3 Reliability Analysis on Dataset D3

For this study, a subset of D3 was double-coded by two annotators for all three components of collaborative argumentation. Collaboration, like argumentation, represents a categorical variable therefore unweighted Cohen kappa was used. Metrics for argumentation and specificity remain unchanged, kappa and qwkappa respectively. Argumentation annotations resulted in the highest kappa at 0.89, indicating almost perfect agreement. Collaboration was the second highest, with kappa of 0.74, suggesting a substantial agreement level. Specificity followed closely with qwkappa of 0.70, also indicating substantial agreement.

Once an adequate agreement level was achieved, the rest of the discussions were single-coded for all 3 classes. In order to build the gold standard annotation set for D3, disagreements in the double-coded transcripts were resolved through discussion and deliberation between the two annotators. For more details on D3, please refer to [Olshefski et al., 2020].

3.4.4 Reliability Analysis on Dataset D4

For the last inter-rater reliability study on D4, we followed the same protocol and metrics used for dataset D3. Initially, a partial set of transcripts in D4 was double-coded. The annotators achieved high kappa for argumentation (0.971) and qwkappa for specificity (0.813), which

indicates almost perfect agreement for both classes. On collaboration, on the other hand, annotators achieved a lower kappa of 0.578, representing moderate agreement. Given the lower reliability for collaboration, the rest of discussions were also double-coded.

As for D3, the gold standard dataset was built by discussing and resolving disagreements between the two annotators.

3.5 OPPORTUNITIES AND CHALLENGES

Our annotation scheme introduces opportunities for the educational community to conduct further research on the relationship between features of student talk, student learning, and discussion quality. Although Chisholm and Godley [Chisholm and Godley, 2011] and we found relations between our coding constructs and discussion quality, these were small-scale studies based on manual annotations. Once automated classifiers are developed, such relations between student talk and learning can be examined at scale. Also, automatic labeling via a standard coding scheme can support the generalization of findings across studies, and potentially lead to automated tools for teachers and students.

The proposed annotation scheme also introduces NLP opportunities and challenges. Existing systems for classifying specificity and argumentation have largely been designed to analyze written text rather than spoken discussions. This is (at least in part) due to a lack of publicly available corpora and schemes for annotating argumentation and specificity in spoken discussions. The development of an annotation scheme explicitly designed for this problem is the first step towards collecting and annotating corpora that can be used by the NLP community to advance the field in this particular area. Furthermore, in text-based discussions, NLP methods need to tightly couple the discussion with contextual information (i.e., the text under discussion). For example, an argument move from one of the discussions mentioned in dataset D2 stated *"She's saying like free like, I don't have to be, I don't have to be this salesman's wife anymore, your know? I don't have to play this role anymore."* The use of the term *salesman* shows the presence of specificity element (3) (see Section 3.2) because the text under discussion is indeed *Death of a Salesman*. If the students were discussing

another text, the mention of the term *salesman* would not indicate one of the specificity elements, therefore lowering the specificity rating. Thus, using existing systems is unlikely to yield good performance. In fact, we previously [Lugini and Litman, 2017] showed that while using an off-the-shelf system for predicting specificity in newspaper articles resulted in low performance when applied to classroom discussions, exploiting characteristics of our data could significantly improve performance. We have similarly evaluated the performance of two existing argument mining systems [Nguyen and Litman, 2018, Niculae et al., 2017] on transcripts in dataset D1. We noticed that since the two systems were trained to classify only claims and premises, they were never able to correctly predict warrants in our transcripts. Additionally, both systems classified the overwhelming majority of moves as premise, resulting in negative kappa in some cases. Collaboration can be equally very interesting and challenging from the NLP point of view. A target turn is annotated for collaboration with respect to a particular reference turn, therefore for a model to achieve high accuracy it needs to correctly identify the reference. It is then challenging to effectively make use of contextual information outside of the target turn. Let us consider the collaboration annotations in Table 7. Suppose in turn 6, Student 6 voices an extension of turn 5. Because turn 5 itself is an extension of turn 1, they relate to the same core idea. Then, it is easy for a classifier to infer (incorrectly) that turn 6 is an extension of turn 1 instead of turn 5. Likewise, suppose turn 6 was annotated as a challenge to turn 5. Since turns 1 and 5 relate to the same idea, a classifier could infer (incorrectly) that turn 6 is a challenge to turn 1 instead of turn 5. Using our scheme to create a corpus of classroom discussion data manually annotated for argumentation, specificity, and collaboration will support the development of more robust NLP prediction systems.

Finally, we collected the Discussion Tracker corpus (D3), a corpus of American high school English classroom discussions. This corpus consists of 29 multi-party, text-based discussions originating in 3 high schools and from 10 teachers. We annotated the dataset for collaboration, argumentation and specificity as described in this chapter, and publicly released it for free use for research purposes. The dataset is available at `https://discussiontracker.cs.pitt.edu` and a full, detailed description of data collection and annotation procedures can be found in [Olshefski et al., 2020]. Along with the dataset, meta-

data containing the folds used for cross-validation in all experiments on D3 in this thesis is available, in order to facilitate reproducibility of results. The public release of the corpus enables NLP researchers to investigate multiple research inquiries. Much like part of the work in this thesis, the annotations can be used individually to develop machine learning models for automated prediction of the three collaborative argumentation components. Perhaps more interesting research questions can be explored based on the fact that D3 provides simultaneous annotations of multiple dimensions of collaborative argumentation. Like we show in Chapter 7, it is possible for example to investigate whether these dimensions are related and how. If a relation does exist, joint models can be developed for multiple dimensions. Additionally, it is possible to compare and contrast D3 to other publicly available argumentative datasets based on written text to understand the difference between written and verbal arguments (which also introduces the possibility for transfer learning). Lastly, it is possible to compare D3 to web-based multi-party discussions, to understand similarities and differences between online and in-person argumentation.

3.6 SUMMARY

In this chapter we proposed a new annotation scheme for three theoretically-motivated features of student talk in classroom discussion: argumentation, specificity and collaboration. We demonstrated usage of the scheme by presenting an annotated excerpt of a classroom discussion. We demonstrated that the scheme can be used to annotate classroom discussions with high reliability. Finally, we discussed some possible applications and challenges posed by the proposed annotation scheme for both the educational and NLP communities.

4.0 PREDICTING SPECIFICITY FOR CLASSROOM DISCUSSIONS

4.1 INTRODUCTION

Specificity is defined by the Oxford Dictionary as "The quality of belonging or relating uniquely to a particular subject" [1]. Natural language processing (NLP) techniques can be used to facilitate the analysis of classroom discussion and of specificity. Speciteller [Li and Nenkova, 2015] is a popular method for predicting sentence specificity. It was developed by analyzing newspaper articles to distinguish between general and specific sentences. Spoken and written language differ in grammatical structure, contextual influence, and cognitive process and skills [Chafe and Tannen, 1987, Biber, 1988]. As such we believe that using Speciteller as-is on classroom discussions will lead to sub-optimal performance, which we can improve.

In this chapter we propose a method to automatically determine specificity of student argument moves in high school ELA classroom discussions of texts. The contributions of this work are twofold. For the educational community this work will enable the exploration of hypotheses concerning specificity and discussion quality over large datasets, spanning multiple classes and including a large number of students, which would otherwise require a prohibitive amount of work for manually annotating data. Additionally, we develop a set of pedagogically meaningful features which can be used to understand important elements of highly specific discussions. For the NLP community, we make the following contributions: we experimentally evaluate the performance of prior approaches for predicting specificity in a new domain; we compare between different feature sets and algorithms; finally, we provide a model for predicting specificity tailored to spoken dialogue and in an educational setting,

[1] https://en.oxforddictionaries.com/definition/specificity

which outperforms the current state of the art.

4.2 RELATED WORK

To the best of our knowledge, this is the first work to analyze specificity of transcripts of spoken dialogue, and more precisely in classroom discussions. Louis and Nenkova [Louis and Nenkova, 2011] analyzed specificity in news articles and their summarizations. Their proposed method leverages a combination of lexical and syntactic features and annotated data from the Penn Discourse Treebank to train a logistic regression classifier. They used the trained model to analyze differences in specificity between human-written and automatically-generated summaries of news articles. Li and Nenkova [Li and Nenkova, 2015] developed Speciteller, a tool for predicting the specificity score of sentences. Specificity was defined in relation to the amount of details in a sentence. This tool uses a set of shallow features (described in Section 4.3.2) and two dense word vector representations to train two logistic regression models on Wall Street Journal articles. Additionally, they improved classification accuracy by using a semi-supervised co-training method on over thirty thousand sentences from the Associated Press, New York Times, and Wall Street Journal. Our annotation scheme is based on prior educational work in coding specificity [Chisholm and Godley, 2011], and our prediction models will incorporate features used by Speciteller.

Like other machine learning-based methods, Speciteller is highly dependent on its training data. Since our objective is to analyze classroom discussions, we also draw on work that has used Speciteller to analyze data that is more similar to our corpus. Swanson et al. [Swanson et al., 2015] analyzed online forum dialogues for the purpose of argument mining. By performing feature selection they observed that argument quality is highly correlated with specificity as measured by Speciteller across multiple topics. We believe there might be a correlation between specificity and other features used in their work (described in Section 4.3.3) to predict argument quality, therefore we used some of these features in our approach. More recently, Gao et al. [Gao et al., 2019] proposed a model for predicting specificity in social media posts. The model is largely based on the same features used by Speciteller

augmented by named entities, part of speech tags, correctness score, and a set of features for explicitly capturing tweet information (URLs, mentions, emojis). Since specificity is annotated as a numerical value, the proposed features are used to train a regression model. Ke et al. [Ke et al., 2018] analyzed specificity for the purpose of argument persuasion. They modeled specificity using a word-level recurrent neural network. Since specificity represents one of several components of argument persuasion and therefore specificity prediction represents an intermediate step, the model does not give much emphasis to this task and (as noted by the authors) performance can be substantially improved. Since we published our proposed model, it has also been extended by other researchers: in their work Ko et al. [Ko et al., 2019] used a base model inspired by the layout in Figure 2, concatenating a sentence embedding obtained using a recurrent neural network to handcrafted features and using it as input to a classifier, and extended it with the goal of generalizing specificity prediction to domains where little training data is available.

4.3 PROPOSED MODELS

This section provides a description of Speciteller [Li and Nenkova, 2015] and additional features and models that we propose to predict specificity.

4.3.1 Speciteller tool

The baseline for testing our hypotheses consists of using Speciteller out of the box to predict the specificity of each argument move. Speciteller accepts a string as input and outputs a specificity score in the range $[0, 1]$, where 0 indicates general sentences and 1 indicates specific sentences. Since the unit of analysis for the current work is an argument move, which may consist of multiple sentences, we evaluated the performance of Speciteller in several scenarios (e.g. sentence, argument move). We found that the best results are obtained when using the complete argument move as input to Speciteller. In order to convert the numeric specificity score into a specificity class (i.e. low, medium, or high) we set two thresholds t_1 and t_2, then

labeled argument moves with specificity score $s \leq t_1$ as low, those with score $t_1 < s \leq t_2$ as medium, and those with score $s > t_2$ as high. The optimal thresholds were found by starting at 0 and iteratively increasing them by 0.001 at each step, while saving the best results. The values for the optimal thresholds are: $t_1 = 0.02$ and $t_2 = 0.78$. It is important to note that this represents the upper bound for Speciteller's performance. Finding the optimal thresholds is not trivial and in practice it could be done through cross-validation.

4.3.2 Speciteller feature set

The initial set of features we evaluated was that used in Speciteller. We extracted features from each argument move using the source code provided by Speciteller[2]. In their proposed method, Li and Nenkova extracted two categories of features, a shallow feature set and a word embeddings set, and used them for two separate classifiers. In this work, we concatenate both shallow features and word embeddings to form a single feature vector. We will refer to these features as the Speciteller set. Shallow features for each sentence consist of: number of connectives, sentence length (number of words), number of numbers, number of capital letters, number of symbols (including punctuation), average number of characters for the words in the sentence, number of stopwords (normalized by sentence length), number of strongly subjective and polar words (using the MPQA [Wilson et al., 2009] and the General Inquirer [Stone and Hunt, 1963] dictionaries), average word familiarity and imageability (using the MRC Psycholinguistic Database [Wilson, 1988]), average, maximum, minimum inverse document frequency values. Word embeddings features consist of the average of 100-dimensional vectors for each word in the sentence. The embeddings were provided by Turian et al. [Turian et al., 2010] and trained on a corpus consisting of news articles.

4.3.3 Online dialogue features

While extracting arguments from online forum dialogues, Swanson at al. [Swanson et al., 2015] found that Speciteller scores (as a measure of specificity) are highly correlated with argument quality. In addition to Speciteller scores, their model used several feature sets.

[2]https://www.cis.upenn.edu/ nlp/software/speciteller.html

While not explicitly stated by the authors, we believe there might exist a correlation between specificity and the other feature sets. We will add the following sets of features to the features already present in Speciteller.

Semantic features[3] The number of pronouns present in a given argument move. Descriptive statistics for word lengths: minimum, maximum, average, and median length of the words in an argument move. It is worth noting that the average word length differs from the one implemented in Speciteller as this feature keeps punctuation into account. Number of occurrences of words of length 1 to 20: one feature for each word length - words longer than 20 characters will be counted in the feature for length 20.

Lexical features N-gram language models are often powerful features, but one drawback is their dependence on specific domains. Since we plan to build a model for predicting specificity which is able to generalize to multiple topics, we did not use the raw N-gram features. To alleviate this problem, we used the term frequency - inverse document frequency (tf-idf) feature for each unigram and bigram in the corpus with frequency of at least 5. Descriptive statistics of lexical features for each argument move, namely minimum, maximum, and average, were also used.

Syntactic features To mitigate the data sparsity that impacts word n-grams, and to get more generalizable features, we extracted unigrams, bigrams, and trigrams of Parts Of Speech (POS) tags, using the Natural Language Toolkit [Bird et al., 2009a].

4.3.4 Additional feature sets

In addition to the previous feature sets, we also extracted the following feature sets which we believe are able to capture specificity with respect to the educational domain of ELA text-based classroom discussions.

Pronoun features Pronouns are grammatical units that might help us gain useful information about the focus of an argument move. For example, if the pronoun "she" is present in an argument move, the student might likely be referring to one specific character, which is one of the aspects considered when annotating specificity. Therefore we extracted a set of the

[3]The name of the feature set in the original paper is semantic-density features; we use semantic features for brevity.

following pronoun features: binary feature indicating presence/absence of pronouns; total number of pronouns in the argument move[4]; the numbers of first, second, and third person pronouns; the number of singular and plural pronouns; the number of pronouns for each of the following categories: personal, possessive, reflexive, reciprocal, relative, demonstrative, interrogative, indefinite.

Named entities Named entities might give us a sense of characters or places that students discuss, with respect to specificity. For example, saying "I did not like Biff" is more specific than saying "I did not like one of the characters" as it points out which of the characters a student might not like. For this task we used the Stanford Named Entity Recognizer [Finkel et al., 2005] (NER) with the pre-trained 3 class model detecting location, person and organization entities. We extracted the following features: a binary feature indicating the presence/absence of any named entity; a binary feature indicating presence/absence of each of the three named entity classes; the total number of named entities; the total number of named entities per class. We complemented the previous counts by adding a normalized feature, with respect to the length of the argument move, for each of them.

Book features Since our dataset consists of text-based discussions, we might be able to leverage information about the texts (i.e. books) for each discussion to understand how much each argument move is related to the book or its characters. First, a manually-created summary and a list of characters for each book were obtained from the web, using Wikipedia when possible or Sparknotes as an alternative. Then, the following character-related features were extracted from each argument move: a binary feature indicating the presence/absence of a character's name; the number of characters mentioned; the number of characters mentioned normalized by the length of the argument move. A character was counted by matching each word in the argument move to their first name, last name, or their nickname. Additionally the following summary related features were extracted: the number of overlapping words with the argument move; Jaccard similarity between the argument move and the summary; tf-idf based cosine similarity between the summary and the argument move. We extracted the summary related features in two different settings: considering the book summary as a

[4]This feature differs from that described in section 4.3.3: the feature from the online dialogue set only considers deictic pronouns.

single entity; computing the similarity between the argument move and each sentence in the summary, then picking the maximum. All features were extracted after removing stopwords from the argument move and summary.

Embeddings Li and Nenkova [Li and Nenkova, 2015] used sentence embeddings based on word embeddings in order to increase the accuracy of Speciteller. The sentence embeddings were obtained by computing the average of pre-trained word embeddings for each word in the sentence. We believe our method can further benefit from sentence embeddings specifically trained on our corpus and optimized for our target: predicting specificity. We generated embeddings by training a character-level Long-Short Term Memory (LSTM) network [Hochreiter and Schmidhuber, 1997], using it as an encoder on the argument moves from our corpus. Each argument move, which might consist of multiple sentences, represents one sequence (training sample) for the LSTM training. Since punctuation is not very meaningful given that we are analyzing spoken discussions, all characters that are not letters or numbers are ignored. Inputs for the LSTM consist of one-hot (1 X N) encoding of individual characters.

The neural network is trained by using the hidden state of the LSTM unit at the end of the argument move as embedding, feeding it to a softmax classifier for predicting specificity, and back-propagating errors. Cross-entropy was used as the objective function to optimize during training. A disadvantage of neural network models is the fact that their large number of parameters requires extensive amount of data to show their expressive power. Given the size of our training data we try to mitigate this problem by merging the embeddings for an argument move with handcrafted features. Ideally we would combine embeddings with all the features described previously but the resulting model would be far too large for our dataset, therefore we chose to use the *Speciteller* + *Semantic* feature set for this task. The training procedure changes slightly: an argument move is propagated through the LSTM resulting in a fixed size embedding; handcrafted features are extracted from the argument move, concatenated to form a vector, and a fully-connected layer is applied to those; the output of the fully-connected layer is concatenated with the embedding, and given as input to a softmax classifier to predict specificity. A graphical overview of the model is given in Figure 2.

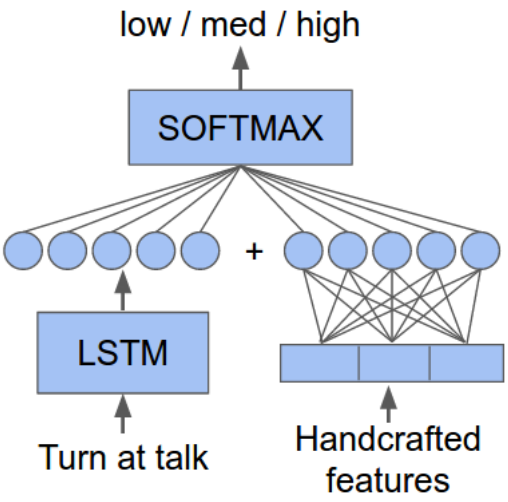


Figure 2: Network setup for training neural network-based embeddings.

It is important to note that the neural network for embeddings and the classifier are jointly trained, therefore the embeddings are specifically tailored to encode information regarding specificity. The Keras library [Chollet et al., 2015a] was used for extracting sentence embeddings as well as for evaluating performance of the softmax classifier.

Pedagogical feature set In addition to maximizing kappa for specificity prediction, an additional objective for this study is to find meaningful features that can help explain different aspects of highly specific discussions. Many of the features described above, like N-grams or tf-idf, might have good predictive power but they are not easily interpretable and bear little relation to our codebook.

When considering NLP techniques applied to the educational domain, there is an increasing interest in developing models that capture important components of the construct to measure. Rahimi et al. [Rahimi et al., 2017], for example, developed a model for automated essay scoring using rubric-based features; Loukina et al. [Loukina et al., 2015] evaluated different feature selection methods to obtain interpretable features in an educational setting.

In order to create an interpretable feature set we started by manually selecting meaningful features from Speciteller (imageability, subjectivity, polarity, and familiarity ratings, number of connectives, fraction of stopwords). At training/test time, this set is combined with

features from the *Pronoun*, *Named entities*, and *Book* feature sets. Since all the features from the last 3 sets are interpretable, we only chose a few features from each set, selecting the ones with highest information gain with respect to specificity. For each fold, we first rank features in the *Pronoun*, *Named entities*, and *Book* sets by information gain, then select the top k (based on the number of features in each respective set), concatenate them to the interpretable Speciteller features and train a logistic regression model. Section 4.4.4 will give examples of selected features.

4.4 EXPERIMENTS AND RESULTS

In this section we provide results for our experiments. All classifiers and feature sets were evaluated using 10-fold cross validation, and using quadratic-weighted Cohen's kappa as the performance metric since it is important to make a distinction between different classification errors (e.g. classifying a low specificity argument move as high should result in bigger error than classifying it as medium). The experiments were performed on dataset D1 (see Chapter 2), and a the distribution of specificity class labels can be seen in Table 8.

Table 8: Dataset D1 statistics.

	Annotation	Total Count	Percentage
Specificity	Low	730	35.49%
	Medium	974	47.35%
	High	353	17.16%
	Total	**2057**	**100.00%**

We used the scikit-learn Python package[5] for training and evaluating classifiers, as well as performing feature selection. Specifically, sections 4.4.1 and 4.4.2 will be used to test our first hypothesis: that by retraining an existing model on our corpus we will obtain an improvement in performance. Sections 4.4.2 and 4.4.3 will be used to test our second

[5]http://scikit-learn.org/stable/

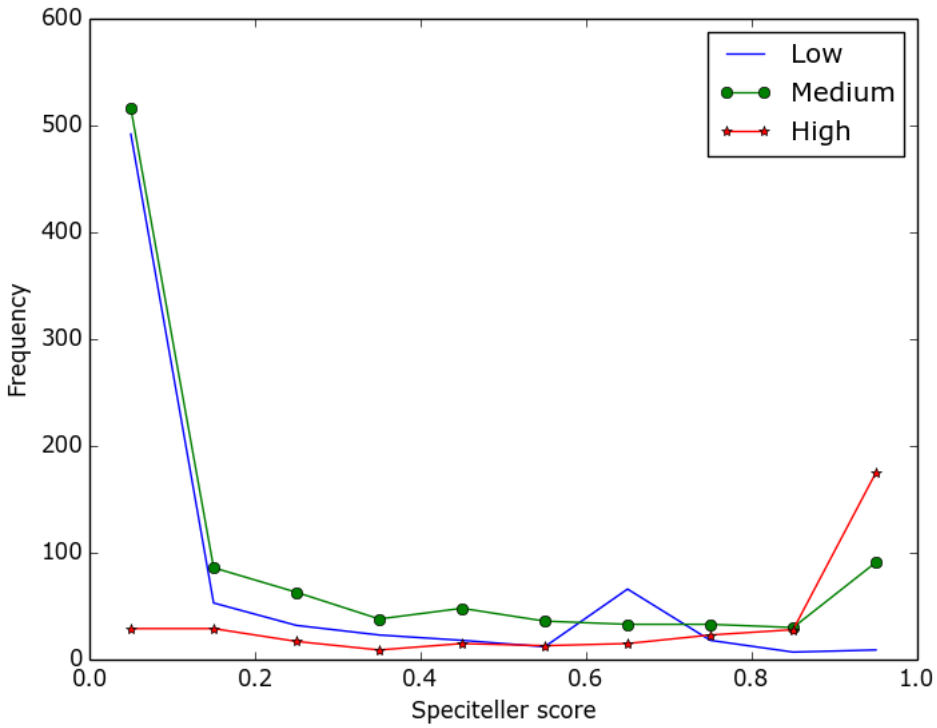


Figure 3: Speciteller scores by specificity class.

hypothesis: that by using features from additional NLP literature we can further improve the performance of a state-of-the-art model. Section 4.4.4 will test our third hypothesis: that the additional features we handcrafted to capture specificity with respect to verbal discussion in an educational setting will lead to better performance.

4.4.1 Baseline using Speciteller off-the-shelf

Since we plan to use Speciteller as a baseline for comparing the performance of our proposed method, we iteratively tested thresholds to find the set which results in the highest quadratic-weighted kappa in all scenarios described in Section 4.3.1. The best result was obtained when the input to Speciteller is the complete argument move, and the resulting quadratic-weighted kappa is 0.495, which represents Speciteller's upper bound performance. Figure 3 shows the frequency distribution of speciteller scores for each specificity class.

From the figure we can see that Speciteller is able to correctly capture specificity for a portion of the argument moves in the dataset, as there is a peak in the low end of the spectrum for the distribution of low specificity scores and a peak in the high end of the spectrum for the distribution of high specificity scores. The medium specificity class seems to be the most problematic one, which has a similar trend as the low specificity class distribution in the low end of the spectrum, and a similar trend to the high specificity class distribution in the high end of the spectrum. Ideally we would expect the medium specificity distribution to have a peak towards the middle of the spectrum but that is not the case. Additionally, the low specificity class distribution shows a peak between 0.6 and 0.7 which will further penalize accuracy.

Table 9 shows the confusion matrix when applying the optimal thresholds in order to get specificity labels from Speciteller scores. As we can see from the confusion matrix the overlap

Table 9: Confusion matrix using Speciteller scores to classify according to the optimal split points.

		predictions		
		low	**med**	**high**
ground truth	**low**	352	360	18
	med	280	565	129
	high	4	139	210

between the low and medium specificity classes and the medium and high specificity classes causes a large number of misclassifications: almost half of the low specificity argument moves are classified as medium, over 40% of the medium specificity argument moves are classified as either low or high, and almost 40% of high specificity argument moves are classified as medium. We believe these errors stem from two main reasons: as with many data-driven approaches, Speciteller is highly dependent on its training corpus. Speciteller was trained on articles from the Wall Street Journal and the New York Times. Articles written by professional writers are inherently different from transcriptions of spoken discussions between high school students. Additionally, for training the model, Speciteller used a binary

general/specific label, while we consider three labels in our work. Since Speciteller has no prior knowledge on medium specificity sentences, it is understandable that most of the misclassifications come from this class.

4.4.2 Training using Speciteller features

Our hypotheses as to why Speciteller does not work effectively out of the box are related to its corpora and the way it was trained. With respect to the features used by Speciteller, we believe they might be useful in classrooms discussion as well. We extracted the shallow feature set and the neural network word embeddings feature sets and combined them to train a logistic regression classifier on our dataset. This classifier was chosen because one of our objectives is to compare the importance of other feature sets in addition to the *Speciteller* one, and in order for this comparison to be fair we decided to use the same classifier Speciteller uses. Additionally, the classifier weights can be used to understand the importance of each feature. It is important to note that, unlike Speciteller, we will be using a single classifier on the combination of all features, and will not be able to leverage semi-supervised co-training.

Table 10 shows the performance of a logistic regression classifier trained on this feature set and others described in the previous section. As we can see from the table, training a

Table 10: Classification performance of different feature sets. * indicates statistically significant improvement over Speciteller features with p-value < 0.001. Statistical significance was tested using a two-tailed paired t-test. Bold font highlights best results.

Feature sets	**QWKappa**
Speciteller	0.5758
Speciteller + Online dialogue	0.6347*
All: Speciteller + Online dialogue + Pronoun + NE + Book	0.6360*
Speciteller + Semantic + Embeddings	**0.6550***
Pedagogical	0.5886

classifier using the Speciteller feature set on our corpus results in a considerable increase

in performance, with QWKappa of 0.5758 which represents a 16% relative improvement over the 0.495 QWKappa obtained using Speciteller out of the box. This confirms our first hypothesis that Speciteller's performance, like many other methods, is highly dependent on its training corpus and using this model out of the box would give sub-optimal results.

4.4.3 Speciteller and online dialogue features

To test whether features from Section 4.3.3 are useful, we combined the Speciteller features with the Semantic, Lexical, and Syntactic features and trained a logistic regression classifier based on the concatenated feature vectors. Table 10 confirms our hypothesis that the 4 feature sets combined result in statistically significant (statistical significance was computed using a two-tailed paired t-test since we used the same folds for all experiments) higher kappa than using only Speciteller features. When combining Speciteller with each of the 3 other feature sets individually, kappa increases but not with statistical significance. We evaluated additional classifiers (Support Vector Machine, decision tree, random forest, Naive Bayes) but none of them outperformed logistic regression. Since the number of features is over 7000, we also tried using Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) for feature selection/reduction, but neither improved performance.

4.4.4 Additional features

To the feature set described in the previous section, we added the features described in Section 4.3.4. We then tested our third hypothesis by evaluating the performance of a logistic regression model trained with these features.

We can see from Table 10 that all additional feature sets yield better performance than the *Speciteller* feature set by itself. This result confirms our third hypothesis: the additional feature sets are able to capture aspects of specificity with respect to verbal discussion and the educational domain. In particular the feature set containing neural network-based sentence embedding achieved the best kappa measure of 0.6550, which suggests that sentence embeddings are also domain-dependent. Compared to using Speciteller off-the-shelf this method improves kappa by 32%. While the size of the neural network was constant

during training/test (not optimized for each fold), we experimented with several numbers of hidden nodes (ranging from 50 to 200) for the LSTM and fully-connected layers, which resulted in kappa values in the range $0.6283 - 0.6550$.

The Pedagogical feature set is also able to marginally outperform the Speciteller feature set. Compared to the best result, the loss in kappa when using the Pedagogical set is 11%. At the expense of a slightly lower accuracy we gain the ability to use only informative features, which can be used to better understand highly specific versus general classroom discussions. The use of logistic regression also makes this possible: the model's coefficients give us an indication of how important features are. Table 11 shows the top 12 features in the Pedagogical feature set ranked by the magnitude of the model's coefficients.

Table 11: Pedagogical feature set and respective logistic regression coefficients. Italic font shows features developed in this study (Section 4.3.4).

Feature	**Coefficient**
Number of connectives	1.9168
Cosine similarity – whole summary	0.9293
MRC imageability	0.8172
Number of characters	0.6931
MPQ subjectivity	-0.5440
Fraction of stopwords	-0.4087
MRC familiarity	0.3986
Number of possessive pronouns	0.2035
Number of named entities normalized	0.1865
Number of 3rd person pronouns	0.1755
Word overlap – whole summary	0.1585
Number of personal pronouns	0.1476

The table shows the results of a model trained on the complete dataset. The number of connectives seems to be the most important feature for predicting high specificity. This seems straightforward, as more connectives translates into more clauses, which provide more

information. While the annotators did not look for connectives during coding, one of the aspects they analyzed was the presence/absence of a chain of reasoning, and the number of connectives might capture that aspect. The cosine similarity between the argument move and the book summary (considered as one entity) is another important feature in the model: higher similarity between the summary and what a student says means that they are using terms from the book. This feature seems to capture another aspect in our codebook, the use of book-specific vocabulary. We can use the information provided by these features to understand specificity, and to give feedback to teachers and students: if for example a student tends to produce low specificity argument moves and the number of connectives used is generally low, that might be an indication that they should elaborate more on their statements. Conversely, if the number of connectives used is high but the number of characters mentioned is low, that might be an indication that the student should reference specific characters more often.

4.5 SUMMARY

In this chapter we proposed several models for predicting specificity and evaluated them on text-based, high school classroom discussion data. We showed that an existing general-purpose system achieves significantly better performance when its features are used for re-training on educational data. We also showed that performance can be further improved by using additional features from the NLP literature [Swanson et al., 2015], especially when combined with neural network embeddings and other new features tailored to text-based classroom discussion. Finally we proposed a subset of pedagogical features which, even though slightly less performing, provide the ability to interpret the features, which is especially important for the educational community.

The findings reported in this chapter were published in [Lugini and Litman, 2017].

5.0 ARGUMENT COMPONENT CLASSIFICATION FOR CLASSROOM DISCUSSIONS

5.1 INTRODUCTION

Although there is no universally agreed upon definition, argument mining is an area of natural language processing which aims to extract structured knowledge from free-form unstructured language. In particular, argument mining systems are built with goals such as: detecting what parts of a text express an argument component, known as argument component identification; categorizing arguments into different component types (e.g. claim, evidence), known as argument component classification; understanding if/how different components are connected to form an argumentative structure (e.g. using evidence to support/attack a claim), known as argument relation identification. The development and release to the public of corpora and annotations in recent years have contributed to the increasing interest in the area.

One domain in which argument mining is rarely found in the literature is educational discussions. With the increasing importance of argumentation in classrooms, especially in student-centered discussions, automatically performing argument component classification is a first step for building tools aimed at helping teachers analyze and better understand student arguments, with the goal of improving students' learning outcomes.

Many current argument mining systems focus on analyzing argumentation in student essays [Stab and Gurevych, 2014, Stab and Gurevych, 2017, Nguyen and Litman, 2015, Nguyen and Litman, 2018], online dialogues [Swanson et al., 2015, McLaren et al., 2010, Ghosh et al., 2014, Lawrence and Reed, 2017], or in the legal domain [Ashley and Walker, 2013, Palau and Moens, 2009]. A key difference between these studies and our work consists

in the source of linguistic content: although we analyze written transcriptions of discussions, the original source for our corpora consists of spoken, multi-party, educational discussions, and the difference in cognitive skills and grammatical structure between written and spoken language [Biber, 1988, Chafe and Tannen, 1987] introduces additional complexity.

5.2 IMPROVING THE PERFORMANCE OF AN EXISTING ARGUMENT MINING MODEL

Our work and previous research studies on student essays share the trait of analyzing argumentation in an educational setting. However, while student essays are typically written by an individual student, in classroom discussions arguments are formed collaboratively between multiple parties (i.e. multiple students and possibly teachers). While our work shares the multi-party setting in which arguments are made with research aimed at argument mining in online dialogues, prior online dialogue studies have not investigated the educational domain.

Given these differences, we believe that argument mining models for student essays and online dialogues will perform poorly when directly applied to educational discussions. However, since similarities between the domains do exist, we expect that features exploited by such argument mining models can help us in classifying argument components in classroom discussions. Moreover, unlike the other two domains, we have access to labels belonging to a different (but related) class, specificity, which we can try to incorporate in argumentation models to boost performance. Our contributions are as follows. We evaluate features proposed in an existing argument mining system, wLDA, to understand their usefulness in a different, educationally-related setting. We then extend wLDA with additional feature sets developed for analyzing arguments in an online, multi-party setting. We finally evaluate two neural network models in several different scenarios pertaining to their input granularity (words vs. characters) and propose a hybrid model which includes both neural network embeddings and handcrafted features.

5.3 RELATED WORK

With respect to the educational domain, previous studies in argument mining were largely aimed at student essays. Persing and Ng [Persing and Ng, 2015] studied argument strength with the ultimate goal of automated essay scoring. Stab and Gurevych [Stab and Gurevych, 2014] performed argument mining on student essays by first jointly performing argument component identification and classification, then predicting argument component relations. Nguyen and Litman [Nguyen and Litman, 2015] developed an argument mining system for analyzing student persuasive essays based on argument words and domain words. While domain words are used only in a specific topic, argument words are used across multiple topics and represent indicators of argumentative content. They later proposed an improved version of the system [Nguyen and Litman, 2016b], which we will refer to as wLDA, by exploiting features able to abstract over specific essay topics and improve cross-topic performance. While our current work is also aimed at developing argument mining systems in the educational setting, we focus on educational discussion instead of student essays. Our work also differs in the argument component types used: we analyze claims, evidence, and warrants, while prior studies mostly focused on claims and premises. The inclusion of warrants is particularly important to explicitly understand how students use them to connect evidence to claims. As such, we do not expect prior models to work well on our corpus, although some of the features might still be useful. Also, while some of the previously proposed systems address multiple subproblems simultaneously, e.g. argument component identification and argument component classification, we only focus on argument component classification.

Swanson et al. [Swanson et al., 2015] developed a model for extracting argumentative portions of text from online dialogues, which were later used for summarizing the multiple argument facets. Misra et al. [Misra et al., 2015] analyzed dyadic online forum discussions to detect central propositions and argument facets. Habernal and Gurevych [Habernal and Gurevych, 2017] analyzed user-generated web discourse data from several sources by performing micro-level argumentation mining. While these prior works analyze multi-party discussions, the discussions are neither originally spoken nor in an educational setting. Furthermore they are based on different annotation scheme, and perform both argument com-

ponent identification and argument component classification, which translates in different units of analysis.

Like other areas of natural language processing, argument mining is experiencing an increase in the development of neural network models. Niculae et al. [Niculae et al., 2017] used a factor graph model which was parametrized by a recurrent neural network. Daxenberger et al. [Daxenberger et al., 2017] investigated the different conceptualizations of claims in several domains by analyzing in-domain and cross-domain performance of recurrent neural networks and convolutional neural networks, in addition to other models. While trying to improve performance of classifiers for rhetorical analysis of texts, Lauscher et al. [Lauscher et al., 2018] developed a recurrent neural network model for argument component classification. Chakrabarty et al. [Chakrabarty et al., 2019] developed a model for argument component and relation classification in online discussions based on a transformer model. We also propose the use of a neural network in our work, however unlike these previous studies it will be only part of the final model which will also include handcrafted features.

5.4 ARGUMENT COMPONENT CLASSIFICATION MODELS

In this section we outline an existing argument component classification system that will serve as a baseline for our experiments, then propose several new models that use features extracted from neural networks and hand-crafted features.

5.4.1 Existing Argument Mining System

The wLDA[1] system was developed for performing argument component identification, classification, and relation extraction from student essays. We chose to use this system as baseline since it was developed for an educational application and we have complete access to the source code and can therefore not only run it as-is but also use it to extract its features sets to re-train the model and make appropriate modifications/improvement. We also tried a

[1]The original name of wLDA+4 stands for "with LDA supported features and expanded with 4 features sets" compared to their previous system. We use wLDA for brevity.

second argument mining system [Niculae et al., 2017], however we were not able to re-train the system on our corpus. A preliminary evaluation of the pre-trained system showed that it overwhelmingly predicted argument moves as claims (the system was trained to classify claims and premises), therefore we decided not to use it in further experiments.

For the purpose of this study, we only consider the argument component classification subsystem of wLDA. The model is based on a support vector machine classifier which exploits features able to improve cross-topic performance. The feature set consists of four main subsets: lexical features (argument words, verbs, adverbs, presence of modal verbs, discourse connectives, singular first person pronoun); parse features (argumentative subject-verb pairs, tense of the main verb, number of sub-clauses, depth of parse tree); structural features (number of tokens, token ratio, number of punctuation signs, sentence position, first/last paragraph, first/last sentence of paragraph); context features (number of tokens, number of punctuation signs, number of sub-clauses, modal verb in preceding/following sentences) extracted from the sentences before and after the one considered; four additional features for abstracting over essay topics.

Since the model was trained on essays annotated for major claim, claim, and premise, but not on warrants, in our evaluation we did not take into account misclassification errors for argument moves in our dataset labeled as warrants. The pre-trained system performs argument component identification using a multiclass classification approach, such that each input will be classified as non argumentative, major claim, claim or premise. Since our goal is to evaluate performance related to the component classification problem, we ignored all the argument moves classified as non argumentative by wLDA. Considering the definitions of premise and evidence in the Toulmin model [Toulmin, 1958], we made the assumption of the two labels being equivalent for this study, i.e. if the predicted class for an argument move is premise and its gold standard label in our dataset is evidence, we consider the prediction correct. In the same way we consider both claim and major claim labels as equivalent to claims in our dataset.

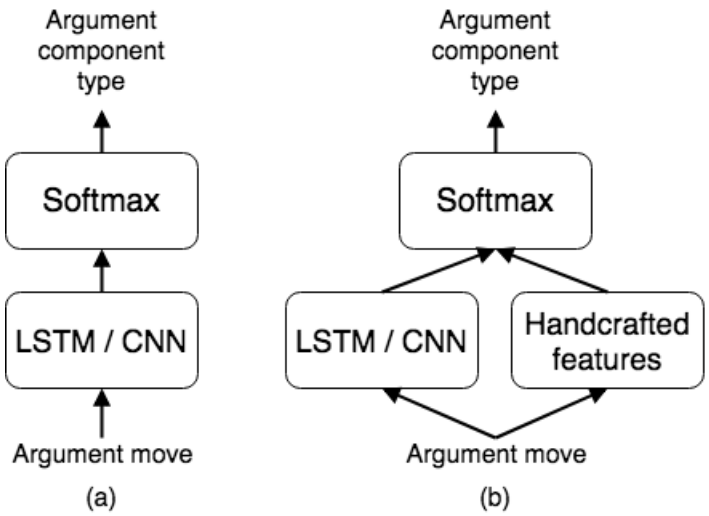


Figure 4: Neural network models used in this study: neural network only setup (a); model incorporating neural network and handcrafted features (wLDA and online dialogue sets) (b).

5.4.2 Neural Network Models

Since the pre-trained model did not work well on our dataset, and the features it is based on show a large gap in performance compared to the original work (see Section 5.5), we decided to use neural networks, and evaluate their ability to automatically extract meaningful features. The proposed models consist of variations of two basic neural network types, namely Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models. All the choices regarding the models were made in order to keep complexity and the number of weights at a minimum, since neural network models require in general a large amount of training data, while we have a limited size dataset. The CNN model is based on a model proposed by Kim [Kim, 2014b] and already used for argument mining in the past [Daxenberger et al., 2017], with a difference in the number of convolutional/pooling layers. In particular, our model uses 3 convolutional/max pooling layers instead of 6, and only one fully connected layer after the convolutional ones, followed by a softmax layer used for classification. This choice resulted from observing significant overfitting when increasing the number of convolutional layers due to the increase in the number of model weights and the limited dataset size. Figure 4 shows diagrams for the different neural network setups used in our experiments.

The RNN model consists of a single Long Short-Term Memory (LSTM) network [Hochreiter and Schmidhuber, 1997]. After propagating a complete argument move through the LSTM network, the resulting hidden state is the feature vector used as input to a softmax layer which outputs the predicted label. Recurrent neural networks have also been used in the field of argument mining [Daxenberger et al., 2017, Niculae et al., 2017]. We set the size of the hidden state to 75 based on several factors. Following Bengio [Bengio, 2012], we decided to have an overcomplete network, i.e. one in which the size of the hidden state is bigger than the size of the input. Since the dimensionality of our character-based encoding is 37 and that for word-based embeddings is 50, we chose a hidden state with size greater than 50 (we use the same hidden state size for both models). Increasing the size introduced overfitting even quicker than the CNN model, given that the number of weights increases more quickly for our LSTM model.

When using text as input to a neural network, we can generally view an argument move as either a sequence of characters, or as a sequence of words. Unlike previous neural network-based argument mining models, each of our models was evaluated under both conditions: for character-based models we used a one-hot encoding (one-out of n) for each letter and number - special characters were filtered since they don't hold particular meaning in speech, and we cannot be sure of transcription conventions; for word-based models we used Global Vectors (GloVe) [Pennington et al., 2014] with dimensionality of 50. An important aspect to consider is that, while word-based models have some prior knowledge encoded in the word embeddings, character-based models do not.

Since neural network models usually require a large amount of training data to be effective, and we have relatively fewer number of argument moves compared to number of model weights, we also tested hybrid models in which a neural network output is combined with handcrafted features before the final softmax classification layer, as shown in Figure 4 (b). Both CNN and LSTM models used categorical cross-entropy as loss function, and the number of epochs was automatically selected at training time by monitoring performance on a validation set consisting of 10% of the training set for each fold.

5.4.3 Online Dialogue Features

Since our dataset is based on multi-party discussion, it shares similarities with prior argumentation work in multi-party online dialogues. Therefore we experimented with features from [Swanson et al., 2015], organized into three main subsets: semantic-density features (number of pronouns, descriptive word-level statistics, number of occurrences of words of different lengths), lexical features (tf-idf feature for each unigram and bigram, descriptive argument move-level statistics), and syntactic features (unigrams, bigrams and trigrams of part of speech tags). The only difference between the original features and the ones we implemented consists in the use of Speciteller [Li and Nenkova, 2015]. As we previously observed [Lugini and Litman, 2017], applying Speciteller as-is to domains other than news articles results in a considerable drop in performance. Therefore, instead of including the specificity score obtained by directly applying Speciteller to an argument move, we decided to use Speciteller's features.

5.5 EXPERIMENTS AND RESULTS

This section provides our experimental results. In Section 5.5.1 we will test our first hypothesis: using an argument mining system trained in a different domain will result in low performance, which can be improved by re-training on classroom discussions and by adding new features. Section 5.5.2 will be used to test our second hypothesis: neural network models can automatically extract important features for argument component classification. Our third hypothesis will be tested in Section 5.5.3: adding handcrafted features (i.e. online dialogue features, wLDA features) to the ones automatically extracted by neural networks will result in an increase of performance.

The data used for the experiments described in this section consists of dataset D2 introduced in Chapter 2. As we can see from the label distribution shown in Table 12, students produced a high number of claims, while warrant is the minority class. We can also observe a class imbalance for specificity labels, though the ratio between majority and minority classes

Table 12: Distribution of class labels for argument component type in dataset D2.

	Annotation	Total Count	Percentage
Argumentation	Claims	1034	50.51%
	Evidence	655	32.00%
	Warrants	358	17.49%
	Total	**2047**	**100.00%**

is lower than that for argument component labels.

The unit of analysis for our work is argument move, which consists of a segment of text containing an argumentative discourse unit (ADU) [Peldszus and Stede, 2013], hence a turn can potentially be segmented into multiple argument moves. Turn segmentation effectively corresponds to argument component identification, and it is carried out manually.

Our experiments evaluate every model using a leave-one-transcript-out cross validation: each fold contains one transcript as test set and the remaining 72 as training set. Cohen kappa, and unweighted precision, recall, and f-score were used as evaluation metrics.

The following python libraries were used for implementing and testing the different models: Scikit-learn [Pedregosa et al., 2011], Tensorflow [Abadi et al., 2015], Keras [Chollet et al., 2015b], NLTK [Bird et al., 2009b].

Given that in our dataset warrants appear much less frequently than claims and evidence, data imbalance is a problem we need to address. If trained naively, the limited amount of training data and the unbalanced class distribution lead the neural network models to specialize towards claims and evidence, with much weaker performance on warrants. This is also the case for non neural network models, although the impact on performance is lower. To combat this phenomenon we decided to use oversampling [Buda et al., 2017] in order to create a balanced dataset, hoping to further reduce the performance gap between the different classes [2]. After computing the class frequency distribution on the training set, we randomly sampled moves from the two minority classes and added them to the current training set,

[2]We also tried setting class weights during training to influence the loss function, though it only improved results marginally.

repeating the process until the class distribution was completely balanced (i.e. until the number of argument moves for each class equals the number of moves in the majority class)

Table 13 shows the results for all experiments. The statistical significance results in the table use the system in row 3 as the comparison baseline, as wLDA represents a system specifically designed for argument component classification (among other tasks). Additional statistical comparisons are provided in the text as well.

5.5.1 Using wLDA Off the Shelf and wLDA Features

Since not all the argument moves were considered when computing results for the pre-trained out of the box wLDA model (see Section 5.4.1), the results in row 2 are not directly comparable to others and are not used for statistical significance tests. Nonetheless they show the upper bound in performance of the pre-trained model, and we can see that it is comparable to a majority baseline which always predicts the majority class in each fold. This result shows that claims and evidence expressed in written essays and classroom discussions have very little in common. While the overall performance of the pre-trained model is comparable to a majority baseline, the individual F-scores for claims and evidence give us an insight into the usefulness of its features: the F-score for evidence is in line with that of several neural network-based models. This is clearer when we look at improvement obtained training a logistic regression model[3] using the same wLDA features on our dataset (row 3), which outperforms the pre-trained wLDA in all metrics (row 2), and indicates that the wLDA features are still able to somewhat distinguish between claims and evidence while performing considerably worse on warrants. Additionally, if we add to this model the online dialogue feature set, the resulting model improves all results and obtains the best kappa overall (row 4). This confirms our hypothesis: given the similarity that exists between our domain and online dialogues, features developed for analyzing argumentation in online dialogues are also useful in classroom discussions.

[3]We also experimented with random forest, naive Bayes and support vector machines, but they provided inferior results compared to logistic regression.

Table 13: Results obtained with the baseline model/features and the proposed neural network models using different feature sets. Each line represents the average of a transcript-wise cross validation. Best results are in bold. *, †, and ‡ indicate statistical significance at the 0.1, 0.05, and 0.01 levels respectively, compared to the model in row 3. The three right-most columns represent per-class F-score for evidence, warrants, and claims respectively.

Row	**Models / Features**	**Kappa**	**Prec**	**Rec**	**F-score**	$\boldsymbol{F_e}$	$\boldsymbol{F_w}$	$\boldsymbol{F_c}$
1	Majority baseline	0.068	0.265	0.406	0.314	0.109	0.004	0.532
2	Pre-trained wLDA	0.077	0.289	0.350	0.269	0.351	N/A	0.456
3	Logistic Regression (wLDA features)	0.142	0.412	0.394	0.379	0.390	0.211	0.540
4	Logistic Regression (wLDA features + online dialogue)	**0.283**	**0.508**	**0.500**	**0.480**	**0.479**	0.222	**0.693**
Character level NN models								
5	LSTM	-0.002	0.062	0.253	0.082	0.007	0.242	0.013
6	LSTM + wLDA + online dialogue	0.034	0.217	0.304	0.150	0.080	$0.272^{\ddagger}$	0.090
7	CNN	0.143	0.439	0.423	0.393	0.372	0.218	0.574
8	CNN + wLDA + online dialogue	$0.241^{\star}$	0.482	0.475	0.450	0.449	0.236	0.637
Word level NN models								
9	LSTM	0.069	0.408	0.399	0.218	0.161	0.198	0.295
10	LSTM + wLDA + online dialogue	0.181	0.462	0.447	0.391	0.362	$\mathbf{0.279}^{\ddagger}$	0.522
11	CNN	0.125	0.410	0.404	0.378	0.370	0.231	0.526
12	CNN + wLDA + online dialogue	$0.241^{\star}$	$0.492^{\star}$	0.488	$0.455^{\dagger}$	0.468	$0.276^{\ddagger}$	0.622

5.5.2 Neural Network Models Alone

Our second hypothesis is validated by the results in Table 13 by comparing row 3 with rows 7 and 11, where we can see that the CNN models achieve performance comparable to a classifier trained on features specifically developed for argument component classification. This indicates that convolutional neural network models are able to extract useful features. Additionally, when comparing the best of these models (row 12, with respect to f-score) to the best performing model based only on handcrafted features (row 4), the difference in performance is not statistically significant for any of the metrics in Table 13.

Looking more closely at the results obtained using neural network models alone we can see two different trends. While LSTM models show performance comparable to random chance (e.g. row 5, with kappa close to zero and lower than the majority baseline), one of our two CNN models (row 7) performs as well as or better than the wLDA based model (row 3) (except for F_e in row 7), while performance for the second CNN model (row 11) is considerably close to the wLDA model. Overall, under the same conditions CNN models almost always outperform LSTM models. One interesting difference between the two models is that the prior knowledge introduced by word embeddings in word-based models is essential for improving performance of LSTMs (e.g. row 5 vs row 9), while this is not the case for CNN models (e.g. row 7 vs row 11). The length of sequences (i.e. argument moves) for character-based models makes it extremely hard for LSTMs to capture long-term dependencies, especially with limited amount of training data. Convolutional models, on the other hand, learn kernels that effectively function as feature detectors and seem to be able to better distinguish important features, and do not always benefit from word level inputs.

5.5.3 Adding wLDA Features and Online Dialogue Features

It is clear from Table 13 that almost all neural network models benefit from additional handcrafted features. This is not surprising, given that neural networks require a large amount of data to be trained effectively, and although random oversampling helped, we still have a limited amount of training data. Even when including additional features the two architectures show slightly different trends: CNN usually outperform LSTM, however LSTM models

benefit more from the additional features. This is at least in part due to LSTMs initially having lower performance without handcrafted features. We analyzed the importance of different subsets of the online dialogue features through a feature ablation study. For CNN models, removing any subset of features resulted in a decrease in performance, except for the *syntax* subset in the *character level CNN + wLDA + online dialogue* model. For LSTM models, however, not all feature subsets contributed to increasing performance.

5.6 SUMMARY

In this chapter we evaluated the performance of an existing argument mining system developed for a different educational application (i.e. student essays) on a corpus composed of spoken classroom discussions. Although the pre-trained system showed poor performance on our dataset, its features show promising results when used in a model specifically trained on classroom discussions. We extracted additional feature sets based on related work in the online dialogue domain, and showed that combining online dialogue and student essay features achieves the highest kappa on our dataset. We then developed additional models based on two types of neural networks, showing that performance can be further improved. Lastly, we provided an experimental evaluation of the differences between convolutional networks and recurrent networks, and between character-based and word-based models.

The findings described in this section were published in [Lugini and Litman, 2018].

6.0 SPEAKER-DEPENDENT CONTEXTUAL INFORMATION FOR ARGUMENT COMPONENT CLASSIFICATION

6.1 INTRODUCTION

While the models proposed in the previous chapter do consider contextual information, such information is used in a very limited way. The contextually-informed features of wLDA were developed for essays written by a single person. They consist of four features (number of punctuation signs, number of tokens, number of sub-clauses and presence/absence of modal verbs) which are extracted from the preceding and following argument move relative to the one under consideration. Therefore, both the number of features and the span of the context are relatively small compared to the rest of the features. Additionally, the neural network portion of our previous model only considers the current argument move, completely disregarding context. In this section we will propose multiple sets of features that aim at capturing contextual information in discussion and analyze their impact on argument component classification.

6.2 RELATED WORK

Stab and Gurevych [Stab and Gurevych, 2014] claim that context plays an important role in identifying argument components, and propose context features extracted from the sentence containing the ADU. Nguyen and Litman [Nguyen and Litman, 2016a] developed a context-aware model for argumentative relation mining of argumentative essays. They proposed two sets of contextual features which extract information related to writing topics

(topic-context) and from surrounding sentences of a source and target sentence, to help predict if the source and the target are connected by an argumentative relation and how (e.g. support, attack). The context-aware model proved to outperform models without contextual information. However, the contextual features cannot be incorporated in our model for argument component classification for two main reasons: *(i)* some of these features require two sentences, a source and a target, whereas our model operates on one single argument move; *(ii)* other features are extracted from components at the sub-sentence level and aligning them with argument moves is not straightforward. Habernal and Gurevych [Habernal and Gurevych, 2017] used a subset of features from their argument component identification/classification model to represent context. Persing and Ng [Persing and Ng, 2016] also used contextual features extracted from the sentences preceding and following the ADU and performing ACI/ACC jointly. Aker et al. [Aker et al., 2017] evaluated different ACI/ACC joint models which use, among others, contextual features. These works share three main limitations: *(i)* a single context configuration is chosen (typically either prior or prior/following sentences) and in some cases one dimension is optimized (typically context size); *(ii)* they only extract a subset of features from context as compared to features for the target ADU, which could potentially reduce the impact of contextual information; *(iii)* even in multi-party discussions, context is solely based on the relative position of each ADU, not considering additional sources of information (e.g. speaker/author ID). Optiz and Frank [Opitz and Frank, 2019] analyzed a previous argument mining system and found that, for its predictions, it relies on context more than it does on the ADU content. Their results showed that in some cases obscuring the target ADU leads to better performance for argument component classification and relation extraction. Chakrabarty et al. [Chakrabarty et al., 2019] developed an argument mining system in which context is indirectly modeled while training the argument component classifier. A BERT model is fine-tuned to predict the next sentence (i.e. the context). Like earlier work, however, context is limited to a fixed size without extensive evaluation.

Other related work was aimed at capturing contextual information in neural models and in multi-party conversations. Memory networks are capable of learning long-term dependencies [Weston et al., 2014, Sukhbaatar et al., 2015], therefore they may be a viable option for

capturing contextual information in our discussion transcriptions. Mohtarami et al. [Mohtarami et al., 2018] developed an end-to-end memory network for the Fake News Challenge (https://www.fakenewschallenge.org). The task addressed was to predict the stance of an article with respect to a given input claim (i.e. agree, disagree, discuss, unrelated), and is somewhat related to argument relation extraction since the proposed model focuses on claims and evidence. The memory network model outperformed conventional convolutional neural networks, recurrent neural networks, or combinations thereof. While achieving good results, their proposed model contains over 100 million parameters and would surely overfit in our case given that we have a few thousand training samples. An interesting example of modeling contextual information in multi-party conversations was recently introduced by Meng et al. [Meng et al., 2018]. When considering the task of speaker classification, they proposed a model which captured both content and temporal contextual information. Content information refers to the actual utterances for each student, while temporal information relates to the order of the utterances. For both types of information, the context is captured with respect to each speaker, i.e. the target variable: each speaker is modeled based on their previous utterances at any point in time. If we were to consider a similar approach we would have to model context separately for each argument component type, and this would be extremely challenging for warrants given the highly imbalanced class distributions in our datasets. Additionally, their modelling of content-context assumes that each of the speakers' utterances is equally important regardless of how or when it happened in the conversation, while more recent utterances are likely to be more important in argument component classification.

In developing a system for dialogue generation Li et al. [Li et al., 2017] modeled contextual information as the concatenation of the previous two utterances with the current one. Ortega and Vu [Ortega and Vu, 2017] analyzed different neural network models for including contextual information in multi-party dialogues in order to predict dialogue acts. They experimented with context size between 2 and 5 utterances, meaning that only local context has an impact on the prediction of the current dialogue act. Neither of these models actually accounts for speakers. They are only focused on extracting information from local context, i.e. previous 2-5 utterances, regardless of who spoke such utterances. They also concatenate

context utterances and the current utterance in a single sequence, which is processed by an single recurrent neural network simultaneously. In contrast to these previous directions we propose to: (1) separately model "local" (immediate vicinity of the target ADU) and "global" (beyond the immediate vicinity of the target ADU, potentially reaching the beginning of the discussion) context; (2) consider the identity of speakers when including previous argument moves; (3) instead of concatenating context argument moves with the current one, we process them separately in order to reserve specific portions of the final feature vector for the two components. Lison and Bibauw [Lison and Bibauw, 2017] analyzed how training instances "(context, response)" can be paired with numerical weights and incorporated into the training of a neural conversational model for domain adaptation. Similarly to our proposed model, the current utterance is processed separately from the context and each has their own portion of the feature vector. However, the context only includes one utterance, and each training instance consists of only two utterances from two speakers, which means that the model is not able to capture long-term contextual dependencies and is therefore inadequate for argument component classification in classroom discussions.

6.3 SPEAKER-DEPENDENT CONTEXT MODEL

We propose to extend the models described in Section 5.4, *hybrid baseline*, by expanding the contextual components and incorporating contextual information appropriately designed for multi-party discussions. Figure 5 shows a diagram of how we incorporated the new contextual features into the model. The main idea behind the proposed model is to significantly expand the feature vector which is used as input to the final softmax classifier with contextual information that the models from the previous chapter largely or entirely ignore.

In order to test whether the same strategy is effective for other argument component classification models, we implemented an additional baseline. Given the recent success of Transformer neural networks [Vaswani et al., 2017] and pretrained models in NLP tasks [Devlin et al., 2019, Wolf et al., 2019] the *BERT baseline* consists of a BERT pretrained model to generate word embeddings of dimensionality 768; average pooling is used to aggregate all

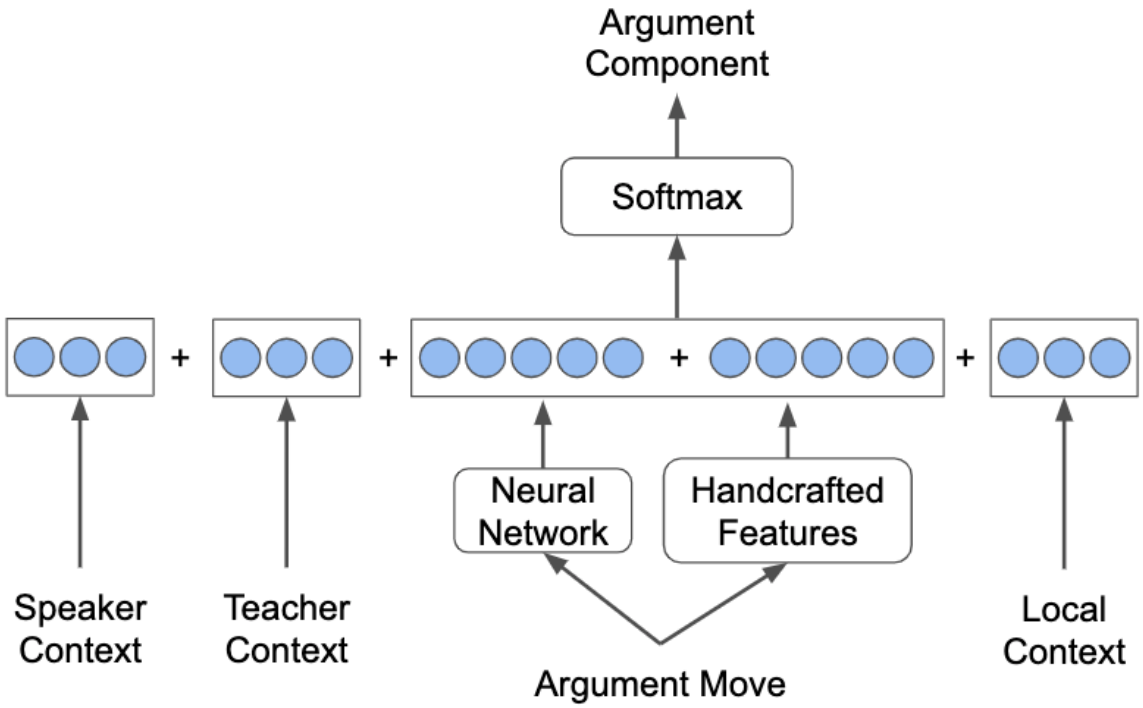


Figure 5: Proposed model incorporating speaker-dependent context features.

word embeddings in an ADU, and a softmax layer is used as classifier.

6.3.1 Local Context

The first component, named *Local Context*, has the objective of extracting information from the immediate vicinity of target ADU in order to capture short-term argumentative patterns. For example, it may be highly likely to have a piece of evidence right after a claim. Likewise, a warrant is highly likely to appear right after a claim and evidence pair. We define local context as ADUs preceding and/or following the target ADU, regardless of the speaker voicing them. In this setting, context size is measured in terms of complete ADUs used to extend the baseline, and context position refers to the relative position of these ADUs to the target one (i.e. preceding, following or both). We address the main limitations of prior work on local context in three ways:

1. We analyze the impact of varying both context location and size on system performance instead of arbitrarily picking a context position and only changing size;
2. We model contextual information the same way we model the target ADU;
3. We propose a method for automatically optimizing context size/position.

For item (1), we consider context sizes from 1 to 6, and 3 positions: *preceding*, where all context ADUs come before the target ADU; *following*, where context follows the target ADU; *both*, where context is centered around the target ADU. The choice of maximum context size 6 was made based on results showing diminishing returns and the consideration that increasing it further would go beyond "local" context. With respect to item (2), for each of the two baselines, the complete model is used to generate ADU embeddings for all ADUs in the context. They are then concatenated with the target ADU embedding to form the final feature vector, which is followed by a softmax classifier. In this case the dimensionality of the final feature vector grows linearly with context size. This would be a problem for the hybrid baseline since it already has 7000+ features. To reduce overfit, we simplify the hybrid baseline model by reducing the handcrafted feature set to the Speciteller feature set consisting of 114 features - reducing the dimensionality to 2514 (2400 for the neural network part and 114 for handcrafted features). The BERT baseline is less prone to overfitting since its dimensionality is only 768. Item (3) is implemented by changing the way context ADU embeddings are processed. Context size is set to the maximum size (6 in this case) and an attention layer [Luong et al., 2015] learns weights to compute a weighted average and aggregate the whole context into a single embedding. By learning attention weights, the model will automatically decide the importance of ADUs at each position and therefore optimize context size/position.

6.3.2 Speaker Context

The second component, named *Speaker Context*, has the objective of capturing information related to the student who is the source of the current argument move. This module analyzes all the student's contributions to the discussion so far, and it is aimed at capturing the student's propensity towards certain argument components, e.g. whether the student

tends to provide many unsubstantiated claims, or whether they tend to explain their evidence at all. Table 14 shows the distribution of argument components for each student participating in one of the discussions in dataset D3. Among the six students participating

Table 14: Argument component class distributions for a discussion about the book "Into the wild".

Student	**Argument Move %**		
	Claim	**Evidence**	**Warrant**
3	100	0.0	0.0
6	91.1	8.9	0.0
7	91.8	6.1	2.1
9	80.0	17.5	2.5
10	84.6	15.6	0.0
14	90.0	10.0	0.0

in the discussion, four of them did not provide warrants. Student 3, for example, always provided unsubstantiated claims throughout the discussion. Students 9 and 10 stand out when considering the amount of evidence given. Even among students 7 and 9, the only ones who occasionally provide warrants, there is a wide gap in the amount of evidence given. By considering individual student propensities towards argument components, the classifier can make more informed predictions. The speaker context features will be particularly helpful in small groups discussions, since the average number of turns per student is typically higher than whole class discussions, providing more information for these features to capture.

In our datasets we typically have access to speaker ID, however the ground truth labels of a speaker's previous ADUs are not available when making predictions for a discussion. Therefore, this information needs to be extracted from the student's ADU text. In order to produce the output feature vector, the speaker context module is composed of two parts as shown in Figure 6: each of the student's previous argument moves is converted into an ADU embedding, and a recurrent neural network (RNN) produces the final speaker context vector by analyzing the sequence of embeddings.

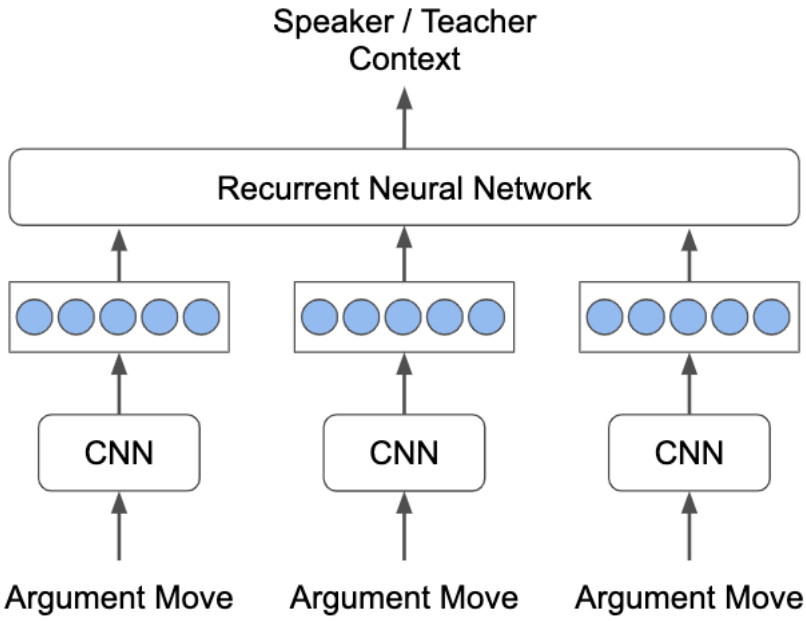


Figure 6: Architecture of the *Speaker Context* and *Teacher Context* components.

Given the speaker ID for the target ADU, the speaker context module performs the following steps: (1) gather the speaker's previous ADUs from the discussion; (2) convert each ADU into an embedding; (3) aggregate them into a single, fixed-size feature vector and concatenate it with the baseline. Step 1 is achieved by simply filtering out ADUs based on speaker ID, which is readily available in each discussion. In step 2 for the hybrid baseline we use a CNN to generate ADU embeddings. Since the number of parameters in the RNN for next step is highly dependent on its input size (i.e. the output dimensionality of the CNN from this step), we decided to implement an additional, simpler CNN just for encoding a student's previous ADUs. The CNN is based on the same structure as the one in the hybrid baseline model, but with the number of filters reduced from 16 to 4. This resulted in a 200-dimensional vector. For the BERT baseline the same embedding - average pooling model was used in this step, given that we are already using the BERT model with smaller dimensionality. Step 3 was accomplished using a Long Short-Term Memory (LSTM) network [Hochreiter and Schmidhuber, 1997], though Gated Recurrent Unit (GRU) [Bahdanau et al.,

2014], as well as bidirectional versions of the two were tested and showed similar results. Similarly to local context, we also implemented an attention-based mechanism for optimizing context size. In this scenario, we set the speaker context size to its maximum (40 in this case) and learn attention weights that decide the importance of each ADU in the context.

6.3.3 Teacher Context

The third component relates to information that we have disregarded entirely so far: teacher talk. We observed during data collection and analysis that the involvement of the teacher can vary greatly between different discussions. This is in part due to the organization of the discussion: most teachers believe they should keep their contributions to a minimum, intervening only if the discussion gets off topic and not influencing the flow of the discussion, in a Socratic seminar style. While this belief is shared among most teachers, it is hard to operationalize in practice. Several other factors can influence the frequency of teacher talk: the size of the class and discussion group, the presence/absence of dominating students in the discussion, and (perhaps the most important) whether students feel comfortable expressing their ideas in front of others. In such cases the teacher has to step in to guide the discussion, asking questions and possibly asking individual students for clarification, thereby significantly contributing towards the discussion as a whole. Table 15 shows an excerpt of another discussion in our corpus.

At the beginning of the discussion, student 8 is mainly making claims but struggles to provide evidence, therefore the teacher has to constantly probe for more information. This back-and-forth between student and teacher continues for 14 turns, after which other students join the discussion. Later in the discussion, the teacher intervenes with the explicit intention of getting students to provide more evidence, with probes such as: *"Can you give me an example of something that he does or that he says to prove that he's anxious?", "Can you give me an example in real life, St 4, of when that would be the case? How you can understand situations better by just observing not participating?", "Other examples of when it would be good to be a wallflower?", "What was happening?"*. Overall, the teacher contributed to the discussion with 43% of the turns. By incorporating teacher talk in our

Table 15: Excerpt from a discussion on the novel "The perks of being a wallflower".

Turn	Student	Talk
3	Teacher	Yes, his sister and her boyfriend. And why does he feel bad for them?
4	8	Because he knows they're not in a good relationship.
5	Teacher	Why not?
6	8	Well for one, that he's just not a good guy or-
7	Teacher	Why?
8	8	Well, for one example is he hit her, and then he just ... Like you said, he's the guy who's going to be throwing up in the bushes at the party house and stuff like that and just uncontrollable.
9	Teacher	So if he's not a good guy, like you said St 8, why does Charlie feel bad for him?
10	8	Because he has to deal with himself. You know?
11	Teacher	You mean the boyfriend?
12	8	Yeah. Like he feels bad for him because that's his own self that he has to deal with. Like when his sister breaks up with him, that's going to be the last that he sees of him. But himself, he has to live with being a bad person, unless he changes.
13	Teacher	How do you know the sister breaks up with him though?
14	8	Well if they do. You know what I mean?
15	Teacher	But if they didn't, then would he still feel bad for him? Would Charlie still feel bad for him?
16	8	Yeah, probably.

model we can capture many text-related concepts that our previous work overlooked (e.g. Charlie feeling bad, a breakup, being anxious).

Similarly to the speaker context module, the *Teacher Context* module will capture the sequence of statements made by the teacher from the beginning of the conversation to the current argument move. The teacher context features can be extracted in the same way as the speaker context features: generating an embedding for each teacher turn and feeding the embeddings in a recurrent neural network (see Figure 6). In order to keep the computational complexity to a minimum and reduce the risk of overfitting the parameters of the neural networks for speaker context and teacher history can be shared entirely (when using both context types).

After examining the results obtained by modeling teacher talk through a neural network,

we found that the small performance benefit may not be worth the increase in computational complexity (the number of parameters in the model increases from 33 thousand to 155 thousand). We then tried replacing the neural teacher context model with a simple, concise set of handcrafted features aimed at capturing teacher contributions at both local (i.e. within a few ADUs of the target ADU) and global level (i.e. going all the way back to the start of the discussion). They mainly focus on capturing two aspects, amount and frequency of teacher talk. The handcrafted features consist of:

- binary feature capturing if the teacher spoke right before the target ADU;
- number of teacher turns in the previous 3 turns;
- number of teacher words in the previous 3 turns;
- number of ADUs since the last teacher turn;
- number of words in the last teacher turn;
- number of teacher turns so far;
- number of teacher words so far.

This feature set is simply concatenated to the other handcrafted features in the hybrid baseline model.

6.4 EXPERIMENTS AND RESULTS

In order to test different research hypotheses we carried out multiple experiments. The components described above were first individually tested in order to understand what contextual information is actually needed for argument component classification. The results of each experiment will be compared to its appropriate baseline (hybrid baseline and BERT baseline) described in the previous section. Datasets D3 and D4 (see Chapter 2) were used in the following experiments. Class distributions for D3 and D4 are shown in Tables 16 and 17.

Table 16: Distribution of class labels for argument component type for dataset D3.

	Annotation	Total Count	Percentage
Argumentation	Claims	2047	65.30%
	Evidence	762	24.30%
	Warrants	326	10.40%
	Total	**3135**	**100.00%**

Table 17: Distribution of class labels for argument component type for dataset D4.

	Annotation	Total Count	Percentage
Argumentation	Claims	1402	72.19%
	Evidence	345	17.77%
	Warrants	195	10.04%
	Total	**1942**	**100.00%**

6.4.1 Local Context

In the first experiment we want to understand the impact that the local context module described above has on the performance of the argument component classification system.

Figure 7 shows results obtained by extending the two baselines with the local context module.

Table 18 shows the best results obtained for each experimental setting. We report unweighted Cohen kappa along with macro precision, recall and F-score.

From the plots in Figure 7 (a) we can make several observations about the hybrid model. Firstly, all models containing any contextual information outperform the context-free baseline. With respect to context size, there seems to be an inflection point after which we see diminishing returns, though the specific value is different for all three position settings. The overall best results were obtained by including both preceding/following local context of

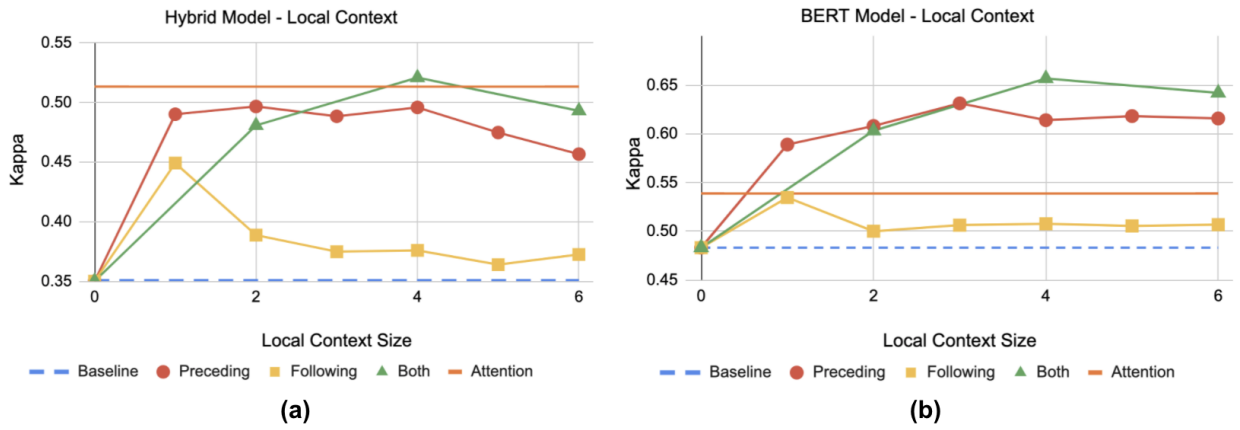


Figure 7: Result plots for adding local context to the hybrid baseline and BERT baseline on dataset D3.

Table 18: Local context results for different experimental settings. Each row shows the best results for the corresponding settings when varying context size. Bold font shows the best results for each model.

Row	Model	Local Context	Kappa	Precision	Recall	F-score
1	Hybrid Baseline	-	0.350	0.535	0.531	0.509
2		Preceding	0.497	0.646	0.691	0.657
3		Following	0.449	0.593	0.641	0.604
4		Both	**0.521**	**0.657**	**0.727**	**0.676**
5		Attention	0.513	0.644	0.706	0.659
6	BERT Baseline	-	0.483	0.620	0.669	0.632
7		Preceding	0.631	0.740	0.775	0.752
8		Following	0.535	0.655	0.697	0.667
9		Both	**0.657**	**0.759**	**0.787**	**0.769**
10		Attention	0.539	0.657	0.704	0.672

size 4; in this setting we obtained significantly better precision and f-score (p-value < 0.05) compared to the same setting with context size 2 (green line in the figure). With respect to position, all models including either preceding or both context positions perform significantly better than the baseline (p-value < 0.01). For following context position the performance improvement is more sensitive to context size: a statistically significant improvement is obtained for context size 1, but for larger context sizes the improvements are not statistically significant. Given that model complexity increases linearly with context size, larger datasets may benefit more from larger context sizes. It is important to note that all the findings discussed so far are also valid for the BERT baseline, which is evident by comparing Figures 7 (a) and (b). Additionally, for each comparable experimental setting, BERT models achieve higher performance than the hybrid model. We also argue that both context position and size should be optimized, unlike prior works which picked a single position and optimized context size. Context position may arguably be even more important than size: Figure 7 shows that the difference between each line (i.e. for different context positions) is bigger than differences observed within each line (i.e. for different context sizes in a given position).

One finding that is not consistent across the two models is the use of attention to optimize context size/position. In the hybrid model, the attention weights learned during training are able to effectively aggregate all context ADUs. This is evidenced by the results showing that the attention model achieves performance that is not statistically significantly different than the best overall results. For the hybrid model, therefore, context size and position can be automatically optimized during training by trading a marginal performance penalty. For the BERT model, however, the attention model is not effective and results show it only outperforms the baseline and context including only ADUs after the target one. In both scenarios where any ADUs before the target one are included in context, the performance of the attention model is significantly worse than those with static context size/position (p-value < 0.05), though still significantly better than the baseline (p-value < 0.05). We speculate that this is due to the BERT model already being entirely based on attention mechanisms. Word embeddings in BERT are dependent on all other words within an ADU, and by using the average pooling layer, this model produces ADU embeddings that are more similar compared to the hybrid model. The ineffectiveness of the attention mechanism in the

BERT model is clear when looking at the actual attention scores. In an effective attention model we would expect the attention score to vary considerably across the context ADUs, so that the weighted average reflects difference in importance of those ADUs. In one of the experiments we conducted, we analyzed the standard deviation of attention scores across context ADUs for both models. We found that, on average, the attention score standard deviation in the attention model was 0.25 for the BERT model, and 19 for the hybrid model. In synthesis, the attention layer for the BERT model produces very similar scores for all context ADUs, acting almost as a simple, unweighted average. For these reasons we believe that, in models that are heavily attention-based, context position and size are hyperparameters to optimize during training.

6.4.2 Speaker Context

In the second experiment we will extend the baseline models with the speaker context module, to understand if taking into consideration individual ADUs for each speaker will be helpful for argument component classification. We define speaker context size K as the K closest prior ADUs to the target ADU, voiced by the target student. In this experimental setting we varied context size between 5 and 40, effectively capturing all of the student's prior ADUs since the beginning of the discussion.

Figure 8 shows results obtained by extending the two baselines with the speaker context module, and the best results for each experimental setting are displayed in Table 19.

Both the hybrid and the BERT model share several findings. Adding speaker context, of any size, significantly increases precision, recall and f-score (p-value < 0.05) over the context-free baseline. With increase of context size, we either see a diminishing return effect or a plateau effect, suggesting that in general a speaker's most recent ADUs are more important than earlier ones. By comparing the purple lines in Figure 8 to the green lines in Figure 7 we can see that, individually, adding local context benefits models more than adding speaker context. This is consistent with our expectations given that we are analyzing multi-party discussions, and at any point in time students try to argue with respect to what other students have just said beforehand, rather than to what they have previously said

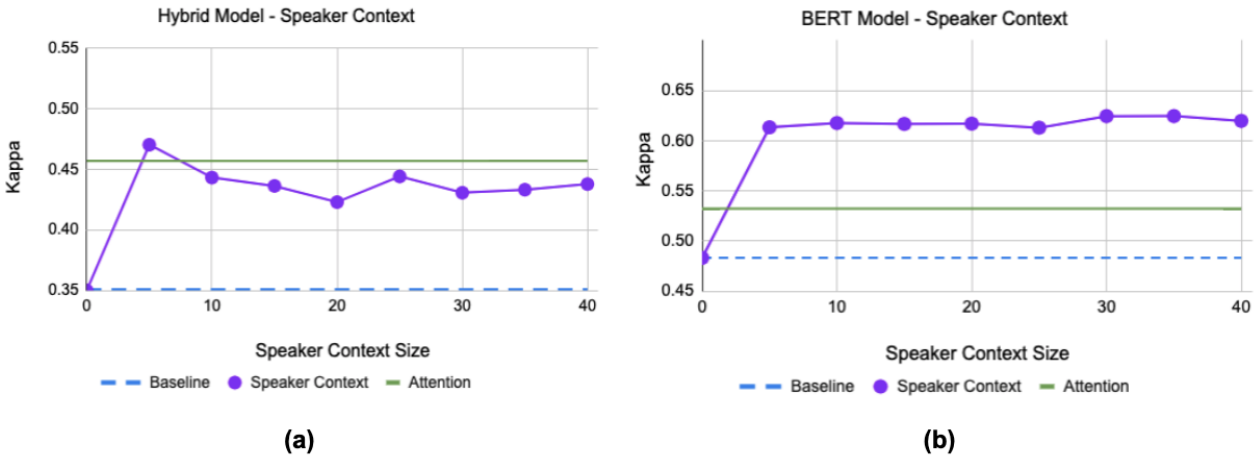


Figure 8: Result plots for adding speaker context to the hybrid baseline and BERT baseline on dataset D3.

Table 19: Speaker context results for different experimental settings. Each row shows the best results for the corresponding settings when varying context size. Bold font shows the best results for each model.

Row	Model	Context	Kappa	Precision	Recall	F-score
1	Hybrid Baseline	-	0.350	0.535	0.531	0.509
2		Speaker Context	**0.470**	**0.626**	**0.682**	**0.636**
3		Attention	0.457	0.605	0.627	0.603
4	BERT Baseline	-	0.483	0.620	0.669	0.632
5		Speaker Context	**0.625**	**0.733**	**0.794**	**0.751**
6		Attention	0.532	0.649	0.706	0.663

themselves. It is again worth pointing out that modeling speaker context in this way is beneficial for two distinct argument component classification models. Additionally, for each respective experimental setting the BERT model always outperforms the hybrid model.

Diverging findings emerge when considering attention mechanisms for automatically optimizing speaker context size. In the hybrid model the attention mechanism works as intended, with results that are not statistically significantly different than the best results, albeit slightly worse. In this case, again, at the expense of a marginal performance loss we can let the model choose which context ADUs are more important than others when aggregating them. For the BERT model on the other hand, the attention mechanism cannot effectively understand the relative importance of each context ADU and the resulting model performance are significantly worse than any fixed-size speaker context. At the same time, however, the attention model is still statistically significantly better than the baseline (p-value < 0.05).

6.4.3 Teacher Context

In this experiment we will examine the importance of teacher talk in our argumentation model. Like the previous experiment, we changed the size of teacher context from a minimum of 5 to a maximum of 40. Only the hybrid baseline was used in this experiment.

Figure 9 shows results obtained by extending the baseline with the teacher context module, and the best results are highlighted in Table 20.

As Figure 9 shows adding the speaker context module always outperforms the baseline result, and we see a similar diminishing return effect when increasing context size as for other context types. Unlike previous results though, none of the improvements achieved through the neural teacher attention models are statistically significantly better than the baseline. This is somewhat disappointing given the increase in model size that this module requires. We hypothesized that this may be due to the variation in teacher talk across discussions: the average percentage of teacher talk in dataset D3 is 24.22, while standard deviation is 18.05 (in terms of raw number of teacher turns per discussion, the average is 41.10 and standard deviation is 38.79). We ran additional cross-validation experiments where, for each fold, we only kept in the test set discussions with percentage of teacher talk above average. Regrettably this evaluation did not show statistically significant improvements either.

We then replaced the neural teacher context modules with the handcrafted teacher con-

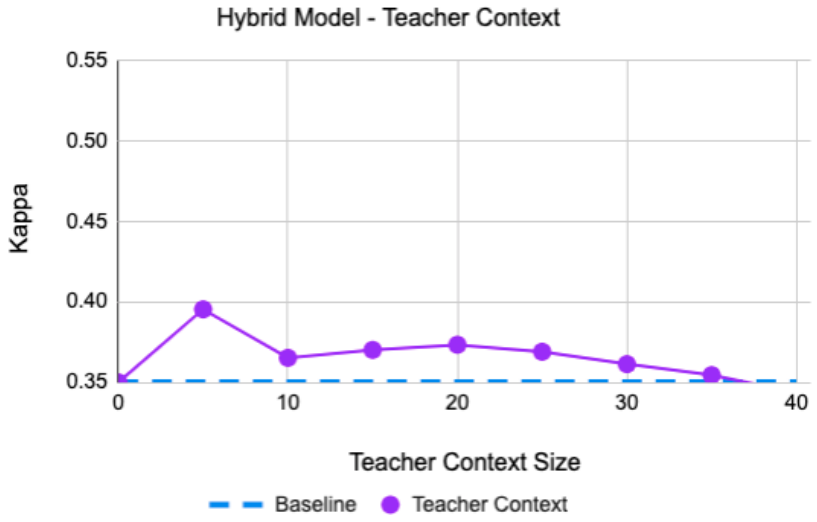


Figure 9: Result plots for adding teacher context to the hybrid baseline on dataset D3.

Table 20: Teacher context results for different experimental settings. Each row shows the best results for the corresponding settings when varying context size. Bold font shows the best results.

Row	Model	Context	Kappa	Precision	Recall	F-score
1	Hybrid Baseline	-	0.350	0.535	0.531	0.509
2		Teacher Context	0.395	0.558	0.575	0.552
3		Teacher Features	**0.441**	**0.597**	**0.633**	**0.594**

text feature set and repeated the cross-validation experiment (the complete cross-validation experiment with all discussions in each test set). In this setting we obtained better performance than both the baseline and the neural teacher context, as shown in row 3 of Table 20, and the improvements over the baseline are statistically significant (p-value < 0.05).

6.4.4 Combining Context Types

In the fourth experiment, we combined the different context modules to test whether our argumentation models can benefit from multiple sources of context simultaneously.

We started by combining teacher context with local context in the hybrid model. We extended the best local context setting (i.e. ADUs preceding and following the target, with context size 4), which achieved Kappa 0.521 and F-score 0.676, with the neural teacher context module. When varying teacher context size, we observed a drop in performance with Kappa ranging from 0.448 to 0.481, and in some cases the difference was statistically significant compared to local context only (p-value < 0.01). Replacing the neural teacher context module with teacher context features did not improve performance in this case: the result was Kappa 0.451 and F-score 0.626, significantly worse than local context alone (p-value < 0.05).

We then combined speaker context and teacher context. We extended the best performing speaker context setting, where Kappa $= 0.470$ and F-score $= 0.636$, with both teacher context modules and observed Kappa between 0.424 and 0.455 and F-score between 0.604 and 0.629, some of which are significantly worse than speaker context alone.

These findings lead us to believe that the performance loss when combining teacher context with other context types is not only due to the increase in model complexity (remember that speaker context and teacher context share the same model/weights). Therefore, although teacher talk can be beneficial for argument component classification when other context information is not available, it should be given lower priority than other context types.

Lastly, we combined local context and speaker context. First we picked the best local context configuration, and extended it with the speaker context module varying speaker context size. We then repeated the experiment with other local context configurations, i.e. picking the best local context size for each position.

By comparing the purple and grey lines in Figure 10 (a) we can see that combining both context types always outperforms speaker context alone. All improvements are statistically significant (p-value < 0.05) for speaker context size > 5. Same trends were observed when

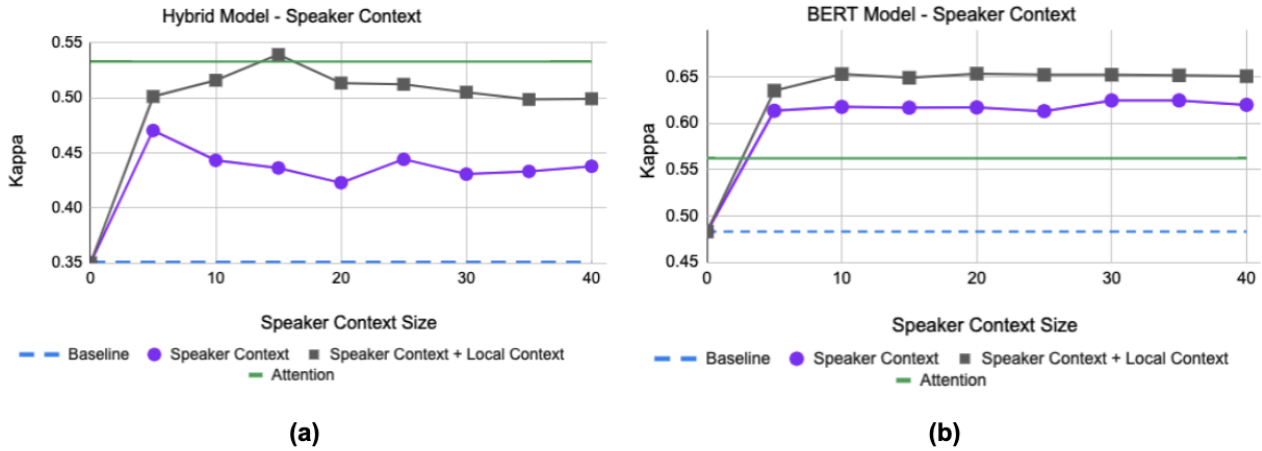


Figure 10: Result plots for adding local context and speaker context to the hybrid baseline and BERT baseline on dataset D3. For the "Speaker Context + Local Context" line, local context position was set to "both" and size to 4.

considering other local context settings, where improvements over speaker context only were also statistically significant (p-value < 0.05). Similar observations can be made with respect to local context only, though the performance improvement of the two combined context types is not as high, and only occasionally statistically significant. We also combined both attention-based models, where both local and speaker context are automatically optimized. It is encouraging to see that difference between the attention model and the best overall result is relatively small and not statistically significant. Additionally, this attention-based model significantly outperforms the speaker context attention-based model (p-value < 0.005).

Similar observations hold true for the BERT models as shown in Figure 10 (b), though differences between individual context and two context models are smaller compared to the hybrid model, and less frequently statistically significant. In particular, the best overall Kappa for this model was obtained with local context only, while both local and speaker context performed equally or better for precision, recall and F-score. As for individual context models, automatically optimizing local and speaker contexts with attention mechanisms

Table 21: Local Context and Speaker Context results for different experimental settings. Each row shows the best results for the corresponding settings when varying context size. Bold font shows the best results for each model. Rows for baseline results and individual context type results are repeated from previous sections.

Row	Model	Context	Kappa	Precision	Recall	F-score
1	Hybrid Baseline	-	0.350	0.535	0.531	0.509
2		Local	0.521	0.657	**0.727**	0.676
3		Speaker	0.470	0.626	0.682	0.636
4		Local + Speaker	**0.539**	**0.674**	**0.727**	**0.689**
5		Attention	0.533	0.664	0.725	0.681
6	BERT Baseline	-	0.483	0.620	0.669	0.632
7		Local	**0.657**	**0.759**	0.787	0.769
8		Speaker	0.625	0.733	0.794	0.751
9		Local + Speaker	0.653	**0.759**	**0.810**	**0.774**
10		Attention	0.562	0.673	0.722	0.686

results in significantly lower performance compared to the best overall (p-value < 0.05).

One important consideration when jointly modeling local context and speaker context relates to model robustness. We observed that it is easier to consistently achieve higher results, on average, when modeling both context types compared to including either context type separately in a model. In other words, optimization of local context position, size, and speaker context size is less influential on performance. As an example, using the hybrid model it is much easier to achieve a Kappa > 0.5 with both context types, in several configurations, compared to either context separately: local context alone is only able to achieve this in the best possible setting, and it is not even possible with speaker context alone.

6.4.5 Cross-dataset Experiment

In the last experiment on context information for argument component classification, we are interested in analyzing the performance of the context modules in a cross-dataset setting. We experimented with the BERT model on datasets D3 and D4.

We first performed a 10 fold cross-validation experiment on dataset D4 to see if our earlier findings are valid for a different dataset, and to obtain baseline performance figures on D4 itself since this is a much smaller dataset than D3 (1942 ADUs compared to D3's 3135).

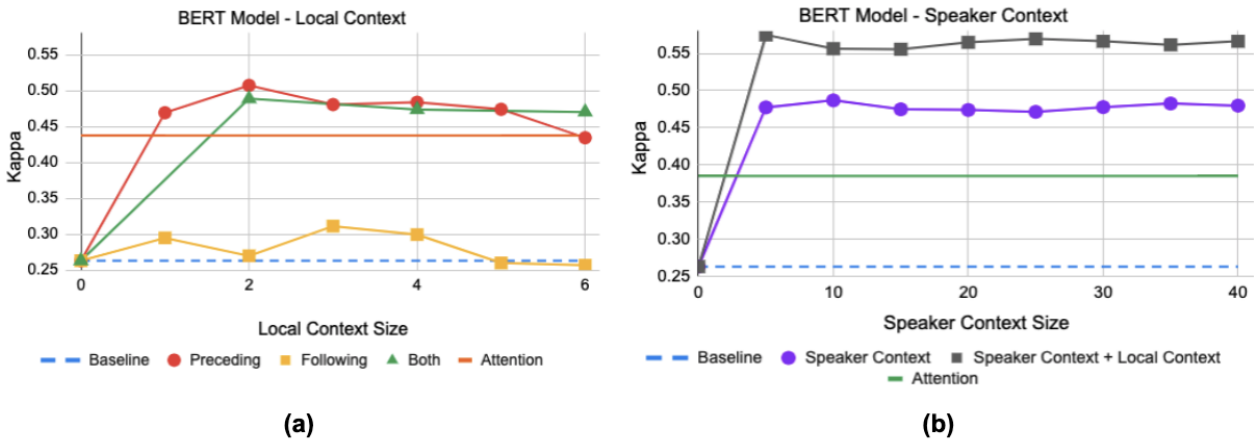


Figure 11: Result plots for adding local context and speaker context to the BERT baseline on dataset D4. For the "Speaker Context + Local Contex" line, local context position was set to "both" and size to 4.

Figure 11 shows that findings on D4 are largely consistent with those on D3. If we compare respective experimental settings, we find that on average the results for D4 are 0.1 to 0.15 lower than those on D3. We believe this is within expectation given the large difference in number of ADUs. The best performance for both local and teacher context on D4 is achieved with smaller context size compared to D3. This is also expected since bigger context size results in larger model, which may require more data to properly train. In general, though, the performance gains of our context models over the baseline are larger

Table 22: Cross-dataset experimental results.

Row	Evaluation	Context	Kappa	Precision	Recall	F-score
1	D4 cross-validation	-	0.263	0.514	0.539	0.501
2		Local + Speaker	0.575	0.753	0.721	0.718
3	Training D3, test D4	-	0.375	0.552	0.661	0.571
4		Local + Speaker	0.599	0.731	0.805	0.755
5	D3 cross-validation	-	0.483	0.620	0.669	0.632
6		Local + Speaker	0.653	0.759	0.810	0.774
7	Training D4, test D3	-	0.401	0.616	0.572	0.586
8		Local + Speaker	0.572	0.703	0.767	0.716

for D4, which indicates that our proposed models are effective in cases when less labeled data is available.

For cross-dataset experiments, we first evaluated the BERT baseline when training on D3 and testing on D4 (and vice versa). We then chose the context settings with the best results in D3 cross-validation, and tested its performance on D4 (and vice versa).

A number of interesting observations can be made from Table 22. First, within-dataset performance is higher than cross-dataset performance when testing on the larger dataset D3, while the smaller dataset D4 can benefit from models being trained on larger corpora. Second, when testing on D4 the performance gap between baseline models (row 1 vs. 3) is on average larger than that between models with context information (row 2 vs. 4). This may be a signal that our context models make it easier to achieve better performance even for smaller dataset, though more experimentation with additional datasets is required to properly confirm this hypothesis. Third, our proposed context models almost always outperform the baseline even in cross-dataset scenarios.

6.5 SUMMARY

In this section, we proposed and evaluated context-aware extensions of our previous models for argument component classification for classroom discussions. We implemented three context types: *(i)* local context, to model the immediate vicinity of the target ADU at any point in the discussion; *(ii)* speaker context, to capture individual speaker behaviors with respect to their own prior ADUs in the discussion; *(iii)* teacher context, to capture teacher/student turn taking patterns and the content of teacher talk. We first evaluated each context individually and found that all contextualized models outperform the baseline, context-free model. After thorough performance evaluation with respect to each context type's parameters, we proposed extensions that use attention mechanisms to automatically optimize context size and/or position. We then combined the different context types to train more robust argument component classification models. Finally, we analyzed the performance of contextualized models on an additional dataset and in cross-dataset settings.

The findings described in this section were published in [Lugini and Litman, 2020].

7.0 JOINT LEARNING OF DIFFERENT ASPECTS OF CLASSROOM DISCUSSIONS

7.1 INTRODUCTION

Multi-task learning has shown potential for improving robustness of neural network models in Computer Vision community [Girshick, 2015], and it is becoming increasingly popular in NLP applications as well. The core premise of multi-task learning is to exploit information from multiple, related tasks in order to better train neural network models that will outperform individual models trained on the respective individual tasks. Collobert and Weston [Collobert and Weston, 2008] proposed a general architecture based on a convolutional neural network and used multi-task learning to jointly train the model on several tasks: Semantic Role Labeling, Part of Speech tagging, Chunking, Named Entity Recognition, Semantically Related Words, and Language Modeling. They showed that models trained jointly on multiple tasks were always able to outperform single-task models, and in several cases they required less training time. Dong et al. [Dong et al., 2015] showed that jointly training a neural machine translation model on multiple languages simultaneously can significantly outperform models learned independently on individual language pairs. Liu et al. [Liu et al., 2015] trained a deep neural network jointly on two tasks, semantic classification (query classification) and information retrieval (ranking for web search). The proposed model consists of shared lower-level representations connected to task-specific representations, which are then used as input to task-specific classifiers. They found that multi-task learning has a regularization effect and is crucial in learning more general representations and avoiding overfitting to a specific task. More recently, Schulz et al. [Schulz et al., 2018] investigated multi-task learning in the context of argumentation mining. They proposed a neural network

in which the primary task of argument component identification and classification (posed as a sequence tagging problem) is augmented by additional tasks. The additional tasks consist of the same task as the primary task when training on different datasets. This approach can essentially be viewed as a transfer learning approach framed as multi-task learning since the learning gains come from a difference in the source domain of the data rather than difference in tasks to perform. It is then not surprising that multi-task models outperformed individual models, especially when limited amount of in-domain data is available for the primary task compared to secondary tasks. Lauscher et al.[Lauscher et al., 2018] additionally showed that argumentation can be used as an auxiliary task to improve other primary tasks. They evaluated the performance of neural network models on several tasks related to analyzing the rhetoric of scientific writing: discourse role classification, subjective aspect classification, citation context identification, and summary relevance classification. Their claim that rhetoric analysis can be more effective when including argumentative information is backed up by their experimental results: each of the four tasks benefited from having argument component identification and classification information as a secondary task. The main difference between primary task and secondary tasks resides in the loss function, which is typically optimized with respect to the primary task, essentially prioritizing it over all secondary tasks. The fact that argument component classification performance did not increase in any of the four multi-task experiments is evidence that argumentation was strictly an auxiliary task to which the models assigned considerably less importance during training.

7.2 PILOT STUDY ON DATASET D2

Motivated by the related works described in the previous section, we investigated the impact of multi-task learning on the analysis of classroom discussions. The multi-task models we envision differ from prior works in the following aspects:

- unlike Schulz et al. [Schulz et al., 2018], which showed that additional data containing the same labels can be effectively used through multi-task learning, our objectives are: to *(i)* make use of additional labels (for related tasks) for the same dataset, without increasing

the overall dataset size; *(ii)* to model argument component classification together with tasks beyond argumentation mining through multi-task learning;

- unlike Lauscher et al. [Lauscher et al., 2018], our main objective is to improve performance of all tasks in multi-task settings, therefore we do not make the distinction between primary and secondary tasks (which typically depends on the definition of the loss function); additionally, we consider more than two tasks simultaneously.

Figure 12 shows the setup of the different multi-task models for the experiments carried out in this chapter. In all the models an ADU is processed through an ADU encoder (the

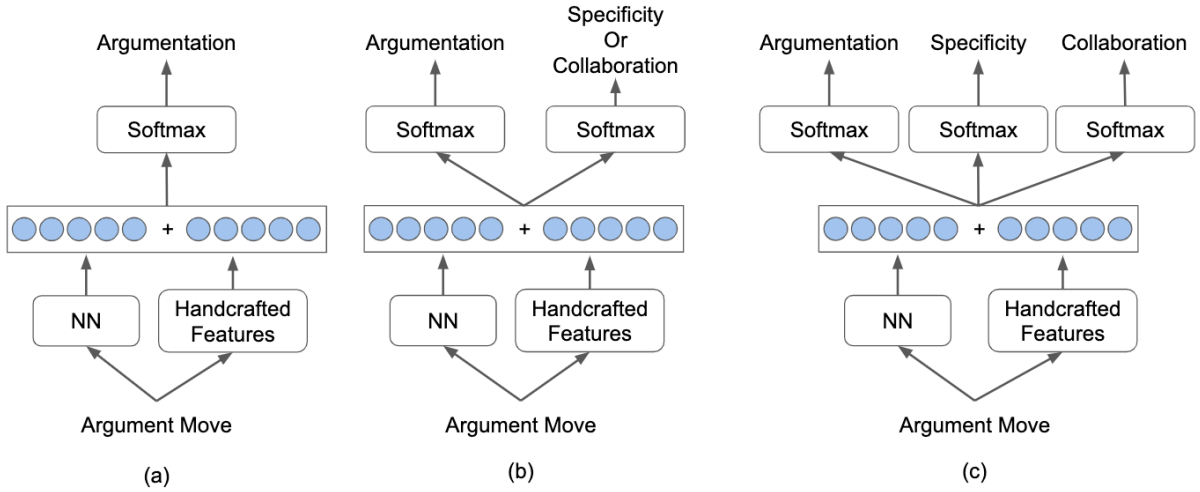


Figure 12: Configurations of multi-task models: baseline single-task model (a); two-tasks model (b); three-tasks model (c).

neural network blocks in the figure, labeled "NN") and the handcrafted features described in Section 6.3.1 are extracted simultaneously. The two are concatenated and form a feature vector which is shared among all tasks in the multi-task models. In the single-task model in Figure 12 (a) the feature vector is given as input to a single softmax classifier.

As a first experiment we will add a secondary task to the model (Figure 12 (b)), which will consist of predicting specificity. In a second experiment, we will test the performance of a single multi-task model trained jointly on all three tasks. Argumentation and specificity are based on the same unit of analysis, while collaboration is coded at the turn level and will

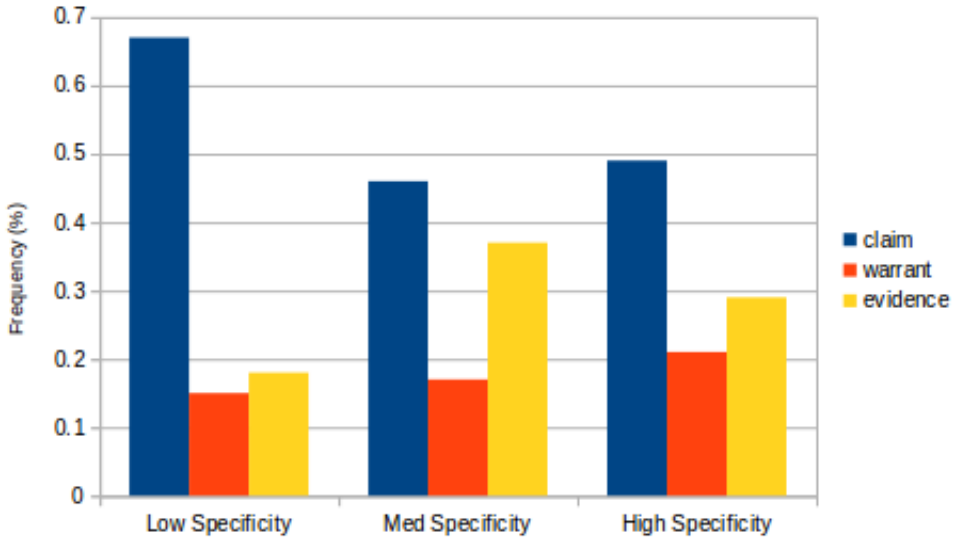


Figure 13: Distribution of argument components by specificity level.

require pre-processing. We addressed this problem by applying BIO tags instead of using the raw collaboration labels, so as to have a label for each ADU and be aligned with the other two tasks: if an ADU is at the "beginning" of a turn, we add the prefix "B-" to its class label (e.g. B-extension), otherwise we add the "inside" prefix "I-" (e.g. I-extension); the "O-" prefix is never used.

Preliminary Experiments and Findings We conducted a preliminary experiment on multi-task learning with the objective of improving performance of an argument component classification system on the dataset D2. We first analyzed the distribution of argument components for different specificity levels, and Figure 13 shows that the frequency of each argument component varies considerably between low, medium, and high specificity ADU.

The amount of information shared between the two classes can also be analyzed through the normalized mutual information (NMI) metric, which can be used as an indicator of performance gain obtained in multi-task settings [Bjerva, 2017]. The metric ranges between 0 an 1, where 1 means the two classes are perfectly correlated and 0 represents no correlation at all; larger gains in performance can be expected from class pairs with higher NMI. The

normalized mutual information between argument components and specificity in dataset D2 is 0.039.

We implemented multi-task learning as depicted in Figure 12 (b), where the feature vector for the two tasks is entirely shared, and is directly given as input to two softmax classifiers. Table 23 shows the argument component classification results for models trained using multi-task learning and specificity as second task (rows 13-20) along with the respective results obtained by the single-task models (rows 5-12).

Our findings are in line with the literature in other domains, with results showing that models trained on argumentation and specificity labels almost always outperform the ones trained only on argumentation. LSTMs benefit from the multi-task setup more than CNN models: among all combinations of LSTM models, the only one able to achieve kappa greater than 0.2 and F-score greater than 0.4 is a multi-task one. Additionally, the word-level CNN model using wLDA and online dialogue feature sets and trained using multi-task learning is the only model able to achieve F-score greater than 0.3 for warrants. Overall these preliminary findings on D2 are encouraging because the multi-task models outperformed the single-task ones even with a very modest NMI between the two tasks.

7.3 IMPROVING ARGUMENT COMPONENT CLASSIFICATION THROUGH MULTI-TASK LEARNING

Looking at the multi-task learning models described in the previous section, we can see that the portion of weights that is truly entirely shared between tasks is the neural network. Therefore we decided to remove most of the handcrafted features and only keep the Speciteller feature set and only focus on CNN, effectively using the same hybrid baseline model described in Section 6.3.1. We also decided to use dataset D3 for our experiments since it has 3 classes (argument component, specificity and collaboration), unlike D2. Class distributions for dataset D3 are shown in Table 4.

Before running new experiments, we computed the normalized mutual information scores between pairs of classes:

Table 23: Argument component classification results of multi-task learning models on dataset D2 (with specificity as second task). Each line represents the average of a transcript-wise cross validation. Best results are in bold. The three right-most columns represent per-class F-score for evidence, warrants, and claims respectively. For easier comparison rows 5-12 for single-task models are repeated here from Table 13.

Row	Models / Features	Kappa	Prec	Rec	F-score	F_e	F_w	F_c
Character level NN models								
5	LSTM	-0.002	0.062	0.253	0.082	0.007	0.242	0.013
6	LSTM + wLDA + online dialogue	0.034	0.217	0.304	0.150	0.080	0.272	0.090
7	CNN	0.143	0.439	0.423	0.393	0.372	0.218	0.574
8	CNN + wLDA + online dialogue	0.241	0.482	0.475	0.450	0.449	0.236	0.637
Word level NN models								
9	LSTM	0.069	0.408	0.399	0.218	0.161	0.198	0.295
10	LSTM + wLDA + online dialogue	0.181	0.462	0.447	0.391	0.362	0.279	0.522
11	CNN	0.125	0.410	0.404	0.378	0.370	0.231	0.526
12	CNN + wLDA + online dialogue	0.241	0.492	0.488	0.455	0.468	0.276	0.622
Multi-task character level NN models								
13	LSTM	0.060	0.408	0.399	0.208	0.134	0.203	0.287
14	LSTM + wLDA + online dialogue	0.117	0.379	0.375	0.287	0.362	0.279	0.522
15	CNN	0.166	0.444	0.437	0.407	0.399	0.220	0.586
16	CNN + wLDA + online dialogue	0.259	0.506	0.488	0.468	0.474	0.262	0.640
Multi-task word level NN models								
17	LSTM	0.093	0.379	0.364	0.276	0.298	0.252	0.378
18	LSTM + wLDA + online dialogue	0.232	0.497	0.482	0.440	0.419	0.299	0.583
19	CNN	0.164	0.351	0.443	0.441	0.476	0.249	0.598
20	CNN + wLDA + online dialogue	**0.276**	**0.521**	**0.512**	**0.485**	**0.484**	**0.312**	**0.638**

- 0.041 for argument component - specificity;
- 0.314 for argument component - collaboration;
- 0.047 for specificity - collaboration.

All of the results are higher than the one obtained for dataset D2, but in particular the relationship between argument component and collaboration seems the most promising. We additionally performed a sanity check to understand the maximum performance improvement we can expect. First, we added the true collaboration labels as input to the softmax classifier for the baseline argumentation model. Cohen Kappa jumped from 0.350 to 0.563, and F-score improved from 0.509 to 0.664. We repeated the experiment with a reduced collaboration label set, new turn vs. rest, in which we combined extensions, challenges and agreements in one label. It should theoretically be easier for a classifier to learn the reduced collaboration label set, and the performance gains obtained in this setting were: Kappa reached 0.573 and F-score 0.671.

7.3.1 Turn-level Collaboration Classifier

We first evaluated the performance of multi-task learning with the model architecture shown in Figure 12. In this setting the classifiers consist of the hybrid model described in the previous section, and is entirely shared between the 3 tasks. Given the limited performance improvement obtained with this multi-task architecture, shown in Table 24, and the proven potential gain with gold standard collaboration labels, we hypothesized that the current limiting factor is the inaccuracy of the predicted collaboration labels. With many attempts to improve the collaboration classifier, it became clear that there are at least two reasons why the current models struggle, with Kappa ranging between 0.2 and 0.3: *(i)* collaboration is annotated at the turn level, and when turns are segmented into multiple ADUs the collaboration model is missing valuable information; *(ii)* when annotating for a collaboration label, the annotators also provide a reference turn (within the 4 prior turns), meaning that a collaboration label represents the relationship between two turns.

After several iterative extensions of the hybrid model we reached a compromise between model performance and limitations, and the resulting collaboration model is shown in Figure 14. The main limitation of the model is the assumption that the reference turn is known beforehand. We tried relaxing this constraint and adding a context module to capture information from the 4 prior turns, however since the target turn could potentially be related

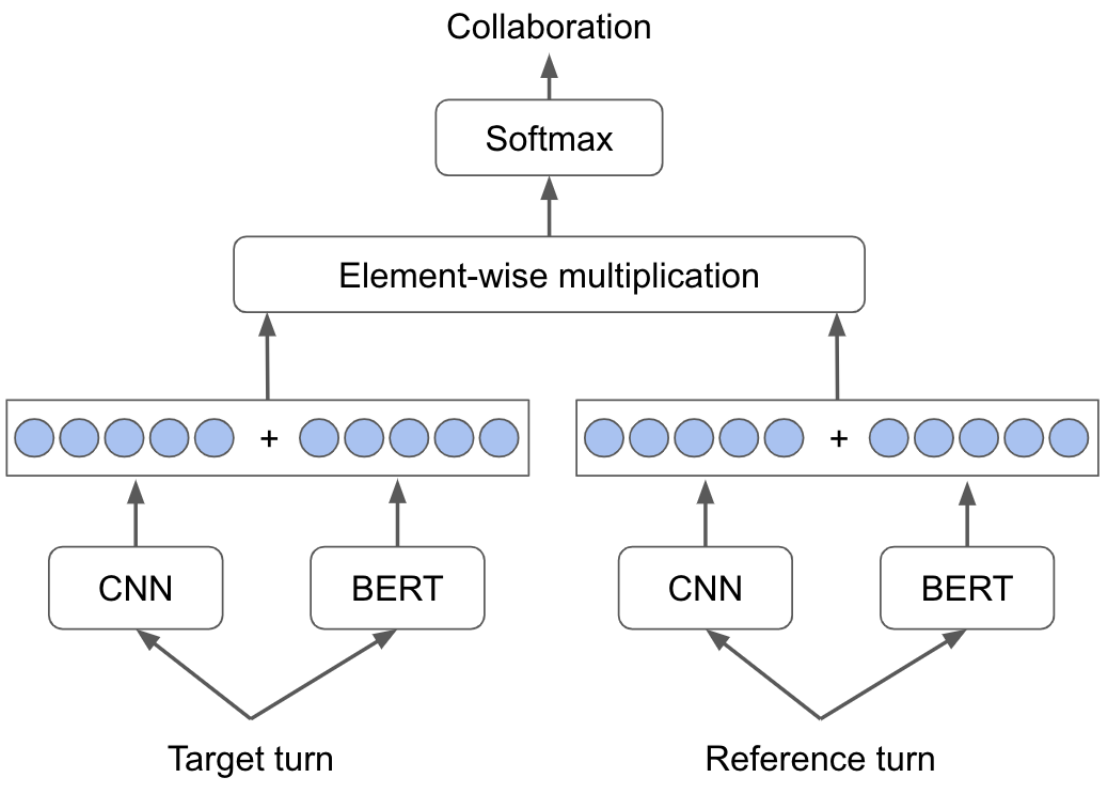


Figure 14: Configuration of the enhanced collaboration classifier which produces one collaboration output for each turn, without the use of BIO tags.

to multiple context turns the simple context model was not able to discern the correct reference and therefore the correct collaboration label. In order to generate turn embedding we use a combination of CNN and BERT embeddings: the CNN is needed as that is the part of the model that will be shared among tasks, while BERT is needed to improve performance. We also found that simply concatenating the target embedding and reference embedding did not work well: the element-wise multiplication explicitly models an approximate similarity between the two embeddings.

7.4 EXPERIMENTS AND RESULTS

In this section we report on several experiments carried out to understand if multi-task learning can improve performance of any of the 3 classifiers, and which ones can benefit from multi-task learning the most. This is achieved by not distinguishing between primary and secondary tasks (i.e. assigning equal weight to each task in the loss function), in the hope that all tasks can simultaneously achieve higher performance compared to their respective single-task models. We evaluated both collaboration models (baseline and turn-level) for the full set of collaboration labels as well as a reduced set described in the previous section.

7.4.1 Pairwise Multi-task Learning

In the first experiment we tested the effect of jointly modeling two tasks, by implementing the models described in Figure 12 (b), where all classifiers produce ADU-level outputs and collaboration uses BIO tags.

Table 24: Results for multi-task models when combining 2 tasks. *Col Set* refers to either full or reduced set of collaboration labels. Per-class Cohen Kappa (QWK for specificity) and macro F-score are displayed. Best results with respect to each collaboration label setting are highlighted in bold. The Baseline Collaboration results refer to row 1 of Table 18.

Row	Tasks	Col Set	K_a	F_a	K_s	F_s	K_c	F_c
1	Baseline	full	0.350	0.509	**0.750**	**0.689**	**0.199**	0.201
2	Arg, Spec	full	0.364	0.529	0.742	0.684	-	-
3	Arg, Col		**0.384**	**0.542**	-	-	**0.199**	**0.213**
4	Spec, Col		-	-	0.739	0.686	0.196	0.200
5	Baseline	reduced	0.350	0.509	0.750	0.689	0.177	0.319
6	Arg, Spec	reduced	0.364	0.529	0.742	0.684	-	-
7	Arg, Col		**0.387**	**0.546**	-	-	**0.243**	**0.370**
8	Spec, Col		-	-	**0.752**	**0.690**	0.223	0.357

From the results in Table 24 we can make several observations. For argumentation and collaboration the multi-task models always outperform the single-task baseline, in both full and reduced collaboration labels. For specificity, on the other hand, there is no large difference in single-task vs. multi-task performance. Additionally, while argumentation and collaboration (for reduced set) can benefit from specificity as second task, the bigger performance gain is obtained between collaboration and argumentation themselves. This was within expectation given the NMI scores in the previous section. Unfortunately the improvements on argumentation are not statistically significant, while those on collaboration for the reduced set are (p-value < 0.05).

7.4.2 Three-task Multi-task Learning

In the second experiment we will use multi-task learning by implementing the model in Figure 12 (c) which makes use of all three tasks simultaneously, to test whether jointly training a model on three tasks is more effective (for some or for all tasks) than training models on two tasks and in single-task setting. Even in this experiment all classifiers produce ADU-level outputs and collaboration uses BIO tags.

Table 25: Results for multi-task models when combining all three tasks. *Col Set* refers to either full or reduced set of collaboration labels. Per-class Cohen Kappa (QWK for specificity) and macro F-score are displayed. Best results with respect to each collaboration label setting are highlighted in bold.

Row	Tasks	Col Set	K_a	F_a	K_s	F_s	K_c	F_c
1	Baseline	full	0.350	0.509	**0.750**	0.689	0.199	0.201
2	Arg, Spec, Col	full	**0.397**	**0.554**	0.745	**0.694**	**0.206**	**0.213**
3	Baseline	reduced	0.350	0.509	**0.750**	**0.689**	0.177	0.319
4	Arg, Spec, Col	reduced	**0.394**	**0.553**	0.741	0.687	**0.241**	**0.368**

Table 25 shows similar trends to the previous experiment: argumentation and collaboration seem to benefit from multi-task learning, while specificity does not. In both ex-

perimental settings with different collaboration label sets, the improvement in F-score for argumentation is statistically significant (p-value < 0.05). For the reduced collaboration set, collaboration performance is significantly boosted by multi-task learning (p-value < 0.05), while the difference for full collaboration set is not statistically significant.

By comparing results between this experiment and the previous one, we can also observe that if a multi-task model already uses argumentation and collaboration labels, adding specificity improves only the argumentation results.

7.4.3 Turn-level Collaboration Classifier

In this experiment we try to improve the performance of the collaboration classifier by testing the model in Figure 14, and see if better collaboration model results in better multi-task results for argumentation.

Table 26: Results for multi-task models when using the new collaboration classifier. *Col Set* refers to either full or reduced set of collaboration labels. Per-class Cohen Kappa (QWK for specificity) and macro F-score are displayed. Best results with respect to each collaboration label setting are highlighted in bold.

Row	Tasks	Col Set	K_a	F_a	K_s	F_s	K_c	F_c
1	Baseline	full	0.350	0.509	**0.750**	**0.689**	0.519	0.482
2	Arg, Spec, Col	full	**0.419**	**0.575**	0.744	0.683	**0.551**	**0.510**
3	Baseline	reduced	0.350	0.509	**0.750**	**0.689**	0.707	0.852
4	Arg, Spec, Col	reduced	**0.424**	**0.581**	0.747	0.685	**0.731**	**0.864**

We can clearly see from Table 26 the same trends we observed in Section 7.4.2, though the magnitude of improvements from single-task to multi-task models for argumentation and collaboration increased. For argumentation, the improvement on recall between rows 2 of Tables 25 and 26 (respectively 0.575 and 0.606) is statistically significant (p-value < 0.05), while that for F-score approaches significance (p-value $= 0.06$). If we compare row 4 across the two tables we also find that increase in recall for argumentation (from 0.565 to 0.601)

is statistically significant (p-value < 0.05). With respect to the single-task baseline, all argumentation metrics in row 2 significantly improved (p-value < 0.01): precision increased from 0.535 to 0.576, and recall from 0.531 to 0.606. We also found statistically significant improvements across all argumentation metrics when comparing row 4 to its baseline in row 3 (p-value < 0.05): precision increased from 0.535 to 0.584, and recall from 0.531 to 0.601.

A large performance improvement was obtained on collaboration, with the turn-level collaboration model significantly outperforming the earlier, non-BERT based, in single-task configuration (p-value < 0.0001). While we are pleased with the performance gain for collaboration, we were expecting the increased accuracy to bring a bigger improvement to argumentation in multi-task setting (the improvement in Kappa was 0.022 and 0.03 for the two collaboration code settings). Overall we are still quite far from the best possible results achievable (Kappa of 0.563 and 0.573), but all the multi-task experiments have shown that the relationship between argumentation and collaboration can be exploited for significant performance benefits.

7.4.4 Local Context and Multi-task Learning

For the last experiment in this section we wanted to explore the possibility of jointly developing a context model from Section 6.3 and multi-task learning. For the context model, we decided to focus on the Local Context and consider the position in which context ADUs are centered around the target ADU. Since the hybrid model was also used for multi-task learning experiments, we extended it with the turn-level collaboration classifier, along with a specificity classifier. In this setting therefore: the argumentation model makes use of local context; the specificity model uses only the text in the target ADU; the collaboration model consists of the model in Figure 14 and uses the full collaboration label set. We first trained the model multiple times with different local context sizes, then combined multi-task learning with the attention-based local context model.

As we can see from Figure 15, the combination of local context (with variable size) and multi-task learning (green line) does not improve performance over local context only (red line). All differences between the two are not statistically significant. When looking at

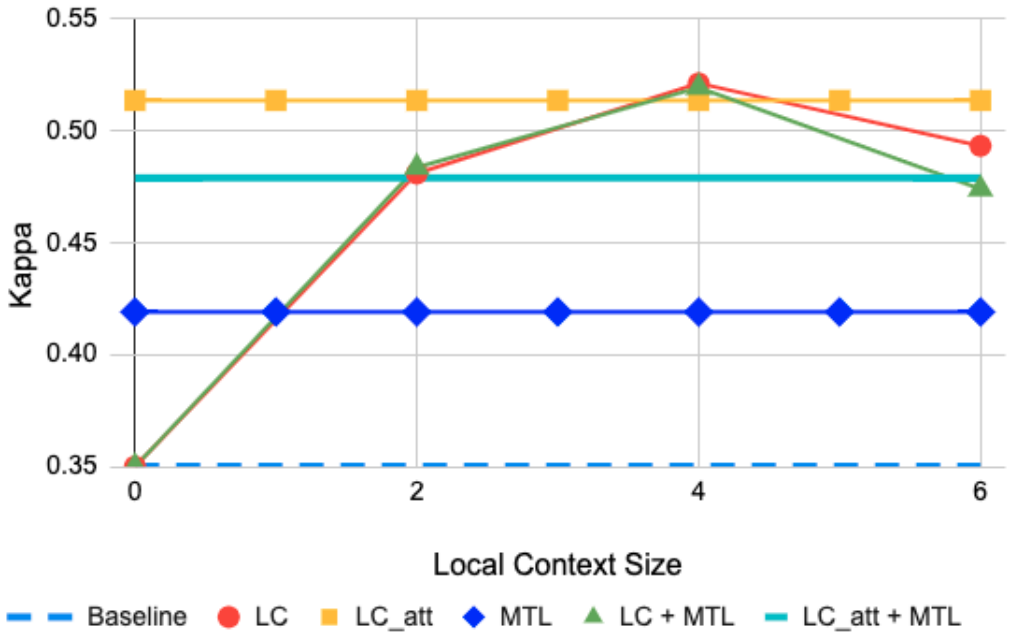


Figure 15: Results obtained combining context models with multi-task learning. Each line shows results for local context (LC), attention-based local context (LC_att), multi-task learning (MTL), local context and multi-task learning (LC+MTL) and local context attention combined with multi-task learning (LC_att+MTL).

attention-based models, on the other hand, introducing multi-task learning in the model actually results in a performance drop (yellow and aqua lines). Additionally, all models including local context always outperform the multi-task learning model (blue line).

These results warrant further investigation to understand why two models which separately give significant performance improvement do not seem to work well together. We speculate that this effect is due to the way the CNN is shared for multiple purposes. In the local context module, the same CNN is used for converting the target ADU into an embedding, as well as converting all context ADUs into embeddings. During training, then, for each ADU the CNN will receive multiple gradient updates. Coincidentally the CNN is also the portion of the model that is shared between tasks in multi-task learning. The number of gradient updates that happen through collaboration is much lower than the one through

argumentation (collaboration is turn-based and the number of turns is always less than or equal to the number of ADUs), therefore the shared portion of the model may focus more on the argumentation task. This is somewhat confirmed by the fact that when introducing local context into the multi-task learning model, Kappa on the collaboration task drops on average by 0.05 and F-score by 0.04. In the joint model with attention mechanism, Kappa for collaboration drops from 0.551 to 0.300 and F-score from 0.510 to 0.279.

7.5 SUMMARY

In this chapter we investigated the use of multi-task learning as a way of jointly training models to predict multiple related aspects of classroom discussions: argument component, specificity and collaboration. The proposed multi-task models differ from the ones proposed in argumentation mining literature in the type of tasks to perform, and they take full advantage of our annotation framework proposed in Chapter 3.

We first implemented pairwise multi-task models which jointly solve two of the three available classification tasks to better understand which classroom discussion components are more or less related to each other. Argumentation and collaboration showed higher affinity and potential for joint modeling, while specificity seems to be almost independent from the other two tasks. We then implemented a multi-task learning model to jointly train classifiers on all three tasks simultaneously and found that, in the presence of argumentation and collaboration, introducing the specificity task only improves argumentation performance. We then developed a more advanced collaboration model that was able to achieve more than double the performance of the previous one, albeit introducing some limitations consisting in needing additional inputs (e.g. collaboration references). The introduction of this model into the multi-task training process resulted in higher performance for the argumentation task. The last experiment we carried out was intended to explore the combination of context information for argumentation and multi-task learning. Results from this experiment tell us that naively combining the two approaches does not result in better models, and we may have to develop better ways of sharing parameters between tasks to take advantages of both.

8.0 CLASSROOM ANALYTICS DEPLOYMENT

8.1 INTRODUCTION

With the goal of addressing challenges in teaching and analyzing collaborative argumentation, we developed Discussion Tracker, a system to provide teachers analytics regarding their classroom discussions. Unlike prior research which developed tools for the classroom by focusing on frequency of participation, teacher questions, instructor/student talk ratio and student turn patterns [Chen et al., 2014, Blanchard et al., 2016, Gerritsen et al., 2015], we decided to build our analytics tool based on the aspects of collaborative argumentation discussed in Section 3.2: argumentation, specificity and collaboration.

8.2 DISCUSSION TRACKER

Now at its second iteration in the development cycle, Discussion Tracker organizes data visualizations across three main tabs:

1. Current Discussion: this tab contains all of the information and analytics pertaining to the current discussion to be analyzed. This tab is in turn divided into 4 minor tabs:
    a. Overview (Figure 16): displays basic information such as teacher name, date, discussion topic, along with percentages of teacher talk, percentage of speaking students, average turns per speaking student, and finally three pie charts with distributions of argumentation, specificity and collaboration.
    b. Annotated transcript (Figure 17): displays the transcript of the entire discussion

sorted by turn number, along with student ID and the three collaborative argumentation components for each turn/ADU.

c. Collaboration map (Figure 18): 2-dimensional visualization of how students collaborate together; new turns stretch out horizontally while instances of extensions, challenge, agree develop the graph vertically.

d. Help page: contains definitions for all terms used throughout the system.

2. Discussion history (Figure 19): after repeatedly using Discussion Tracker, teachers can keep track of frequency distributions of the three annotated components over time; additionally they can check for each discussion whether they achieved the goal they had set beforehand.

3. Plan next discussion (Figure 20): displays the main strengths and weaknesses of the current discussion, and prompts the teacher to select one of them to address for next time; after selecting a weakness, the teacher is prompted to three instructional resources targeting that weakness; strengths and weaknesses are automatically identified applying handcrafted rules to frequency distributions for collaborative argumentation labels.

Discussion Tracker was initially built as a standalone desktop application using Python and Tkinter, though we are in the process of developing a web-based version of the application to improve accessibility.

We used Discussion Tracker in the process of collecting dataset D4 in the Spring semester of 2020 as well as performing usability study and assessing the performance of NLP classifiers. In the absence of a downstream task for extrinsic evaluation of the computational models proposed in this thesis so far, this study serves at least as a point of reference in understanding if teachers would find the collaborative argumentation aspects beneficial[1]. Furthermore, by collecting and annotating the dataset, it provided a benchmark for us to evaluate the NLP classifiers previously trained on dataset D3 and estimate whether they could be used in practice.

[1]This represents an upper bound since we used manual annotations in the study, which do not reflect the actual performance of the NLP models we developed for collaborative argumentation.

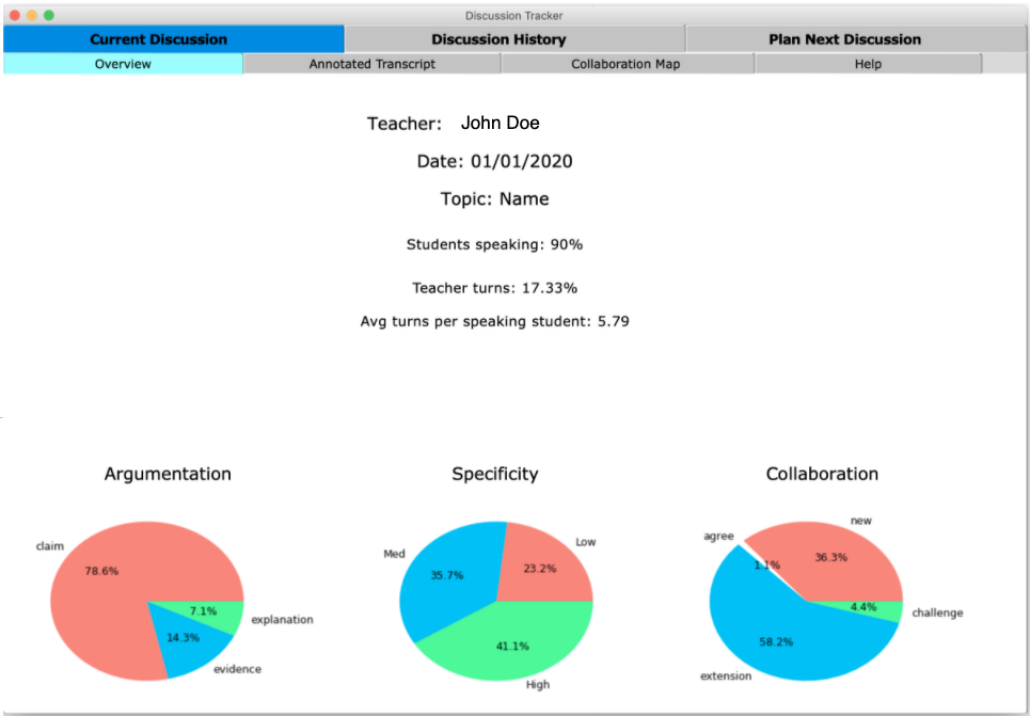
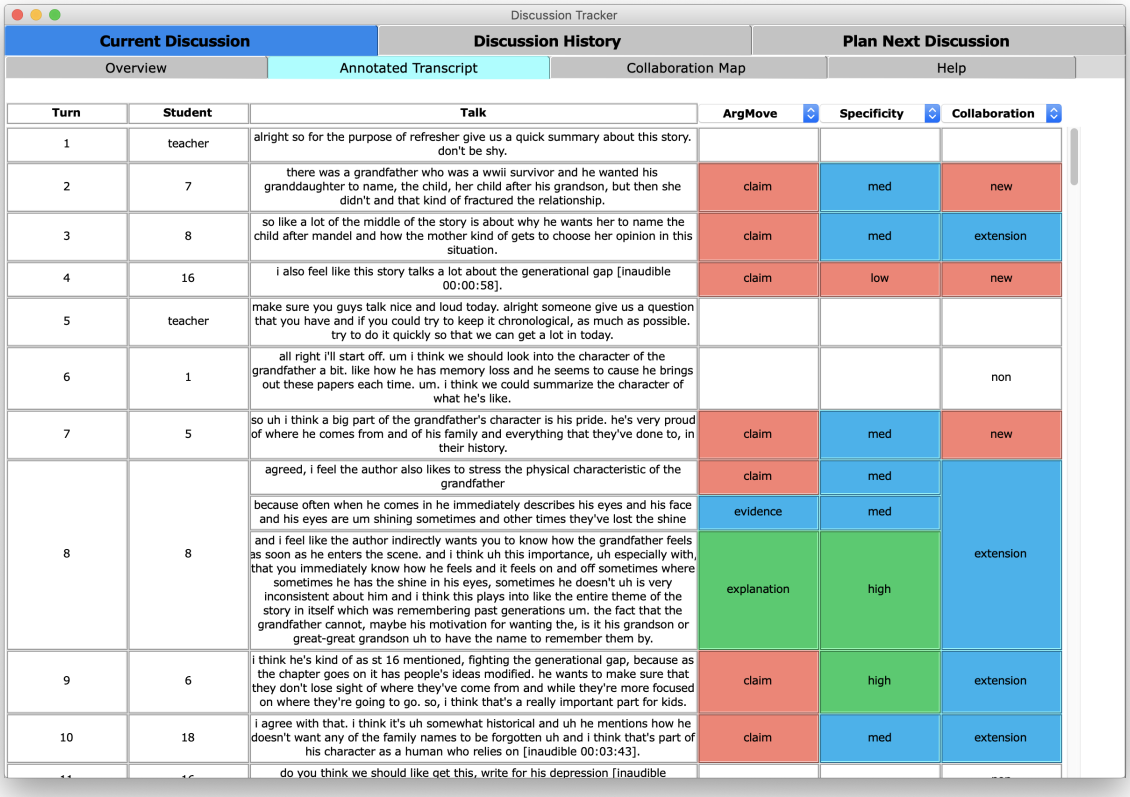


Figure 16: Screenshot of Discussion Tracker overview page.

8.3 EVALUATION

The procedure for deployment of Discussion Tracker was as follows:

- a 40 minute text-based literary discussion was recorded in a classroom;
- the teacher completed an online survey the same day;
- the discussion was transcribed and manually coded by experienced annotators;
- within 2 weeks from the initial recording, a researcher carried out a 45 minute interview with the teacher; during this time the teacher used Discussion Tracker to explore the discussion analytics provided by the system (based on manual annotations);
- the teacher completed a second online survey the same day with questions regarding their experience in using Discussion Tracker.



Turn	Student	Talk	ArgMove	Specificity	Collaboration
1	teacher	alright so for the purpose of refresher give us a quick summary about this story. don't be shy.			
2	7	there was a grandfather who was a wwii survivor and he wanted his granddaughter to name, the child, her child after his grandson, but then she didn't and that kind of fractured the relationship.	claim	med	new
3	8	so like a lot of the middle of the story is about why he wants her to name the child after mandel and how the mother kind of gets to choose her opinion in this situation.	claim	med	extension
4	16	i also feel like this story talks a lot about the generational gap [inaudible 00:00:58].	claim	low	new
5	teacher	make sure you guys talk nice and loud today. alright someone give us a question that you have and if you could try to keep it chronological, as much as possible. try to do it quickly so that we can get a lot in today.			
6	1	all right i'll start off. um i think we should look into the character of the grandfather a bit. like how he has memory loss and he seems to cause he brings out these papers each time. um. i think we could summarize the character of what he's like.			non
7	5	so uh i think a big part of the grandfather's character is his pride. he's very proud of where he comes from and of his family and everything that they've done to, in their history.	claim	med	new
8	8	agreed, i feel the author also likes to stress the physical characteristic of the grandfather	claim	med	extension
		because often when he comes in he immediately describes his eyes and his face and his eyes are um shining sometimes and other times they've lost the shine	evidence	med	
		and i feel like the author indirectly wants you to know how the grandfather feels as soon as he enters the scene. and i think uh this importance, uh especially with, that you immediately know how he feels and it feels on and off sometimes where sometimes he has the shine in his eyes, sometimes he doesn't uh is very inconsistent about him and i think this plays into like the entire theme of the story in itself which was remembering past generations um. the fact that the grandfather cannot, maybe his motivation for wanting the, is it his grandson or great-great grandson uh to have the name to remember them by.	explanation	high	
9	6	i think he's kind of as st 16 mentioned, fighting the generational gap, because as the chapter goes on it has people's ideas modified. he wants to make sure that they don't lose sight of where they've come from and while they're more focused on where they're going to go. so, i think that's a really important part for kids.	claim	high	extension
10	18	i agree with that. i think it's uh somewhat historical and uh he mentions how he doesn't want any of the family names to be forgotten uh and i think that's part of his character as a human who relies on [inaudible 00:03:43].	claim	med	extension
		do you think we should like get this, write for his depression [inaudible			

Figure 17: Screenshot of Discussion Tracker annotated transcript page.

Discussion Tracker was deployed in 18 discussions, with 18 different teachers from 4 high school in the Pittsburgh area.

8.3.1 Results on Discussion Tracker Usability

During the second survey, each teacher answered 13 questions on a 5-point Likert scale ("Strongly Disagree" to "Strongly Agree") about usability of the system and usefulness of the information provided. Table 27 shows the average rating for each of the 13 questions (value of 1 means they "Strongly Disagree" with the question, while value of 5 means they "Strongly Agree" with it).

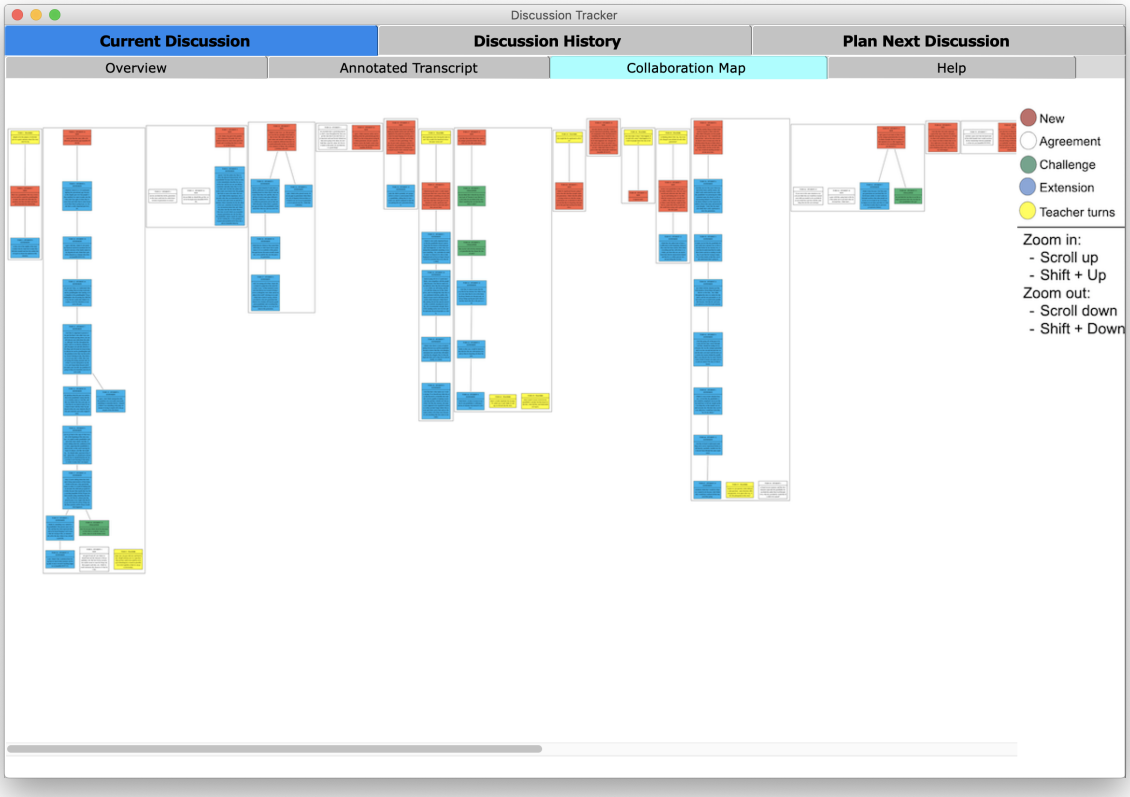


Figure 18: Screenshot of Discussion Tracker collaboration map page.

It is encouraging to see that all responses averaged above 4. One of the items of most interest to us is "Overall, Discussion Tracker is helpful for my teaching of literature discussion", which received one of the highest mean scores.

8.3.2 NLP Classifiers

We evaluated a classifier for each of the 3 collaborative argumentation aspects. For argumentation, the model is based on the BERT baseline combined with local context, described in Section 6.3.1. For specificity, the model consisted of the same BERT baseline, but without context information. For collaboration, the classifier consisted of a model very similar to the

Figure 19: Screenshot of Discussion Tracker discussion history page.

one from Section 7.3.1, but we removed the CNN component (because we are not training multi-task learning models). The choice of models was dictated by including components which achieved good results and showed high efficiency in terms of computation time and complexity (e.g. local context for argumentation). On the other hand, components that significantly increased computational complexity while not consistently improving performance were left out of this experiment (e.g. multi-task components for collaboration).

The models were trained on dataset D3 and tested on D4. When needed (i.e. for choosing context size and position for the argumentation model), hyperparameter optimization was carried out through cross-validation on D3, then the model was trained on the complete dataset D3. Results are reported in Table 28.

For the specificity model, kappa indicates substantial agreement with manual labels. For

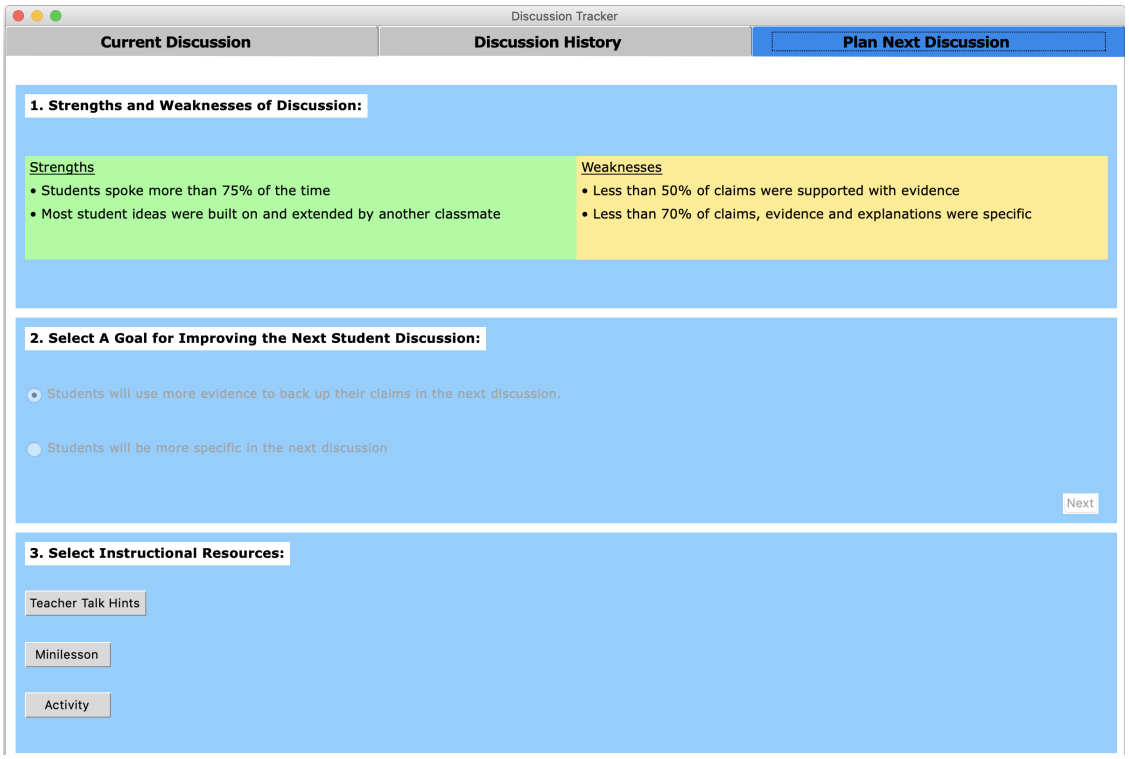


Figure 20: Screenshot of Discussion Tracker plan next discussion page.

argumentation and collaboration kappa indicates moderate agreement. It is worth pointing out that inter-rater agreement between the two human annotators for collaboration (with respect to the collaboration labels only, excluding turn references) over the whole dataset was measured in kappa = 0.578, not far from the result obtained by the classifier. On the other hand, inter-rater agreement between annotators was measured in Kappa = 0.971 for argumentation and quadratic-weighted kappa = 0.813 for specificity. The larger performance gap between these two kappas and the ones obtained from the classifiers indicate that there is much room for improvement in our proposed models for automated prediction of specificity and argumentation. The collaboration classifier also stands out when looking at the difference between macro F-score and micro F-score. The high difference is due to the fact that the NLP model never predicts "agree", and very rarely predicts "challenge". As for argumentation and specificity, the differences between macro F-score and micro F-score,

Table 27: Teacher survey items and Likert score means.

Question	**Mean**	**Question**	**Mean**
The overview of the discussion is helpful.	4.67	I find the system easy to use.	4.11
The pie charts of different features of the student discussion are helpful.	4.78	The system helps me to recognize my students' strengths during discussion.	4.72
The annotated transcript of student discussion is helpful.	4.89	The system helps me to recognize my students' weakness during discussion.	4.72
The collaboration diagram is helpful.	4.22	The system gives me more insight into student learning than I usually get from thinking about the discussion.	4.67
The system-generated strengths and weaknesses are helpful.	4.44	The system encourages me to make more changes to my facilitation of discussion than I usually do.	4.28
The goal-setting is helpful.	4.56	Overall, Discussion Tracker is helpful for my teaching of literature discussions.	4.72
The instructional resources are helpful.	4.17		

Table 28: Results of the three classifiers on dataset D4.

Model	**Kappa**	**Macro F-score**	**Micro F-score**
Argumentation	0.574	0.730	0.789
Specificity	0.727	0.688	0.679
Collaboration	0.566	0.439	0.775

though not negligible, are within expectation if we consider imbalance between class labels (particularly for argumentation).

8.4 SUMMARY

In this section we described Discussion Tracker, a discussion analytics system designed to help teachers analyze their classroom discussions. Discussion Tracker shows teachers transcripts of classroom discussions along with analytics based on collaborative argumentation annotations we described in Chapter 3. We deployed the system in 18 classrooms and collected usability results via surveys. Teachers reported that they found the system easy to use, and that the analytics helped in analyzing collaborative argumentation. Using the dataset collected in this study as well as manual annotations, we carried out experiments to evaluate NLP models for argumentation, specificity and collaboration. The NLP models evaluated in in this chapter consist of a subset of the models developed in Chapters 5, 6, and 7. They represent a compromise between performance and complexity: they include select components that proved to be well performing in our experiments while at the same time not drastically increasing computational complexity. This represents a simulation of the classifiers that would have been deployed, had we shown teachers automated predictions instead of manual annotations. Results showed the classifiers to be in moderate to substantial agreement with labels provided by human annotators.

9.0 CONCLUSIONS

9.1 SUMMARY OF CONTRIBUTIONS

In this thesis, we proposed to analyze collaborative argumentation for text-based classroom discussions. First, we developed a framework for annotating transcripts of classroom discussions. The scheme utilizes turns and argument moves (ADUs) as units of analysis and introduces guidelines on annotating three important aspects for collaborative argumentation: argument component, specificity and collaboration. Several datasets were annotated using the proposed scheme and reliability analyses showed that discussion transcripts can be reliably annotated by human coders.

Second, we proposed a computational model for automated prediction of specificity in classroom discussions. We showed that the proposed models can effectively combine handcrafted features and neural networks and outperform previously proposed specificity prediction models. At the same time we investigated the use of a pedagogically meaningful, interpretable feature set and the importance each individual feature carries: at the expense of a slight decrease in accuracy, using this feature set could potentially enable us to give students very detailed feedback on the specificity of their utterances.

Third, we proposed several computational models for automatically predicting argument components in transcripts of classroom discussions. We initially improved the performance of an existing argument mining system by extending its feature set with features previously used in the analysis of online dialogues. We also combined the handcrafted features with multiple types of neural networks and were able to obtain additional performance gains. Since these models include a very limited amount of contextual information, we proposed to extend the models by considering local context, teacher- and speaker-dependent context. We

demonstrated that adding contextual information will improve the performance of argument component classifiers, and we performed extensive experiments to understand in detail how important each context type (e.g. local vs. teacher vs. speaker) is in multi-party discussion.

Finally, since argument component, specificity and collaboration are related aspects of collaborative argumentation, we developed a set of joint models which are trained simultaneously on these tasks, in order to capture potential relationships existing between the three classes.

With respect to the research hypotheses stated in Chapter 1, we provide support for H1.1 by showing that transcripts of classroom discussions can reliably be annotated by trained human annotators; at the same time we provide support for H1.2 with teacher surveys showing these annotations to be useful in analyzing discussions. We also support hypotheses H2.1 and H2.2 where the performance of existing models for predicting specificity and argument component, respectively, can be improved by extending the models with additional handcrafted features and pairing them with neural networks. We proved hypothesis H2.3 by showing that contextual information can substantially enhance argument component classification models. Our experiments on contextual models for two baselines, one CNN-based and one transformer-based, indicates that the proposed contextual models can be effective for different neural network models, supporting hypothesis H2.4. Hypothesis H3 was partially confirmed: while it is true that multi-task learning can lead to better argument component classification models, the improvements are not always statistically significant. Multi-task learning can also improve performance of collaboration models, while it does not have a significant effect on specificity. Furthermore, the combination of context models and multi-task learning needs more extensive exploration.

Our contributions with respect to corpora are also of note. We collected two datasets consisting of 29 and 18 classroom discussions over the span of a year and a half. We then annotated both datasets with the annotation scheme proposed in Chapter 3. We also made the first dataset (D3) publicly available and free for research use, and are working towards releasing D4. While we also annotated datasets D1 and D2, they cannot be publicly released for copyright reasons.

9.2 LIMITATIONS AND FUTURE DIRECTIONS

Though many efforts went into developing, validating, and experimenting with the proposed features and models, we want to acknowledge that there are several limitations to our current work.

In Chapters 4, 5, 6 and 7 every model makes the assumption that turns are already segmented into multiple ADUs, and disregards the possibility of having non-argumentative text. A model to be deployed as part of an automated analytics system cannot make such assumptions, therefore the task of automating turn segmentation should be explored, along with the task of filtering-out non-argumentative content.

Likewise, in Chapter 7 when we developed a new collaboration classifier, in order to achieve significantly higher performance, we had to impose the restriction of having reference turns as an additional input. In practice this would require manual pre-processing which would render any classroom tool considerably less useful. Additionally, when developing the multi-task learning models we limited ourselves to three argumentative collaboration tasks. There are potentially many additional tasks that would be interesting to explore from a research standpoint (e.g. argument relation classification, discourse relations) as well as potentially beneficial to overall classifier performance (e.g. language modeling, named entity recognition) which we have not explored.

In Chapter 6 we discovered that using an attention mechanism to automatically optimize context size and position does not work well for different classifiers. In particular it does not work well for the classifier with currently the best performance on argument component classification in classroom discussions. For such model we can apply typical hyperparameter optimization strategies, however it is worth exploring additional ways of automating context optimization that in general work well for several computational models.

Some of the context models in Chapter 6 are based on speaker ID (whether it be individual speaker or teacher) though in a real-world scenario, for a completely automated system this would require accurate speaker diarization, which may not perform well in scenarios with lower audio quality.

Maybe an even bigger limitation of all our current approaches lies in the fact that we

are completely reliant on textual information derived from audio transcriptions. However, there are additional components that we are omitting: interpersonal dynamics based on audio-visual cues are integral parts of face-to-face discussions. Regrettably we do not yet have access to audio and video signals, though it would be interesting to explore additional models that can take advantage of these additional modalities.

Of particular note are limitations with respect to the classroom analytics deployment described in Chapter 8. Most teachers found a strength in the fact that Discussion Tracker could enable them to give detailed hard evidence to students as feedback on how they build their arguments and how they collaborate with others, which is not feasible with their hand-written notes. On the other hand, we don't have a definitive answer on whether the teacher themselves notice new discussion aspects or details when using Discussion Tracker. Other immediate limitations we gathered from teachers cognitive interviews are directly linked to the annotation framework and predictive models developed throughout this thesis. While carefully reading through the annotated transcript page in Discussion Tracker some teachers raised the issue of disagreement with respect to particular ADU/turn labels. Since the cognitive interview was conducted in-person by a researcher on our team, such disagreements were discussed on a case-by-case basis. This is only feasible during the system development phase as the ultimate goal is for teachers to use Discussion Tracker independently without supervision. There are two relatively straightforward avenues to address this limitation and increase teacher trust in the automated models. We can augment each prediction with a confidence score, so that teachers know which annotations to be mindful of when reading through an annotated transcript. This solution is easy to implement since most machine learning models (certainly the ones discussed in this thesis) can produce confidence scores as additional output. However, it requires teachers to understand what confidence scores are, along with estimating thresholds beyond which we might not want to trust the system, and does not provide details on why the system made such prediction. The second solution consists in employing explainable prediction models (e.g. the specificity classifier described in Section 4.3.4 which uses a pedagogical feature set). This approach has the benefit of providing detailed reasons compared to a single confidence score, but it may be counterproductive if it results in information overload for the teacher (an explanation can be generated

for each label, and there may be hundreds of labels in a single discussion).

Another particular aspect for which we received episodic feedback from teachers relates to the Discussion Tracker collaboration map. In general, teachers liked the graphical structure of the collaboration map, where it is easy to identify highly collaborative moments by looking for parts of the graph that extend vertically. What is not clear is how important it is for teachers to know precisely how students are collaborating. This relates directly to the two collaboration classifiers from in Section 7.3.1: one which produces all four possible collaboration labels, and one that only produces two (new vs. rest). Future work needs to address the issue of whether it is enough for teachers to know if a particular turn contains a new idea or relates to a prior idea: if this hypothesis holds true, collaboration classifiers can be considerably more accurate, as experiments in Chapter 7 showed.

Additional feedback we gathered from teachers can be categorized as human-in-the-loop, where they would like to be involved (to some degree) in the decisions made by the system instead of being passive users. Some teachers inquired about the possibility to specify new class labels for one of the three collaborative argumentation components (e.g. argumentation) and have Discussion Tracker automatically learn to identify it in the future (perhaps after providing a few examples). Other teachers raised questions about adjusting the specificity labels since they are defined on a scale: for example someone thought the specificity labels were "too generous", whereas they would label certain ADUs as medium specificity instead of high. In a few cases teachers also raised the possibility of using Discussion Tracker for different tasks than it was designed to. While we explicitly designed the system as a tool to help teachers understand how students collaboratively build arguments, some teachers would like to use it as tool to help them in grading individual students for each discussion. At a system level, these problems require usability analyses to understand how teachers would actually use Discussion Tracker in real-world scenario without supervision. From the NLP point of view, they would require that our models be able to accomplish few-shot or zero-shot learning. Recent developments in few-shot learning (in particular with the use of large language models) point to promising future avenues, though there are two considerations to be made: model size and computational complexity might increase significantly; the ability to add new class labels should be carefully managed in order to not deviate too much

from the original label set in Discussion Tracker.

Overall there is much room for progress, and we can highlight three main lines of research that would be beneficial for better understanding of collaborative argumentation. First, for the educational community it may be interesting to perform evaluations of systems like Discussion Tracker at large scale to better estimate what teachers need in order to improve classroom discussions. This includes going beyond the three aspects of collaborative argumentation analyzed in this work, and going beyond text-based discussions.

Second, large-scale studies would also be of interest for the NLP community - in no small part due to the datasets they would generate, which educational problems often lack. A clear research direction consists of developing better models, but we also argue that an interesting area of exploration would be models capable of giving concrete, actionable feedback to teachers and students.

Finally, with respect to "systems" area, much work is needed to produce an end-to-end, completely automated analytics platform: performing ASR and speaker diarization, automating turn segmentation, implementing systems that perform in or close to real-time.

BIBLIOGRAPHY

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Aker et al., 2017] Aker, A., Sliwa, A., Ma, Y., Lui, R., Borad, N., Ziyaei, S., and Ghobadi, M. (2017). What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96.

[Applebee et al., 2003] Applebee, A. N., Langer, J. A., Nystrand, M., and Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school english. *American Educational Research Journal*, 40(3):685–730.

[Ashley and Walker, 2013] Ashley, K. D. and Walker, V. R. (2013). Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 176–180. ACM.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bengio, 2012] Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer.

[Biber, 1988] Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

[Bird et al., 2009a] Bird, S., Klein, E., and Loper, E. (2009a). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[Bird et al., 2009b] Bird, S., Klein, E., and Loper, E. (2009b). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[Bjerva, 2017] Bjerva, J. (2017). Will my auxiliary tagging task help? estimating auxiliary tasks effectivity in multi-task learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 216–220.

[Blanchard et al., 2016] Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., and D'Mello, S. K. (2016). Identifying teacher questions using automatic speech recognition in classrooms. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 191.

[Buda et al., 2017] Buda, M., Maki, A., and Mazurowski, M. A. (2017). A systematic study of the class imbalance problem in convolutional neural networks. *arXiv preprint arXiv:1710.05381*.

[Carlile et al., 2018] Carlile, W., Gurrapadi, N., Ke, Z., and Ng, V. (2018). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 621–631.

[Chafe and Tannen, 1987] Chafe, W. and Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology*, 16(1):383–407.

[Chakrabarty et al., 2019] Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., and Hwang, A. (2019). AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China.

[Chen et al., 2014] Chen, G., Clarke, S., and Resnick, L. (2014). An analytic tool for supporting teachers' reflection on classroom talk. In *Learning and becoming in practice: The International Conference of the Learning Sciences (ICLS) 2014*. International Society of the Learning Sciences.

[Chisholm and Godley, 2011] Chisholm, J. S. and Godley, A. J. (2011). Learning about language through inquiry-based discussion: Three bidialectal high school students' talk about dialect variation, identity, and power. *Journal of Literacy Research*, 43(4):430–468.

[Chollet et al., 2015a] Chollet, F. et al. (2015a). Keras. https://github.com/fchollet/keras.

[Chollet et al., 2015b] Chollet, F. et al. (2015b). Keras. https://keras.io.

[Collobert and Weston, 2008] Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

[Daxenberger et al., 2017] Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (2017). What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066. Association for Computational Linguistics.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota.

[Dong et al., 2015] Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

[Engle and Conant, 2002] Engle, R. A. and Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, 20(4):399–483.

[Fiacco and Rosé, 2018] Fiacco, J. and Rosé, C. (2018). Towards domain general detection of transactive knowledge building behavior. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pages 1–11.

[Finkel et al., 2005] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

[Gao et al., 2019] Gao, Y., Zhong, Y., Preotiuc-Pietro, D., and Li, J. J. (2019). Predicting and analyzing language specificity in social media posts. *Thirty-Third AAAI Conference on Artificial Intelligence*.

[Gerritsen et al., 2015] Gerritsen, D., Zimmerman, J., and Ogan, A. (2015). Exploring power distance, classroom activity, and the international classroom through personal informatics. In *Proceedings Sixth International Workshop on Culturally-Aware Tutoring Systems*, pages 11–19.

[Ghosh et al., 2016] Ghosh, D., Khanam, A., Han, Y., and Muresan, S. (2016). Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 549–554.

[Ghosh et al., 2014] Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., and Mitsui, M. (2014). Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48.

[Girshick, 2015] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

[Graham and Perin, 2007] Graham, S. and Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3):445–476.

[Grossman et al., 2014] Grossman, P., Cohen, J., Ronfeldt, M., and Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, 43(6):293–303.

[Gweon et al., 2013] Gweon, G., Jain, M., McDonough, J., Raj, B., and Rosé, C. P. (2013). Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *International Journal of Computer-Supported Collaborative Learning*, 8(2):245–265.

[Habernal and Gurevych, 2017] Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hong and Scardamalia, 2014] Hong, H.-Y. and Scardamalia, M. (2014). Community knowledge assessment in a knowledge building environment. *Computers & Education*, 71:279–288.

[Ke et al., 2018] Ke, Z., Carlile, W., Gurrapadi, N., and Ng, V. (2018). Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI*, pages 4130–4136.

[Keefer et al., 2000] Keefer, M. W., Zeitz, C. M., and Resnick, L. B. (2000). Judging the quality of peer-led student dialogues. *Cognition and Instruction*, 18:53–81.

[Kelly et al., 2018] Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., and D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464.

[Kim, 2014a] Kim, I.-H. (2014a). Development of reasoning skills through participation in collaborative synchronous online discussions. *Interactive Learning Environments*, 22(4):467–484.

[Kim, 2014b] Kim, Y. (2014b). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

[Klebanov et al., 2016] Klebanov, B. B., Stab, C., Burstein, J., Song, Y., Gyawali, B., and Gurevych, I. (2016). Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75.

[Ko et al., 2019] Ko, W.-J., Durrett, G., and Li, J. J. (2019). Domain agnostic real-valued specificity prediction. *Thirty-Third AAAI Conference on Artificial Intelligence*.

[Krippendorff, 2004] Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, 38:787–800.

[Lampert et al., 2010] Lampert, M., Beasley, H., Ghousseini, H., Kazemi, E., and Franke, M. (2010). Using designed instructional activities to enable novices to manage ambitious mathematics teaching. In *Instructional explanations in the disciplines*, pages 129–141. Springer.

[Lauscher et al., 2018] Lauscher, A., Glavaš, G., Ponzetto, S. P., and Eckert, K. (2018). Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3326–3338.

[Lawrence and Reed, 2017] Lawrence, J. and Reed, C. (2017). Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates. In *Proceedings of the 4th Workshop on Argument Mining*, pages 108–117.

[Lee, 2006] Lee, C. D. (2006). 'every good-bye ain't gone': analyzing the cultural underpinnings of classroom talk. *International Journal of Qualitative Studies in Education*, 19(3):305–327.

[Li et al., 2017] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

[Li and Nenkova, 2015] Li, J. J. and Nenkova, A. (2015). Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, pages 2281–2287.

[Li et al., 2016] Li, J. J., O'Daniel, B., Wu, Y., Zhao, W., and Nenkova, A. (2016). Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

[Lison and Bibauw, 2017] Lison, P. and Bibauw, S. (2017). Not all dialogues are created equal: Instance weighting for neural conversational models. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 384–394. Association for Computational Linguistics; Stroudsburg, PA.

[Liu et al., 2015] Liu, X., Gao, J., He, X., Deng, L., Duh, K., and Wang, Y.-Y. (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval.

[Louis and Nenkova, 2011] Louis, A. and Nenkova, A. (2011). General versus specific sentences: automatic identification and application to analysis of news summaries. Technical Report MS-CIS-11-07, University of Pennsylvania.

[Louis and Nenkova, 2012] Louis, A. and Nenkova, A. (2012). A corpus of general and specific sentences from news. In *LREC*, pages 1818–1821.

[Loukina et al., 2015] Loukina, A., Zechner, K., Chen, L., and Heilman, M. (2015). Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–19.

[Lugini and Litman, 2017] Lugini, L. and Litman, D. (2017). Predicting specificity in classroom discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61.

[Lugini and Litman, 2018] Lugini, L. and Litman, D. (2018). Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67.

[Lugini and Litman, 2020] Lugini, L. and Litman, D. (2020). Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1475–1480, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Lugini et al., 2018] Lugini, L., Litman, D., Godley, A., and Olshefski, C. (2018). Annotating student talk in text-based classroom discussions. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 110–116.

[Luong et al., 2015] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

[McHugh, 2012] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

[McLaren et al., 2010] McLaren, B. M., Scheuer, O., and Mikšátko, J. (2010). Supporting collaborative learning and e-discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence in Education*, 20(1):1–46.

[Meng et al., 2018] Meng, Z., Mou, L., and Jin, Z. (2018). Towards neural speaker modeling in multi-party conversation: The task, dataset, and models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Michaels et al., 2008] Michaels, S., O'Connor, C., and Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297.

[Michaels et al., 2010] Michaels, S., O'Connor, M. C., Hall, M. W., and Resnick, L. B. (2010). Accountable talk sourcebook: For classroom conversation that works. *Pittsburgh, PA: University of Pittsburgh Institute for Learning*.

[Misra et al., 2015] Misra, A., Anand, P., Fox Tree, J., and Walker, M. (2015). Using summarization to discover argument facets in dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[Mohtarami et al., 2018] Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., and Moschitti, A. (2018). Automatic stance detection using end-to-end memory networks. *arXiv preprint arXiv:1804.07581*.

[NGA & CSSO, 2010] NGA & CSSO (2010). Common core state standards initiative.

[Nguyen and Litman, 2015] Nguyen, H. and Litman, D. (2015). Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28.

[Nguyen and Litman, 2016a] Nguyen, H. and Litman, D. (2016a). Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1127–1137.

[Nguyen and Litman, 2016b] Nguyen, H. and Litman, D. J. (2016b). Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics. In *FLAIRS Conference*, pages 485–490.

[Nguyen and Litman, 2018] Nguyen, H. V. and Litman, D. J. (2018). Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

[Niculae et al., 2017] Niculae, V., Park, J., and Cardie, C. (2017). Argument Mining with Structured SVMs and RNNs. In *Proceedings of ACL*.

[Olshefski et al., 2020] Olshefski, C., Lugini, L., Singh, R., Litman, D., and Godley, A. (2020). The discussion tracker corpus of collaborative argumentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1033–1043, Marseille, France. European Language Resources Association.

[Opitz and Frank, 2019] Opitz, J. and Frank, A. (2019). Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy.

[Ortega and Vu, 2017] Ortega, D. and Vu, N. T. (2017). Neural-based context representation learning for dialog act classification. *arXiv preprint arXiv:1708.02561*.

[Palau and Moens, 2009] Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Peldszus and Stede, 2013] Peldszus, A. and Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Persing and Ng, 2015] Persing, I. and Ng, V. (2015). Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 543–552.

[Persing and Ng, 2016] Persing, I. and Ng, V. (2016). End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394.

[Rahimi et al., 2017] Rahimi, Z., Litman, D., Correnti, R., Wang, E., and Matsumura, L. C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, pages 1–35.

[Reznitskaya and Gregory, 2013] Reznitskaya, A. and Gregory, M. (2013). Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist*, 48(2):114–133.

[Reznitskaya et al., 2009] Reznitskaya, A., Kuo, L.-J., Clark, A.-M., Miller, B., Jadallah, M., Anderson, R. C., and Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education*, 39(1):29–48.

[Richey et al., 2016] Richey, C., D'Angelo, C., Alozie, N., Bratt, H., and Shriberg, E. (2016). The sri speech-based collaborative learning corpus. In *INTERSPEECH*, pages 1550–1554.

[Samei et al., 2014] Samei, B., Olney, A., Kelly, S., Nystrand, M., D'Mello, S. K., Blanchard, N., Sun, X., Glaus, M., and Graesser, A. C. (2014). Domain independent assessment of dialogic properties of classroom discourse. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 233–236.

[Schulz et al., 2018] Schulz, C., Eger, S., Daxenberger, J., Kahse, T., and Gurevych, I. (2018). Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41. Association for Computational Linguistics.

[Sohmer et al., 2009] Sohmer, R., Michaels, S., O'Connor, M., and Resnick, L. (2009). Guided construction of knowledge in the classroom. *Transformation of knowledge through classroom interaction*, pages 105–129.

[Stab and Gurevych, 2014] Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

[Stab and Gurevych, 2017] Stab, C. and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

[Stone and Hunt, 1963] Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, pages 241–256. ACM.

[Sukhbaatar et al., 2015] Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

[Swanson et al., 2015] Swanson, R., Ecker, B., and Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226.

[Toulmin, 1958] Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

[Turian et al., 2010] Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Wachsmuth et al., 2016] Wachsmuth, H., Al Khatib, K., and Stein, B. (2016). Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691.

[Weston et al., 2014] Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.

[Wilson, 1988] Wilson, M. (1988). Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods*, 20(1):6–10.

[Wilson et al., 2009] Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

[Wolf et al., 2019] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

[Zhang et al., 2013] Zhang, J., Chen, M.-H., Chen, J., and Mico, T. F. (2013). Computer-supported metadiscourse to foster collective progress in wu knowledge-building communities. In *Proceedings of the International Conference of Computer-supported Collaborative Learning*.

APPENDIX A

CODING MANUAL

This chapter contains the complete coding manual used for annotating discussions.

Collaboration Coding

Unit of Analysis: Turns at Talk
Description: Collaboration coding involves the way each turn at talk functions in relation to previous talk. E.g., does the turn introduce a new idea, repeat a previous idea, add to a previous idea, or contradict a previous idea? We code for four different types of turns at talk; new ideas, agreements, extensions, and challenges.

1.
   - Variable: Initiating Turn / New Idea.
   - Code: N.
   - Definition: An initiating turn is the expression of a new idea in the discussion. This does not have to be a new topic, but should be a new idea, concept, or perspective. It usually does not reference ideas in prior turns at talk, or it does so only superficially (as in Student 4 example) . Turns that build on ideas in previous turns at talk are coded as "extension."
     New student questions posed to the whole class that do not probe or question a previous answer are uncoded [ex. "What real-world problem can be related in terms of acceptance and steps towards change to the racial problem described in the text?"].
     Turn Reference: Even though new ideas do not reference any prior turns at talk, copy and paste the disc ID into the turn reference column (if coding a turn labeled "T126.EAGER.2.Mockingbird.2", copy "T126.EAGER.2.Mockingbird.2" in the turn reference column).
   - Example: Teacher: would you consider that an example of satire?... Yes? Any comments?
     Student 22: Uh I would say yes because, the whole story's kind of making fun of the social requirements, I guess.
     Teacher: What other things did the colleagues talk about when they talked about going to this funeral?
     Student 6: They felt like they were required to go but even though um they felt that it was required, they still believed that um they were more relieved that they were,

they weren't dead, that he was.

Both student comments above are coded as initiation because they introduce new ideas that do not refer back to previous student talk.

Student 3: Well, um, I have another quote on page 13, "His wife, especially the youngest, lived in perpetual fear of his fiery temper, and so did his little children." So that shows that he really does rule in his household and that everyone under him is treated like, they're not treated with enough respect in my point of view.

Student 4: And I **also** believe that men are **also** providers, like um yams are like an important crop in the Igbo culture. And I have a quote, and um it says, "yams stood for manliness, and he who could feed his family on yams from one harvest to another was a very great man indeed."

Explanation: *Student 4's turn is coded as an initiating turn because they raise the new idea of men being providers of food while the previous turns at talk have focused on men's violence and domination. Even though they use the word "also" their turn introduces a new idea that is not an explicit extension of previous ideas.*

2.
- Variable: Extension.
- Code: E.
- Definition: A turn is an extension if it builds off another student's ideas. Extension turns must extend one of the preceding four codeable student turns unless a turn prior to those 4 is specifically referenced.

  Extension turns include at least 2 key ideas or terms that were voiced by another student. Key ideas/terms may be textual, topical or conceptual terms. Textual terms may include characters and places from a text under discussion (like "Macbeth" or "Birnam Wood"), but do not include titles of texts. Topical terms may include disciplinary topics (like theme, metaphor, symbol, etc.). Conceptual terms may include abstract ideas (like "culture," "domination," "regret").

  Extensions sometimes (but not always) include terms like "also, another, too"; or indicators of agreement/alignment (such as, "like X said. . . ")

  Extensions can also include a self extension which is a turn of talk that adds information to or re-words one's own idea that was shared without acknowledging the idea of other speakers in close proximity.

  "Turn Reference" coding: For extension turns, code the "turn reference" as the related idea that was most recently expressed by another student within the past four codeable student turns. However, if a turn coded as "extension" includes an explicit reference to a particular turn (by student name or idea) many turns in the past, code the explicitly mentioned turn as the "turn reference" even if it is many turns in the past.
- Example: Student 1: "angered by his youngest wife, who went to plait her hair at her friend's house, and then I returned early enough to pick the afternoon milk." This shows that men are violent and the boss of their family.

  Student 2: Well I have another quote that shows that men are violent. Like, in page 38, yeah, I think, wait yeah 38, "Okonkwo's second wife merely cut a few leaves off the umm banana tree, to make some food." And she said, "So without um further

agreement Okonkwo gave her a beating." Which means that Okonkwo has like a superior, and he beats whatever he wants to beat.

Explanation: Student 2 above repeats three key words/ideas from Student 1's statement (wives, men, violence)but adds new information (a new quote, a different explanation) so their turn is coded as extension.

3.
   - Variable: Challenge / Probe.
   - Code: C.
   - Definition: Challenge and probe turns challenge or question a prior idea. Challenges and probes should reference another student's turn (which may or may not have been coded) in the preceding four codeable student talk turns. Challenges to points made further back are considered "Initiating Turns/New Ideas" (N).

     A turn is considered a challenge if it includes both (1) key words/concepts from previous turns (such as "culture," "domination," or "regretful") and (2) some indication of disagreement. Note that indications of disagreement can be very subtle (such as "still" or "actually" or "he did tell his sister") or more explicit (such as "I disagree", "No," "but," "however," "though")

     A turn is considered a probe if it challenges or requests more information, detail, elaboration, or clarification/explanation in the form of a question ("Why do you think that?" "You really think Macbeth wasn't crazy?" or "What do you mean?"). Will often include second person pronoun or direct address. Does not include procedural questions like "Wait what was his question?"

     NOTE: Turns sometimes contain what may appear to be indications of disagreement (e.g., "however" "isn't") but are actually referring to ideas within the turn—these would likely fall under the category of extensions.

     "Turn Reference" coding: For challenge turns, code the "turn reference" as the challenged idea that was most recently expressed by another student within the past 4 codeable student turns. However, if a turn coded as "challenge" includes an explicit reference to a particular turn (by student name or idea) many turns in the past, code the explicitly mentioned turn as the "turn reference" even if it is many turns in the past.
   - Example: Challenge statements: Turn C: They don't really care about their families. Turn D: ([C] with Turn C) Actually, they do care, because, um, if he didn't beat his children, then his children wouldn't learn their lessons. So he has a reason to beat them.

     Explanation: turn D is coded as a Challenge to turn C because it includes (1) the key idea of the way men relate to their families, and (2) an indication of disagreement ("actually")

     Challenge questions: Why do you think that? What do you mean?

     Explanation:. Other ways to challenge could be through questions like "why do you think that?" Or "What do you mean?"

4.
   - Variable: Agreeing Turn.
   - Code: A.

- Definition: Turns that either express almost the exact thing in one of the preceding four coded student turns OR affirm the previous statement with a short response like "yeah" or "I agree with what she said."
  When a turn seems like it should be coded as an extension but lacks two clear key terms or ideas, it is likely to be coded as an agreement. (see last example)
- Example: St 1: I think that it's people speak up more, I think if people speak up more about the problem then it'll be better.
  Student 11: I agree
  Explanation: Student 11 turn is coded as agreeing because they simply say "I agree" with Student 1.
  Student 3: His strength and innocence as well as his childlike wonder and hope were lost with his [x]
  Teacher: Wow, say that again.
  Student 3: His strength and innocence, along with his childlike wonder and hope, were lost.
  Student 3's second turn is coded as agreeing/repeating because she repeats what she said in the previous turn only with slightly different wording.
  Student 15: Like I've realized that it's not really his choice to treat Hassan like that. You kinda can't blame him cause he's young, and he's part of society.
  Student 16: He was following what he knows. That's really all he ever knew.
  Student 16's turn affirms Student 15's idea that the main character did what he did not by choice, and that "you can't blame him" since "he's part of society." Student 16 says something similar by saying "He was following what he knows"—This turn cannot be considered extension because it shares only one key term or idea (he), but it is also unsound to code it as a new idea, because it is not exactly a new idea, concept or perspective. Because it affirms what was said above and only has one key term, it is coded as agreement.

Argumentation Coding

Segmenting into Argument Moves

In order to code for argumentation, we first need to consider whether or not a turn at talk is made up of multiple moves. There are three types of argument moves: claims, evidence, and warrants

1.
   - Variable: Claim
   - Code: CL
   - Definition: An arguable statement that presents a particular interpretation of a text or topic.
     DOES: often (but not always) precedes evidence and warrants. States something that can more or less be contested—infers, predicts, hypothesizes, considers possibilities.
     DOES NOT: simply recount details from text that are accessible to all readers (everyone knows Macbeth became king)
   - Example: "Linda Loman is like really just protecting Willy from everything."

2.
   - Variable: Evidence
   - Code: EVI
   - Definition: Talk used to support, justify, or back a claim.
     DOES: includes facts, textual references, anecdotes. Often (but not always) follows a claim. Always proximal to a claim (within 1 or 2 turns)
     DOES NOT: does not exist without a claim.
   - Example: "Like at the end of the book remember how she was telling the kids to leave and never come back."

3.
   - Variable: Warrant
   - Code: WAE
   - Definition: Move that provides explanation for why evidence supports the claim.
     DOES: Always proximal to evidence supporting a claim (almost always follows evidence)
     DOES NOT: It rarely occurs before claim/ evidence that it is explaining (despite Toulmin's [1958] structure).
   - Example: "Like she's not even caring about them, she's caring about Willy."

Specificity Coding

Unit of Analysis: Argument Moves
Description: Coding for specificity involves labeling the content of an argument move based on its degree of detail, clarity, elaboration, and content-related vocabulary. We code for three different degrees of specificity: high, medium, low.

1.
   - Variable: High
   - Code: HI
   - Definition: An argument move that includes more than one of the following: 1) particulars, 2) details, 3) content-language, 4) chain of reasons.
     Definitions:
     Particulars: Argument move contains at least two terms that are particular (e.g., a person and a setting or a setting and an action) rather than a general group or situation such as "you," "everyone" "books." Clichés or overgeneralizations are not particulars.
     Details: Descriptions, explanations or elaborations that make the idea more understandable, contextualized, qualified, substantiated or vivid. ("Detailed" argument moves should avoid or at least explain general terms like good, bad, stupid, i.e., terms with multiple definitions).
     Content language: Uses vocabulary or phrases that are specific to English Language Arts (such as "irony, simile, tragedy," etc.) or the text being discussed (such as quotes or expressions). References to main characters, places, etc. in the text are NOT content language: these are coded as "particulars."
     Chain of reasoning: Phrases or clauses that attempt to rationalize, justify or explain an idea(s). They link or synthesize at least two pieces of information or ideas. Thus, "Because Macbeth is scared" does not include a chain of reasoning because it only has one idea but "Macbeth kills the guards because he is scared" includes two ideas and thus has a chain of reasoning. Chains of reasoning often show cause/effect ("because" and "then"), contrast ("however, although, but") or reasoning ("so that"). The reasoning does not need to be convincing to the coder, nor should a coder assess its logic; presence of a chain of reasoning is enough.
   - Example: "**He**'s (*particulars*) so wrapped up in the thought that it's his destiny to be king and that this is what fate has chosen for him, **that when he finally does get what he wants** (*chain of reasoning*) he's scared and **it just bothers him that the way he became king is deceiving who he was loyal to** (*details*). And **he becomes obsessed with the fact that he doesn't want someone to do what he did to become king**, (*details*) which is really **ironic** (*content language*)."
     Look fors (often): quotes from the book; Multiple expressions of causality; Multiple qualifiers; Highly arguable; Usually longer; Content-specific vocabulary

2.
   - Variable: Medium
   - Code: MED
   - Definition: The statement clearly accomplishes one of the above.
     Clichés cannot be coded as medium.

- Example: "There's this movie about this guy **who can see like forward into the future and he always changes it**. And he always said **that once you see into the future, it changes, because you know what's going to happen and you choose to do it faster.**" (*details, but no particulars, content language or chain of reasoning*)

3.
- Variable: Low
- Code: LOW
- Definition: The statement does not clearly accomplish any of the above criteria Even if statement refers to a character, extremely underqualified statements, gross overgeneralizations, or cliché's can limit statement to low specificity.
- Example: "Like, one man's trash is another man's treasure"
  Look fors: Clichés and overgeneralizations; Unclear subject, time, circumstances; No explanation or qualifiers.

Guidelines for Segmenting Turns

Table 29: Segmentation guidelines.

Guideline	**Turn Example**	**Oversegmentation**	**Suggested Segmentation**
Segment only when there is a clear and indisputable shift in the turn at talk. Don't segment unless the phrase/clause can be taken alone and be understood as qualitatively different from the phrase/ clause before it. Be attentive to moments when talk shifts from opinion to fact and fact to opinion. Err on the side of undersegmenting as opposed to over segmenting	Okay, wait, okay. My idea builds up on X's idea. Because, beause, its not really about if they're gettinng disobeyed or not, its really about like, head. Like it's like, not that she's disrespecting them, like, them like, in Igbo culture they have like a head game, and they like, real like, its not being, about being, about being disrespected, but it's like the principle of it. And like, like while men are the boss, women are weak and inferior. They don't even own their own children. But thats, like, it's not about what respect. They don't own thier own children, and they don't get enough respect, and it's not about respect. It's not about respect from their men. It's about like, about how the Igbo culture and the men, look at how other men look at them. They don't wanna, that's why they have more than one wife and wives barely get to do anything. It's because it says, the Oracle says so.	Okay, wait, okay. My idea builds up on X's idea. Because, beause, its not really about if they're gettinng disobeyed or not, its really about like, head. Like it's like, not that she's disrespecting them, like, them like,	Okay, wait, okay. My idea builds up on X's idea. Because, beause, its not really about if they're gettinng disobeyed or not, its really about like, head. Like it's like, not that she's disrespecting them, like, them like, in Igbo culture they have like a head game, and they like, real like, its not being, about being, about being disrespected, but it's like the principle of it.

		in Igbo culture they have like a head game, and they like, real like, its not being, about being, about being disrespected, but it's like the principle of it.	And like, like while men are the boss, women are weak and inferior.
		And like, like while men are the boss, women are weak and inferior.	They don't own thier own children, and they don't get enough respect, and it's not about respect. It's not about respect from their men. It's about like, about how the Igbo culture and the men, look at how other men look at them. They don't wanna, that's why they have more than one wife and wives barely get to do anything. It's because it says, the Oracle says so.
		They don't even own their own children.	
		But thats, like, it's not about what respect. They don't own thier own children, and they don't get enough respect, and it's not about respect. It's not about respect from their men.	
		It's about like, about how the Igbo culture and the men, look at how other men look at them. They don't wanna,	

		that's why they have more than one wife and wives barely get to do anything. It's because it says, the Oracle says so.	

Coding Procedures

All of the above features are coded using an Excel spreadsheet. Prior to coding, the spreadsheet will represent a classroom discussion (see figure on next page) segmented by turns at talk. Each spreadsheet will include 19 columns, A through S. For each turn at talk, columns A- E appear as one row (they are actually five excel cells that have been merged), while columns F – S will appear as five rows (see figures on p. 8). Below are brief explanations of the columns.

Column A: Disc id. This column provides 5 items of information: The teacher id (T#), the discussion id (D#), the text id (e.g., TFA), the class id (C#), and the turn # within the discussion (#). Thus in the example provided (T1D1TFA.C1.1), we know that the teacher is T1, the first discussion (D1), the text is Things Falls Apart (TFA), the class id is C1, and the turn number is 1. The next discussion on the same book facilitated by the same teacher with the same class would begin with the id T1D2TFA.C1.1. Column B: Speaker id. This column provides the identification of the speaker. "St" indicates a student and T indicates a teacher. Student i.d.s are based on the order of speakers. Column C: Talk. The transcribed talk. For transcription conventions see pp. 8-9. Column D: Talk Summary. In this column please briefly summarize the main idea from the turn at talk. This will help with any disagreements in collaboration coding. Column E: Relation to Collaborative Reasoning. For this column, please write one of the six the corresponding collaboration codes (see pp. 1-3). Column F: Turn Reference. For this column please list the prior turn(s) that informed your collaboration code. For example, if turn TFA1.2 is [E] to the prior turn, then please COPY and PASTE TFA1.1 into Column E. Column G: Argument Move Segmentation. For this column, please segment the turn at talk into corresponding argument moves (as done on p. 4) Columns H –J: CL, EVI, WAR. Please place an X in column that corresponds with the appropriate argument move label. See pp. 4-5 for definitions.

Columns K—M: TEXT, INTER, EXP. Please place an X in column that corresponds with the appropriate domain label. See pp. 5-6 for definitions. Columns O—Q: Particulars, Detail, Content, Chain. These categories refer to the four different characteristics of specificity on pp. 6-7. Please place an X in the columns that apply. Columns R—T: LO, MED, HI. Please place an X in the column that corresponds to the correct level of specificity (see pp. 6-7).

APPENDIX B

TRANSCRIPT EXAMPLE

This section contains a complete transcript of a classroom discussion from dataset D3, on the text *The Legend of Sleepy Hollow*. The transcript shows turn numbers, student ID and argument moves. It also shows labels for the three components of collaborative argumentation: argumentation, specificity and collaboration. For each component we include two labels: the first one is the manual label that can be found in dataset (ground truth), while the second one is the output of our predictive model. For argumentation we used the model described in Section 6.4.4 which consists of the BERT model and includes local context and speaker context. The specificity model has the same architecture as the argumentation model, though without context modules. The collaboration model consists of the turn-level collaboration classifier described in Section 7.3.1. Note that while we have predictions for collaboration labels, the collaboration reference column only includes the ground truth labels as our collaboration model only produces a label and not a turn reference. Turns without annotations represent non-argumentative student talk.

Table 30: Transcript of a classroom discussion.

Turn	**Sp id**	**Collab**	**Ref**	**Argument Move**	**Arg**	**Spec**
1	St 3			Okay how does the how does the excessive use of imagery help or detract from the story?		
2	St 7	N, N	2	I think it actually helps the story. Student ? I'm talking. {overlap with many students} um I think I think that it helps the story cause it like helps create an environment like a movie for like the story	C, C	M, M

				for instance like um it says that "each small brook lies through it with just murmured enough to uh lull lull ones response." Its like its in like the first couple paragraphs	E, E	M, M
				so it basically paints paints like he's describing like uh he's describing the little valley and I think it paints like a picture of what it would look like and helps the readers understand the story more.	W, C	M, L
3	St 10	E, E	2	I think that um it also when they're describing the setting they also describe like the people of the town and it says that, "they're given to all kinds of marvelous beliefs and are subject to trances and visions and frequently see strange sights and hear music and voices in the air."	E, E	M, L
				I think that helps explain the extent of like the strange activity occurring in the town where this story [xx]	C, W	L, L
4	St 11	E, C	2	Okay so going back to what Student 7 said about um how it kind helps paint a picture I think that the picture that it paints is like of nature and how this place is a very like um is like a place that's kind of like enclosed by nature	C, C	M, M
				and how it says that like there were spacious coves um river dominated and it says that um that he was in a grove of tall walnut trees that shades one side of the valley	E, E	M, M
				just like this picture just makes it makes the uh like the place seem very like surrounded by nature and helps for the setting be developed	W, W	L, L
6	St 5	C, E	2	I think it like detracts from the story because although it helps the reader like imagine what's happening in the story it kind of distracts them,	C, C	M, M

				and this is like one of the descriptions he used, he said "he was tall but exceedingly lanky with with narrow sol, with narrow shoulders, long arms, and legs, hands that dangled a mile out of his sleeves, feet that may have served for shovels, and his whole frame was loosely hung together. His head was small and flat at the top with huge ears, large pained glassy eyes and a long sniped nose so that it looked like a leather cup perched upon his spindled neck, could tell which way the wind blew." And that was just like a small part of it	E, E	M, M
				so I think its like by the time he's done describing each person you kind of just forget what's happening because he uses like pages long to describe a single person so you're like multiple times where you have to read before the description to remember what was going on. So I feel like if he still like described the characters it would be helpful but just not to like the extent he did	W, W	H, H
7	St 6	E, E	6	I agree with Student 5	C, C	L, M
				because like when he gets to the point of entering like the Van Tassels house he starts describing like "It was one of those spacious farmhouses, with high-ridged but lowly sloping roofs, built in the style handed down from the first Dutch settlers; the low projecting eaves forming a piazza along the front, capable of being closed up in bad weather. Under this were hung flails, harness, various utensils of husbandry, and nets for fishing in the neighboring river." It still goes on about like describing what the Van Tassels house looks like instead of actually like, like telling what the story's more about.	E, E	H, H

8	St 4			Spacious farmhouses		
9	St 2	E, C	7	Yeah going off of that, I have, I have a really long paragraph, it's really long, and alright, it's it's describing Van Tassel and I guess it's supposed to emphasize his like wealth because it, it just it uh keeps going on about how many birds he has and like and like how, just how many birds he has	E, W	H, H
				and I guess that does like emphasize some a little bit of their characters but I just didn't think it was really necessary to add to the story	C, C	M, M
10	St 6	E, E	6	I think that kinda like Student ? said about like all this exposition is necessary for character character development but I think he put too much in and that took away from like the continuity of the story	C, C	H, H
				cause as a reader it was difficult to like follow along with what was happening. Kinda got lost in all the description.	E, E	M, M
				So I think it was a bit excessive, but maybe would have been like adequate for the time when that's what they expected in writing back then.	W, W	M, M
11	St 9	C, C	10	I agree with Student 6 but I also think that the imagery's important	C, C	M, M
				because it gives perspective of how uh how uh he, like the main character thinks and like how he like thinks like when he meets a character he thinks that like their entire physical appearance and like picks them apart of like what they look like and relates them to like the stories he's read.	E, E	M, M

12	St 8	E, E	11	Uh I would agree with Student 9 that the excessive use of imagery really helps add to the overall story rather than to detract from it. Washington Irving uses imagery to portray his characters and uh setting more vividly and to really give it the literature more like, like uh, it states within a story such as the general [xx] population [xx] which has furnished material for many a wild story in the regional shout out of the specters [xx] at all the country firesides where they didn't really have those horsemen–	C, W	H, M
13	T			I'm just having a hard time understanding you student laughter Student 8 "I'm just trying to get through it" I also wanted to stop and point out to the people recording this that they may never hear Student 8's voice again Student laughter It's like the only time Student 8 speaks student 8 "You said in the paper it was an honor" Student laughter So his name's not, so I'm sorry it's [xx] does anybody know that? [xx] peak again. We'll burn these tapes don't worry. They won't save them. I could, I get the quote but honestly you were speaking so quickly and mumbling Student 8 "I was trying to get through, I didn't want to like, I didn't want to, it was a long quote" Alright, I get the quote now, is that the end of your thought, or go ahead? Student 8 "Alright, okay, alright" Student laughter Student ? "You said it was anonymous" Student 8 "I remember you, I remember using Dutch terms to get the headless horseman into [xx] a dark vibe into get you into the devil [xx]" We're good you got the quote.		

14	St 1	E, E	12	I would agree with Student 8, I was going to use the quote that Student 5 said how it described how Crane was not [xx] to his person and it goes on for an entire paragraph to describe how tall and lanky he was, and it says "one might have mistaken him for the genius of famine descending upon the earth, or some scarecrow eloped from a cornfield."	E, E	M, M
				So I think the lengthy use of imagery is, the purpose of it is to be like very dramatic in all the descriptions so that it, the point is really gotten across to what he looks like and how and like how to give like the overall feeling of like I guess like what vibe he gives off {Student 11 "Going off, go ahead, oh you're not done yet"} and but I would also agree to something some of the other people have said that it is very lengthy	C, W	H, H
15	St 11	E, E	14	Going off of what Student 1 was saying about how kind of it like adds dramatic effect, I think that the author put this extensive, um I thought that Irving put extensive uh, uh imagery in the story because as like we were talking about how like this is kind of like shaping American literature I think that like the reason for this imagery is kind of like um like start what American literature kind of like is, just add to more like qualities of what how like American literature could be like identified so I think that like it was like he did this on purpose cause no one had ever really like done anything like this, so he kind of had like a blank slate as to what he could do and I think that like the imagery like just shapes it yep end of point	C, C	H, H

16	St 2	C, E	15	I think that the imagery also detracts from the story too because it's so much imagery and it's more about the outside person. It doesn't really delve too much into the personality. I mean obviously we know that uh Ichabod was kind of shallow but we don't know like much more about him we don't really know his backstory. It's more superficial details about him that we know rather than like the deep traits of Ichabod	C, W	H, H
17	St 7	C, E	16	I agree with you that like most of the like imagery deals with like, it doesn't delve into like deep thoughts but doesn't that like help the readers like understand like Ichabod's shallow character more? So like wouldn't it like help the progressive like the progressiveness of the story?	C, C	H, M
18	St 1	C, C	17	I would disagree I think that like the extensive imagery kind of builds off the like façade of what like Ichabod looks like or like what the house looks like and it builds the like classic archetype like for the Van Tassels that like they're wealthy and have a lot of money but then beyond that you don't really know a lot about them and I think like you have to really think about um like Ichabod's actions to like pull out a theme of like him being like really greedy cause you don't really realize, I guess I didn't really figure it out until I went back and like thought about it again.	C, C	H, H

19	T			Well I'm still likes, the likes are getting real [xx] Overlapping student talk So when we look at who's the protagonist, Ichabod. Don't we typically want to root for the protagonist? Students "Yes" Do we want to root for Ichabod? Students "No" Not quite, not a great guy. So who's the antagonist? Student "Ron" Ron, do we want to root for Ron? Students "No" Ehh not particularly, right. One of the things that also was introduced in romanticism is there are more more human characters. They, these are like real humans with real life flaws and despite the fact that Ichabod is our protagonist there's nothing really likeable about him. He's driven by the desire to accumulate, right? Even when he describes Katrina, he uses terms that sort, imagery that revolves around what? (...) What are the images are all similar to what, like he uses what, I'm trying to ask it without, what kind of imagery does he use when he describes		
20	St 9			Says like plump as a partridge		
21	T			Right, so he's driven by Student ? "Food" Food. He describes her using all these food terms because her dad owns a farm and again she's beautiful and her dad's rich and this just kind of is what he's looking for. But we've got these two characters and were not really rooting for one or the other, got it?		

22	St ?	E, E	17	I also think that the imagery shows that since it's like from Ichabod's perspective it's like and it's unreliable in a sense where it's like he's not really telling us facts but his own opinion of like his extensive opinion of the imagery and also like he Teacher said, he's like more interested in his desires. He wants food, he wants greed, and he gets that in the farm essentially	C, C	H, H
23	St 3			Are we good? Question 1? Alright, question 2. In what ways did Ichabod's overactive imagination eventually lead to his downfall?		
24	St 8	N, N	24	I feel that he bought [xx] imagination {Student laughter} I feel like Ichabod's imagination leads to his downfall in really two ways. I feel like uh first his imagination leads to him getting completely carried away	C, E	M, M
				about the situation with Katrina such that he thinks his chances are much better than they really are and that he really fantasizes about the future so much that he cannot imagine failing. This [xx] keeps it from making the necessary life changes to become the kind of person that's suitable for her.	E, E	H, H
				Second, I think his great enjoyment of ghost stories about supernatural [xx] that he actually believes because of his strong imagination makes him utterly susceptible to Brom Bones brain	C, C	H, H

				which is stated in the book as that "when he was returning one night from the neighboring village of Sing Sing, he had been overtaken by this midnight trooper; that he had offered to race with him for a bowl of punch, and should have won it too, for Daredevil beat the goblin horse all hollow, but just as they came to the church bridge, the Hessian bolted, and vanished in a flash of fire." {Overlapping student and teacher talk} Uh uh the [xx] Brom Bones only shows Ichabod's gullibility	E, E	H, H
25	St 10	E, E	24	I think that his like love of scary stories and like all the magic that was associated with them. Um when he was being chased it says that "All the stories of ghosts and goblins that he had heard in the afternoon, now came crowding upon his recollection." and like he became more frightened	E, E	H, H
				and I think he played into like into his chasing was more frightened than he was in the fight path and I think like his overactive imagination like lead to his downfall	C, C	M, M

26	T			We, was this regarding? I'm going to pause you. We know, is there actually a headless horseman? No. Who is the headless horseman? Students "Brom" Brom. We all know that, right? Cause Student ? did not know that. He thought, he thought there was a headless horseman Students laughing And I told him someone would have to talk to him about Santa Clause later Students laughing So we get that this is, and I use the air quote purposefully, we get that this is a ghost story, right? It's supposed to be this tongue and cheek, humorous, like not laugh out loud but like humorous ghost story that we get that it's not actually. Do we also, did Katrina have any interest in Ichabod romantically? No. So then why act like she had interest in him romantically? (..) If she didn't, then why act like it?		
27	St 1	N, N		Was she like using him to get lessons out of him or something?	C, C	M, L
28	T			No. No, you got the first part right, I thought you had it and then you threw the lessons thing in		
29	St 2	E, N	27	Was she um using him to make Brom jealous?	C, C	M, M

30	T			Yes! So Brom it says earlier, it says earlier in the story that Brom was kind of dragging his feet on you know, looking to wed Katrina, so Ichabod goes to town, he's new, he's flashy, she starts to flirt with him, Brom gets ticked off and expedites the process, and then at the end of the story they get married, right? And they live happily ever after. It's just like, did Ichabod actually die? (..) No. He, this, this kind of story would not have been punctuated with a bloody death, like Brom's not a murderer, he's just a big ole goon. So he embarrassed Ichabod to get him out of town like whether or not Ichabod actually knew it was Brom Bones like that's kind of ambiguous, there's a lot of ambiguity at the end, but if you look at the vibe of the story, that's the word of the day, vibe. If you look at the feel of the story ending it in a horrific death of the protagonist, it just doesn't fit with like the mood of what we've got goin on. So alright I wanted to make sure we were clear about that, alright go ahead		
31	St 7	E, E	25	Alright so I agree with what Student 10 said uh that like Ichabod's imagination is really powerful	C, C	L, L
				cause like cause like he's so like extremely superstitious like that he like begins to believe like {Teacher "LIKES"} Alright, so uh {student laughter} he begins to believe legends that like he hears {Student laughter} {Teacher "keep goin"} And like to the point that he's like afraid to walk home more like ride home by himself.	E, E	M, M

32	St 4	E, E	31	Going off of what Student 10 said him having an overactive imagination of these stories that he truly believes, when he was on his way home passing up this bridge that, this supposed bridge that Andre was captured, "he summoned up, however, all his resolution, gave his horse half a score of kicks in the ribs, and attempted to dash briskly across the bridge; but instead of starting forward, the perverse old animal made a lateral movement, and ran broadside against the fence."	E, E	H, H
				So him being like afraid for his life because of these imaginative stories made him like scared and the horse is scared too, so now he's fearing more for his life	C, E	H, H
33	St 11	E, E	32	I also think his um ambition, kind of like his um ego kind of also got the best of him	C, C	M, M
				because you know, when he saw like the Katrina and her dad and like everything that they had he kind of like fell, not like fell in love with it, but like fell in love with the idea of having all of that, and the quote where it says that " Ichabod fancied all this, and as he rolled his great green eyes over the fat meadow-lands," like just goes on and on about how that's everything that he always wanted and he just assumed that he could get all this by um because he was kind of like, he was very, he was he was respected in the town to an extent just because he was intelligent and you know he's like a new guy and he has all this knowledge and he can sing and he can teach	E, E	H, H

				and I think that the fact that he was respected in this new little town that he came to kind of made his ego like extremely large and he thought that he could um go in and like have everything but like little did he know that he was kind of, he was kind of being like tricked by Brom and he ended up you know, getting hit with a pumpkin {Student laughter} and run out of town and so yeah I think that his ambition also played a part along with his imagination too.	W, C	H, H
34	St 9	E, C	33	I agree, I think that um because of his imagination he tends to live in his fantasy world instead of reality where he's, he's always like instead of like really seeing what it is, he starts imagining all this fiction in the real world and that leads to him being gullible	C, C	H, H
				when Brom actually hits him with the pumpkin like Student 11 said and then he just uh he just leaves town because of it	E, E	M, M
35	St 5	E, N	34	[xx] imagination ran wild. That kind of like led to his downfall cause he didn't really think logically. [xx] when people see something out of the ordinary they like, we're at first scared but then they like pause for a second and they reason through what it could be but because of his imagination he automatically thought it was something like like supernatural such as a goblin so he became scared for no reason and he kind of just ran but [xx] his imagination he would have been able to realize it was just a prank	C, C	H, H

36	St 1	E, E	33	So I wonder what Student 11 before said that like he's so like he's so wrapped up in like he wants to have lots of wealth and women and lots of food and he, it talks about how he goes from house to house with his students to try to find work to pay off his like rent in return for like food	E, C	H, M
				and I think he lives a life of like he's always getting like food and things handed to him and he like thinks like I'm almost there like I'm living the life but in reality he doesn't really have anything because he's um kind of like house hopping from student to student, so it seems that like he can't, he doesn't really realize what he's doing and he thinks like he's almost made it and Katrina will like set him up for [xx]	C, W	H, H
37	St 6	E, E	36	Yeah I agree with that. His imagination kind of makes him complacent	C, C	L, M
				because um he always saw like this idealized version of him and him and Katrina but uh in the end he never like took an active step like to actually make that his idea value, he always treated the idea as reality and never saw that he didn't have a chance with Katrina because he's ya know, he's like a freeloader and Katrina has all this money and wealth and uh that would never happen in real life, so [xx]	E, E	H, H
38	St 3			Good, alright cool. Question 3, do you believe Ichabod comes across as a sympathetic character and why?		

39	St 10	E, E	37	I think that um he does come across as a sympathetic character like to an extent because he does uh have redeeming qualities, like he is uh smart and bright, he is helpful teaching everybody, but he also does have his flaws like his imagination and like how vain he is, and I think that because he isn't painted like a stereotypical hero, or like a stereotypical protagonist, and like how it says he's exceedingly lank, like he's not like a strong uh person but he's super awkward and he's a bit more relatable than just like a superhero	C, E	H, H
40	St 2	E, E	39	Yeah going off of that, I think he is a relatable person because of just like the situation he was put in like that he wanted something that he didn't realize what it would take and like that's something that everybody has been through like when they wanted something but they don't know like what to do to actually get that, and so I feel like people can relate to him that way	C, C	H, H
41	St 5	C, C	39	I don't think he was sympathetic cause like he just came across as greedy which made people not really like him as a character.	C, C	M, H
				"As Ichabod jogged slowly on his way, his eye, ever open to every symptom of culinary abundance, and as he beheld them, soft anticipations stole over his mind of dainty slapjacks, well buttered, and garnished with honey or treacle," anyways so whenever he sees the farm he just kind of thinks about how those animals would benefit him not really about how they were treated or what will happen to them	E, E	H, H
				which like shows how greedy he is and how he only cares about things that specifically like applied for him	W, W	M, M

42	St 1	E, E	41	I would agree with Student 5 like Student 10 you pointed out some of the points of like he's really really smart and he has good like singing and choir voice and that can make him likeable but I think at the same time he uses those qualities as a way to get people to give him their charity which I find rather annoying and I think like similar to what Student 5 said he's so wrapped up in trying to uh find like his great wealth that he wants that he, I can't really have sympathy for him because he's so greedy and wrapped up in like the more material things, that he's not, like I don't think he realizes what he could actually do with like his um like smarts and choir	C, C	H, H
43	St 11	E, C	42	I think that it's kind of a matter of like he's not so much, you don't really feel sympathy for him because he is like not a great dude but you kind of relate to him almost I think cause like some people were saying that he's kind of like a relatable character just cause like um he's more human than like most heros and you find out cause like nobody's perfect and everybody here knows that like none of us are perfect and that none of those people are perfect than I think that his flaws are just kind of like highlighted, cause he is a character in a story that Irving wanted to like highlight his flaws to make him a little more relatable than the average protagonist in a story	C, C	H, H

44	St 9	E, E	43	I agree that he is more human but I don't think he's sympathetic because through his imaginations and him living in his fantasy world he comes across as like very self centered where he wants, he thinks he has this majestic voice when truly since he's unreliable as a narrator we don't know if he really had a good voice, so {Teacher "Who's unreliable?"} Or like, his, his perspective, like where he thinks he has a majestic voice that he can give people singing lessons	C, C	H, H
45	T			I just don't get the unreliability of that, that's just part of the story that the third person narrator is telling us Student laughter But I stand by the rest of my statement		
46	St 9	E, N	44	I think he's self centered and he thinks that Katrina is like falling in love with him when in reality she's just using him	C, W	M, M
47	St 6	E, E	44	Yeah I think that is like ignorance towards like situation doesn't really make him that sympathetic	C, C	L, L
				cause as the reader you're just more irritated by what he's doing than um sympathetic towards him because he doesn't act to take steps to help himself. That, that was irritating to me as the reader, but he does like draw a lot of sympathy as a character because like you said all the school children stuff he had a lot of uh he had a lot of sympathy for their parents and stuff cause he'd always sing and stuff so he garnered a lot of sympathy with the adults	E, E	H, H
48	T			Student ? close this with something really super smart. (. . .) Student ? close this with something moderately not dumb		

49	St 2	E, E	46	Um I think Ichabod, okay so what do you think truly happened to Ichabod the night after the party, do you think Irving wants us to think one way or another. I think going back to like his imagination that he was just, and just like the situation in general that he was just so scared because of the ghost stories and uh what he heard and also because he was really upset that Katrina had rejected him, and I think those added up together and that's what caused him to like not see that he was [xx] to see this apparition of the headless horseman and that I don't think Irving wants us to think one way or the other	C, C	H, H
				because he gives clues that it was Brom Bones but he also says that like, he said that Brom Bones would laugh anytime uh the story was mentioned and he also said that the old maids say it was the headless horseman.	E, E	H, M
				So he does kind of give both sides of the story [xx]	W, W	M, M

By analyzing the confusion matrices between true and predicted labels we can analyze the behavior of our NLP models. Table 31 shows the confusion matrix for collaboration.

Table 31: Confusion matrix for collaboration labels.

		Predictions			
		New	**Challenge**	**Extension**	**Agree**
Ground truth	**New**	3	0	0	0
	Challenge	0	3	3	0
	Extension	3	4	19	0
	Agree	0	0	0	0

First, we can observe that there are no "agree" labels in this particular discussion. This is not unusual since it is by far the label with least number of occurrences (less than 2%

frequency in D3 and less than 3% in D4). However, for other discussions which do contain agreements, we noticed that our collaboration model never predicts this particular label. Since all the performance metrics we report represent an unweighted average, the low performance on agreements impacts the overall performance considerably. It is not easy to mitigate this effect since the class imbalance is so high, but we believe future studies can investigate whether oversampling (or other sampling procedures) can have a positive effect on performance.

Second, the confusion matrix shows the collaboration model performing much better on extensions: it misses 7/26 extensions, and 19/22 extensions predicted by the model are actual extensions. Overall, the trend that emerges indicates that our proposed model is well equipped for distinguishing between new turns from the remaining ones. On the other hand, if a turn is related to a prior one, as is the case for extensions and challenges, our model can struggle in distinguishing between them.

Table 32: Confusion matrix for argumentation labels.

		Predictions		
		Claim	**Evidence**	**Warrant**
Ground truth	**Claim**	27	3	6
	Evidence	1	19	1
	Warrant	2	0	5

Table 32 shows the confusion matrix for the argument component classifier. The main takeaway in this case is that the model is better at distinguishing between evidence and the remaining two labels (5/13 errors) while there is room for improvement in differentiating claims from warrants (8/13 errors). A considerable improvement in this direction was achieved with the context models discussed in Chapter 6. In particular, local context was beneficial for warrants since it incorporates information on nearby argument components (recall from Chapter 3 that a warrant requires the existence of claim and evidence).

Lastly, Table 33 represents the confusion matrix for the specificity model. A good indicator of model robustness in this case is the absence of low/high specificity misclassifications (0 for this particular transcript, and a low number in general), which are heavily weighted in our reported quadratic-weighted kappa metric. Data imbalance is not an issue for datasets D3 and D4 in particular, which results in homogeneous performance across all three specificity labels. Interestingly, we can see a slight tendency of our model to underpredict specificity: in 7 cases the predicted value was lower than the actual one, while only in 3 cases the model predicted a higher value than the ground truth.

Table 33: Confusion matrix for specificity labels.

		Predictions		
		Low	**Med**	**High**
Ground truth	**Low**	4	2	0
	Med	3	21	1
	High	0	4	29