Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

Discourse Parsing of Contentious, Non-Convergent Online Discussions

Stepan Zakharov,[1] Omri Hadar,[2] Tovit Hakak,[2] Dina Grossman,[2]
Yifat Ben-David Kolikant[2] and **Oren Tsur[1]**
[1]Software and Information Systems Engineering, Ben Gurion University
[2]School of Education, The Hebrew University of Jerusalem
stepanz@post.bgu.ac.il,orentsur@bgu.ac.il
{omri.hadar,tovit.hakak,dina.grossman,yifat.kolikant}@mail.huji.ac.il

Abstract

Online discourse is often perceived as polarized and unproductive. While some conversational discourse parsing frameworks are available, they do not naturally lend themselves to the analysis of contentious and polarizing discussions. Inspired by the Bakhtinian theory of Dialogism, we propose a novel theoretical and computational framework, better suited for non-convergent discussions. We redefine the measure of a successful discussion, and develop a novel discourse annotation schema which reflects a hierarchy of discursive strategies. We consider an array of classification models – from Logistic Regression to BERT. We also consider various feature types and representations, e.g., LIWC categories, standard embeddings, conversational sequences, and non-conversational discourse markers learnt separately. Given the 31 labels in the tagset, an average F-Score of 0.61 is achieved if we allow a different model for each tag, and 0.526 with a single model. The promising results achieved in annotating discussions according to the proposed schema paves the way for a number of downstream tasks and applications such as early detection of discussion trajectories, active moderation of open discussions, and teacher-assistive bots. Finally, we share the first labeled dataset of contentious non-convergent online discussions.

1 Introduction

Online discussions are often perceived as polarized and unproductive. Empirical evidence, gained from field experiments in collaborative learning, suggests that raising the awareness of the participants to the dynamics of the discussion influences their discursive behavior, ultimately making the discussion more productive (Wise and Chiu(2011); Yoon(2011); Chen et al.(2018)Chen, Chang, Ouyang, and Zhou). However, reflecting on conversational dynamics in real time is challenging and often requires experienced human moderators. In order to mitigate this challenge, we had developed a novel annotation scheme for discourse parsing. Our annotation scheme, inspired by the Bakhtinian principles of *dialogism*, is explicitly designed to meets two objectives: (i) capturing strategic discursive moves that impact the trajectory of a discussion, and (ii) acknowledging that a successful and productive discussion does not necessarily mean that participants converge to agreement.

Existing schemes, for the most part, rely on *convergence* as a measure for the success (productiveness) of a discussion. Success is declared if a convergence to consensus is achieved, e.g. (Chiu(2000); Diziol et al.(2010)Diziol, Walker, Rummel, and Koedinger; Schwarz et al.(2018)Schwarz, Prusak, Swidan, Livny, Gal, and Segal). Consequently, computational work addressing conversational discourse typically aim at modeling discussions in which convergence/agreement is the declared goal. Two examples of this setting are customer service log calls (Jurafsky, Shriberg, and Biasca(1997)), and problem solving of a math assignment in a collaborative learning environment (Schwarz et al.(2018)Schwarz, Prusak, Swidan, Livny, Gal, and Segal).

Many discussions, however, are often polarized, contentious, and *non-convergent* in nature. We argue that existing schemes for discourse analysis do not properly capture the dynamics exhibited in fruitful, yet non-convergent discussions, and therefore cannot be used effectively to evaluate or improve the quality of online discourse.

In order to address this problem, we redefine the success of a discussion in a way that accounts for polarized, non-convergent discussions. We further introduce a new annotation scheme for non-convergent discussions, and define a novel discourse parsing task. Our annotation scheme is inspired by the Bakhtinian concepts of Internally Persuasive Discourse (IPD) and Dialogic Agency – a crucial capacity in a modern, liberal, multi-cultural society. Another key factor guiding the development of the annotation schema is the observation that conversations unfold sequentially, thus a speaker's utterance is triggered only by previous utterances. While this observation may appear trivial, most previous work on discourse parsing analyze conversations as a whole. In this work we take a different approach. First, we focus on the discursive evolution of an ongoing conversation, rather than on an isolated utterance or on the complete conversation. Accordingly, our proposed tagset and the annotation process, were tailored to capture conversational strategies used by speakers (and perceived by participants and annotators) at the time of the utterance, and not in hindsight, after the conversation is complete. We elaborate on the theoretical background and the novel parsing tasks in Sections 2 and 3.

In line with the motivation stated above, the proposed annotation scheme poses two main computational challenges: (i) it aims at discourse parsing at *real time*, as the discussion evolves, thus the available context is limited, and (ii) the tagset contains 31 tags and each utterance can (sometimes should) be tagged with a set of tags, rather than a single tag, making sequence optimization hard. We address the computational challenges in Section 5.

We use discussions from Change My View (CMV) – a unique forum (subreddit) on Reddit – as an ideal test case due to its abundance with polarized discussions on multiple topics. We carefully studied a few dozen discussions in developing the tagset and the annotation schema. Consequently, we have created a dataset of thousands of manually annotated utterances from hundreds of CMV discussions. A detailed description of the data and the annotation procedure is provided in Section 4.

Contribution The contribution of this work is threefold: (i) We propose a novel framework (annotation scheme, tagset, task) for conversational discourse parsing, especially designed for non-convergent discussions, so common online, (ii) We share the first labeled conversational dataset annotated with the proposed scheme, and (iii) we demonstrate the learnability of the proposed tagset and provide an extensive analysis of the relevant features and the way different levels of discourse interact.

2 Related Work

2.1 Theoretical Framework

Oftentimes, a successful discussion is conceptualized in terms of convergence, an increment in the overlap of participants shared knowledge (Teasley et al.(2008)Teasley, Fischer, Dillenbourg, Kapur, Chi, Weinberger, and Stegmann). The convergence metaphor is viable for discussions that revolve around a problem or a task participants need to solve or achieve together, e.g. (Chiu(2000); Diziol et al.(2010)Diziol, Walker, Rummel, and Koedinger). It captures the gradual process through which a consensual answer is achieved (Lu, Chiu, and Law(2011)), a problem is resolved by a support agent (Bhuiyan et al.(2018)Bhuiyan, Misra, Tripathy, Mahmud, and Akkiraju; Oraby et al.(2017)Oraby, Gundecha, Mahmud, Bhuiyan, and Akkiraju) or collaborative learning in which students solve math problems (Schwarz, Neuman, and Biezuner(2000)).

Nonetheless, social media is replete with non-convergent discussions, for example discussions about morality, historical events, politically charged debates, and so forth. Consensual answer is not expected in non-convergent discussions, yet these discussions may serve as a fruitful venue for the development of dialogic agency (Parker(2006)). Dialogic agency is related to the Bakhtinian concept of Internally Persuasive Discourse (IPD), a discursive regime, where participants examine and refine their vested truth – ideas, consciousness, anthology – in light of alternatives and critique provided by their peers, and even transcend beyond them (Bakhtin(1981); Bakhtin(1986)). The quality of such non-convergent discussions, therefore, depends on participants' *responsiveness* – their ability and will to consider their interlocutor's point of view. Taking into account the interlocutor's perspective does not necessarily imply agreement with her. Rather, it is about acknowledging the essence of the message conveyed in one's response (Matusov and von Duyke(2009)).

Inspired by Bakhtin's concept of dialogicity, we posit that a non-convergent discussion is productive as long as discussants exchange new ideas, while being responsive to alternatives and critiques offered by other interlocutors. The discourse annotation schema developed in Section 3 is designed to cover a range of discursive moves that reflect the speakers' responsiveness and the way it sets the trajectory of a discussion.

2.2 Conversational Discourse Analysis

Bakthin's dialogical stance has already been employed in the analysis of discourse in the field of collaborative learning, e.g., (Trausan-Matu, Dascalu, and Rebedea(2014); Hennessy et al.(2016)Hennessy, Rojas-Drummond, Higham, Márquez, Maine, Ríos, García-Carrión, Torreblanca, and Barrera). However, these applications were not sensitive to non-productive interactions – cases in which the discussion has deteriorated, which are common in online discourse. Recognizing potential "bad" discursive moves is critical for improving the productivity of a discussion. Moreover, we conceptualize importance and productivity, in terms of *potential* impact on the unfolding conversation. A comprehensive review of other annotation schemas discussions in educational settings can be found in (Noroozi et al.(2018)Noroozi, Kirschner, Biemans, and Mulder). We are interested in a system (and a coding scheme) that captures the evolution of a discussion in an online manner – allowing intervention when needed.

Zhang et al. (Zhang et al.(2018)Zhang, Chang, Danescu-Niculescu-Mizil, Dixon, Hua, Taraborelli, and Thain) analyzed the role of politeness (or lack thereof) in keeping conversations on track. Among other methods, they asked annotators to intuitively predict which conversation will get derailed, with emphasis on personal attacks and aggressiveness as causes. This work is the closest to ours, as it tackles early signs of deterioration in an online manner. However, Zhang et al. focus on politeness while we propose a comprehensive framework for discourse annotation and analysis. Another relevant work is that of (Cheng et al.(2017)Cheng, Bernstein, Danescu-Niculescu-Mizil, and Leskovec), who found that trolling behavior is not confined to a vocal and antisocial minority, but rather ordinary people can engage in such behavior as well. Such behavior is not merely due the individual's mood, but also the surrounding context of the discussion.

Two important lines of work include general discourse parsing, e.g., (Core and Allen(1997); Jurafsky, Shriberg, and Biasca(1997); Stolcke et al.(2000)Stolcke, Ries, Coccaro, Shriberg, Bates, Jurafsky, Taylor, Martin, Ess-Dykema, and Meteer;

Khanpour, Guntakandla, and Nielsen(2016); Ji, Haffari, and Eisenstein(2016)) and argumentation mining, e.g., (Suthers(2003); Erkens and Janssen(2008); Rosé et al.(2008)Rosé, Wang, Cui, Arguello, Stegmann, Weinberger, and Fischer; Pinkwart et al.(2009)Pinkwart, Ashley, Lynch, and Aleven; McLaren, Scheuer, and Mikšátko(2010); Klebanov et al.(2016)Klebanov, Stab, Burstein, Song, Gyawali, and Gurevych; Bar-Haim et al.(2017)Bar-Haim, Bhattacharya, Dinuzzo, Saha, and Slonim), among others. These works differ from ours in a number of fundamental ways. First, we are not interested in convergence or persuasion. Second, we are interested in capturing responsiveness in the context of IDP. That is, we are not interested in labeling the grammatical function of an utterance (such as 'a question'), nor merely the type of the argument, but rather to label how it relates to previous utterances (such as 'a request for clarification of previous statement'). Our focus is on *discursive moves* rather than on argument types[1].

2.3 Non-conversational Discourse Analysis

Discourse is also analyzed within a single textual unit[2], modeling its coherence and topic drift (Grimes(1975); HOBBS(1985); Barzilay and Lee(2004)). Recent work aims at modeling the relations between consecutive sentence pairs from the same document. The Penn Discourse Tree Bank (PDTB) 2.0 is composed of sentence pairs from five sections of the WSJ news stories, annotated with a set of discourse markers such as *but*, *and*, *as*, and *if* (Prasad et al.(2008)Prasad, Dinesh, Lee, Miltsakaki, Robaldo, Joshi, and Webber). State-of-the-art results are achieved using a bi-LSTM coupled with a BERT Transformer (Nie, Bennett, and Goodman(2019)).

Naturally, the two levels of discourse – the in-utterance sentence-pair level and the conversational level, complement each other. We therefore use the model proposed by Nie et al. (Nie, Bennett, and Goodman(2019)) to enrich the representation of the conversational data by adding PDTB tags to each utterance.

3 Discourse Annotation Scheme

We have developed a discourse annotation schema that captures the dynamics in a non-convergent discussion, with emphasis of speakers' responsiveness to previous utterances. This schema reflects the dynamics of an *ongoing* discussion, lending itself to sequential annotation, in which an annotator labels each utterance as she reads it (and not in an ad-hoc manner).

We define potentially[3] productive discursive moves in light of Gricean principles (Grice(1975)): (i) The interaction should revolve around the topic of discussion. Hence, discursive moves that shift the topic entirely are potentially unproductive, (ii) Utterances adding knowledge (or a different perspective) potentially contribute to the discussion and are likely to encourage the continuity of the conversation, whereas repetition may hasten the termination of the discussion or steer it into a loop, (iii) Mere addition of new knowledge, even if relevant and viable, is insufficient. Rather, the interlocutors are expected to be responsive to each other, taking into account the (essence of the) preceding utterances. Ignoring, distorting, practicing pettiness or referring only to insignificant elements of the previous utterance has the potential to harm the discussion, and (iv) Politeness and aggression impact the responsiveness of the conversants. Some of these maxims are echoed implicitly by (Barron(2003); Zhang, Culbertson, and Paritosh(2017); Cheng et al.(2017)Cheng, Bernstein, Danescu-Niculescu-Mizil, and Leskovec; Zhang et al.(2018)Zhang, Chang, Danescu-Niculescu-Mizil, Dixon, Hua, Taraborelli, and Thain). These Grician principles serve in determining the level of responsiveness. We wish to stress that the term *responsiveness* is used in the Bakhtinian sense of dialogic agency and IPD, as presented in Section 2.1, and not as mere counts of utterance-response turn-taking.

Guided by these principles, we thoroughly analyzed dozens of discussions from Reddit's CMV (see Section 4), identifying significant discursive moves that reflect high and low levels of responsiveness. This qualitative analysis served as the basis for the new annotation scheme, introduced below (Section 3.1). The tagset was revised and refined through an initial annotation process, having a group of experienced meta-annotators discuss the annotation schema and the various discursive strategies observed.

3.1 The Tagset

The new tagset is presented in Table 1, and excerpts from two annotated conversations are presented in Table 2. The tags fall under four main categories, corresponding to a range of (potentially) un/productive discursive moves.

The first two categories capture *responsiveness quality* – knowledge exchange or the lack of it. For example, `CounterArgument`, `RequestClarification`, or `Extension` (of an argument made beforehand).

The third category corresponds to *style and tone*. We distinguish between positive (e.g., appreciation of the merit of one's response even if disagreeing with it, using humor, attempting to lessen the tension, predicting objections and tackling them), and negative attitudes (e.g. sarcasm, blaming the other interlocutor, expletive language).

The forth category – *Disagreement Strategies* captures a finer grained counter argumentation. The need for these nuanced subcategories is supported by a rich body of empirical literature, suggesting that different disagreement strategies intensify or soften aggression in argumentative discussions, e.g., (Felton, Garcia-Mila, and Gilabert(2009); Locher(2010); Angouri and Tseliga(2010); Lu, Chiu, and Law(2011); Shum and Lee(2013); Cheng et al.(2017)Cheng, Bernstein, Danescu-Niculescu-Mizil, and Leskovec; Cougnon, Coppin, and Figueroa(2019)). For

[1] This is consistent with our interest in non-convergent discussions vs. dialogues striving for agreement or debates culminating in a 'victory'.

[2] The terms 'discourse analysis' and 'discourse parsing' often used to refer to non-conversational units.

[3] We refer to the *potential* of the act since we predict its contribution in an online manner. At the time of the utterance the actual significance is not yet known.

Description	Tag
1. Discursive moves that potentially promote the discussion	
Moderating /regulating, e.g. "let's go back to the discussion topic"	Moderation
Request for clarification	RequestClarification
Attack on the validity of the argument ("can I see the evidence?")	AttackValidity
Clarification of previous statement (utterance)	Clarification
Informative answer of a question asked (rather than clarifying)	Answer
A reasoned disagreement, a refutation. Can be accompanied by disagreement strategies	CounterArgument
Building/extending previous argument (w/o negation or disagreement). The speaker takes the idea of the previous speaker and extends it.	Extension
A viable transformation of the discussion topic	ViableTransformation
Personal statement "this happened to me")	Personal
2. Moves with low responsiveness	
A severe case of low responsiveness, such as continuous squabbling and quarreling	BAD
Repeating previous argument without any new detail or substantive variation	Repetition
Response to ancillary /distract the discussion	NegTransformation
Negation/disagreement without reasoning	NoReasonDisagreement
Convergence towards previous speaker	Convergence Agreement
Announcement that the issue is not solvable yet there is legitimacy to the other's voice	AgreeToDisagree
3. Tone and style	
3.1 Negative tone and style	
Aggressive and Blatant (+ intensifying qualifiers) , e.g., "this is stupid"	Aggressive
Ridiculing the partner (or her argument)	Ridicule
Addressing (complaining about) a negative approach e.g. "you were rude to me"	Complaint
Sarcasm/ cynicism /patronizing	Sarcasm
3.2 Positive tone and style	
Attempts to reduce tension: respectful, flattering , socially unifying tone	Positive
Weakening qualifiers e.g. "I'm not an expert in this topic..."	WQualifiers
4. Disagreement strategies (applied in addition to CounterArgument)	
4.1 Easing tension	
Softening the blow; Saying something nice before or after conveying the disagreement	Softening
Partial agreement yet disagreement , e.g. "I disagree only with one part of your text"	AgreeBut
Explicitly taking into account other participants' voices	DoubleVoicing
Using an external source (a URL, academic reference) to support a claim	Sources
4.2 Intensifying tension	
Reframing or paraphrasing the previous comment in a way that changes the original meaning/sheds light on the limitations of the argument	RephraseAttack
Critical question, phrasing the (counter) argument as a question	CriticalQuestion
Offering an alternative without directly refutation or disagreement	Alternative
Direct disagreement ("I disagree", "this is simply not true")	DirectNo
Refutation focuses on the relevance of previous claim	Irrelevance
Breaking previous argument to pieces without real coherence	Nitpicking

Table 1: Tagset for non-convergent discourse parsing. The tags are divided to four discourse categories.

example, consider the intuitive difference between the statement "that's not true!" (`DirectNo`) and a discursive move that 'softens the blow' by preceding the counter argument with a compliment like "that's a good point, but..." (`Softening`) – the former is potentially destructive, while the latter has the potential to ease tension. Other disagreement strategies, such as `Nitpicking` emerged as we qualitatively analyzed the initial data. Accordingly, the tags under category 4.1 reflect strategies that ease potential tension; whereas those in category 4.2 reflect strategies that tend to intensify it.

3.2 Annotation Procedure

The unit of analysis is a full utterance. Annotators were asked to classify each utterance in terms of the central response made vis-a-vis the preceding utterances (and only the preceding utterances). Annotators were asked to code the essence of the utterance, considering its potential in shaping further utterances, rather than to dissect and annotate each and every part of the response. This gestalt viewpoint on responses is well established from a theoretical perspective (Gunawardena, Lowe, and Anderson(1997)).

We allow multiple labels to be assigned to a single utterance. Tag collocation captures fine grained argumentation along with the interlocutor's tone and allows a responder to address a specific critique and extending a previous argument (see Table 2 for examples).

Although annotator agreement is not forced, annotators have a consolidation session after the annotation of each discussion. Consolidation was found to increase the quality of the annotation. See more on agreement in Section 4.

3.3 Examples: Labels and Discourse Trajectory

We now provide a brief analysis of two excerpts from two longer discussions taken from Reddit's CMV subreddit (A detailed description of the CMV and the dataset is provided in Section 4). These examples demonstrate the effectiveness of the annotation schema in capturing (sequences) of un/productive discursive moves.

Table 2 (top) presents a three-turn sequence extracted from a longer discussion about 'racist words'. Speaker A's discursive strategy (turn 7) is to ignore the sarcastic tone (`Sarcasm`) of speaker B (turn 6), only providing the requested clarification (`Clarification`). Consequently, B responds more positively, respectfully and on point by strategically adding positive tone (`Softening` and `WQualifiers`) to his main move – `CounterArgument`. In essence, A's choice to ignore B's tone promotes the complex discursive move taken by B to put the discussion back on track.

Unfortunately, not all discussions recover. Iterating between `NegTransformation` ([petty] distraction), `Ridicule` and `Aggressive` signals that the conversation may not be going well (`BAD`). Table 2 (bottom) presents an excerpt from a discussion about the "appropriateness of bringing meat to a potluck with a vegetarian". This sequence of discursive moves pushes the discussion from bad to worse, essentially deeming it unproductive – for dozen or so discursive turns (not in the table) the dialogue is off topic, oscillating the semantic "can" vs. "want" argument, all embedded in insults. This is the case in which a conversational discourse parser could be used to produce a gentle warning, alert a human moderator or prompt the participants with a suggestion for a different discursive strategy that will break the vicious cycle.

The two excerpts in Table 2 provide an important observation – the deterioration of a conversation is a process, reflected by a *sequence* of moves. Both cases started with a potentially "bad" move (`Sarcasm` and `Ridicule`), yet one recovered, whereas the other deteriorated. As the discussion on the top demonstrates, one bad move, one blatant, aggressive move, does not always bring about deterioration.

4 Data and Annotations

Reddit and r/ChangeMyView Change My View, commonly known as CMV, is a unique forum in the Reddit platform. The premise of CMV is to promote a well reasoned discussion on contentious issues. Users initiate a discussion with a provocative statement, backing it by some initial argumentation, prompting the community to "change the author's initial viewpoint". Good argumentation is voted up and convincing arguments are awarded a 'delta'. The community is strictly moderated thus trolling and ad-hominem attacks are deleted, ideally creating a sterile debate atmosphere. However, discussions often turn heated, presenting a heavy use of sarcasm and aggressive language in spite of the moderation. CMV is, therefore, an ideal arena for stimulating open-ended and non-convergent discussions, rich enough in the number of users, variety topics discussed, argumentation styles and productivity of discussion.

Discussions as trees and branches A discussion unfolds in a tree structure. Given a post, a user can reply directly to the post. Branching occurs if two or more users reply directly to the same post. Branching happens for reasons ranging from a naive chaining mistakes to conscious decisions to "take the discussion to another direction". Modeling the reasons for branching and their effect on the discussion is beyond the scope of this paper. For the sake of simplicity, in this paper we view a branch – an ordered sequence of posts (root to leaf), as a discussion. Multiple branches of the same tree are considered separate discussions on the same topic, although some of them may share the branch prefix (all of them share the root – the original prompt). Indeed, a manual examination of a sample of trees shows that it is reasonable to consider branches of the same tree as separate discussions if the shared prefix is not dominant (in length) compared to the mutual exclusive suffixes of the branches.

The Annotated Dataset[4] Sixteen thousands discussions that took place between April 2017–January 2018 were extracted from Reddit/r/CMV and converted to conversation trees. 1,946 branches from 100 trees were annotated by a group of trained annotators. Each discussion was annotated by two annotators. We consider annotations as valid only if both annotators agreed on a label. Agreement was reached in 95% of the instances (lowest agreement for tag type was 91%).

The dataset contains a total of 10,559 posts out of which 9,620 were labeled with 17,964 tags (31 unique tags). Average number of users per tree, average length of a post and further details are provided in Table 3.

Tag collocation measured through Point-wise Mutual Information (PMI) values. `Ridicule` and `RephraseAttack` have high PMI, while `Ridicule` and `Clarification` have low PMI. Transition probabilities between tags in consecutive utterances are presented in Figure 1. As expected, we observe that

[4]The Annotated dataset and guidelines are available at https://www.naslab.ise.bgu.ac.il/projects

Turn	Speaker	Text	Labels
6	B	So why are we even discussing this right now if it's not your responsibility to change my mind. Shouldn't you be somewhere else not interfering?	`RequestClarification`, `Sarcasm`
7	A	I never said not interfering. I do not think it is worth my time trying to change racists. It is not worth my time to prevent them from speaking. It is worth my time to signal to others that the person ideas are racist. I am not going to change David Duke and I can't get him to shut up, by shutting his ideas I can help send the message that we do not tolerate his ideals and if you espouse them we will not tolerate you.	`Clarification`
8	B	There's a time and a place for outright blockading of the views of others if they can't find a healthy way to express them, but there are absolutely opportunities to help individuals progress towards changing their beliefs and behaviors. If there weren't then essentially it would be hopeless for gang members, sex offenders, or other criminals to ever change. There might be professionals out there who facilitate the change process best, but if it's impacting our society we absolutely have a responsibility to try and change those behaviors if confronted with them, in addition to preventing others from modeling them.	`CounterArgument`, `Positive`, `Softening`
2	A	You don't have to eat these shared group meals. If you voluntarily choose of your own free will to cook and eat a group lunch for a specific group of people then of course the meal should be something everyone in the group can eat. Otherwise you aren't cooking for a group, you're cooking for a portion of the group	`CounterArgument`
3	B	"something everyone in the group *can* eat" Unless they're been bitten by a Lone Star Tick, there's nothing physically preventing them eating meat.	`NegTransformation`, `Ridicule`
4	A	Sure, if you want to be an abrasive aggressive asshole who doesn't exist within society but instead on the outskirts who hates people and is purposefully antagonist and rude, you could say that yes, it is physically possible for vegans to eat meat. But the rest of us here in society who aren't purposefully trying to be an asshole, understand that "what everyone can eat" isn't about literal physical ability.	`CounterArgument`, `Aggressive`
5	B	" 'what everyone can eat' isn't about literal physical ability". Then use your words: "the meal should be something everyone in the group ~can~*wants to* eat". "if you want to be an abrasive aggressive asshole" You're the one flinging insults man.	`CounterArgument`, `Complaint`

Table 2: Excerpts from two CMV discussions – a productive one (top) and an unproductive discussion (bottom).

`RequestClarification` is relatively likely to be followed by `Clarification`, which in turn is likely to be followed by `CounterArgument` or `CriticalQuestion`.

A careful examination of the collocations and the transition probabilities reveals intricate conversational dynamics, e.g., the surprising likelihood of `Extension` to follow `Convergence`, as if one discussant is not happy with the convergence of ideas and tries to keep the discussion going. The collocation and transition probabilities are leveraged as we model the discourse parsing task as a sequence model problem. The computational details are presented in the next section.

5 Computational Approach

5.1 Task Definition

Formal Definitions The discourse parsing task can be broadly defined in the following manner: find the sequence L that maximizes $P(L|T)$ where $L = l_1, ..., l_n$ denotes a sequence of labels, $l_i \in \mathcal{L}$, a predefined tagset. The sequence L corresponds to a sequence of texts/utterances $T = t_1, ..., t_n$.

We note that L actually encode a sequences of *sets* of labels ($l_i = \{l\}_1^k \subset \mathcal{L}$) as each utterance can be assigned more than one label (e.g. <*CounterArgument, Positive, Softening*> – its discursive move, tone, disagreement strategy, see Turn 8 (line 3) in Table 2, top).

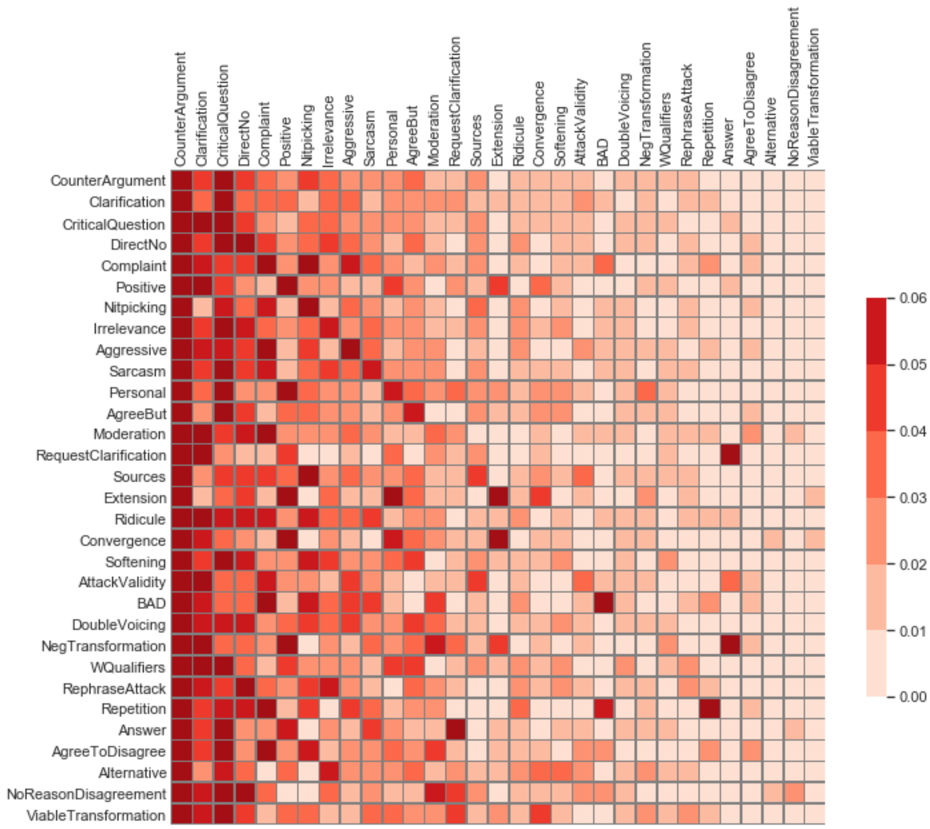

Figure 1: Tag Transition probabilities.

Variable	Value	STD.
Num of trees	100	–
Total users	1610	–
Total branches	1946	–
Total nodes	10559	–
Total labeled nodes	9620	–
Total labels for all nodes	17964	–
Total tokens	1143777	–
Avg. branches per tree	19.5	10.4
Avg. nodes per tree	105.6	60.8
Avg. branch length	8.3	7.2
Avg. tree depth	20.9	16.6
Avg. users per branch	3.3	1.1
Avg. users per tree	22.8	12.4
Avg. nodes per user	6.4	11.9
Avg. tokens per post (node)	108.3	139.9

Table 3: Dataset statistics.

Naive Classification The most straight forward approach is to apply supervised learning, using each textual unit (utterance) as the basic unit for label prediction. Utterances like "My view is that this is all nonsense", "but that's just my two cents" or "I disagree", intuitively suggest that discourse markers are part of the text and could be learned efficiently. In other words, we want to maximize $P(L|T)$ under the naive assumption that the order of the texts in T is insignificant: $P(L|T) = \Pi_{t_{i=1...|T|} \in T} P(l_i|t_i)$. A similar approach is considered by (Hirschberg and Litman(1993)).

Sequence Labeling: Dialogical acts tend to correspond, at least in a cooperative conversation (and under Gricean framework described in Section 3). Intuitively, a question is likely to draw an answer and an aggressive comment is likely to draw an aggressive response. This assumed conversational structure, supported by the collocation and transition probabilities presented in Figure **??** and 1, suggests looking at the annotation task as a structured prediction problem, in which labels are not assigned based on single textual units but depend on other labels in the sequence. The general formulation of the annotation-optimization task is to maximize L in $P(L|T) = \Pi_{t_{i=1...|T|} \in T} P(l_i|t_i, L^{-i})$, where L^{-i} denotes the labels predicted in the sequence except the i-th label.

5.2 Feature Types and Representation

In order to assign labels to texts we stack a number of binary classifiers - training a model for each label. Each classifier learns a model for a single label.

The input text is represented by a combination of features. Feature types are presented below and results are reported in Section 6.

Bag of Words (BOW) The most basic way to represent a piece of text is by the words it contains. We considered different dimensions for the BOW vectors: 300, 500, 1000 & 1500. We experimented with both a TF-IDF weighing (Salton and Buckley(1988)) and standard binary vectors of the most frequent words.

Embeddings: GloVe, Doc2vec Two types of embeddings were considered: a 300 dimension GloVe word representations (Pennington, Socher, and Manning(2014)), averaged for all the words in a post, and a another vector created by doc2vec (Le and Mikolov(2014)), instead of averaging.

LIWC Categories The Linguistic Inquiry and Word Counts (LIWC) dictionary is used to assign words to cognitive and emotion categories (Pennebaker, Francis, and Booth(2001)). We expect open-ended argumentative discourse to be emotionally charged. Utterances are represented by a vector of LIWC categories – each entry reflects the weight of the corresponding LIWC category in that text.

PDTB discourse relations While this work is focused on conversational structure, the discursive structure within a single document or utterance is also of interest (see Section 2.3). We hypothesize that the inner structure complements the conversational structure, therefore can serve as a powerful signal in modeling the latter. We use a Bi-LSTM finetuned over BERT as described in (Nie, Bennett, and Goodman(2019)) to label the inner structure (semantic relation between each pair of consecutive sentences) of each post with the PDTB2.0 tags (Prasad et al.(2008)Prasad, Dinesh, Lee, Miltsakaki, Robaldo, Joshi, and Webber). We then represent each post by two vectors: one of the counts of the PDTB tags found in the post, and the other of the counts of bigrams of PDTB tags. These vectors are used as an additional layer of features representing a post.

Conversational Sequence The main objective in the defined discourse parsing task is to model dialogicity. A text t_i is a direct response to a previous comment t_{i-1} and may also be effected (argumentatively, stylistically, emotionally) by previous comments t_{1-2}, t_{i-3}... made by one or more interlocutors. We therefore expect the text of previous comments to carry some relevant signal. We experiment with history length $\in \{0, 1, 2, 3\}$.

Label Sequence In essence, discourse parsing is a sequence prediction task as the labels of the current text are expected to be related to previous labels. Developing a sequence prediction of multi-class-multi tag is beyond the scope of this paper. However, in order to demonstrate the learnability of the proposed tagset we take the gold standard labels of previous texts as given. Labels of a previous text t' are represented as a binary vector in the dimension of the tagset. We experiment with using the labels of the 1,2 & 3 previous posts, as well as no previous labels at all.

Label Collocation The proposed annotation scheme allows a text to be labeled with more than a single label (e.g. labels for function, tone and argumentation strategy, see Section 3.1). Label collocation is, therefore, similar to the sequential nature of the task – a specific label may be more likely to appear with other labels. We consider the use of the collocated labels as yet another feature type.

Combining Feature Types Each feature type is a vector. Combining feature types is done by concatenating the respective vectors. For example, when using features of previous posts (either BOW, GloVe, LIWC or labels) the vectors of the texts are simply appended. That is, if $\overrightarrow{v_i}$ and $\overrightarrow{v_{i-1}}$ are the GloVe vectors representing texts t_i and t_{i-1} respectively, and we only use the text of the previous post t_{i-1} along with the text of the current post t_i for predicting the label/s of t_i, the feature vector representation of t_i is $\overrightarrow{v_i.v_{i-1}}$.

5.3 Classification Models

We experimented with the following classifiers: Logistic Regression, Naive Bayes, Decision Trees, SVM, Feed Forward Neural Network, and a BERT Transformer[5].

The input for all classifiers, except the BERT Transformer, was a feature vector composed of a concatenation of some or all the feature categories described above. The BERT transformer, on the other hand, takes a textual sequence as an input, thus the the feature vectors we used for other classifiers could not be used. The transformer was fed by the text of the current post. Since the input sequence of the BERT Transformer is limited to 512 tokens we opted for not using the text of preceding posts. Similarly, the text of the current post was truncated if it exceeded the allowed length. However, only $\sim$2.4% of the posts exceeded that length. We did manipulate the BERT input in order to introduce label sequence and collocation. This was done by appending the text of each post to "dummy" sentences listing the collocated and preceding tags in a simplified grammar[6].

We note that a bi-directional neural architecture could not be used due to the "online" definition of the task – annotating the discussion as it unfolds, rather than annotating a full discussion.

6 Results and Analysis

A set of 15 trees was set aside for testing. This test set is composed of 321 discussion branches with a total of 1835

[5]We used the sklearn package (Pedregosa et al.(2011)Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay) for the LR, SVM, NB and DT algorithms and Roberta (Liu et al.(2019)Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov) for the BERT transformer.

[6]Input structure example: 'previous tags: tag t_1 tag t_2 tag t_3, current post: <text of the current post>'. White spaces were added to tags composed of multiple words in order to have match the tag to the BERT dictionary (e.g., `CounterArgument` $\rightarrow$ 'counter argument').

Tag	Prec	Rec	F1-score	Prior
1. Promoting discussion				
CounterArgument	0.927	0.951	0.939	0.585
Clarification	0.872	0.768	0.817	0.100
RequestClarification	0.831	0.653	0.731	0.035
Extension	0.517	0.585	0.549	0.021
Answer	0.857	0.375	0.522	0.013
AttackValidity	0.500	0.520	0.510	0.026
Moderation	0.739	0.293	0.420	0.033
Personal	0.304	0.570	0.396	0.043
ViableTransformation	0.089	0.733	0.158	0.009
macro avg.	0.626	0.606	0.560	0.096
w.avg	**0.846**	**0.842**	**0.833**	**0.414**
2. Low responsiveness				
Convergence	0.632	0.511	0.565	0.026
NegTransformation	0.722	0.283	0.406	0.022
NoReasonDisagreement	0.455	0.357	0.400	0.010
AgreeToDisagree	0.131	0.423	0.200	0.013
Repetition	0.313	0.109	0.161	0.015
BAD	0.066	0.419	0.114	0.017
macro avg.	0.386	0.350	0.308	0.017
w. avg.	**0.429**	**0.362**	**0.335**	**0.019**
3. Tone & style				
Complaint	0.300	0.400	0.343	0.059
Positive	0.325	0.348	0.336	0.053
Aggressive	0.133	0.234	0.170	0.047
Sarcasm	0.123	0.243	0.164	0.044
WQualifiers	0.071	0.345	0.118	0.022
Ridicule	0.061	0.580	0.110	0.027
macro avg	0.169	0.358	0.207	0.042
w. avg	**0.198**	**0.345**	**0.233**	**0.046**
4. Disagreement strategies				
4.1 Easing tension				
Sources	0.884	0.884	0.884	0.042
Softening	0.330	0.446	0.379	0.053
DoubleVoicing	0.106	0.565	0.179	0.027
AgreeBut	0.056	0.867	0.106	0.024
macro avg	0.344	0.690	0.387	0.036
w. avg	**0.403**	**0.663**	**0.442**	**0.040**
4.2 Intensifying tension				
Nitpicking	0.730	0.860	0.790	0.056
CriticalQuestion	0.786	0.668	0.722	0.118
DirectNo	0.260	0.258	0.259	0.073
Irrelevance	0.116	0.333	0.172	0.054
Alternative	0.082	0.366	0.133	0.017
RephraseAttack	0.042	0.514	0.077	0.020
macro avg	0.336	0.500	0.359	0.056
w. avg	**0.476**	**0.533**	**0.477**	**0.077**
Total				
macro avg.	0.399	0.499	0.382	0.055
w.avg.	**0.614**	**0.663**	**0.610**	**0.237**

Table 4: Best results for each tag grouped by categories (each group ordered by F1-score), allowing a different model for each tag. Macro average given for reference.

T^C	T_2	L_1	B_1	PD_1^B	PD_2^U	**w.P**	**w.R**	**w.F1**	**m.F1**
				✓	✓	0.281	0.233	0.253	0.061
		✓				0.278	0.538	0.314	0.119
			✓			0.387	0.356	0.367	0.181
		✓	✓			0.397	0.360	0.374	0.183
		✓	✓	✓		0.388	0.353	0.366	0.181
		✓	✓	✓	✓	0.394	0.366	0.376	0.185
	✓					0.311	0.260	0.279	0.083
	✓	✓				0.306	0.495	0.348	0.146
	✓	✓		✓		0.324	0.443	0.354	0.157
	✓	✓		✓	✓	0.330	0.412	0.354	0.153
	✓	✓	✓			0.404	0.381	0.389	0.190
	✓	✓	✓		✓	0.414	0.390	0.398	0.200
	✓	✓	✓	✓		0.405	0.375	0.387	0.186
	✓	✓	✓	✓	✓	0.416	0.390	0.398	0.195
✓	✓					0.474	0.438	0.452	0.235
✓	✓	✓				0.439	**0.568**	0.474	0.265
✓	✓	✓			✓	0.454	0.536	0.476	0.273
✓	✓	✓		✓	✓	0.473	0.529	0.488	0.275
✓	✓	✓	✓			0.549	0.500	0.517	0.292
✓	✓	✓	✓		✓	0.553	0.501	0.519	0.287
✓	✓	✓	✓	✓	✓	0.551	0.499	0.518	0.283
✓	✓	✓	✓	✓		**0.555**	0.509	**0.526**	**0.297**

Table 5: Results of ablation test for average (weighted) over all labels. Macro F1 given for reference. Notation: B: BOW; L: LIWC; PD: PDTB labels; PD^B: bigrams of PDTB labels; Subscript index denotes the context length, *including* current post. T^C: tag collocation of current post. Note: T_j: context length, current post *not* included (note the exception).

nodes annotated with 3412 labels. Training was done in a 5-fold cross validation manner on the other 85 trees. This procedure was repeated in multiple settings differing in the classification model and combinations of features types used. For the sake of brevity we report the best result for each tag and category (Table 4) and ablation results for the single model that performed best averaged over all tags (Table 5). All results are reported on the 15 trees in the held-out set.

Model per tag: We achieved a weighted F-score for all 31 tags is 0.61. The F-score for each individual tags varies from 0.939 (`CounterArgument`, prior 0.585) to 0.077 (`RephraseAttack`, prior 0.02). The average F-score for the first category (Promoting Discussion) is 0.833, significantly higher than the other categories. While `CounterArgument` has the highest prior (0.585), the high F-score for the first category cannot be attributed to higher priors as tags with low priors ranging from 0.1 to 0.013 achieved impressive F-scores ranging from 0.817 to 0.522 (e.g., `Clarification`, `Extension`, and `Answer`). A significant improvement over the prior was achieved for all the tags.

Analysing the performance of the different classifiers we find that the BERT classifier, adopted for label sequence and collocation, achieved nine of the top results in Table 4, including the impressive F-scores for `CounterArgument`, `Sources`,`Clarification`, `Nitpicking`, and `RequestClarification`, most of which fall under Category 1 (Promoting Discussion). Surprisingly though, the BERT classifier performed poorly on all other tags. Further analysis of the variance of the BERT performance is beyond the scope of this paper and will be addressed in future work. Top results for the other tags were obtained using a different classifier and feature set for each tag. Each of the feature types listed in Section 5.2 proved useful in achieving top results for at least one of the tags, with BOW, LIWC, label sequence and collocation being the most stable features across labels and classifiers.

Single model for all tags: Allowing only a single classification model for all tags, we achieve a weighted F-score of 0.526, using a feed-forward network with three hidden layers, using a tanh activation and L2 regularization with min-max scaling. Nevertheless, while using a single model results in a significant drop of the weighted F-score, it is still impressive, given the number of classes (31) and the complexity of the task. Using a single model allows us to explore the contribution of the different feature types to the parsing task. Some ablation results are reported in Table 5. For example, using only the text (BOW) of the current post (B_1) achieves an F-score of 0.367, while using both the text and the LIWC features of the current post (B_1 and L_1) yields a small improvement to 0.374. Top results for a single classifier were achieved by using a feature vector combining the BOW (B_1), LIWC (L_1), PDTB discourse bi-grams (PD_1^B), collocation of tags in the current post (T^C), and the tag sequence of the two preceding posts (T_2).

Noisy sequences and collocations: In order to simulate a sub-optimal prediction of the preceding sequence we performed a set of experiments in which we introduce noise to the input by masking or randomly replacing 10%, 20% or 50% of the collocated and preceding tags, or by adding tags at random (by priors). An F-score of 0.463 was achieved upon a random substitution of 50% of the tags – the largest drop observed. Masking only 10% of the tags achieved an F-score of 0.523, and randomly replacing 10% of the tags resulted in an F-score of 0.518, comparing to the 0.526 achieved without distorting the input.

The results and analysis presented here demonstrate the learnability of the complex and nuanced schema we proposed. Moreover, it supports the theoretical framework of Dialogic Agency, IPD, and Responsiveness, presented in Section 2, thus promoting further research of conversational dynamics.

7 Conclusion

Many online discussions are non-convergent. We have argued that the productivity of a discussion should not be measured through convergence and that a new discourse annotation scheme is required in order to capture the productivity of online discussions. Inspired by theory of dialogism, we have proposed a new discourse annotation schema that can be used for early detection of the trajectories a discussion takes, and ultimately improve discussion quality. We have explored a number of classification algorithms and an array of feature types, demonstrating the learnability of the proposed schema. Making the annotated data public will allow the research community to further explore conversational dynamics – a much needed undertaking in this era of highly polarized contentious discourse.

References

Angouri, J.; and Tseliga, T. 2010. "you HAVE NO IDEA WHAT YOU ARE TALKING ABOUT!" From e-disagreement to e-impoliteness in two online fora. *Journal of Politeness Research* 6(1): 57–82.

Bakhtin, M. M. 1981. The dialogic imagination: Four essays, ed. *Michael Holquist, trans. Caryl Emerson and Michael Holquist (Austin: University of Texas Press, 1981)* 84(8).

Bakhtin, M. M. 1986. Speech genres and other late essays (VW McGee, Trans.; C. Emerson & M. Holquist, Eds.). *Austin, TX: University of Texas Press.(Original work published 1979)* .

Bar-Haim, R.; Bhattacharya, I.; Dinuzzo, F.; Saha, A.; and Slonim, N. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 251–261.

Barron, B. 2003. When smart groups fail. *The journal of the learning sciences* 12(3): 307–359.

Barzilay, R.; and Lee, L. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Bhuiyan, M.; Misra, A.; Tripathy, S.; Mahmud, J.; and Akkiraju, R. 2018. Don't get Lost in Negation: An Effective Negation Handled Dialogue Acts Prediction Algorithm for Twitter Customer Service Conversations. *arXiv preprint arXiv:1807.06107* .

Chen, B.; Chang, Y.-H.; Ouyang, F.; and Zhou, W. 2018. Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education* 37: 21–30.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1217–1230.

Chiu, M. M. 2000. Group Problem-Solving Processes: Social Interactions andIndividual Actions. *Journal for the theory of social behaviour* 30(1): 26–49.

Core, M. G.; and Allen, J. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA.

Cougnon, L.-A.; Coppin, J.; and Figueroa, V. G. 2019. A Mixed Quantitative-Qualitative Approach to Disagreement in Online News Comments on Social Networking

Sites. *Social Media Corpora for the Humanities (CMC-Corpora2019)* 32.

Diziol, D.; Walker, E.; Rummel, N.; and Koedinger, K. R. 2010. Using intelligent tutor technology to implement adaptive support for student collaboration. *Educational Psychology Review* 22(1): 89–102.

Erkens, G.; and Janssen, J. 2008. Automatic coding of dialogue acts in collaboration protocols. *International journal of computer-supported collaborative learning* 3(4): 447–470.

Felton, M.; Garcia-Mila, M.; and Gilabert, S. 2009. Deliberation versus dispute: The impact of argumentative discourse goals on learning and reasoning in the science classroom. *Informal Logic* 29(4): 417–446.

Grice, H. P. 1975. Logic and conversation. *1975* 41–58.

Grimes, J. E. 1975. *The thread of discourse*, volume 207. Walter de Gruyter.

Gunawardena, C. N.; Lowe, C. A.; and Anderson, T. 1997. Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of educational computing research* 17(4): 397–431.

Hennessy, S.; Rojas-Drummond, S.; Higham, R.; Márquez, A. M.; Maine, F.; Ríos, R. M.; García-Carrión, R.; Torreblanca, O.; and Barrera, M. J. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction* 9: 16–44.

Hirschberg, J.; and Litman, D. 1993. Empirical studies on the disambiguation of cue phrases. *Computational linguistics* 19(3): 501–530.

HOBBS, J. 1985. On the Coherence and Structure of Discourse. *CSLI* 85(37).

Ji, Y.; Haffari, G.; and Eisenstein, J. 2016. A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. In *Proceedings of NAACL-HLT*, 332–342.

Jurafsky, D.; Shriberg, E.; and Biasca, D. 1997. Switchboard-damsl labeling project coder's manual. *Technická Zpráva* 97–02.

Khanpour, H.; Guntakandla, N.; and Nielsen, R. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2012–2021.

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)*, 70–75.

Le, Q.; and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Locher, M. A. 2010. *Power and politeness in action: Disagreements in oral communication*, volume 12. Walter de Gruyter.

Lu, J.; Chiu, M. M.; and Law, N. W. 2011. Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. *Computers in Human Behavior* 27(2): 946–955.

Matusov, E.; and von Duyke, K. 2009. Bakhtin's notion of the internally persuasive discourse in education: Internal to what? *Perspectives and limits of dialogism in Mikhail Bakhtin* 174.

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.

Nie, A.; Bennett, E.; and Goodman, N. 2019. DisSent: Learning Sentence Representations from Explicit Discourse Relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4497–4510. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1442. URL https://www.aclweb.org/anthology/P19-1442.

Noroozi, O.; Kirschner, P. A.; Biemans, H. J.; and Mulder, M. 2018. Promoting argumentation competence: Extending from first-to second-order scaffolding through adaptive fading. *Educational Psychology Review* 30(1): 153–176.

Oraby, S.; Gundecha, P.; Mahmud, J.; Bhuiyan, M.; and Akkiraju, R. 2017. How May I Help You?: Modeling Twitter Customer ServiceConversations Using Fine-Grained Dialogue Acts. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 343–355. ACM.

Parker, W. C. 2006. Public discourses in schools: Purposes, problems, possibilities. *Educational Researcher* 35(8): 11–18.

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.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001): 2001.

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)*, 1532–1543.

Pinkwart, N.; Ashley, K.; Lynch, C.; and Aleven, V. 2009. Evaluating an intelligent tutoring system for making legal arguments with hypotheticals. *International Journal of Artificial Intelligence in Education* 19(4): 401–424.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. K.; and Webber, B. L. 2008. The Penn Discourse TreeBank 2.0. In *LREC*.

Rosé, C.; Wang, Y.-C.; Cui, Y.; Arguello, J.; Stegmann, K.; Weinberger, A.; and Fischer, F. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning* 3(3): 237–271.

Salton, G.; and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24(5): 513–523.

Schwarz, B. B.; Neuman, Y.; and Biezuner, S. 2000. Two wrongs may make a right... if they argue together! *Cognition and instruction* 18(4): 461–494.

Schwarz, B. B.; Prusak, N.; Swidan, O.; Livny, A.; Gal, K.; and Segal, A. 2018. Orchestrating the emergence of conceptual learning: A case study in a geometry class. *International Journal of Computer-Supported Collaborative Learning* 13: 189–211.

Shum, W.; and Lee, C. 2013. (Im) politeness and disagreement in two Hong Kong Internet discussion forums. *Journal of Pragmatics* 50(1): 52–83.

Stolcke, A.; Ries, K.; Coccaro, N.; Shriberg, E.; Bates, R.; Jurafsky, D.; Taylor, P.; Martin, R.; Ess-Dykema, C. V.; and Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26(3): 339–373.

Suthers, D. D. 2003. Representational guidance for collaborative inquiry. In *Arguing to learn*, 27–46. Springer.

Teasley, S.; Fischer, F.; Dillenbourg, P.; Kapur, M.; Chi, M.; Weinberger, A.; and Stegmann, K. 2008. Cognitive convergence in collaborative learning .

Trausan-Matu, S.; Dascalu, M.; and Rebedea, T. 2014. PolyCAFe – automatic support for the polyphonic analysis of CSCL chats. *International Journal of Computer-Supported Collaborative Learning* 9(2): 127–156.

Wise, A. F.; and Chiu, M. M. 2011. Analyzing temporal patterns of knowledge construction in a role-based online discussion. *International Journal of Computer-Supported Collaborative Learning* 6(3): 445–470.

Yoon, S. A. 2011. Using social network graphs as visualization tools to influence peer selection decision-making strategies to access information about complex socioscientific issues. *Journal of the Learning Sciences* 20(4): 549–588.

Zhang, A.; Culbertson, B.; and Paritosh, P. 2017. Characterizing online discussion using coarse discourse sequences .

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361.