12-2020

Argumentation Stance Polarity and Intensity Prediction and its Application for Argumentation Polarization Modeling and Diverse Social Connection Recommendation

Joseph Winstead Sirrianni
University of Arkansas, Fayetteville

Argumentation Stance Polarity and Intensity Prediction and its Application for Argumentation Polarization Modeling and Diverse Social Connection Recommendation.

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Computer Science

by

Joseph W Sirrianni
University of Arkansas
Bachelor of Science in Computer Science, 2016

December 2020
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Xiaoqing "Frank" Liu, Ph.D.
Committee Co-Chair

Brajendra Nath Panda, Ph.D.
Committee Co-Chair

Douglas Adams, Ph.D.
Committee Member

Xintao Wu, Ph.D.
Committee Member

Susan E. Gauch, Ph.D.
Committee Member

Abstract

Cyber argumentation platforms implement theoretical argumentation structures that promote higher quality argumentation and allow for informative analysis of the discussions. Dr. Liu's research group has designed and implemented a unique platform called the Intelligent Cyber Argumentation System (ICAS). ICAS structures its discussions into a weighted cyber argumentation graph, which describes the relationships between the different users, their posts in a discussion, the discussion topic, and the various subtopics in a discussion. This platform is unique as it encodes online discussions into weighted cyber argumentation graphs based on the user's stances toward one another's arguments and ideas. The resulting weighted cyber argumentation graphs can then be used by various analytical models to measure aspects of the discussion. In prior work, many aspects of cyber argumentation have been modeled and analyzed using these stance relationships.

However, many challenging problems remain in cyber argumentation. In this dissertation, I address three of these problems: 1) modeling and measure argumentation polarization in cyber argumentation discussions, 2) encouraging diverse social networks and preventing echo chambers by injecting ideological diversity into social connection recommendations, and 3) developing a predictive model to predict the stance polarity and intensity relationships between posts in online discussions, allowing discussions from outside of the ICAS platform to be encoded as weighted cyber argumentation graphs and be analyzed by the cyber argumentation models. In this dissertation, I present models to measure polarization in online argumentation discussions, prevent polarizing echo-chambers and diversifying users' social networks ideologically, and allow online discussions from outside of the ICAS environment to be analyzed using the previous models from this dissertation and the prior work from various researchers on the ICAS system.

This work serves to progress the field of cyber argumentation by introducing a new analytical model for measuring argumentation polarization and developing a novel method of encouraging ideological diversity into social connection recommendations. The argumentation polarization model is the first of its kind to look specifically at the polarization among the users contained within a single discussion in cyber argumentation. Likewise, the diversity enhanced social connection recommendation re-ranking method is also the first of its kind to introduce ideological diversity into social connections. The former model will allow stakeholders and moderators to monitor and respond to argumentation polarization detected in online discussions in cyber argumentation. The latter method will help prevent network-level social polarization by encouraging social connections among users who differ in terms of ideological beliefs. This work also serves as an initial step to expanding cyber argumentation research into the broader online deliberation field. The stance polarity and intensity prediction model presented in this dissertation is the first step in allowing discussions from various online platforms to be encoded into weighted cyber argumentation graphs by predicting the stance weights between users' posts. These resulting predicted weighted cyber augmentation graphs could then be used to apply cyber argumentation models and methods to these online discussions from popular online discussion platforms, such as Twitter and Reddit, opening many new possibilities for cyber argumentation research in the future.

Acknowledgements

I would like to thank my doctoral advisor Dr. Xiaoqing “Frank” Liu for his leadership and guidance throughout my graduate studies. He provided many insights, suggestions, and assistance throughout my studies and was an invaluable asset to this dissertation.

I would also like to thank Dr. Douglas Adams for the many discussions we shared about this research. He brought valuable outside perspective on our research projects that led to their success.

I would also like to thank my co-workers and lab-mates, particularly my fellow graduate students that worked on the ICAS project, Md Mahfuzer Rahman Limon and Najla Althuniyan. I am also thankful to senior co-workers, now Ph.D.’s, Md Rakib Shahriar and S M Nahian Al Sunny for their research consultation.

Lastly, I would like to thank Dr. Brajendra Nath Panda for agreeing to become my co-chair and co-advisor for my last semester. Without him agreeing to take on those roles, I may not have made it this far. I would also like to thank my committee members Dr. Susan Gauch and Dr. Xintao Wu for their valuable input and suggestions about the research in this dissertation. Thank you all so much!

Table of Contents

Table of Figures

Chapter 1: Introduction

Research in cyber argumentation aims to develop online discussion platforms that are designed to facilitate and analyze online discussion and debate better than more conventional platforms by providing formal argumentation structures to their systems. Cyber argumentation platforms are designed to promote effective deliberation. Many platforms have focused on improving discussion and debate quality by providing discussion visualization or teaching productive deliberation behaviors to their users [see [1, 2, 3, 4, 5] for examples], while other system focus on measure the productive outcomes of discussions using built-in analytics [6,7,8].

Dr. Liu's research group has been researching and developing a crowdsourcing, large-scale cyber argumentation platform for over a decade. The current iteration of the system is called the Intelligent Cyber Argumentation System (ICAS). The goal of ICAS is to effectively facilitate high-quality massive online discussions and provide built-in analytical models for assessing the outcomes of the discussions. ICAS implements an argumentation framework to structure online discussions to capture decision rational effectively [9, 10, 11, 12], provide decision making support [13, 14, 15, 16], and analysis of the discussion and individual contribution [17]. In ICAS, users discuss topics (called issues) and subtopics (called positions) by posting arguments where they defend or attack the positions or other arguments. The key innovation of ICAS is that the platform allows users to explicitly express partial agreement or disagreement with other participant's arguments or ideas [8, 9]. ICAS allows its users to explicitly express both the polarity (Supporting, Opposing, or Neutral) and intensity of their opinion or stance toward an argument, idea, or opinion. This expression of partial agreement/disagreement is encoded as a floating-point number between -1.0 and +1.0, which we call an "agreement value," where the sign indicates the stance polarity (negative is opposing, positive is supporting, and 0 is neutral) and the magnitude indicates the

intensity of the stance. With this stance information, ICAS structures each discussion in a weighted cyber argumentation graph, which encodes the relationships between users, issues, positions, and arguments with the stance information. These weighted cyber argumentation graphs allow ICAS to perform uniquely specific opinion analysis on the discourse data. Prior work with the platform has developed a fuzzy logic reduction engine that can approximate each user's opinion toward each position they discussed based on the stance information associated with each argument they have posted [8, 14]. These approximated opinions serve as input to several downstream analytical user opinion models. Prior research by Dr. Liu's research group has demonstrated that the partial stance information and the weighted cyber argumentation graphs in ICAS can be used to analyze collective opinion [8], the credibility of arguments [18], detect conflicting opinions [16], identify outlier opinions [19], opinion factions [20], and predict user opinion on discussions which they have yet to participate [21].

However, even with all these previously mentioned analytical models and methods, many challenging problems remain in cyber argumentation. In this dissertation, I address three of these problems: 1) modeling and measuring argumentation polarization in cyber argumentation discussions, 2) preventing online echo chambers and increasing ideological diversity among users' social networks by injecting ideological diversity into social connection recommendations, and 3) developing a predictive model to predict the stance polarity and intensity relationships between two posts in an online discussion with a reply-to relationship. These predicted relationships will allow discussions from outside the ICAS platform to be encoded as weighted cyber argumentation graphs and enable them to be analyzed by the cyber argumentation models. The framework for this dissertation is presented in Figure 1-1. Discussions in ICAS are encoded as weighted argumentation graphs. These argumentation graphs are then fed as input into the fuzzy logic engine

[8, 14], which derives each user's opinion on each of the discussed positions into user opinion vectors. These user opinion vectors serve as input into this dissertation's first research output, the argumentation polarization model.

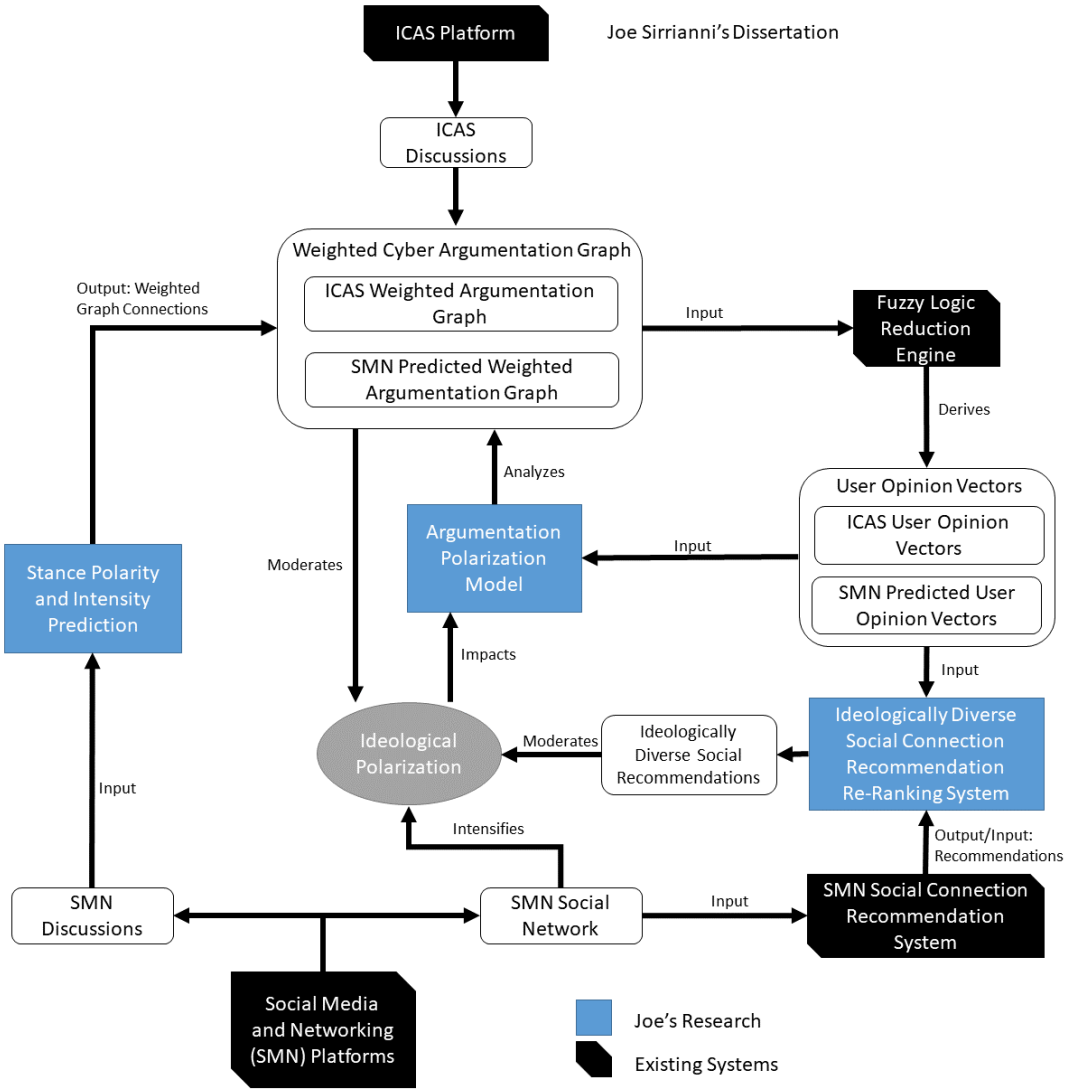


Figure 1-1 Disseration Framework

The argumentation polarization model measures how polarized users in each discussion are in terms of their stance toward the discussion topic. According to argumentation theory, online discussions ought to moderate ideological polarization [22, 23]. However, polarization may instead increase as a result of discussions, depending on the deliberation quality of the discussion

[24, 25, 26, 27]. Thus, there is a need to monitor the level of polarization present in a discussion at a given time. Prior work for measuring polarization in a cyber argumentation context used fuzzy clustering to group users based on their opinions across multiple different discussions and measured the distances between the clusters [28, 29]. However, this method does not quantify the level of polarization, nor can it be used to measure the development of argumentation polarization over time. Thus, there is a need for an argumentation polarization model that can quantify the level of polarization present in an online discussion at any given time in the discussion. Such a model would allow stakeholders and moderators to track the development and evolution of argumentation polarization in a cyber argumentation discussion. In the first task of this dissertation, I present an argumentation polarization model that measures the total amount of polarization present in a discussion in cyber argumentation using the weighted cyber argumentation graph encoding of a discussion.

In addition to argumentation polarization present in online discussions, there is also interest in reducing polarization caused by social networks by discouraging the formation of online echo chambers. In many social networking environments, users self-sort into ideological clusters that typically leads to isolation from diverse viewpoints [30] and leads to extreme opinions [31]. Recent updates to ICAS have included social networking features, which should encourage participation in the system, but also invite the possibility of echo chambers forming in the social network. To prevent these online echo chambers, I introduce a novel diversity enhanced social connection recommendation re-ranking system to promote ideological diversity into social connection recommendations. This diversity enhancing system re-ranks the output recommendations from a native social-connection recommendation system to prioritize users with differing opinions from the subject user. The system uses the user opinion vectors derived from the weighted

argumentation graphs output by discussions in ICAS (or predicted using the stance polarity and intensity prediction model) to determine the difference in opinion to re-rank the recommendations. Thus, the resulting re-ranked recommendations encourage users to make social connections whose opinions differ with the user's opinions, which will result in a more diverse local social network.

The previously described models, both those described in prior work and this dissertation, serve to analyze online discussions for various cyber argumentation phenomena. However, for many of these models to operate, they must have discussion data encoded as weighted cyber argumentation graphs, which is currently only implemented in the ICAS platform. There is interest in applying cyber argumentation models to online discussions outside of individual cyber argumentation platforms, such as popular social media and networking platforms such as Twitter and Reddit. However, without the necessary data to encode their discussions into cyber argumentation graphs, applying cyber argumentation models is not possible. Klein et al. (2017) [32] proposed a method for encoding online discussion in Reddit by hand into weighted social network graphs that are similar in design to ICAS's weighted cyber argumentation graph. However, they used an arbitrary weighting scheme to weight their network connections based on the users' argumentation behavior, and they had to manually annotate the weighted connections, which does not scale to larger datasets. In this task, I propose a stance polarity and intensity prediction model to predict the stance relationship between two posts with reply-to relationships based on their textual information. This model can predict the stance relationships between online posts, and the predicted values can be used to weight the connections between the arguments to encode the discussions as weighted argumentation graphs. Thus, this model will allow online deliberation data outside of the ICAS platform to be encoded as weighted argumentation graphs, which can then be further processed for argumentation polarization or other analysis.

Each of the three presented models in this dissertation, the argumentation polarization model, the diversity enhanced social connection recommendation re-ranking system, and the stance polarity and intensity prediction model, were all developed and evaluating used empirical data collected using the ICAS system. Over three years, our research group has gathered empirical data from three different empirical studies and have generated a combined dataset of over 22,000 arguments from over 900 participants. This dataset, described in Chapter 2, serves as the testing and training data for the models presented in this dissertation.

In this chapter, I have presented the introduction of the research dissertation. Section 1.1 provides background information by describing the current ICAS system, which is used for data collection in our empirical studies and serves as the focal point of this research dissertation. Chapter 2 describes the three large scale empirical studies conducted from Fall 2017 to Spring 2019 that is used to train and validate the models presented in this dissertation. These studies collected the data that serves as the basis for all the research presented in this dissertation. Beginning in Chapter 3, I present my novel contributions in this dissertation. These works have been published or are being reviewed for publication in peer-reviewed conferences or journals. Chapter 3 presents our work in developing a model for analyzing the argumentation polarization in large scale online deliberation; that work was accepted for publication in IEEE Transactions on Computational Social Systems, and is an extension of a conference paper published in the 2018 IEEE International Conference on Cognitive Computing (ICCC). Chapter 4 presents our work in developing a social recommendation re-ranking method for encouraging opinion diversity among social recommendations in social network-enabled online deliberation platforms, that work was published in 2019 IEEE International Conference on Cognitive Computing (ICCC). Chapters 5 and 6 presents my work in developing a stance polarity and intensity prediction model that can

predict the agreement value associated with a post from its text. Chapter 5 presents our work of introducing the stance polarity and intensity prediction problem and adapting prior state-of-the-art stance detection models for our introduced problem. This work was published in the 2020 annual conference of the Association for Computational Linguistics. Chapter 6 presents our work fine-tuning the BERT [33] language understanding model for stance polarity and intensity prediction and has been submitted for review to IEEE Transaction on Computational Social Systems.

The contributions expected from this research are as follows:

- The development of a novel argumentation polarization model for discussions in cyber argumentation. This model is the first designed specifically for argumentation polarization and reflects both intragroup cohesion and intergroup heterogeneity.
- The development of an opinion diversity enhanced social connection re-ranking method to promote opinion diversity in social recommendations on social networking enabled platforms.
- The development of a stance polarity and intensity prediction model to predict the agreement values associated with unlabeled text posts in online discussions outside of the ICAS environment. This model will enable the utilization of the technologies associated with ICAS's agreement values outside of the ICAS platform.
- Application of the stance polarity and intensity prediction model to encode online discussion data from outside the ICAS platform into a weighted cyber argumentation graph. The resulting graph will be used for applying the argumentation polarization model for polarization analysis.

1.1 Background

This section will provide background information relating to the ICAS platform, that have been developed in prior work.

1.1.1 ICAS Platform

The Intelligent Cyber Argumentation System (ICAS), is the core platform on which the research in this dissertation centers. ICAS is designed to facilitate large scale argumentation among many users. The platform is an updated version of the argumentation system first developed by Dr. Liu's research team over many interactions [8-20]. The current version, ICAS, has a modern user interface along with many additional features, including social networking capabilities.

ICAS, like many cyber argumentation systems, seeks to improve online discourse compared to popular online and social media and networking platforms where the bulk of public discourse is currently located. However, unlike other cyber argumentation systems, ICAS seeks to encourage participation using social networking capabilities. The key features of ICAS are as follows:

- Our platform is highly structured in a way that organizes discussions by issue and idea, instead of by time or social connection. This structure prevents discussions from becoming fragmented, as all related content is grouped in the same place.
- Our platform allows users to express partial agreement/disagreement toward other's arguments and ideas. Unlike other platforms, which usually only allow full approval (likes in Facebook, upvotes in Reddit, etc.) or disapproval (dislike in YouTube, downvote in Reddit, etc.), our platform allows users to express their

opinion in a more nuanced way. Allowing partial agreement/disagreement gives a more accurate picture of each user's true feelings in a discussion and assists in evaluating a user's contribution in a discussion.

- Our platform has an artificial intelligence-based backend that contains reasoning and analytic models to analyze large-scale discourse. These models can offer users a perspective on what is occurring in these discussions.
- Our platform contains many social networking features that encourage user participation.

1.1.2 ICAS Design

ICAS is designed as an issue centered deliberation platform, meaning that all discussions in ICAS relate to a specific issue. ICAS employs the IBIS framework [33] for structuring its discussions. In ICAS, deliberation is structured in discussion trees. At the root of a discussion tree is the topic issue. Directly under the root issue, there are one or many positions. A position is a solution, stance, or idea that addresses or resolves the parent issue. Under each position is a discussion subtree, where all of the discussion and debate takes place. In ICAS, users argue for or against positions that address the root issue by posting arguments in the system. An argument can be posted under its parent position or another argument.

In the system, arguments have two components: the argument text and a level of agreement. The argument text is a description of the user's argument and rational. The level of agreement is an author selected label, which indicates the user's opinion stance toward the position or argument they are addressing. ICAS, unlike other deliberation platforms, allows users to express partial agreement or disagreement with other ideas using the agreement values in their arguments. The level of agreement selected by the user for their argument is a continuous value from the range -

1.0 to +1.0, where the sign of the value (positive or negative) indicates the author's stance (agreeing or disagreeing) with the parent post and the magnitude of the value (0 to 1.0) indicates the intensity of the stance. For example, a value of +1.0 would represent complete agreement, while an agreement value of +0.4 would represent only moderate agreement. The user selects the agreement value using a sliding bar interface, where they select a value at 0.2 intervals that correspond to different semantic descriptions, such as "Completely Agree", "Strongly Agree", "Moderately Agree", etc. These agreement values provide explicit argumentation relationships between arguments that are visible to the reader in the system. Figure 1-2 shows an example structure of the ICAS discussion structure, and a screenshot of the actual ICAS user interface is shown in Figure 1-3.

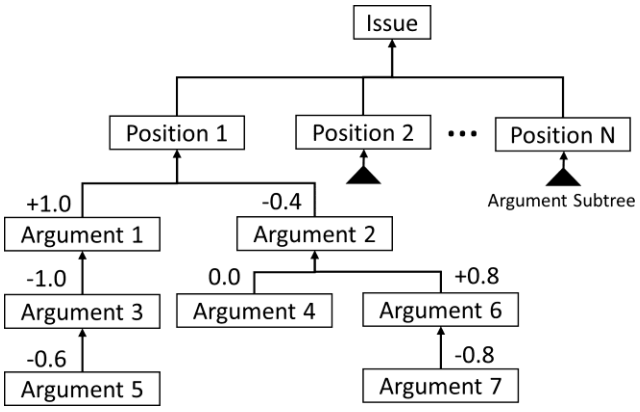


Figure 1-2 Example of the discussion structure in ICAS.

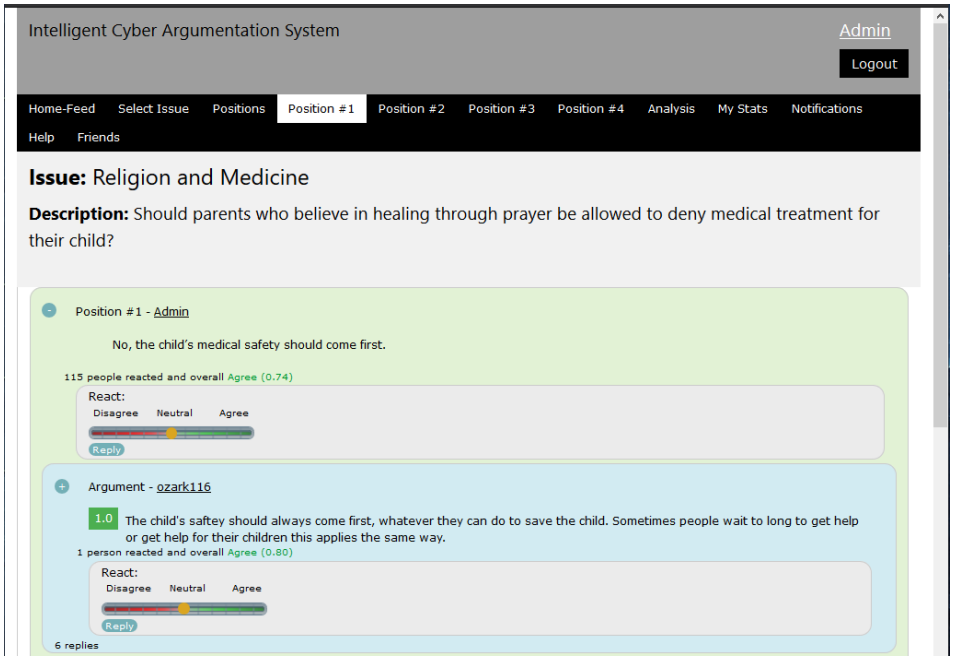


Figure 1-3 A screen shot of the ICAS system. An Issue (top), Position (middle) and Argument (bottom) are all displayed in a cascading format.

1.1.3 Fuzzy Logic Reduction Engine

This section briefly outlines the fuzzy logic reduction engine integrated into ICAS. The fuzzy logic engine was developed by other researchers in Dr. Liu's research group [8-16]. The fuzzy logic reduction engine approximates the overall opinion of each argument toward their parent position by reducing their associated agreement values from representing the relationship with their parent node to representing the argument's relationships with the root position. Using these reduced agreement values, the overall opinion expressed in the discussion of the position can be evaluated, along with the opinions of each participant toward the position. These reduced values serve as the core data processing step for many analytical models built around ICAS, including the two models described in this dissertation.

The argument reduction method uses fuzzy logic and 25 inference rules to reduce an argument's agreement value from any level of the argument tree to relate to the parent position. The reduction process works as follows:

- Given an argument A, with parent argument B, and grand-parent C, where VA and VB are the agreement values of A and B respectively, the method will update the value of VA to V'A reflect the relationship between A and C, instead of A and B. First, the inference rules identify the logical relationships between A and C. The relationship describes how A addresses C in the argumentation. For example, if A attacks B, and B attacks C, then A indirectly supports C. While these relationships are not always certain, they offer a useful heuristic for determining the relationship between A and C.
- Based on the relationships identified between A and C, VA's sign (agree or disagree) is updated. Then the value of VA is updated to reflect the partial agreement relationship between B and C using trapezoidal fuzzy logic rules between VA and VB. After the value of VA is updated to VA', we assign this value to argumentation relationships between A and C, effectively reducing A down the tree one level.
- Steps 1 and 2 are repeated until A is reduced to relate to the parent position. The final value of VA is used as A's relationship toward the position.

Figure 3-1 shows an example of an argument reduction in the system. For a more in-depth explanation of the fuzzy logic argument reduction method, please refer to [8, 10, 11, 12, 13]. While the fuzzy logic reduction system offers an estimation of an argument's agreement towards the root

position, several case studies have shown that this method achieves reasonable accuracy [10, 11, 15].

1.2 References

[1] S. Vesic, M. Ianchuk, and A. Rubtsov, “The Synergy: A Platform for Argumentation-Based Group Decision Making,” presented at the COMMA, 2012, pp. 501–502.

[2] C. Reed and G. Rowe, “Araucaria: software for argument analysis, diagramming and representation,” Int. J. Artif. Intell. Tools, vol. 13, no. 04, pp. 961–979, Dec. 2004, doi: 10.1142/S0218213004001922.

[3] S. Shum, “The Roots of Computer Supported Argument Visualization,” in Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, London: Springer-Verlag, 2003, pp. 3–24.

[4] C.-Y. Tsai, C.-N. Lin, W.-L. Shih, and P.-L. Wu, “The effect of online argumentation upon students’ pseudoscientific beliefs,” Computers & Education, vol. 80, pp. 187–197, Jan. 2015, doi: 10.1016/j.compedu.2014.08.018.

[5] T. Krauthoff, C. Meter, M. Baurmann, G. Betz, and M. Mauve, “D-BAS - A Dialog-Based Online Argumentation System,” Frontiers in Artificial Intelligence and Applications, pp. 325–336, 2018, doi: 10.3233/978-1-61499-906-5-325.

[6] M. Klein, “How to Harvest Collective Wisdom on Complex Problems : An Introduction to the MIT Deliberatorium,” 2011.

[7] A. Gürkan, L. Iandoli, M. Klein, and G. Zollo, “Mediating debate through on-line large-scale argumentation: Evidence from the field,” Information Sciences, vol. 180, no. 19, pp. 3686–3702, Oct. 2010, doi: 10.1016/j.ins.2010.06.011.

[8] X. (Frank) Liu, E. Khudkhudia, L. Wen, V. Sajja, and M. C. Leu, “An Intelligent Computational Argumentation System for Supporting Collaborative Software Development Decision Making,” in Artificial Intelligence Applications for Improved Software Engineering Development: New Prospects, F. Meziane and S. Vadera, Eds. IGI Global, 2010, pp. 167–180.S. Sigman and X. F. Liu, “A computational argumentation methodology for capturing and analyzing design rationale arising from multiple perspectives,” Information and Software Technology, vol. 45, no. 3, pp. 113–122, Mar. 2003, doi: 10.1016/S0950-5849(02)00187-8.

[9] R. S. Arvapally and X. (Frank) Liu, “Empirical Evaluation of Intellligent Argumentation System for Collaborative Software Project Decision Making,” in 5th Annual ISC Research Symposium, Rolla, Missouri, Apr. 2011, p. 6.

[10] X. Liu, R. Wanchoo, and R. S. Arvapally, “Empirical study of an intelligent argumentation system in MCDM,” in 2011 International Conference on Collaboration Technologies and Systems (CTS), May 2011, pp. 125–133, doi: 10.1109/CTS.2011.5928674.

[11] N. Chanda and X. F. Liu, “Intelligent analysis of software architecture rationale for collaborative software design,” in 2015 International Conference on Collaboration Technologies and Systems (CTS), Jun. 2015, pp. 287–294, doi: 10.1109/CTS.2015.7210436.

[12] X. F. Liu, S. Raorane, Man Zheng, and Ming Leu, “An Internet Based Intelligent Argumentation System for Collaborative Engineering Design,” in International Symposium on Collaborative Technologies and Systems (CTS’06), May 2006, pp. 318–325, doi: 10.1109/CTS.2006.14.

[13] X. Liu, S. Raorane, and M. C. Leu, “A Web-based Intelligent Collaborative System for Engineering Design,” in Collaborative Product Design and Manufacturing Methodologies and Applications, W. D. Li, C. McMahon, S. K. Ong, and A. Y. C. Nee, Eds. London: Springer London, 2007, pp. 37–58.

[14] X. Liu, E. Khudkhudia, and Ming Leu, “Incorporation of evidences into an intelligent computational argumentation network for a web-based collaborative engineering design system,” in 2008 International Symposium on Collaborative Technologies and Systems, May 2008, pp. 376–382, doi: 10.1109/CTS.2008.4543954.

[15] X. Liu, E. C. Barnes, and J. E. Savolainen, “Conflict Detection and Resolution for Product Line Design in a Collaborative Decision Making Environment,” in Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, New York, NY, USA, 2012, pp. 1327–1336, doi: 10.1145/2145204.2145402.

[16] X. Liu, M. Satyavolu, and M. C. Leu, “Contribution based priority assessment in a web-based intelligent argumentation network for collaborative software development,” in 2009 International Symposium on Collaborative Technologies and Systems, May 2009, pp. 147–154, doi: 10.1109/CTS.2009.5067475.

[17] R. Arvapally and X. (Frank) Liu, “Analyzing credibility of arguments in a web-based intelligent argumentation system for collective decision support based on K-means clustering algorithm: Knowledge Management Research & Practice: Vol 10, No 4,” Knowledge Management Research & Preactice, vol. 10, no. 4, pp. 326–341, 2012, doi: https://doi.org/10.1057/kmrp.2012.26.

[18] R. S. Arvapally, X. F. Liu, F. F.-H. Nah, and W. Jiang, “Identifying outlier opinions in an online intelligent argumentation system,” Concurrency and Computation: Practice and Experience, vol. n/a, no. n/a, p. e4107, 2017, doi: 10.1002/cpe.4107.

[19] R. S. Arvapally, X. Liu, and W. Jiang, “Identification of faction groups and leaders in Web-based intelligent argumentation system for collaborative decision support,” in 2012 International Conference on Collaboration Technologies and Systems (CTS), May 2012, pp. 509–516, doi: 10.1109/CTS.2012.6261098.

[20] M. M. Rahman, J. Sirrianni, X. F. Liu, and D. Adams, “Predicting opinions across multiple issues in large scale cyber argumentation using collaborative filtering and viewpoint correlation,” presented at the The Ninth International Conference on Social Media Technologies, Communication, and Informatics, 2019.

[21] M. Wojcieszak, “‘Don’t talk to me’: effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism,” New Media & Society, vol. 12, no. 4, pp. 637–655, Jun. 2010, doi: 10.1177/1461444809342775.

[22] J. K. Lee, J. Choi, C. Kim, and Y. Kim, “Social Media, Network Heterogeneity, and Opinion Polarization,” J Commun, vol. 64, no. 4, pp. 702–722, Aug. 2014, doi: 10.1111/jcom.12077.

[23] J. Stromer-Galley and P. Muhlberger, “Agreement and Disagreement in Group Deliberation: Effects on Deliberation Satisfaction, Future Engagement, and Decision Legitimacy,” Political Communication, vol. 26, no. 2, pp. 173–192, May 2009, doi: 10.1080/10584600902850775.

[24] M. E. Wojcieszak and V. Price, “Perceived Versus Actual Disagreement: Which Influences Deliberative Experiences?,” J Commun, vol. 62, no. 3, pp. 418–436, Jun. 2012, doi: 10.1111/j.1460-2466.2012.01645.x.

[25] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig, “The ‘Nasty Effect:’ Online Incivility and Risk Perceptions of Emerging Technologies,” J Comput Mediat Commun, vol. 19, no. 3, pp. 373–387, Apr. 2014, doi: 10.1111/jcc4.12009.

[26] B. T. Gervais, “Incivility Online: Affective and Behavioral Reactions to Uncivil Political Posts in a Web-based Experiment,” Journal of Information Technology & Politics, vol. 12, no. 2, pp. 167–185, Apr. 2015, doi: 10.1080/19331681.2014.997416.

[27] R. S. Arvapally, X. F. Liu, and D. C. Wunsch, “Fuzzy c-Means Clustering Based Polarization Assessment in Intelligent Argumentation System for Collaborative Decision Support,” in 2013 IEEE 37th Annual Computer Software and Applications Conference, Jul. 2013, pp. 59–64, doi: 10.1109/COMPSAC.2013.12.

[28] R. S. Arvapally and X. (Frank) Liu, “Polarisation assessment in an intelligent argumentation system using fuzzy clustering algorithm for collaborative decision support,” Argument & Computation, vol. 4, no. 3, pp. 181–208, Sep. 2013, doi: 10.1080/19462166.2013.794163.

[29] E. Bakshy, S. Messing, and L. A. Adamic, “Exposure to ideologically diverse news and opinion on Facebook,” Science, vol. 348, no. 6239, pp. 1130–1132, Jun. 2015, doi: 10.1126/science.aaa1160.

[30] S. Flaxman, S. Goel, and J. M. Rao, “Filter Bubbles, Echo Chambers, and Online News Consumption,” Public Opin Q, vol. 80, no. S1, pp. 298–320, Jan. 2016, doi: 10.1093/poq/nfw006.

[31] M. Klein, A. Gruzd, and J. Lannigan, “Using Deliberation-Centric Social Network Analysis to Measure Balkanization,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2914554, Feb. 2017. Accessed: Dec. 17, 2018. [Online]. Available: https://papers.ssrn.com/abstract=2914554.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[33] W. Kunz and H. W. J. Rittel, “Issues as elements of information systems,” presented at the Working Paper No. 131, Berkeley, 1970, vol. 131.

Chapter 2: Empirical Study Datasets

Our research group has used ICAS to conduct three separate empirical studies in Fall 2017, Spring 2018, and Spring 2019, which makes up the core data used in the existing work and for the proposed research tasks. This data was collected by Dr. Frank Liu, Dr. Douglas Adams, Md Mahfuzer Rahman, Najla Althuniyan, Zheng Hu, and myself in a joint research effort. These studies were approved by the IRB (protocol #1710077940). In each study, undergraduate students in an entry-level sociology class were offered extra credit to participate in discussions related to their course work using the ICAS platform. Each student was asked to discuss the four following Issues:

- Healthcare: Should individuals be required by the government to have health insurance?
- Same Sex Adoption: Should same sex married couples be allowed to adopt children?
- Guns on Campus: Should students with a concealed carry permit be allowed to carry guns on campus?
- Religion and Medicine: Should parents who believe in healing through prayer be allowed to deny medical treatment for their child?

The issues were selected based on their controversial nature and their relatedness with the topics covered in the coursework. Under each issue, were four pre-defined positions, constructed such that each issue had one strong conservative position, one moderately conservative position, one moderately liberal position, and one strong liberal position. They are listed below:

- Healthcare: Should individuals be required by the government to have health insurance?

- H1: No, the government should not require health insurance.
- H2: No, but the government should provide help paying for health insurance.
- H3: Yes, the government should require health insurance and help pay for it, but uninsured individuals will have to pay a fine.
- H4: Yes, the government should require health insurance and guarantee health coverage for everyone.

- Same Sex Adoption: Should same sex married couples be allowed to adopt children?
  - S1: No, same sex couples should not be allowed to legally adopt children.
  - S2: No, but adoption should be allowed for blood relatives of the couple, such as nieces/nephews.
  - S3: Yes, but same sex couples should have special vetting to ensure that they can provide as much as a heterosexual couple.
  - S4: Yes, same sex couples should be treated the same as heterosexual couples and be allowed to adopt via the standard process.
- Guns on Campus: Should students with a concealed carry permit be allowed to carry guns on campus?
  - G1: No, college campuses should not allow students to carry firearms under any circumstances.
  - G2: No, but those who receive special permission from the university should be allowed to concealed carry.
  - G3: Yes, but students should have to undergo additional training.
  - G4: Yes, and there should be no additional test. A concealed carry permit is enough to carry on campus.

- Religion and Medicine: Should parents who believe in healing through prayer be allowed to deny medical treatment for their child?
    - R1: Yes, religious freedom should be respected.
    - R2: Yes, but only in cases where the child's life is not in immediate danger.
    - R3: No, but may deny preventative treatments like vaccines.
    - R4: No, the child's medical safety should come first.

These positions were used in each study (except for Fall 2017, which had the following position:

> "Yes, the government should require health insurance and should punish anyone who does not have it."

instead of positions G2 and G3). In the study, each student was given extra credit for posting 10 arguments for each issue. So, a student who completed their tasks posed forty arguments total, ten under each issue. Each study took place in the last month of the semester. Table 2.1 has a breakdown of the total participation for each empirical study.

Table 2.1: Date and Participation information for each empirical study.

Study	Start Date	End Date	Total Days	Total Number of participants	Total Number of Arguments posted
Fall 2017	11/20/2017	12/13/2017	23	318	5722
Spring 2018	4/10/2018	5/5/2018	27	335	10,573
Spring 2019	4/2/2019	5/3/2019	31	251	6384
Spring 2020	3/30/2020	4/30/2020	31	129	1476

Chapter 3: Quantitative Modeling of Polarization in Online Intelligent Argumentation and Deliberation for Capturing Collective Intelligence

3.1 Abstract

Cyber argumentation platforms offer specially designed environments for users to discuss and debate their stances and viewpoints on important issues. However, argumentation polarization often occurs in discussions and debates on these cyber argumentation platforms. Several researchers investigated argumentation polarization qualitatively in the past, but none have developed a quantitative model for measuring the degree of argumentation polarization. We addressed this important and challenging issue by developing an innovative argumentation polarization model to measure argumentation polarization by incorporating four important attributes of argumentation polarization: 1). The total number of argumentation poles, 2) The population size of the argumentation poles, 3) similarity within argumentation poles, and 4) the dissimilarity between argumentation poles. Its baseline model was derived from an economic polarization model proposed by Esteban and Ray, which measures polarization using three features: 1) Homogeneity within each group, 2) Heterogeneity across groups, and 3) a small number of significant groups. We adapted their model by incorporating population sizes for poles, normalizing for population size, and normalizing for parameter selection, to fit our formulation of argumentation polarization. This model was evaluated using an empirical study conducted using our cyber argumentation platform, the Intelligent Cyber Argumentation System (ICAS). This model was evaluated with two other distribution-based opinion polarization models that were applied in online discussion contexts, both analytically and empirically, since there are no existing argumentation polarization models. The analytical and empirical evaluations indicate that our

model performs more effectively in terms of the definition of argumentation polarization and the four attributes.

3.2 Introduction

Cyber argumentation platforms are specialized online discussion and debate platforms that encourage productive deliberation through the platform's design and provide analysis on the deliberation process and outcomes [1]–[7]. These platforms are alternatives to popular online platforms, such as social media and networking platforms, which serve as the de facto public forums for online deliberation and debate on important topics [8]. Those platforms, while popular, often result in undesirable deliberation outcomes, such as echo chambers [9]–[13], where only like-minded users discuss ideas with one another which isolates them from diverse opinions, viewpoints, and ideas, often leading to extreme opinions [11], animosity toward out-group members [14], and a more polarized environment [12]. Likewise, discussions in online social media and networking platforms also tend to have very low deliberation quality [15]–[19].

Thus, many researchers in online deliberation and cyber argumentation have investigated various argumentation structures and features to promote higher quality online deliberation [19]–[21]. Many research groups have proposed specialized cyber argumentation platforms that use a variety of approaches to promote more productive deliberation and provide analysis on the deliberation process and outcomes [1]–[7]. However, providing structure and features alone is not enough to ensure a productive discourse environment. One major factor that can detrimentally affect online discussions and debate is polarization.

Polarization has been shown to have many detrimental effects on discussion in both offline [22] and online [13], [23] settings. In online settings, polarization has shown to lower the quality

of the crowd-wisdom from discussions [13], increased tension between ideological groups [14], contributed to decreased civility in public discourse [24], and polarize attitudes at the individual level [25]. Much attention has been paid to opinion polarization in online settings, such as social media [9], [23], [26]–[30], among political blogs [31], [32], and in comment sections [15]. However, polarization in online argumentation has not received much research attention.

Polarization in discussions in cyber argumentation, which we refer to as argumentation polarization, differs in some key ways from opinion polarization in other contexts, such as in social media and network platforms. We define argumentation polarization as the degree to which the discussion participants in cyber argumentation form distinct, internally consistent argumentation poles or groups of significant size that conflict to some degree with one another based on the degree of their agreement or disagreement with the discussion topic. Discussions in cyber argumentation are typically centralized, meaning all of the posts in a discussion are located in the same place and are available to all participants. This centralization of the discussions means that participants are exposed to a wide variety of viewpoints and opinions from their peers and are encouraged to engage with them. In cyber argumentation, the participants must defend their stances and attack opposing users' stances by making arguments. Thus, opinions formed through argumentation are more informed through consideration of alternative views, ideas, and experiences [33]–[37], compared to unjustified opinions collected through polling or surveying, which do not reflect these considerations [38]. Furthermore, because discussions in cyber argumentation are centralized, users cannot easily isolate their views within an echo chamber, as often happens in social media and networking platforms. Social media and networking platforms focus on user interaction through social connections, while cyber argumentation focuses on interaction through discussion and debate in a shared space. Thus, the network approaches taken

to model polarization in social media, which rely on social connections as the main mechanism for opinion influence [23], [39], [40], do not apply to argumentation polarization, where social connections are de-emphasized.

Developing a polarization model specifically for argumentation polarization is important because the opinion data produced for cyber argumentation has specific qualities that need to be considered. Furthermore, because the opinions in cyber argumentation are justified through the argumentation, the opinions reported through cyber argumentation are more specific and nuanced than opinions collected without justification. As such, the polarization model ought to consider even small differences in opinions between users as somewhat significant and adjust its measurement of polarization accordingly. An argumentation polarization model must place more emphasis on the differences between user opinion, even when those differences are small, when measuring the polarization in cyber argumentation.

Some prior research has been done to investigate argumentation polarization. Avrapalley et al. [41] proposed a method of assessing polarization in cyber argumentation for collaborative decisions using Fuzzy c-means clustering. They clustered discussion participants based on their opinions in several debates under a shared issue. They asserted that each cluster represented a polarized group and each participant's fuzzy membership with the clusters represented the extent to which they as individuals are polarized. Klein et al. [42] proposed a method of assessing cyber balkanization (which is related to polarization [28]) in Reddit forums by encoding the discussions into signed interaction graphs and analyzing the different graph connections between the users. However, these investigations only focus on the presence of argumentation polarization and do not quantitatively model argumentation polarization to quantify the degree of argumentation polarization.

In this work, we address this important issue by developing an innovative argumentation polarization model to measure the degree of argumentation polarization among the participants in cyber argumentation. To develop this model, we identify four important attributes of argumentation polarization that characterize its presence in cyber argumentation, based on the distribution of the participants' opinion stance (agreement or disagreement) toward the discussion topic. Given the distribution of the participants' opinion stances, our four identified attributes that characterize argumentation polarization are: 1) The total number of argumentation poles, 2) the population size of the argumentation poles, 3) the similarity within the argumentation poles, and 4) the dissimilarity between argumentation poles.

Given these four attributes, we develop a novel argumentation polarization model by adapting a multi-modal economic polarization model proposed by Esteban and Ray [43]. Their original model used three features to measure polarization among different income groups: 1) homogeneity within each group, 2) heterogeneity across groups, and 3) a small number of significant groups, which are similar to our attributes 1, 3, and 4. However, we still need to incorporate attribute 2 into their model to make it suitable for argumentation polarization. Their original model measures polarization as a result of conflict between any two individuals, regardless of whether they are in a pole or not. Argumentation polarization, on the other hand, requires that polarization may only occur as the result of conflicts between two significant groups/poles of users. Thus, we adapted their model to consider the population sizes of each user group (attribute 2) by introducing a minimum pole strength threshold requirement to the model. This threshold ensures that only users in a sufficiently strong pole will produce polarization, and users who are not in a pole will not produce polarization. Additionally, we also normalized the model by total discussion population size to ensure that discussions containing more participants are not always more

polarized than discussions containing fewer participants (as is the case in their original model) and we normalize the index based on selected parameters in the model to ensure the polarization index output by the model is between 0 and 1. These adaptations ensure that the resulting modified argumentation polarization model (MAP) considers all four important attributes of argumentation polarization and is normalized by the population size.

We evaluate our model on an empirical dataset of online discussions collected using our cyber argumentation platform, the Intelligent Cyber Argumentation System (ICAS). ICAS is equipped with a powerful fuzzy logic argument reduction system, which was used to approximate each user's overall agreement stance toward the discussion topic, based on their discussion posts. The resulting distribution of user agreement (i.e. their opinion stance toward the discussion topic) in a discussion was fed into the model as input to evaluate the degree of argumentation polarization in the agreement distribution.

To our knowledge, we are the first to present a model of argumentation polarization, so we are not able to directly compare our model to other existing argumentation polarization models. Instead, we compare our model to two other distribution-based polarization models that have been applied to online discussions, Flache and Macy's model (FM) [44] and Morales et al.'s model (MBLB) [27]. These models have been applied in contexts similar to cyber argumentation, and demonstrate two popular approaches to modeling polarization: a variance-based approach and a bi-modal based approach respectively. We compare these models to our model, both analytically and on the empirical dataset, in terms of their ability to capture the four attributes of argumentation polarization. Our analytical and empirical results indicate that our model performs more effectively in terms of the definition of argumentation polarization and our four attributes.

We further justify our model's multi-modal approach to argumentation polarization by examining the topic selection made by the participants in their arguments in the discussion. The results of this analysis show that a bi-modal assumption does not sufficiently capture the important differences between participants and a multi-modal approach is more effective.

This article makes the following contributions:

1) We present a novel model of argumentation polarization that measures the polarization within an agreement distribution in cyber argumentation. We identify four key attributes of polarization within the agreement distribution that must be considered when modeling polarization, which our model captures.

2) We compare our model to two other polarization models used in online discussion and online deliberation literature, both theoretically and on empirical data. Our results indicate that the other models do not sufficiently capture our four attributes of argumentation polarization, which results in some unorthodox polarization values in the empirical data.

3) We justify our model's multi-modal approach based on our analysis of frame and topic selection in discussions from our empirical data

3.3 Related Work

In the broader polarization literature, there are typically three approaches to modeling and measuring opinion polarization: 1) A survey-based approach, 2) A network-based approach, and 3) A distribution-based approach.

The survey-based approach is the most common and uses the difference between user responses on surveys to measure the total amount of polarization between groups [26], [45]– [50]. Polarization itself is measured by combining the survey responses using some relevant measurement, spanning from simple measurements [45], [47], such as averages, variance, and differences, correlation analysis between responses [46], regressions [26], [48], and prediction models [29]. Survey approaches, however, have the weakness that the opinion expressed by the users may not be well considered or well-informed [38].

The network-based approach models polarization as the degree to which a social network is clustered into distinct, conflicting groups [23], [31], [32], [39], [40], [51]. This approach to polarization is popular among those studying polarization in social media [23], [39], [40] or other online content like blogs [31] and those modeling polarization using agent-based network simulations [51]–[53]. Typically these models approach opinion formation as a result of social interactions through network connections [23], [39], [40] and measure polarization as the degree of group cohesion in the network [23], [31], [32], [54].

The distribution-based approaches models polarization as a function of a single or multiple variable distributed across a population [27], [43], [44]. This approach conceptualizes polarization as a division of a population into distinct, conflicting, internally-cohesive groups along some axis. While some studies used basic metrics [55], such as the number of peaks in the distribution [56], to measure polarization, often studies present their own unique models of measuring polarization on a given distribution [27], [43], [44], depending on the type of polarization being modeled.

Of these approaches, only the distribution-based approach applies to argumentation polarization. Survey-based approaches gather opinion information from surveys, which do not require users to justify their opinions, as argumentation polarization does. Network-based

approaches require the subjects to be arranged in a social network, where users formulate their opinions based in part on their network connections. However, cyber argumentation does not require users to form a social network and instead encourages opinion formulation through deliberation. Thus, a distribution-based approach, using the user's agreement or disagreement toward the discussion topic as the distribution variable, is most applicable to argumentation polarization.

Work by Bramson et al. breaks down distribution-based polarization measurements into nine different, pairwise independent senses of polarization [57], [58]. These nine senses of polarization, spread, dispersion, coverage, regionalization, community fragmentation, distinctness, and group divergence, are designed to help examine polarization models to determine which senses of polarization are and are not captured by the model. The authors note that different types and contexts of polarization can focus on different subsets of their proposed senses. We discuss our model's coverage and the coverage of other relevant models of these nine senses in later sections.

3.4 Deriving Participant's Level of Agreement with ICAS

Argumentation systems often contain cognitive computing components, which use AI techniques to automatically derive information about the participants that are based on their participation. Our argumentation tool, ICAS, uses a built-in fuzzy logic engine to derive each participant's agreement toward a position. The following section gives a brief outline of ICAS.

3.4.1 Fuzzy Logic Agreement Reduction

Since each argument and reaction has an agreement value, a user's overall agreement on a position can be determined by examining the agreement values of all the user's arguments and reactions under the position. However, we must first reconcile the arguments and reactions a user

has made further down the argument tree. This issue is resolved by using the fuzzy logic engine built into ICAS (refer to Chapter 1.1.3 for a description of the reduction process).

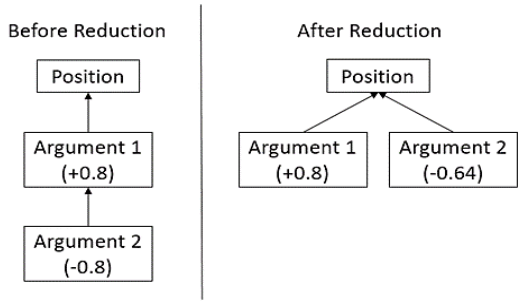


Figure 3-1 Example of a fuzzy logic reduction.

Once the fuzzy logic engine has reduced all arguments for a given position (as shown in an example in Figure 3-1), we can determine each user's agreement level towards that position by averaging the sentiment of all their arguments and reactions together. Using the average ensures that user agreement will always be bounded by -1 and +1. Thus, we get a distribution for user agreement on a position at any point in time.

3.5 Argumentation Polarization Formulation

To model argumentation polarization as a function of the agreement distribution of the users in a discussion, we must first identify the key attributes of the agreement distribution that characterize polarization.

3.5.1 Polarized vs Non-Polarized Argumentation Distributions

The first consideration to make in investigating argumentation polarization is to identify distributions that are at the extremes of polarization. The most polarized scenario for a distribution is widely considered to occur when the entire population is split evenly among the most two most extremes [27], [43], [44]. In this scenario, both poles have a maximum distance between one another, collectively contain the entire population (no outliers), are entirely internally consistent,

and have equal strength (in terms of population and similarity). This is the only scenario for maximum polarization.

For a distribution that contains no polarization, there are two scenarios. The first scenario is when only one pole forms in the distribution. If all the users in the distribution agree with one another, there is only one pole, and there are no rival poles to cause polarizing tension. Likewise, even if some individual users dissent from the pole, they are treated as outliers and have no impact on the polarization, unless they coalesce into a like-minded pole with a substantial enough population size to rival the original pole.

The second scenario is when all the users disagree with one another, to the extent to which that is possible, resulting in the uniform distribution. If each user is uniformly distributed across all agreement values, then there are no groups of similar users of significant size, and thus there are no poles. Every section of the distribution has the same population and similarity as every other section of the same size. Since no conflicting groups of significant size form in the uniform distribution, according to our definition of argumentation polarization, no polarization can occur in the distribution. Thus, a uniform distribution yields no polarization.

3.5.2 Attributes of Argumentation Polarization

From our examination of the maximally and minimally polarized distributions, we can observe four main attributes of the distribution that determine the degree of argumentation polarization. These attributes can be seen as the key determinants of polarization in the agreement distribution in argumentation.

Attribute 1: The total number of argumentation poles. The biggest indicator of argumentation polarization is the total number of poles in the distribution. As discussed in the

previous subsection, if a distribution does not contain at least two poles, then no polarization can exist. So, the distribution needs at least two argumentation poles for argumentation polarization to occur.

However, in the maximal polarization distribution, there are only two poles. The addition of more than two poles would lower the argumentation polarization because the population would be split between all of them, which would weaken each individual pole. From this observation, we can say that the population of an argumentation pole is an important factor in the pole's overall strength, which leads to attribute 2.

Attribute 2: The population size of the argumentation poles. The population size within an argumentation pole is the main component of its strength. Poles with more users within them are stronger than poles with fewer users. If more users are grouped into conflicting poles, more polarization should occur. Likewise, the ratio between two poles' population size is important when determining the amount of conflict between them. If one pole has significantly more population in it than the other, then the stronger pole overwhelms the weaker pole, resulting in lower polarization. If the poles have a similar population size then they are more equal in strength and thus produce more polarization. However, population size is not the only component of a pole's strength, which leads to attribute 3.

Attribute 3: The similarity within the argumentation poles. In addition to the population, the strength of the agreement pole is also related to the agreement similarity of the users within it. The more unified the users in the pole are, the stronger the pole. In our distribution with maximum polarization, each pole was entirely internally consistent. If the poles were not entirely consistent and had some internal variance, they would be weaker poles and the overall argumentation

polarization would be lower. Thus, the internal distances between the users in the pole impact the strength of the pole and the overall argumentation polarization.

Attribute 4: The dissimilarity between the agreement poles. The polarizing tension created between argumentation poles is measured by their difference in agreement. The larger the difference between two poles the more polarization is occurring as a result. In our distribution with maximum polarization, the poles were at a maximum distance from one another. If they were to move closer to each other, then the overall polarization would lower.

These four attributes characterize the presence and intensity of argumentation polarization in the agreement distribution of a discussion in cyber argumentation. In general, polarization is greater when there are fewer poles (but more than two) (attribute 1), when more of the population is a member of a pole (attribute 2), when argumentation poles are internally similar (attribute 3), and when distinct argumentation poles are externally dissimilar (attribute 4).

From Bramson et al's examination of distribution-based polarization measurements, they identify nine pair-wise independent senses that describe various aspects of polarization [57], [58]. These nine senses were designed to be broad enough to cover several types of polarization in a variety of contexts, so all nine senses may not apply to every context. While our attributes were developed independently from these senses (we were unaware of Bramson et al.'s work at the time), they map neatly to four of the nine senses that are most indicative of argumentation polarization: community fracturing, group divergence, group consensus, and size parity. The other five senses do not directly reflect the absence or presence of argumentation polarization in the context of cyber argumentation and so are not covered by attributes. For example, the spread of the distribution may not be indicative of argumentation polarization if extreme views are held by outlier users in the distribution (which was common in our empirical data). Thus our model does

not explicitly cover these senses, though these senses are captured implicitly in the model due to its formulation, which we outline in the next section.

3.5.3 Argumentation Polarization Model

The argumentation polarization model will take in a distribution of user agreement toward a position as input and calculate the total amount of argumentation polarization according to our four identified attributes. Our model is adapted from a distribution-based polarization model by Esteban and Ray [43] for measuring economic polarization. Their model is designed using an axiomatic approach based on three basic features they observed about economic polarization: 1) homogeneity within each group, 2) heterogeneity across groups, and 3) a small number of significantly sized groups. While designed for economic polarization, this model has been applied to other domains to measure polarization within a distribution [59]–[62]. As we can see, their features are very similar to our attributes 1, 3, and 4, which makes this model a good fit for argumentation polarization. But, we still need to adapt the model to consider attribute 2 (population sizes of poles). Additionally, we want to ensure that the polarization index produced from the model is normalized by the population size so that discussions with more population do not always result in increased polarization, so we also normalized the model by population size.

Their original extended model is shown in (1) and (2). The model takes in a distribution (π, y), which is a set of user-value pairs (π_i , y_i), where π_i is the total number of users with value y_i in the distribution. Their model measures polarization as the linear representation of the distances between each user along the distribution, weighted by the mass of similarly clustered users. The max function ($max(|y_i - y_j| - D, 0)$) calculates the distance between users. The parameter D, is a threshold value that determines if two users are similar ($|y_i - y_j| \leq D$) (do not

produce polarization) or are dissimilar ($|y_i - y_j| > D$) (produces polarization) with each other. The identity function (I_i) for a user is the total amount of similarity a user has with their near-by neighbors. Here, identity is equivalent to the internal strength of the user's pole (i.e. the more similar others are, the more the user identifies them as in their pole). The function ($w(y_j)$) is a user-defined function that determines how the degree of similarity is weighted in the identity function. The variable α acts as an identity intensifier and determines how important identity (i.e. a pole's internal strength) is in the model. Esteban and Ray prove that α must be bounded by $\alpha \in (0, \alpha^*]$ where $\alpha^* \cong 1.6$, to satisfy their axioms.

$$P(\pi, \vec{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i \pi_j I_i^{\alpha} \max\{|y_i - y_j| - D, 0\} \quad (1)$$

$$I_i \equiv \sum_{j:|y_i - y_j| \leq D} \pi_j w(y_j) \quad (2)$$

Their model captures their three basic features: 1) homogeneity within each group, 2) heterogeneity across groups, and 3) a small number of significantly sized groups, which are analogous to our attributes 1, 3, and 4. The first feature is captured through the identity function. If a user's neighbors are more similar to them (homogeneous), in terms of distance along the distribution, then their identity value will be larger, which will result in more polarization. The second feature is captured through the max distance function. If two groups of users are further away from one another (heterogeneous), their total distance along the distribution will increase, which will result in a greater amount of polarization. The third feature is captured through the product of the distance function and identity function. Since the population size is finite, if more of the population is distributed among a smaller number of poles, each pole's identity value will be larger, resulting in greater polarization value. For brevity, we will not outline how the model

performs on specific theoretical cases, and instead will refer to their original paper [43] for these cases and a more in-depth discussion of the base model. Our modifications do not change the fundamental operation of the model relating to these features, so their examples and discussion are still is applicable to our modified model.

Their original model does not explicitly consider the population sizes of poles (attribute 2). Instead, the model implicitly considers population size by weighing the polarization produced from a user by their identity function (I_i) which is the sum the similarity of their neighbors. So, users with more neighbors (i.e. are in a bigger pole) will produce more polarization than users with fewer neighbors. This approach is sufficient for economic polarization because their distribution domain is unbounded (from 0 to infinity). So, for example, the polarization for a uniform distribution in an unbounded domain would have each user in their own pole ($I_i = 1$) and would yield a small polarization value (assuming we normalize by population size). However, our agreement distribution domain is bounded (from -1 to +1). So, a uniform distribution in our distribution domain would not yield small identity values for the users and instead would measure a substantial amount of polarization. From our formulation of argumentation polarization, a uniform distribution does not contain polarization, so the model should output a polarization index value of zero when given a uniform distribution, which their model, in its original form, does not.

Thus, we adapt the model by defining a threshold value for a user's internal pole strength (i.e. their identity value (I_i)). If a user's identity value is greater than the threshold, T, they are considered a member of a significant pole and can produce polarization. This threshold checks if a user is in a valid pole or not, as shown in (3). This threshold ensures that users are only considered a member of an agreement pole if their pole is sufficiently strong enough to produce polarization.

$$pole(\pi_i, y_i) = \begin{cases} True & if\ I_i > T \\ False & if\ I_i \leq T \end{cases} \quad (3)$$

Since we are using the uniform distribution as the baseline for a non-polarized distribution, we can set the value of T to be the expected identity value from a uniform distribution. If a user does not have an identity value greater than T, then they are not identifying more than they would in a uniform distribution, and thus are not considered part of a pole. Using this definition, we set T as a function of y_i (the user's agreement value) and the parameter D. The function $T(y_i)$ shown in (4) calculates the expected identity value of the user at y_i if they were in a uniform distribution, bounded by [-1, +1]. The second term ($|y_i| + D > 1$) accounts for the distribution boundary.

$$T(y_i) = \begin{cases} D/2 & |y_i| + D \leq 1 \\ \dfrac{2D^2 - (|y_i| + D - 1)^2}{4D} & |y_i| + D > 1 \end{cases} \quad (4)$$

In addition to the threshold T, we also want to normalize the polarization model to the total population size. Esteban and Ray's original model does not weight by population, so adding more users in the distribution will always increase polarization, regardless of where they are in the distribution. Therefore, we normalize the model by the total population size N. The resulting adapted model is shown in (5) and (6).

$$P(\pi, \vec{y}, N) = \sum_{i:pole(\pi_i, y_i)}^{n} \sum_{j:pole(\pi_j, y_j)}^{n} \frac{\pi_i \pi_j}{N} I_i^{\alpha} \max\{|y_i - y_j| - D, 0\} \quad (5)$$

$$I_i = 1 + \sum_{|y_i - y_j| \leq D} \left(\frac{\pi_i}{N} * \left(1 - \frac{|y_i - y_j|}{D} \right) \right) \quad (6)$$

For the identity function (6), we implemented the $w(y_j)$ function as the linear distance away from the target user. Notice that the identity function is normalized by population size. We

add one to the summation to ensure the identity value is always greater than one, which is important to ensure that parameter α operates as an identity intensifier and behaves as intended.

The resulting adapted model shown in (5) now explicitly considers the population size of an argumentation pole (attribute 2). If a user does not have an identity value greater than T, then they are not within a pole with sufficient population and similarity to be considered a valid argumentation pole. In this model, a uniform distribution would not produce any polarization.

However, depending on the parameter values selected for variables D and α, the maximum value of (5) may be greater than or less than one. To fix our model's output range to be between zero and one, we normalize the model using min-max normalization. The maximum value of P, given the parameters, can be calculated by providing the maximally polarized distribution: $\left[\left((N/2),-1\right),\left((N/2),+1\right)\right]$. Since the model normalizes the distribution by population size, we can derive P_{max} as a function of D and α as shown in (7).

$$P_{max} = \left(1 - \frac{D}{2}\right)(1.5^{\alpha}) \tag{7}$$

Using this maximum (and since we know the minimum value is zero), P can be normalized with respect to the input parameters is shown in (8). We define the value of P_{norm} as the polarization index value in our final model.

$$P_{norm}(\pi, y, N) = \frac{P(\pi, y, N)}{P_{max}} \tag{8}$$

Our adapted model still satisfies Esteban and Ray's original axioms used to develop their model, if we assume that all of poles in the axioms are above threshold T (except pole p in Axiom 4, which they describe as insignificant) [43].

In terms of our four attributes of opinion polarization in cyber argumentation, our model also accounts for each of these attributes as well. The similarity within each pole (attribute 3) is modeled by the identity function I_i, which increases the polarization produced by a pole as a factor of its total internal similarity. The number of poles (attribute 1) is also modeled by the identity function by increasing the amount of polarization produced by a pole based on the proportion of the population contained within it. More poles mean each pole has a smaller portion of the population and thus a lower identification value, which in aggregate lowers the total polarization. The dissimilarity across different poles (attribute 4) is modeled by the max function. The difference in population size (attribute 2) is modeled by the threshold T and by weighing the polarization produced by each pole by their proportion of the population; maximum polarization between two poles occurs when they contain equal population proportionality.

The model has a runtime complexity of $O(n^2)$ (identity values can be computed in advance) where n is the number of unique (π_i , y_i) pairs. The total number of pairs depends on the total number of unique agreement values, and since the distribution is bounded from -1.0 to +1.0, the total possible number of pairs is bounded by the precision of the agreement values. Small differences in agreement (< 0.01) have very little impact on the polarization value, so the number of unique pairs can be limited by rounding the values to a lower precision. For example, rounding the values to two decimal places will ensure that the number of unique pairs will be less than or equal to 201.

Referring back to Bramson et al.'s nine senses of polarization [57], [58], this model covers four senses explicitly (from the attributes) and five senses implicitly. Community fracturing (attribute 1), group divergence (attribute 4), group consensus (attribute 3), and size parity (attribute 2) are all covered explicitly by the four attributes. Spread, dispersion, coverage, regionalization,

and distinctness are captured implicitly, but only under some conditions. The biggest condition being that our model tries to ignore outliers using the threshold parameter T. So, revisiting our previous example, the spread of the distribution could be ignored in some cases where the extreme users in the distribution are treated as outliers (i.e. non-poles) by the model. Due to our formation of a single combined polarization model, each sense cannot be examined independently of one another. However, in most common scenarios these five senses are implicitly captured.

3.6 Argumentation Polarization Model Parameters

In the previous section, we introduced the polarization model. In this section, we will discuss the two user-defined parameters, D, and α, their role in modeling opinion polarization, and recommended value to use for measuring polarization.

3.6.1 Parameter α

Polarization The parameter α is designed as an identity intensifier and determines how important internal pole strength is in the model. An α value of zero means internal pole strength does not impact the polarization at all, making the model more similar to inequality measures [43], while an α value of 1.59 (near maximum) would indicate that the model will strongly weigh internal pole strength when measuring polarization. The α parameter chiefly affects the importance of the internal similarity within the agreement poles in the model (attribute 3). Figure 3-2 shows how various α values affect the polarization index value for a simulated bi-modal distribution with different standard deviations within the poles (higher standard deviation creates less internal similarity). The greater the value of α, the more sensitive the model is to changes in the internal similarity of the poles.

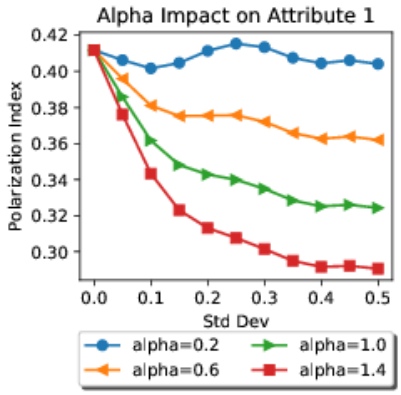


Figure 3-2 Different α values on the polarization index for a simulated bimodal distribution with different standard deviations within the poles. Parameters: $D = 0.3, T = 0$.

In practice, we recommend larger values of α. Argumentation polarization describes the degree to which the users have formed internally similar groups based on their agreement with the topic, making internal identification very important in the modeling of argumentation polarization. An α value that is too low will not take internal similarity of the poles into account, thus we recommend an α value between 0.8 and 1.59.

3.6.2 Parameter D

The parameter D acts as the maximum agreement distance threshold that determines if two users are similar to one another (i.e. are in the same pole) or not. For example, if user A has an overall agreement value of +0.5, and user B has an agreement value of +0.68, then these two users would identify with one another if D was greater than or equal to 0.18. This parameter acts as the model's consideration of the similarity between two user agreement values. If the D value is set to a small value, then the model will consider smaller differences between agreement more important. Likewise, for larger values for D, the model will only consider larger differences important.

Given the bounds of the agreement distribution, we recommend D values between 0.2 and 0.5. If the value of D is too large, then the model will group users who have large differences in

agreement, which will affect the total number of poles the model will consider. For example, if D = 1.0, it is not possible to have three non-overlapping poles in the distribution. However, D values that are too small will cause the model to exaggerate the minor differences in user agreement in its measurement of argumentation polarization.

3.7 Experiments and Comparison with Other Polarization Models

In this section, we compare our proposed modified argumentation polarization model (MAP) with two other distribution based polarization models that have been applied to online deliberation research. We compare the models both theoretically on their design in terms of both our four identified attributes of argumentation polarization and using Bramson et al.'s nine senses of polarization [57], [58]. We also compare the models' performance on an empirical dataset of cyber argumentation discussions collected using our cyber argumentation platform ICAS. In that comparison, we demonstrate how the different approaches taken by each model affect how polarization is reported for three illustrative discussions in our ICAS empirical dataset.

3.7.1 Flache and Macy's Model (FM)

In 2014, Gabbriellini and Torroni presented an agent-based dialogue simulation model for argumentative reasoning [63]. As part of their analysis, they measure the polarization of the simulated dialogs using a polarization measurement presented by Flache and Macy [44]. Flache and Macy's model (FM) is relatively straightforward. Given a distribution of user opinions on a bounded scale from -1 to +1, the distance between two users' opinions is dij. The level of polarization of the population N is the variance of the distribution of all distances between every user as shown in (9).

$$P = \frac{1}{N(N-1)} \sum_{i \neq j}^{i=N, j=N} \left(d_{ij} - \bar{d}\right)^2 \quad (9)$$

Where $\bar{d}$ is the average opinion distance across all pairs of opinions (excluding self-distances).

3.7.2 Morales, Borondo, Lasada, and Benito's Model (MBLB)

Morales, Borondo, Losada, and Benito's polarization model (MBLB) [27] was originally used to measure polarization in a Twitter discussion about the late Venezuelan present, Hugo Cháves, in 2015.

Their model assumes a bi-modal distribution. Given a distribution of opinion values of the range -1 to +1, polarization is measured by calculating the center of gravity (average) of each side of the distribution, both positive (range: (0,+1]) and negative (range: [-1,0)), and subtracting each center of gravity from one another. Then the value is scaled by the maximum possible distance between the centers of gravity and by the difference in both sides' population sizes. The model is shown in (10).

$$P = \left(1 - \frac{\left|N_{pos} - N_{neg}\right|}{N}\right) \frac{\left|gc^{+} - gc^{-}\right|}{2} \quad (10)$$

Where N is the total population size, N_{pos} is the population size with opinion greater than 0, N_{neg} is the population size with opinion less than 0, and gc^{+} and gc^{-} are the averages of the opinion values in populations, N_{pos} and N_{neg} respectively.

3.7.3 Theoretical Comparison of the Models

We compare the assumptions and approaches made the two models introduced in the previous section in terms of our definition and attributes of argumentation polarization, and the nine senses of polarization proposed by Bramson et al. [57], [58].

The MBLB model is most distinct from the other two in that it takes a bi-modal approach to model polarization. Bi-modal approaches are very common in analysis of polarization [29], [39], [40], [49]. This approach assumes that polarization is the result of a population dividing into two groups that are in opposition to one another. The direct analog to argumentation polarization would be to assign the participants to agree and disagree groups based on their agreement polarity. However, a consequence of grouping participants in this way is that it strips out the strength of their agreement or disagreement in the model.

Research in political psychology has suggested that careful consideration of opposing arguments surrounding an issue can cause ambivalence among the participants [64]–[66], resulting in a weak commitment to attitudes about an issue [64], [65] Likewise, group polarization research suggests that some users become more extreme in their opinion as a result of group discussion, especially when users mostly engage with their like-minded peers [67]–[70] and ignore the arguments and viewpoints with which they disagree. Since participation in online argumentation may produce different outcomes for different types of users, it is somewhat misleading to group the participants based only on whether they agree or disagree with a position, regardless of their agreement strength. In cyber argumentation, more ambivalent users likely have more in common with one another, even if they are on the other “side” of the agreement spectrum than with more extreme users.

This approach that assumes the group boundary exists at the origin may not be conducive to cyber argumentation, since it does not consider the ambivalence of the participants. Furthermore, this model does not consider intra-group similarity (our attribute 3) or Bramson et al.'s sense of group consensus [57], [58]. As previously discussed, this sense/attribute is important to argumentation polarization, as the strength of the user agreement/disagreement is very consequential in group discussions. In addition to group consensus, MBLB does not consider the spread, regionalization, or coverage senses, which indicates that this model is unable to detect the presence (or absence) of distinct groups in various regions of the distribution.

The FM model characterizes polarization as the variance of the distribution of distances between the users' agreement. Using the variance is a common polarization modeling technique [45], [71], [72]. This approach does not explicitly consider the formation of poles and instead only focuses on the distances between the user agreement.

In terms of Bramson et al.'s senses of polarization [57], [58], this model only explicitly covers dispersion and does not cover the other eight senses, including community fracturing, distinctness, group divergence, and size parity which we explicitly identify in our four attributes as important for capturing argumentation polarization. As a result, the FM model may register polarization being present in a distribution that contains zero poles. Consider our previous example of a uniform distribution. As previously discussed, the uniform distribution does not contain any poles, since there are no significant clusters of users, and as such there can be no polarization. However, the FM model does not consider the formation of clusters or poles, so it will measure a non-zero polarization value.

As previously discussed, the presence or absence of poles is important to argumentation polarization. Argumentation polarization is not merely a lack of consensus, instead it is explicitly

characterized by the formation of internally similar groups in the agreement distribution. In terms of our attributes, FM ignores the number of poles (attribute 1).

3.7.4 Empirical Comparison of the Models

In the previous subsection, we outlined the major theoretical differences between our model and two other distribution-based models used in polarization analysis of discussions in terms of our definition and attributes of argumentation polarization. In this section, we build on that analysis by comparing how the models measure polarization on empirical discussion data collected using our cyber argumentation platform ICAS.

We compare the polarization results for each model on three discussions from two of the four issues in the empirical dataset. Since each of the polarization models only considers the distributions of the user agreement, not discussion topics themselves, and all of the participants were the same across all issues, we can compare the discussions of the two issues together without issue.

These three discussions were selected because they best demonstrate how the approaches of the models affect how polarization is calculated from the distribution. By comparing the polarization results from each model to the underlying user agreement distribution, we illustrate how the FM and MBLB models can produce polarization values that run counter to our conceptualization of argumentation polarization outlined in our four attributes of argumentation polarization, while our MAP model produces results that are consistent with the attributes.

3.7.4.1 Empirical Dataset

The dataset comes from a twenty-four-day exercise where students from an introductory level sociology class participated in an online discussion of various topics using ICAS. The

discussion was split into four different issues, with each issue having three or four predetermined positions for the participants to discuss. The issues were preselected to be hot button issues in the current public debate. The participants were offered extra credit for participating in at least ten arguments under each issue. If they did not want to participate, they were offered an alternative assignment. This research used only the Fall 2017 empirical dataset (see Chapter 3).

The study contained 308 users (N=308, Gender: 40% Male, 60% Female, Race: 79% White, 21% Non-White) who posted a total of 10,573 arguments under the four issues. On average, users posted 2.6 arguments per discussion. The study was completed with IRB approval (Protocol #1710077940).

Using the agreement reduction method described in Section 3.4.1, each user was assigned an overall agreement value for each position discussion. If the user did not participate in the discussion they were not included in the distribution. These overall agreement values were used as the input agreement distribution for the polarization models.

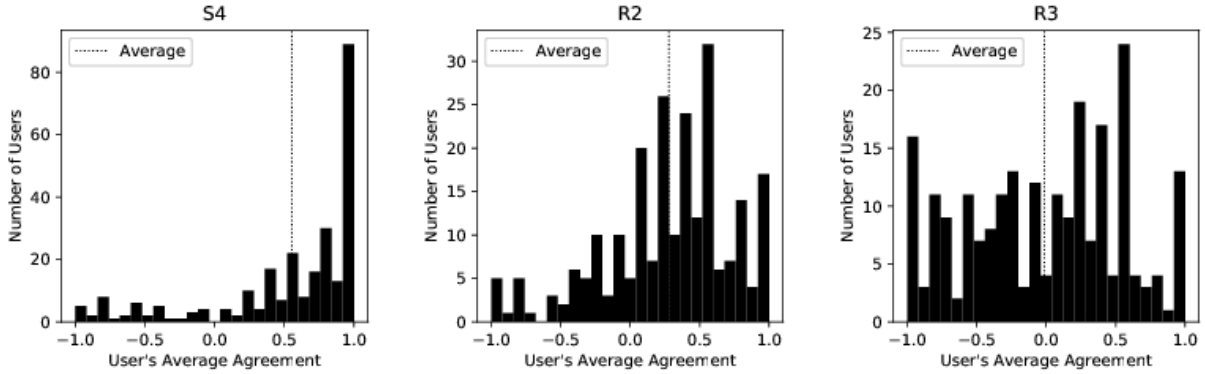
3.7.4.2 Comparison of Models on Empirical Data

Table 3.1 shows the polarization index value for each of the models on each of the discussions of the positions. For brevity, we will examine only the Same Sex Adoption and Religion and Medicine issues. The labels for the positions are assigned as such: the first letter reflects the issue the position is under (S = Same Sex Adoption, R = Religion and Medicine), and the number represents the ideological tilt of the position (1 = Strong Conservative, 2 = Moderately Conservative, 3 = Moderately Liberal, 4 = Strong Liberal). Figure 3-3 shows the agreement distributions for each of the positions that are assigned the highest polarization value by each model. Our model, MAP calculated that R2 was the most polarized position (0.0297), FM

calculated S4 (0.2914), and MBLB calculated R3 (0.4662). MAP was run with parameters $D = 0.5$, $\alpha = 1$.

Table 3.1: The rank and polarization value for all fo the positions in the Relgion and Medicine and Same Sex Adoption issues for each polarization model.

Rank	MAP (Score)	FM (Score)	MBLB (Score)
1	R2 (0.0297)	S4 (0.2914)	R3 (0.4662)
2	R1 (0.0062)	S1 (0.2437)	S3 (0.4633)
3	R4 (0.0048)	R3 (0.2183)	R1 (0.3000)
4	S2 (0.0029)	S3 (0.2133)	S2 (0.2674)
5	S3 (0.0006)	S2 (0.2102)	S1 (0.2269)
6	R3 (0.0002)	R1 (0.1781)	S4 (0.2032)
7	S4 (0)	R4 (0.1741)	R2 (0.1994)
8	S1 (0)	R2 (0.1652)	R4 (0.1496)



(a) Agreement Distribution for Position S4 (b) Agreement Distribution for Position R2 (c) Agreement Distribution for Position R3

Figure 3-3 Histograms of users by their overall average agreement for each position.

FM's most polarized discussion is S4 with a polarization value of 0.2914. From our attributes of argumentation polarization, S4 is not very polarized since almost all of its population concentrated between agreement +0.5 and +1.0, creating only one major pole (attribute 1). Intuitively, positions R2 and R3 are more polarized than S4, as the users in these distributions are more spread out. However, FM's approach to modeling polarization does not consider the formation of poles, and instead only considers the variance between the distances between users. The variance of the distance distribution in S4 is greater than R2 or R3 because the concentration

of users in the agreement distribution at +1.0 creates greater distance with every other user. In R2 and R3, the users are more evenly distributed from one another, resulting in distances that are closer to the average distance. Since FM only considers the variance of the distances and does not consider the formation of poles in the agreement distribution (attributes 1 and 2), its measurements on the empirical study produce results that are inconsistent with our definition of argumentation polarization. Similar scenarios can be observed in positions S1, R4, and G4 (see Appendix Figure A-1).

MBLB's most polarized discussion is R3 with a polarization value of 0.4662. MBLB's definition of polarization focuses on the concentration of user agreement at the extremes, as does the model's approach. MBLB primarily gives higher polarization values to distributions that are evenly split across the origin and are closer to the extremes. This is true of position R3's distribution, however, MBLB does not explicitly look for poles, it instead assumes they are there. Thus, R3's distribution is assumed to be in two poles (one on the positive side and one on the negative), even when the distribution, as shown in Figure 3-3c, has its population relatively evenly spread. The population in R3 is evenly distributed (48% negative, 52% positive) with either side's center of gravity (i.e. average agreement value) roughly in the middle of the distribution (negative center: -0.51, positive center: +0.45). In the MBLB model, this distribution is interpreted as decently polarized, threshold. however, when looking at the actual distribution in Figure 3-3c, the distribution more closely resembles a uniform distribution than a bi-modal distribution. A similar scenario can be observed in position H4.

Our model MAP, for comparison, measured a polarization index value of 0.0002 for position R3, one hundred times lower than for position R2. This low polarization value was due mostly to only very few of the users in R3 being above the pole strength threshold T. Figure 3-4

shows the identity values of the users in the distribution compared to the threshold T. The majority of R3's population was distributed too evenly to have an identity value greater than the threshold. MAP's most polarized discussion was position R2. Position R2 had the most users with an identity value that was greater than T that were far enough away from one another (a distance greater than D=0.5) to generate polarization. Position S4, by contrast, only had users in poles that were between +0.5 and +1.0, which were not far enough away from one another to conflict. Even though position R2 had the most users in conflicting poles, the total distance between the conflicting pole users was fairly low, resulting in a low polarization index value 0.0297 for R2.

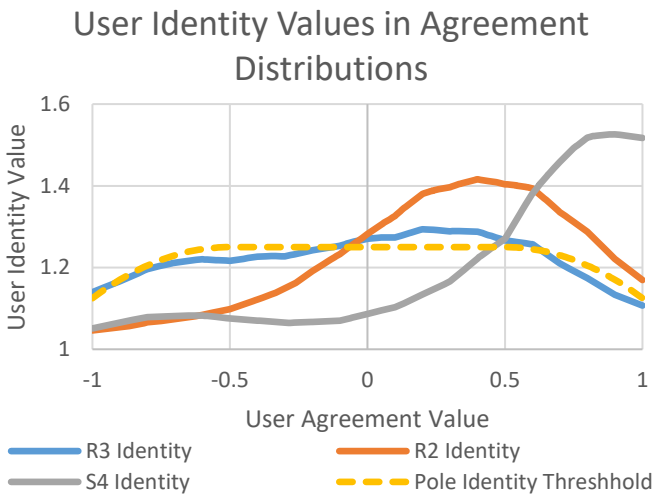


Figure 3-4 The population pole sizes for poles centered at each agreement value in positions S4, R2, and R3. The dashed line is the uniform distribution threshold.

3.8 Justifying a Multi-Modal Approach using Topic Modeling

Previously, we discussed the limitations of a bi-modal approach to polarization because it excludes ambivalence and the reservations that users have concerning their agreement toward a position in cyber argumentation. Instead, our model uses a multi-modal approach. While this assumption of a distinction between ambivalent users and users near the extremes is supported

theoretically, we want to confirm this difference empirically from our dataset to present a stronger argument.

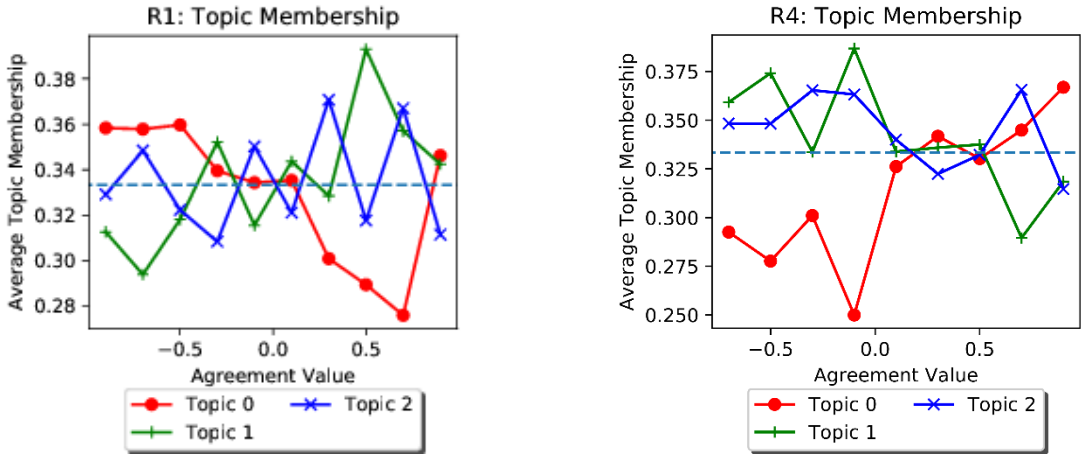
Thus, we examined the agreement groups in terms of topic and framing that users selected during argumentation. Framing and argument topic selection indicate what an author selects as important to discuss [73]. Discussions containing various topics and framings indicate which aspects associated with the issues that the participants are focusing on, and thus indicate their underlying values and thoughts. Topic modeling techniques have shown to be an effective approach to capturing framing in online discussions [30].

We used an LDA model in MALLET [74] to perform topic modeling over all of the arguments under the issues in the empirical data. The number of topics per issue was selected based on the highest coherence score; we tested two to five topics per issue. For brevity, we will only examine the Religion issue.

The Religion issue had three topics, summarized in Table 3.2. Figure 3-5 shows the topic distributions for R4 (the most liberal position) and R1 (the most conservative position). Figure 3-5 shows that Topic 0, which focused on prioritizing health over religion (a traditionally liberal argument), tended to be more popular among more liberal participants (i.e. those who disagreed with R1 and agreed with R4).

Table 3.2: The Topic descriptions for each of the topics in the Religion and Medicine issue.

Topic Number	Topic Description
0	Focuses on arguing that the child's health is more important the parent's religious freedom.
1	Focuses on autonomy of children and who should be allowed to make decisions for whom.
2	Religion's coexistence with Medicine; How religion should play into life and death decisions.



(a) The topic membership for position R1. (b) The topic membership for position R4.

Figure 3-5 The topic membership for users at different overall agreement values.

On the other hand, Topic 1, which focused on the autonomy of children versus parental authority (a more conservative argument), tended to be more popular with more conservative participants (i.e. those who agreed with R1 and disagreed with R4). Topic 2, which discussed the coexistence between religion and medical treatment, was not preferred by either liberal or conservative users.

The tendencies of users to discuss these topics were not uniformly consistent across all user agreement groups. Users near the extremes tended to favor one topic over another. Those who disagreed with R4 favored Topics 1 and 2 over Topic 0, those who more strongly agreed with R4 favored Topics 0 and 2 over Topic 1, those who more strongly agreed with R1 favored Topics 1 and 2 over Topic 0, and those who strong disagreed R1 favored Topics 0 and 2 over Topic 1. On the other hand, more moderate or ambivalent users were more likely to discuss all three topics instead of only one or two.

Taken as a whole, the topic distribution by agreement value indicates that users at various levels of agreement, even within the same "side" of the distribution, have different concerns

surrounding the issue. Furthermore, it suggests that more moderate users are discussing a wider variety of topics than those at the extremes, which reflects the literature suggesting that ambivalence is caused in part by considering many arguments. Thus, the approach of assuming a bi-model distribution, based only on whether the user agrees or disagrees, inadvertently combines groups of users who behave differently, discuss different topics, and have different underlying values. Thus, we assert that a multi-modal approach is more suitable for detecting this divide between users than a bi-modal approach.

3.9 Discussion

Polarization is a broad concept that can be examined from many different perspectives. For example, polarization can be examined as the quality of respect and attitude from one group toward another (affect polarization) [47], or it can be examined as phenomena that push people toward extreme ideologies (group polarization) [75]. It is unlikely that any one model or definition can cover every facet of polarization.

In this work, we focus on covering polarization from the argumentation perspective. Argumentation polarization focuses on the polarization of the users' agreement toward a discussion topic in cyber argumentation. Unlike other types of polarization, argumentation polarization assumes that the users' attitudes have been developed through informed, thoughtful deliberation. Argumentation requires the participants to carefully consider, defend, and formulate their stance on an issue, which leads to opinions and stances that are better informed through consideration of alternative views, ideas, and experiences [33]–[37] than unjustified opinions [38].

The argumentation process may have different outcome effects on different users. As previously discussed, some users may moderate their opinions and become more ambivalent as

they carefully consider arguments from the opposing sides of the debate [64]–[66], while other users may ignore or dismiss the arguments and posts made by those they disagree with and engage mostly with their like-minded peers, becoming more extreme in their stance [67], [68]. Our analysis of topic selection seems to support both of these theories; users on the extremes of the agreement distribution tended to focus on one or two topics while ignoring the topic favored by their opposition, while more moderate users tended to consider all of the topics more equally.

Our model's multi-modal approach is able to make a distinction between users on the extreme and users who are more ambivalent, and pays more attention to the disagreement between users, even when they are on the same side of the distribution. The model's parameters can be adjusted to make the model more sensitive to smaller differences in the user agreement or increase the importance of the formation of argumentation poles.

Polarization as a concept differentiates itself from other phenomena, such as inequality or a lack of consensus, in that it requires the formation of distinct, strong poles that conflict with one another. Our comparison with the FM and MBLB models highlights the importance of how each model handles the presence or absence of poles. FM did not consider poles at all, which resulted in odd modeling behavior in our empirical results. MBLB, on the other hand, always assumes there are only two poles on either side of the distribution, even if the underlying distribution does not match that assumption, which is reflected in the empirical results. Our model does not make prior assumptions about the number and locations of poles in the distribution and instead tries to identify users in poles by proposing a threshold value, T, based on the user's identity with their neighbors. This approach handles different types of distributions better than the bi-modal assumption of MBLB, while still considering the importance of poles, which is key to our definition of argumentation polarization.

Our empirical results showed that argumentation polarization in the discussions in ICAS was lower than one might have expected due to the controversial nature of the issues discussed. Many discussions resulted in either near consensus (such is the case in S4) or near-uniform distribution (such as for R3). For the distributions for each position, please refer to the supplemental material. While these results are only among undergraduate students in a controlled setting, it does suggest that for these issues, users are not polarized as a result of argumentation discussion.

3.10 Conclusion

Online deliberation through cyber argumentation platforms offers an environment for users to discuss and debate their opinions, viewpoints, and stances on important issues. These well-structured debates help users develop well-informed opinions and stances. However, argumentation polarization often arises in discussions of controversial topics. Thus, we present an argumentation polarization model, adapted from an economic polarization model, to measure the polarization among users in terms of their agreement with the debate topic in cyber argumentation. We discussed how argumentation polarization manifests itself in the distribution of user agreement toward the discussion topic and identified four key attributes of argumentation polarization that a model must capture to effectively measure it. We presented a model that measures the argumentation polarization of an agreement distribution derived from discussions in our cyber argumentation platform. We justified our model's design and adaptations, in terms of our four attributes of argumentation polarization, and compared it to two other distribution-based polarization models. Those models' approaches: treating polarization as the variance of distance between users, and treating polarization as bi-modal phenomena. Both produced results on our empirical data that did not align with our definition and key attributes of argumentation

polarization. Our model's multi-modal approach more closely aligns with our observations of argumentation polarization in the empirical data, better adapts to the behaviors of the participants, as shown in our topic modeling analysis, and allows different parameter values to increase or decrease the granularity of the measurement. Thus, our proposed model is effective at measuring argumentation polarization in terms of our definition and attributes of argumentation polarization.

3.11 References

[1] J. Conklin and M. L. Begeman, "gIBIS: A tool for all reasons," Journal of the American Society for Information Science, vol. 40, no. 3, pp. 200–213, 1989. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28198905%2940% 3A3%3C200%3A%3AAID-ASI11%3E3.0.CO%3B2-U

[2] N. Karacapilidis and D. Papadias, "Computer supported argumentation and collaborative decision making: the HERMES system," Information Systems, vol. 26, no. 4, pp. 259–277, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306437901000205

[3] C. Reed and G. Rowe, "Araucaria: software for argument analysis, diagramming and representation," International Journal on Artificial Intelligence Tools, vol. 13, no. 4, pp. 961–979, 2004. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218213004001922

[4] M. Tzagarakis, G. Gkotsis, M. Hatzitaskos, N. Karousos, and N. Karacapilidis, "CoPe it!: argumentative collaboration towards learning," in Proceedings of the 9th international conference on Computer supported collaborative learning - CSCL'09, vol. 2. Association for Computational Linguistics, 2009, pp. 126–128. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1599503.1599546

[5] X. F. Liu, E. Khudkhudia, L. Wen, V. Sajja, and M. C. Leu, "An intelligent computational argumentation system for supporting collaborative software development decision making," in Artificial Intelligence Applications for Improved Software Engineering Development: New Prospects, ser. Advances in Computational Intelligence and Robotics, F. Meziane and S. Vadera, Eds. IGI Global, 2010, pp. 167 – 180. [Online]. Available: http://services.igi-global.com/ resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-758-4

[6] M. Klein, P. Spada, and R. Calabretta, "Enabling deliberations in a political party using large-scale argumentation: A preliminary report," in Proceedings of the 10th international conference on the design of cooperative systems, p. 17.

[7] A. C. B. Garcia and M. Klein, “Making sense of large-group discussion using automatically generated RST-based explanations,” SSRN Electronic Journal, 2015. [Online]. Available: http://www.ssrn. com/abstract=2554838

[8] Z. Papacharissi, “The virtual sphere: The internet as a public sphere,” New Media & Society, vol. 4, no. 1, pp. 9–27, 2002. [Online]. Available: https://doi.org/10.1177/14614440222226244

[9] E. Bakshy, S. Messing, and L. A. Adamic, “Exposure to ideologically diverse news and opinion on facebook,” Science, vol. 348, no. 6239, pp. 1130–1132, 2015. [Online]. Available: http://science.sciencemag. org/content/348/6239/1130

[10] A. Bessi, F. Zollo, M. Del Vicario, M. Puliga, A. Scala, G. Caldarelli, B. Uzzi, and W. Quattrociocchi, “Users polarization on facebook and youtube,” PLoS ONE, vol. 11, no. 8, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4994967/

[11] S. Flaxman, S. Goel, and J. M. Rao, “Filter bubbles, echo chambers, and online news consumption,” Public Opinion Quarterly, vol. 80, pp. 298–320, 2016. [Online]. Available: https://academic.oup.com/poq/ article/80/S1/298/2223402

[12] W. Quattrociocchi, A. Scala, and C. R. Sunstein, “Echo chambers on facebook,” 2016. [Online]. Available: https://papers.ssrn.com/abstract= 2795110

[13] S. Hong and S. H. Kim, “Political polarization on twitter: Implications for the use of social media in digital governments,” Government Information Quarterly, vol. 33, no. 4, pp. 777–782, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/ pii/S0740624X16300375

[14] C. R. Sunstein, #Republic : Divided Democracy in the Age of Social Media. Princeton University Press, 2017. [Online]. Available: http://0-search.ebscohost.com.library.uark.edu/login.aspx?direct= true&db=nlebk&AN=1431815&site=ehost-live&scope=site

[15] K. Coe, K. Kenski, and S. A. Rains, “Online and uncivil? patterns and determinants of incivility in newspaper website comments,” Journal of Communication, vol. 64, no. 4, pp. 658–679, 2014. [Online]. Available: https://academic.oup.com/joc/article/64/4/658/4086037

[16] I. Rowe, “Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms,” Journal of Broadcasting & Electronic Media, vol. 59, no. 4, pp. 539–555, 2015. [Online]. Available: http://0-search.ebscohost.com.library.uark.edu/login.aspx?direct= true&db=cms&AN=111289777&site=ehost-live&scope=site

[17] N. J. Stroud, J. M. Scacco, A. Muddiman, and A. L. Curry, “Changing deliberative norms on news organizations’ facebook sites,” Journal of Computer-Mediated Communication, vol. 20, no. 2, pp. 188–203, 2015. [Online]. Available: https://academic.oup.com/jcmc/article/20/2/ 188/4067559

[18] R. Medaglia and D. Zhu, “Public deliberation on government-managed social media: A study on weibo users in china,” Government Information Quarterly, vol. 34, no. 3, pp. 533–544, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0740624X16301903

[19] K. Esau, D. Friess, and C. Eilders, “Design matters! an empirical analysis of online deliberation on different news platforms,” Policy & Internet, vol. 9, no. 3, pp. 321–342, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.154

[20] D. Friess and C. Eilders, “A systematic review of online deliberation research,” Policy & Internet, vol. 7, no. 3, pp. 319–339, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.95

[21] P. Aragon, V. Gómez, and A. Kaltenbrunner, “Detecting platform ´ effects in online discussions,” Policy & Internet, vol. 9, no. 4, pp. 420–443, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/ abs/10.1002/poi3.158

[22] P. S. Hart and E. C. Nisbet, “Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies,” Communication Research, vol. 39, no. 6, pp. 701–723, 2012. [Online]. Available: https://doi.org/10.1177/0093650211416646

[23] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini, “Political polarization on twitter,” in Fifth International AAAI Conference on Weblogs and Social Media, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ ICWSM11/paper/view/2847

[24] S. Sobieraj and J. M. Berry, “From incivility to outrage: Political discourse in blogs, talk radio, and cable news,” Political Communication, vol. 28, no. 1, pp. 19–41, 2011. [Online]. Available: https: //doi.org/10.1080/10584609.2010.542360

[25] N. J. Stroud, “Polarization and partisan selective exposure,” Journal of Communication, vol. 60, no. 3, pp. 556–576, 2010. [Online]. Available: https://academic.oup.com/joc/article/60/3/556/4098564

[26] J. K. Lee, J. Choi, C. Kim, and Y. Kim, “Social media, network heterogeneity, and opinion polarization,” Journal of Communication, vol. 64, no. 4, pp. 702–722, 2014. [Online]. Available: https: //onlinelibrary.wiley.com/doi/abs/10.1111/jcom.12077

[27] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito, "Measuring political polarization: Twitter shows the two sides of venezuela," Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 25, no. 3, p. 033114, 2015. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.4913758

[28] C.-h. Chan and K.-w. Fu, "The relationship between cyberbalkanization and opinion polarization: Time-series analysis on facebook pages and opinion polls during the hong kong occupy movement and the associated debate on political reform," Journal of Computer-Mediated Communication, vol. 22, no. 5, pp. 266–283, 2017. [Online]. Available: https://academic.oup.com/jcmc/article/22/5/266/4666427

[29] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, "Exposure to opposing views on social media can increase political polarization," Proceedings of the National Academy of Sciences, vol. 115, no. 37, pp. 9216–9221, 2018. [Online]. Available: https://www.pnas.org/content/115/37/9216

[30] D. Demszky, N. Garg, R. Voigt, J. Zou, M. Gentzkow, J. Shapiro, and D. Jurafsky, "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings," in 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019. [Online]. Available: https://nlp.stanford.edu/pubs/demszky2019analyzing.pdf

[31] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in Proceedings of the 3rd International Workshop on Link Discovery, ser. LinkKDD '05. ACM, 2005, pp. 36–43, event-place: Chicago, Illinois. [Online]. Available: http://doi.acm.org/10.1145/1134271.1134277

[32] P. C. Guerra, W. M. Jr, C. Cardie, and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," in Seventh International AAAI Conference on Weblogs and Social Media, 2013. [Online]. Available: https: //www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6104

[33] V. Price, J. N. Cappella, and L. Nir, "Does disagreement contribute to more deliberative opinion?" Political Communication, vol. 19, no. 1, pp. 95–112, 2002. [Online]. Available: https://doi.org/10.1080/ 105846002317246506

[34] J. Barabas, "How deliberation affects policy opinions," American Political Science Review, vol. 98, no. 4, pp. 687–701, 2004. [Online]. Available: https: //www.cambridge.org/core/journals/american-political-science-review/ article/how-deliberation-affects-policy-opinions/ 60BBEAE885EB4EF99D81913375176743

[35] S. Iyengar, R. C. Luskin, and J. S. Fishkin, "Deliberative preferences in the presidential nomination campaign: Evidence from an online deliberative poll," 2005.

[36] D. C. Mutz and J. J. Mondak, “The workplace as a context for cross-cutting political discourse,” The Journal of Politics, vol. 68, no. 1, pp. 140–155, 2006. [Online]. Available: https://www.journals.uchicago. edu/doi/abs/10.1111/j.1468-2508.2006.00376.x

[37] K. Gronlund, K. Strandberg and, and S. Himmelroos, “The challenge ¨ of deliberative democracy online – a comparison of face-to-face and virtual experiments in citizen deliberation,” Information Polity, vol. 14, no. 3, pp. 187–201, 2009. [Online]. Available: https://www.medra.org/ servlet/aliasResolver?alias=iospress&doi=10.3233/IP-2009-0182

[38] J. S. Fishkin and R. C. Luskin, “Experimenting with a democratic ideal: Deliberative polling and public opinion,” Acta Politica, vol. 40, no. 3, pp. 284–298, 2005. [Online]. Available: https: //doi.org/10.1057/palgrave.ap.5500121

[39] V. R. K. Garimella and I. Weber, “A long-term analysis of polarization on twitter,” in Eleventh International AAAI Conference on Web and Social Media, 2017. [Online]. Available: https: //www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15592

[40] G. Olivares, J. P. Cardenas, J. C. Losada, and J. Borondo, ´ “Opinion polarization during a dichotomous electoral process,” Complexity, vol. 2019, p. 5854037, 2019. [Online]. Available: https://doi.org/10.1155/2019/5854037

[41] R. S. Arvapally and X. F. Liu, “Polarization assessment in an intelligent argumentation system using fuzzy clustering algorithm for collaborative decision support,” Argument & Computation, vol. 4, no. 3, pp. 181–208, 2013. [Online]. Available: https://doi.org/10.1080/19462166. 2013.794163

[42] M. Klein, “How to harvest collective wisdom on complex problems : An introduction to the MIT deliberatorium,” 2011.

[43] J.-M. Esteban and D. Ray, “On the measurement of polarization,” Econometrica, vol. 62, no. 4, pp. 819–851, 1994. [Online]. Available: https://www.jstor.org/stable/2951734

[44] A. Flache and M. W. Macy, “Small worlds and cultural polarization,” The Journal of Mathematical Sociology, vol. 35, no. 1, pp. 146–176, 2011. [Online]. Available: https://doi.org/10.1080/0022250X.2010.532261

[45] P. DiMaggio, J. Evans, and B. Bryson, “Have american’s social attitudes become more polarized?” American Journal of Sociology, vol. 102, no. 3, pp. 690–755, 1996. [Online]. Available: https: //www.journals.uchicago.edu/doi/abs/10.1086/230995

[46] D. Baldassarri and A. Gelman, “Partisans without constraint: Political polarization and trends in american public opinion,” American Journal of Sociology, vol. 114, no. 2, pp. 408–446, 2008. [Online]. Available: https://www.journals.uchicago.edu/doi/abs/10.1086/590649

[47] S. Iyengar, G. Sood, and Y. Lelkes, “Affect, not Ideology a social identity perspective on polarization,” Public Opinion Quarterly, vol. 76, no. 3, pp. 405–431, 2012. [Online]. Available: https: //academic.oup.com/poq/article/76/3/405/1894274

[48] L. Mason, ““i disrespectfully agree”: The differential effects of partisan sorting on social and issue polarization,” American Journal of Political Science, vol. 59, no. 1, pp. 128–145, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12089

[49] Y. Lelkes, “Mass polarization: Manifestations and measurements,” Public Opinion Quarterly, vol. 80, pp. 392–410, 2016. [Online]. Available: https://academic.oup.com/poq/article/80/S1/392/2223374

[50] E. Suhay, E. Bello-Pardo, and B. Maurer, “The polarizing effects of online partisan criticism: Evidence from two experiments,” The International Journal of Press/Politics, vol. 23, no. 1, pp. 95–115, 2018. [Online]. Available: https://doi.org/10.1177/1940161217740697

[51] A. Matakos, E. Terzi, and P. Tsaparas, “Measuring and moderating opinion polarization in social networks,” Data Mining and Knowledge Discovery, vol. 31, no. 5, pp. 1480–1505, 2017. [Online]. Available: https://doi.org/10.1007/s10618-017-0527-9

[52] F. Schweitzer, T. Krivachy, and D. Garcia, “How emotions drive opinion polarization: An agent-based model,” arXiv:1908.11623 [nlin, physics:physics, q-bio], 2019. [Online]. Available: http://arxiv.org/abs/ 1908.11623

[53] S. Banisch and E. Olbrich, “Opinion polarization by learning from social feedback,” The Journal of Mathematical Sociology, vol. 43, no. 2, pp. 76–103, 2019. [Online]. Available: https: //doi.org/10.1080/0022250X.2018.1517761

[54] A. Gruzd and J. Roy, “Investigating political polarization on twitter: A canadian perspective,” Policy & Internet, vol. 6, no. 1, pp. 28–45, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10. 1002/1944-2866.POI354

[55] J. Schmitt, “How to measure ideological polarization in party systems,” ECPR Graduate Student Conference, p. 31, 2016.

[56] M. Del Vicario, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, “Modeling confirmation bias and polarization,” Scientific Reports, vol. 7, p. 40391, 2017. [Online]. Available: https://www.nature.com/articles/srep40391

[57] A. Bramson, P. Grim, D. J. Singer, S. Fisher, W. Berger, G. Sack, and C. Flocken, “Disambiguation of social polarization concepts and measures,” The Journal of Mathematical Sociology, vol. 40, no. 2, pp. 80–111, 2016. [Online]. Available: http: //www.tandfonline.com/doi/full/10.1080/0022250X.2016.1147443

[58] A. Bramson, P. Grim, D. J. Singer, W. J. Berger, G. Sack, S. Fisher, C. Flocken, and B. Holman, "Understanding polarization: Meanings, measures, and model evaluation," Philosophy of Science, vol. 84, no. 1, pp. 115–159, 2017, publisher: The University of Chicago Press. [Online]. Available: https://www.journals.uchicago.edu/doi/abs/10.1086/688938

[59] B. Apouey, "Measuring health polarization with self-assessed health data," Health Economics, vol. 16, no. 9, pp. 875–894, 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.1284

[60] T. S. Clark, "Measuring ideological polarization on the united states supreme court," Political Research Quarterly, vol. 62, no. 1, pp. 146–157, 2009. [Online]. Available: https://doi.org/10.1177/1065912908314652

[61] Z. Maoz and Z. Somer-Topcu, "Political polarization and cabinet stability in multiparty systems: A social networks analysis of european parliaments, 1945–98," British Journal of Political Science, vol. 40, no. 4, pp. 805–833, 2010.

[62] P. Rehm and T. Reilly, "United we stand: Constituency homogeneity and comparative party polarization," Electoral Studies, vol. 29, no. 1, pp. 40–53, 2010. [Online]. Available: http://www.sciencedirect.com/ science/article/pii/S0261379409000419

[63] S. Gabbriellini and P. Torroni, "A new framework for ABMs based on argumentative reasoning," in Advances in Social Simulation, ser. Advances in Intelligent Systems and Computing, B. Kaminski and ´ G. Koloch, Eds. Springer Berlin Heidelberg, 2014, pp. 25–36.

[64] R. M. Alvarez and J. Brehm, Hard Choices, Easy Answers: Values, Information, and American Public Opinion. Princeton University Press, 2002, google-Books-ID: dWh10zFdjNUC.

[65] T. J. Rudolph and E. Popp, "An information processing theory of ambivalence," Political Psychology, vol. 28, no. 5, pp. 563–585, 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10. 1111/j.1467-9221.2007.00590.x

[66] L. Keele and J. Wolak, "Contextual sources of ambivalence," Political Psychology, vol. 29, no. 5, pp. 653–673, 2008. [Online]. Available: https: //onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9221.2008.00659.x

[67] S. Yardi and D. Boyd, "Dynamic debates: An analysis of group polarization over time on twitter," Bulletin of Science, Technology & Society, vol. 30, no. 5, pp. 316–327, 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0270467610380011

[68] E. J. Olsson, “A bayesian simulation model of group deliberation and polarization,” in Bayesian Argumentation, F. Zenker, Ed. Springer Netherlands, 2013, pp. 113–133. [Online]. Available: http: //link.springer.com/10.1007/978-94-007-5357-0 6

[69] M. D. Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, “Echo chambers: Emotional contagion and group polarization on facebook,” Scientific Reports, vol. 6, no. 1, pp. 1–12, 2016. [Online]. Available: https://www.nature.com/articles/srep37825

[70] C. Proietti, “Understanding group polarization with bipolar argumentation frameworks,” in Frontiers in Artificial Intelligence and Applications, ser. Frontiers in Artificial Intelligence and Applications, vol. 287. IOS Press, 2016, pp. 41 – 52.

[71] J. H. Evans, “Have americans’ attitudes become more polarized?—an update*,” Social Science Quarterly, vol. 84, no. 1, pp. 71–90, 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10. 1111/1540-6237.8401005

[72] D. Baldassarri and P. Bearman, “Dynamics of political polarization,” American Sociological Review, vol. 72, no. 5, pp. 784–811, 2007. [Online]. Available: https://doi.org/10.1177/000312240707200507

[73] M. McCombs, “The agenda-setting role of the mass media,” in Proceedings of the 2002 Conference of Mass Media Economics. London School of Economics, 2002.

[74] A. K. McCallum, “MALLET: A machine learning for language toolkit.” 2002. [Online]. Available: http://mallet.cs.umass.edu

[75] D. G. Myers and H. Lamm, “The group polarization phenomenon,” Psychological Bulletin, vol. 83, no. 4, pp. 602–627, 1976.

Chapter 4: An Opinion Diversity Enhanced Social Connection Recommendation Re-ranking Method based on Opinion Distance in Cyber Argumentation with Social Networking

4.1 Abstract

The quality of crowd wisdom extracted from online communities decreases as the community becomes less ideologically diverse, which is an issue in many online spaces. One cause of this decline is that users tend not to engage with diverse, idea-challenging content that contrasts their prior opinions. However, they do tend to engage with content endorsed by their social connections, even if it goes against their personal opinion. Thus, by increasing the diversity of opinion in a user's social network, they will likely engage with more diverse content. We are developing a cyber argumentation system with social networking and present a social connection recommendation re-ranking method that promotes opinion diversity. We use artificial intelligence and data mining techniques to mine and analyze user opinions from argumentation data on important issues, then use furthest opinion distance to re-rank the recommendations. Our method is designed to easily integrate with existing social connection recommenders, which preserves platform specific criteria. We compare the opinion diversity of recommendations from five types of social connection recommendation methods, with and without our re-ranking method, on a large empirical dataset. Our results show that our method improves the recommended diversity by around 15% for five existing social connection recommendation methods, while only reordering around 50% of the initial social connection recommendations.

4.2 Introduction

Online social platforms, like Facebook and Twitter, offer the unique capability to connect many different types of people together and facilitate interaction between them. As such, they are excellent sources for the measurement and analysis of collective intelligence and crowd wisdom. However, because these social networking sites are oriented around social connections, their design often yields low-quality discussions of complex issues. In these sites, discussions are disorganized because users are fragmented into their social circles that create disconnected, isolated discussions, which makes it very difficult to comprehend the collective opinions of all of the users. Additionally, these systems do not encourage well-considered, factually based argumentation [1] and instead allow misinformation and conspiracy theories to circulate. For these reasons, existing social networking sites are not effective at facilitating large-scale deliberations.

To address this issue, it is desirable to develop an issue-oriented cyber argumentation platform with social networking capabilities. Such a system will center on discussions of issues, encourage high-quality contributions, and allow easier analysis of collective intelligence and crowd wisdom, while also allowing social networking activities, like making connections and sharing content. However, introducing social networking into cyber argumentation raises new issues. One important issue, which could affect the argumentation as a whole, is ideological polarization caused by echo chambers.

Recently, it's been observed that several online social networking spaces have become increasingly less ideologically diverse and more polarized, especially when the content is related to politics. Polarization and the lack of ideological diversity has been shown to detrimentally affect the quality of crowd-wisdom [24], which, given the desire to use crowd wisdom and collective intelligence in policy-related areas (such as e-government), poses a potentially large threat.

Americans in particular have become more ideologically polarized in the past few decades. While this has been observed in both online and offline contexts, many have argued that online social networking has played an outsized role in furthering polarization and tension between ideological groups [2], contributing to decreased civil public discourse [3] and polarized attitudes at the individual level [4]. Many researchers allege that social networking sites have given rise to echo chambers, where users' social communities are mostly populated by like-minded peers, contributing significantly to online opinion polarization and other detrimental phenomena previously mentioned [5, 6]. These echo chambers allow users to easily avoid ideas and opinions that do not fit their previously held worldview and opinion, even though people typically express a desire to hear diverse views [7]. Thus, many researchers have explored different ways to motivate users to engage with more diverse ideas and content.

One approach to solve this problem is to create diversity-enhanced news recommendation systems [13,14], that expose the user to different perspectives and opinions. Some evidence suggests, however, that simply exposing users to diverse content will not necessarily encourage them to engage with it; users' personal preferences typically play a larger role in content selection than algorithmic recommendations [8]. Thus, how to encourage users to engage with diverse ideas is still an open question.

Dynamics in social networks might be the key to increasing ideologically diverse content/idea engagement. Barberá found that weak social ties on Twitter tended to expose users to more diverse content [9]. Messing and Westwood showed that social endorsements were stronger predictors of news selection than partisan sentiment [10]. This evidence suggest that users are more likely to engage with content/ideas endorsed or generated by their social connections. Considering that diverse networks containing weak ties tend to recommend diverse content, it follows that if a

user’s social network is diverse in opinions, then the content and opinions they share will likely be more diverse. Therefore, we argue that diversifying a user’s social network in terms of opinion will reduce the likelihood of being caught in an echo chamber and will expose them to diverse ideas and opinions of their social connections.

In this paper, we propose a social connection recommendation re-ranking method that specifically encourages opinion diversity using a furthest opinion distance approach in cyber argumentation with social networking. First, the system quantifies a user's opinion from argumentation data on several important issues using cognitive computing/artificial intelligence and data mining techniques into opinion vectors. Then, these vectors are used to re-rank the incoming recommendation list using furthest cosine distance with the target user.

Because social connection recommendation is very specific to the context of the online platform, the re-ranking method does a last pass reordering of recommendations made by the platform-specific, native connection recommending system. This allows our method to increase diversity while maintaining the other recommendation criteria. We used empirical data collected using our argumentation platform, the Intelligent Cyber Argumentation System (ICAS), to demonstrate that our re-ranking method increases the opinion diversity of its recommendations, as compared recommender methods without re-ranking. Results show that our method improves the recommended diversity for all the examined recommender systems by around 15%, while only reordering around 50% of the initial recommendations. Our method is designed to be easily integrated into existing recommendation systems, while still providing improvements to the opinion diversity of the recommendations.

Our contributions are as follows:

- We propose a social recommendation re-ranking method that encourages opinion diversity in its recommendations in cyber argumentation with social networking.
- We demonstrate on an empirical dataset that our method increases the opinion diversity of recommendation results for several types of social connection recommendation methods by around 15% by reordering 50% of the recommendations

4.3 Related Work

4.3.1 Social Recommendation Systems

Social media sites often contain an overwhelming amount of user accounts and content that is very difficult for users to navigate. To assist with content and user discovery, Social Recommender Systems (SRS) focus on recommending content and people from social media. Unlike other typical recommender systems, SRS contend with social media data that is often unstructured, sparse, and contains many different types of data (e.g., images, videos, etc) [11]. SRS encapsulates both social media content recommendation systems, like news recommendation systems, and social connection recommendation systems, also called friend recommender systems.

4.3.2 Social Connection Recommendation

Social connection recommenders have been thoroughly researched for many years. The goal of social connection recommendation is to recommend other users with whom the target user will want to connect. The meaning of the connection is very dependent on the function of the social networking platform and how their connections are designed. For example, LinkedIn is a website for professionals and focuses on work related relationships, while Pinterest is an information sharing site for hobbyists. In addition to context, the construction of the connections also affects

use. Twitter, for example, has asymmetric relationships, which encourages users to follow famous people and celebrities; whereas Facebook has symmetric relationships which encourages users to friend people who they know. Both the types of connections and the context of the social media platforms are considered in the social connection recommenders. Therefore, social connection recommenders are very specific to their platform and can consider several thousand features when making recommendations [11].

One popular technique for injecting diversity and novelty into recommendations is to re-rank initial recommendation lists [12], which allows diversity to be incorporated into the system while maintaining the initial recommendation criteria and without interrupting the workflow of the sophisticated pre-existing recommendation system. We adopted this approach for our work in this paper.

4.3.3 Diversity in Social Connection Recommendation

Diversity aware recommendation systems have become a popular research area. The goal is to increase the amount of diversity in the recommendation results so that users are exposed to a wider range of content, which is generally desirable for users [12]. While many diversity enhancing social content systems have been developed [13,14], social connection recommendation systems have not seen as much investigation. Both diversity in terms of interests [15] and information [16] have been applied to social connection recommendation, but, to our knowledge, opinion diversity has not been developed.

4.3.4 Cyber Argumentation Systems

Most online deliberation takes place on social media platforms or in online forums. While these platforms are very popular, they do not effectively facilitate large-scale debates. Often the

discussions are fragmented, difficult to comprehend, and require a lot of effort to analyze. As a result, many researchers have looked at cyber argumentation systems to serve as online debate platforms instead.

Cyber argumentation systems assist the facilitation of large-scale online discussions/debates. These systems enhance deliberation in a variety of ways. As opposed to social media and forums, cyber argumentation systems typically employ explicit argumentation frameworks which provide structure to discussions and lead to higher-quality reasoning and debate. Computer-Supported Argument Visualization (CSAV) systems improve argumentation by presenting the various arguments in a discussion in an intuitive and easy to understand manner [17]. Educational tools seek to teach students different strategies for engaging in productive online argumentation by providing instructional scaffolding, which guide students to produce better reasoning during discussions [18]. More complex tools, such as the Deliberatorium [1] and ICAS, have integrated analytical models that report various phenomena that are occurring in the discussions. Unlike social media analysis, these models are integrated directly into the argumentation systems and leverage the underlying argumentation framework's structure to effectively analyze the different phenomena, such as group-think [19], opinion consensus [20], and position polarization [21]. None of these systems, to our knowledge, have attempted to integrate social networking into them.

4.4 System Architecture

4.4.1 Conceptual Structure of Cyber Argumentation with Social Networking

In this section we briefly describe the conceptual structure behind the argumentation system with social networking. The main idea is to combine issue-centric argumentation

discussions with social networking data, from the argumentation system and/or other social network sites like Facebook or Twitter, as shown in Figure 4-1.

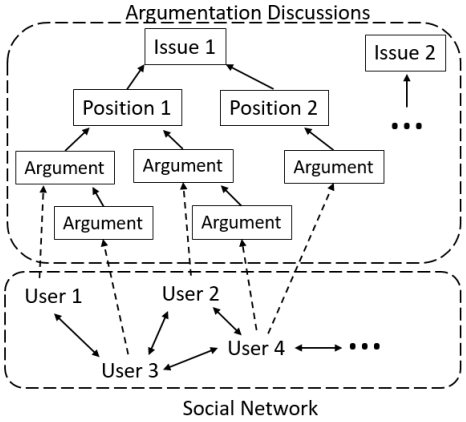


Figure 4-1 Cyber argumentation with social networking conceptual design. The user connections (bottom) comprise the social network, while the top elements are the argumentation discussions. The dotted lines represents an authorship relationship.

In cyber argumentation, users make arguments to argue for or against proposed positions or stances about a given issue. Users are also able to make arguments attacking or supporting other user's arguments. In these systems it is assumed that users do not have any explicit relationships with one another during the discussions. However, in the real world, argumentation often takes place between multiple people with some relation to one another, which affects the underlying debate. Debates between friends are often very different than debates between strangers. Users with certain types of relationships might feel more or less inclined to support or attack one another's ideas. For example, an employee would feel reluctant to attack the ideas of their superior, even if they had a logical reason to do so. Even in an online setting, like social networking sites, the relationships between users often dictate the nature of their interactions with one another. Incorporating social networking into cyber argumentation, by allowing network connections to be made in the system or imported from external sources like Facebook and Twitter, allows us to examine the different effects that social relationships have on the argumentation process.

4.4.2 Diversity Enhanced Social Connection Re-ranking Method

The re-ranking method is designed to fit on top of existing recommendation architectures. This has two benefits: first, this allows for the re-ranking method to be applied to existing systems without major changes, and second, it maintains the quality of the initial recommendation lists by keeping the platform-optimized system in place. This design is useful because if we use an external social networking site to handle the user social network, we can build our diversity model on top of it. Figure 4-2 describes the architecture of the integrated re-ranking system.

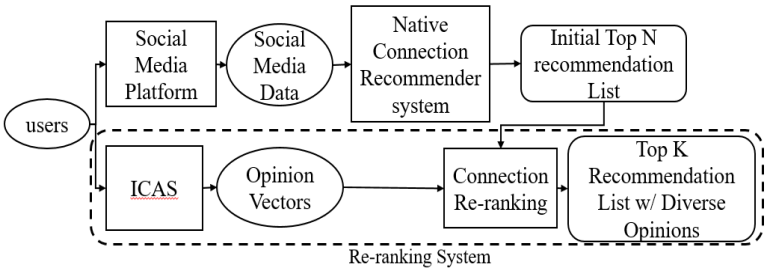


Figure 4-2 Framework for the Re-ranking System.

The top row of the framework make up the native social connection recommendation systems used in practice. For the purpose of this system, the recommendation system will operate as normal to produce a top N list of recommendations using the criteria built into those systems. These systems are assumed to be mature and produce high quality recommendations.

The bottom row of the framework (ICAS, Opinion Vectors, Connection Re-ranking, Top K Recommendation List with Diverse Opinions) make up the added steps for the re-ranking method. The following subsections will describe the different elements in the re-ranking method and how they interact with each other.

4.4.2.1 Mining user opinion vectors:

From the argumentation data, we need to derive the opinion of each user on the important issues they discuss. ICAS has built in models to automatically perform this step. The opinion mining model has been tested in previous research on multiple occasions with reasonable accuracy (see [20, 22] for examples).

Quantifying each user's opinions on the issues has two steps: first the user's opinion on each position is mined, then the opinions are formatted into the user's opinion vector.

4.4.2.2 Deriving User Opinion on a Position:

ICAS uses artificial intelligence and data mining to approximate each user's opinion for a given position in the discussion. This process examines each argument that the user posted under the target position and attempts to mine that user's overall agreement (for/against) towards the position. As mentioned in the overview, each argument in the discussion has a level of agreement associated with it, which indicates the user's opinion toward the parent argument/position that their argument is addressing. ICAS uses these agreement levels to mine a user's opinion toward each position in the discussion, by averaging all of a user's argument's agreement levels in the discussion sub-tree of the position.

Some arguments do not directly address the parent position, instead they argue for or against some other argument further down the position sub-tree. To derive the user's opinion, we need to know how these arguments relate to the root position. This is resolved by ICAS's cognitive computing component, its fuzzy logic reduction engine (see Chapter 1.1.3). An example of fuzzy logic reduction is shown in Figure 4-3.

Once the fuzzy logic engine has reduced all arguments for a given position to the first level of the sub-tree, each user's opinion toward the position can be approximated by averaging the agreement level values of all their arguments. If a user does not have any arguments for a position, their opinion value is defaulted to 0.

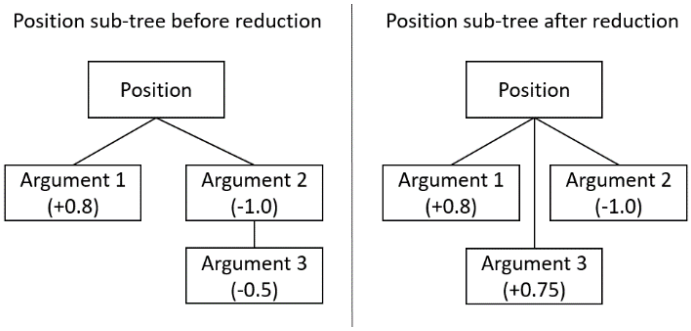


Figure 4-3 Left: A position sub-tree. Right: A position sub-tree after argument 3 has been reduced to the first level of the tree.

4.4.2.3 Forming Opinion Vectors

After a user's opinion on each of the different positions is calculated, they are concatenated together to form an opinion vector. Each element in the opinion vector Vu is the user's opinion on a corresponding position. So, viu is user u's opinion on position i. If issue A has three positions (i = 1, 2, 3), then the opinion vector for user u on that issue would be:

$$V_{u,A}=\{v_1^u,v_2^u,v_3^u\}$$

So, the combined opinion vector for a user across all issues is the concatenation of the user's opinion on each issue's set of positions.

$$V_u=(V_{u,A}, V_{u,B},\ldots)=\{v_1^u,v_2^u,v_3^u, v_4^u, \ldots, v_n^u\}$$

4.4.2.4 Connection Re-ranking

The next step of the re-ranking method is to re-rank the top N initial recommendations from the native recommender using furthest opinion distance from the target user. In our case, distance represents how different the target user's opinion is from the recommended connection.

We use cosine distance as the distance measurement, which measures the angle between two vectors. This is useful for two reasons: First, the orientation between two opinion vectors is more descriptive of differing opinion than a difference in vector intensity. If one user strongly agrees with a position and another only moderately agrees, their vector's intensities will be different, but that doesn't necessarily mean their opinions differ very much. Second, because each opinion vector is the concatenation of all of the issues' positions, it is likely that the dimensionality of the vectors will become very large. Cosine distance works very well in high dimensional space.

The cosine distance between two user's viewpoint vectors, Vu and Vw, is defined in (1).

$$\text{dist}(V_u, V_w) = 1 - \frac{\sum_{i=1}^{n} v_i^u * v_i^w}{\sqrt{\sum_{i=1}^{n} (v_i^u)^2} + \sqrt{\sum_{i=1}^{n} (v_i^w)^2}} \quad (1)$$

Once the distance value has been calculated for each recommendation, the method re-ranks the recommendations by their cosine distance in descending order and recommends the top K users in their sorted list, where K is less than or equal to N (the number of initial recommendations).

4.5 Experiments

We tested the re-ranking method against many fundamental social connection recommendation techniques in terms of opinion diversity on a large empirical dataset.

4.5.1 Empirical Data Description

An empirical study was conducted from April 10th, 2018, to May 4th, 2018, on a group of undergraduate students from an introductory level sociology class. The students were asked to participate in an online discussion of various topics relating to what they were learning in class using ICAS. The students were offered extra credit in the class for posting at least ten arguments

across any of the discussions for each issue. If they did not want to participate, then they were offered an alternative assignment for the extra credit points.

A total of 344 students registered with the system, of which 335 discussed four issues, with each issue having four positions under them. Over the course of the twenty-five days, the participants posted more than 10000 arguments across the 16 positions.

4.5.1.1 Student Social Network:

In addition to the deliberation, we also asked the students to enter in the names of students in the class who they knew. This gave us a sense of the students' social network within the class. Student's names were matched with their accounts to form a social network in the database. Of the 335 students asked, 193 answered the question with a list of students they knew. Of the total 335 users, 101 users did not answer the question nor were listed as a connection. The average degree of the nodes in the graph was 1.74.

The student's physical social network is a symmetric network, where users mutually know one another, similar to social media sites such as Facebook and LinkedIn. This physical network was used in place of an online social network.

4.5.1.2 Positive Interaction Graph:

From the deliberation, we derived a positive interaction network between all the users. This network is an undirected graph with no self-loops. Each edge stores a value greater than 0 representing the number of positive interactions between the users. The graph was constructed such that given two users, u and v, an edge exists between u and v if and only if v makes an argument supporting u (positive level of agreement) or u makes an argument supporting v. Of the 335 users in the data, 299 of them had a positive interaction with another user.

4.5.2 Social Connection Recommenders

We wanted to test our re-ranking method with many different recommendation techniques, but while there are many recommendation methods used in practice, our ability to access the necessary data to implement them was limited. So, we instead tested against many fundamental techniques that make the backbone of social connection recommenders.

We tested our re-ranking method on five social connection recommendation techniques to measure the impact of re-ranking the lists on opinion diversity. Of the algorithms we tested against, the first four were taken from (or adapted to best fit) the examined methods in [23]. These four methods make up the fundamental approaches to connection recommendation.

The following subsections describe the different scoring techniques used by each algorithm. For each described algorithm, all users are scored using the scoring functions and sorted into ranked lists.

4.5.2.1 Friend of Friend (FoF)

The friend of a friend technique recommends connections to users based on their distance in the social network and is very popular among network-based recommenders. In our implementation, the similarity score between two users is the inverse of the shortest distance between the users multiplied by the number of paths (at that distance) between them. Users with a distance of one are already friends and are not counted.

$$S(u,v)=\left(1/D_{u,v}\right) * P_{u,v} \quad (2)$$

Equation (2) describes scoring function S, where Duv is the minimum distance between nodes u and v and Puv is the number of paths at distance D.

Since all users did not have network connections in the social network graph, not all users were able to receive recommendations. In our dataset 44% of users were able to receive a recommendation.

4.5.2.2 Friend of Friend plus Interactions (FoF+I)

This algorithm is an extension of the FoF algorithm. Like the FoF algorithm we find the social network paths between a source user u and their candidate user v, however, we also consider the positive interactions between the users. This method scores two users, u and v, by combining the score from FoF (S(u,v)) and the number of positive interactions between u and v (P(u,v)) in the positive interaction graph using (3).

$$T(u,v)=(S(u,v)+1) * (P(u,v)+1) \quad (3)$$

The method allowed recommendations to be made for users who were connected in either the social network graph or the positive interaction graph. In our dataset, we were able to make at least one recommendation to 73% of users.

4.5.2.3 Content Similarity (CS)

This technique considers the similarity of text posts in the arguments made by users. The idea is that if two users are posting with similar language, then they are likely discussing the same ideas and should be recommended as friends.

First, each user had all their posts combined into a single document. The user's document is then tokenized and weighted using tf-idf, producing a weighted content vector for each user that represented all their posts. To make a recommendation, the target user u would rank their cosine similarity of their content vector, U, with each other user v's content vector, V.

4.5.2.4 Content plus Link (C+L)

In this method, the algorithm incorporates content similarity and network information. This method is a combination of Content similarity and FoF+I. Like in content similarity, each user's weighted content vector is derived using the method previously described. However, when calculating similarity, if a user has a 2-hop connection in the social network graph or if they have an edge connection in the positive interaction graph with a user v, then the cosine similarity value of their content vectors is boosted by 50%. This method is adapted from the CplusL method from [23].

4.5.2.5 Random Match (Rand)

This method randomly recommends connections for users based on no criteria. Each username is placed into an array, then when a recommendation needs to be made, the technique generates a random index value and recommends that user. This method helps serve as a baseline.

4.5.3 Analysis Metrics

We want to analyze our method in two ways. First we want to confirm that our re-ranking method does its basic functionality of finding connections who have different opinions from the user. Second, we want to measure how much the recommendations would improve the network diversity. Since we are in continuous space standard diversity measurements did not apply, instead we used two distance based measurements for the diversity of the recommendation lists: normalized average distance from target (NADT) and normalized average distance from network (NADN).

4.5.3.1 Normalized Average Distance from Target (NADT):

The NADT is the normalized average Manhattan distance between the target user's opinion vector and the opinion vector of each user in the top K recommended list. This measurement tells us whether or not the re-ranking system is able to optimize the initial recommendations using opinion distance, and therefore contributing something to the process. An increase in NADT when the re-ranking method is applied would indicate that the re-ranking step is able to optimize the given list, while no increase would indicate that the re-ranking method is not optimizing for opinion distance.

$$\mathrm{ADT}(U, R)=\frac{1}{|U|}\sum_{\mathbf{u}\in U}\left(\frac{1}{|R_u|}\sum_{i\in R_u} d(i, \mathbf{u})\right) \tag{4}$$

Equation (4) describes the measurement, where U is the set of user opinion vectors, K is the set of top K recommendation sets for each user, Ru is the recommended set for user u, u is user u's option vector, and d is the Manhattan distance function. ADT in the results section is normalized using min-max normalization.

4.5.3.2 Normalized Average Distance from Network (NADN):

The NADN is the normalized average distance a recommendation is to each of the target user's network connections (1-hop). This measurement indicates how much opinion diversity the recommendation would add to the user if they accepted the connection. A recommendation is considered diverse if it is far away from existing network connections in terms of opinion. An increase in NADN for the re-ranked lists would indicate that the re-ranking method is improving the opinion diversity of the user's social network, while no increase would indicate that the opinion diversity is not being affected.

For each recommendation, the average Manhattan distance was calculated against each network connection. Then, that average distance was averaged across all recommendations in the top K recommendations list. Lastly, the averaged distance for the K recommendations was averaged across all users who received at least one recommendation. The average value is normalized using min-max normalization to produce the NADN.

4.6 Results

For each recommender, each user was recommended a list of connections with at most thirty recommendations (N=30). However, due to some of the limitations of the data, not all the recommendation algorithms produced recommendations for each user, and they were not always of length 30. Table 4.1 describes the percentage of users who received at least one recommendation (called hit rate) and the average recommendations lengths for each user who did receive at least one recommendation (called Avg. Recs per hit).

Table 4.1: The hit rate and average recommendations per hit for each recommendation algorithm for N = 30.

Algorithm	**Hit Rate**	**Avg. Recs per hit.**
CS	92%	30.00
C+L	92%	30.00
FoF	44%	26.56
FoF+I	73%	20.70
Rand	100%	30.00

To test the effectiveness of our re-ranking technique, we measured the NADT and NADN on the originally ranked list and the re-ranked recommendation lists for each recommendation algorithm. Each algorithm produced a recommendation list of length 30, then the re-ranking method re-ranked the lists. Then we evaluated the top 5 recommendations for both the original

output lists and the re-ranked lists. Figures 4-4 and 4-5 describe the NATD and NADN for each technique respectably.

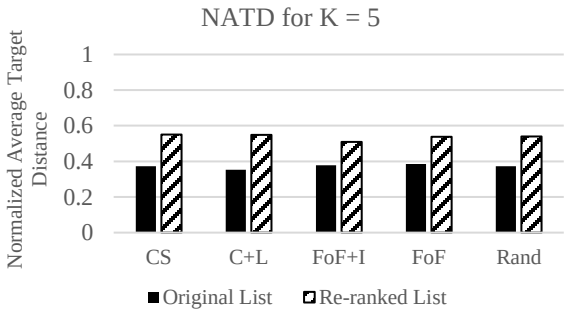


Figure 4-4 NATD for the various recommender algorithms.

Our re-ranked method improved NADT for every recommendation technique. This was the expected result because our re-ranking technique maximizes opinion distance from the target user. This result shows that our re-ranking method does actually affect the output list in terms of distance from the target. The NADT was improved for all of the re-ranked list, regardless of the underlying technique.

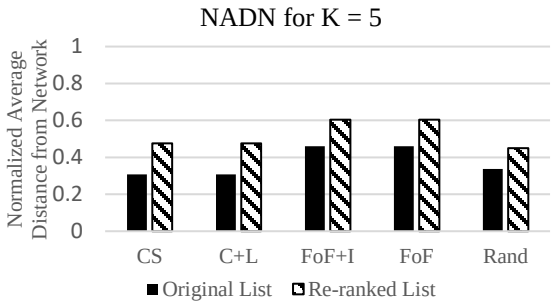


Figure 4-5 NADN for the various recommender algorithms.

Likewise, the re-ranked lists improved the NADC by around 15% for each of the recommenders. This outcome demonstrates that if the user accepts the recommendations, the re-ranked list would improve the diversity of the user's network more so than the un-ranked lists. The improvement in NADC was consistent across all the examined recommender types regardless of their approaches.

Table 4.2. Average Kendall Tau distance between original recommendation lists and the re-ranked list by recommender algorithm (K=30)

Algorithm	Average Kendall Tau Distance
CS	0.504
C+L	0.533
FoF	0.512
FoF+I	0.491
Rand	0.455

Lastly, we want to know how much the re-ranking method changes the original ranked list. We used Kendall's Tau distance to measure how different the re-ranked list is compared to the original list. Table 4.2 shows the average distances for each method for K = 30 lists. On average, the re-ranked lists had 50% of the recommendations reordered. Further examination found that this holds true for other values of k as well.

4.7 Discussion

From the results, we see that our re-ranking method improves the recommendation techniques in terms of distance from the target user and distance from the target user's network opinion by approximately 15%. Both network-based techniques (FoF, FoF+I) and content-based techniques (CS, C+L) improved at around the same rate, which implies that the underlying recommendation technique does not significantly impact the effectiveness of the re-ranking method.

Prioritizing opinion diversity in the recommended connection list increased the diversity of the recommendations but can come at the expense of the criteria of the platform-specific recommendation system. Further study revealed that the longer the recommendation list given to our re-ranking method, the greater the improvement in NADT and NADC. However, the longer the lists, the more likely that the recommendation from further down the original recommendation

list will be moved higher in the re-ranked list. On average, the re-ranking method reorders around half of the recommendations on the list. The exact balance between diversity and platform-specific criteria needs to be carefully managed.

4.7.1 Limitations

The effectiveness of our method relies greatly on the performance of the underlying recommendation systems that provide the initial list. We were unable to measure the user's satisfaction with the diverse recommendations. This research was conducted long after the empirical study had ended, so we were unable to get feedback on the quality of the recommendations. There is reason to believe that the quality would not be significantly altered, since our approach re-ranks the recommendations that are assumed to be already high quality.

Another limitation of the method relates to the composition of the opinion vectors. In this study, we constructed the opinion vectors from 16 positions related to four issues we deemed important. However, this may not be enough information to accurately reflect a user's opinion. Our method does not give any guidance on which issues or topics to include in quantifying the user's opinion and instead leaves that up to the researcher.

4.8 Conclusion

Ideological polarization and echo chambers pose a sizable threat to crowd wisdom quality and usefulness. We argue that by increasing the opinion diversity of a user's social network, they will likely engage with a much wider range of ideas than content recommender systems alone can provide. In this paper, we developed an innovative method for prioritizing diversity in social connection recommendations that improves diversity in terms of opinion distance from the target user and the target user's network. The method is separated into two steps, an opinion vectorization

step and a re-ranking step. The opinion quantification step in our method used artificial intelligence and data mining to mine the user's opinion. The second step re-ranks connections to prioritize opinion distance, which increases diversity and is computationally inexpensive. We tested our method on several recommendation techniques on a large empirical dataset and found it improved network diversity by around 15%. This technique is designed to be easy to integrate into existing recommendation workflows. Adoption of this technique will ideally diversify users' social networks and expose them to diverse and thought-provoking ideas.

4.9 References

[1] M. Klein, "How to Harvest Collective Wisdom on Complex Problems : An Introduction to the MIT Deliberatorium," 2011.

[2] C. R. Sunstein, #Republic : Divided Democracy in the Age of Social Media. Princeton: Princeton University Press, 2017.

[3] S. Sobieraj and J. M. Berry, "From Incivility to Outrage: Political Discourse in Blogs, Talk Radio, and Cable News," Political Communication, vol. 28, no. 1, pp. 19–41, Feb. 2011.

[4] N. J. Stroud, "Polarization and Partisan Selective Exposure," J Commun, vol. 60, no. 3, pp. 556–576, Sep. 2010.

[5] S. Flaxman, S. Goel, and J. M. Rao, "Filter Bubbles, Echo Chambers, and Online News Consumption," Public Opin Q, vol. 80, no. S1, pp. 298–320, Jan. 2016.

[6] Q. V. Liao and W.-T. Fu, "Beyond the Filter Bubble: Interactive Effects of Perceived Threat and Topic Involvement on Selective Exposure to Information," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2013, pp. 2359–2368.

[7] J. Stromer-Galley, "Diversity of Political Conversation on the Internet: Users' Perspectives," J Comput Mediat Commun, vol. 8, no. 3, Apr. 2003.

[8] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," Science, vol. 348, no. 6239, pp. 1130–1132, Jun. 2015.

[9] P. Barberá, “How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S.,” Job Market Paper, New York University, 2014.

[10] S. Messing and S. J. Westwood, “Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online,” Communication Research, vol. 41, no. 8, pp. 1042–1063, Dec. 2014.

[11] I. Guy, “Social Recommender Systems,” in Recommender Systems Handbook, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA: Springer US, 2015, pp. 511–543.

[12] P. Castells, N. J. Hurley, and S. Vargas, “Novelty and Diversity in Recommender Systems,” in Recommender Systems Handbook, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA: Springer US, 2015, pp. 881–918.

[13] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, “SCENE: A Scalable Two-stage Personalized News Recommendation System,” in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2011, pp. 125–134.

[14] S. A. Munson, D. X. Zhou, and P. Resnick, “Sidelines: An Algorithm for Increasing Diversity in News and Opinion Aggregators,” p. 8.

[15] H. Wu, V. Sorathia, and V. K. Prasanna, “When Diversity Meets Speciality: Friend Recommendation in Online Social Networks,” c ASE, p. 9, 2012.

[16] S. Wan, Y. Lan, J. Guo, C. Fan, and X. Cheng, “Informational friend recommendation in social media,” in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’13, Dublin, Ireland, 2013, p. 1045.

[17] S. Shum, “The Roots of Computer Supported Argument Visualization,” in Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, London: Springer-Verlag, 2003, pp. 3–24.

[18] C.-Y. Tsai, C.-N. Lin, W.-L. Shih, and P.-L. Wu, “The effect of online argumentation upon students’ pseudoscientific beliefs,” Computers & Education, vol. 80, pp. 187–197, Jan. 2015.

[19] M. Klein, “The CATALYST Deliberation Analytics Server,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2962524, Nov. 2015.

[20] S. Sigman and X. F. Liu, “A computational argumentation methodology for capturing and analyzing design rationale arising from multiple perspectives,” Information and Software Technology, vol. 45, no. 3, pp. 113–122, Mar. 2003.

[21] J. Sirrianni, X. Liu, and D. Adams, “Quantitative Modeling of Polarization in Online Intelligent Argumentation and Deliberation for Capturing Collective Intelligence,” in 2018 IEEE International Conference on Cognitive Computing (ICCC), 2018, pp. 57–64.

[22] X. (Frank) Liu, S. Raorane, and M. C. Leu, “A Web-based Intelligent Collaborative System for Engineering Design,” in Collaborative Product Design and Manufacturing Methodologies and Applications, W. D. Li, C. McMahon, S. K. Ong, and A. Y. C. Nee, Eds. London: Springer London, 2007, pp. 37–58.

[23] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, “Make New Friends, but Keep the Old: Recommending People on Social Networking Sites,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2009, pp. 201–210.

[24] H. Hong, Q. Du, G. Wang, W. Fan, and D. Xu, “Crowd Wisdom: The Impact of Opinion Diversity and Participant Independence on Crowd Performance,” AMCIS 2016 Proceedings, Aug. 2016.

Chapter 5: Agreement Prediction of Arguments in Cyber Argumentation for Detecting Stance Polarity and Intensity

5.1 Abstract

In online debates, users express different levels of agreement/disagreement with one another's arguments and ideas. Often levels of agreement/disagreement are implicit in the text, and must be predicted to analyze collective opinions. Existing stance detection methods predict the polarity of a post's stance toward a topic or post, but don't consider the stance's degree of intensity. We introduce a new research problem, stance polarity and intensity prediction in response relationships between posts. This problem is challenging because differences in stance intensity are often subtle and require nuanced language understanding. Cyber argumentation research has shown that incorporating both stance polarity and intensity data in online debates leads to better discussion analysis. We explore five different learning models: Ridge-M regression, Ridge-S regression, SVR-RF-R, pkudblab-PIP, and T-PAN-PIP for predicting stance polarity and intensity in argumentation. These models are evaluated using a new dataset for stance polarity and intensity prediction collected using a cyber argumentation platform. The SVR-RF-R model performs best for prediction of stance polarity with an accuracy of 70.43% and intensity with RMSE of 0.596. This work is the first to train models for predicting a post's stance polarity and intensity in one combined value in cyber argumentation with reasonably good accuracy.

5.2 Introduction

This In the digital age, many major online and social media and networking sites, such as Facebook, Twitter, and Wikipedia, have taken over as the new public forum for people to discuss and debate issues of national and international importance. With more participants in these debates

than ever before, the volume of unstructured discourse data continues to increase, and the need for automatic processing of this data becomes more prevalent. One important task in processing online debates is to automatically determine the different argumentative relationships between online posts in a discussion. These relationships typically consist of a stance polarity (i.e. whether a post is supporting, opposing, or is neutral toward another post) and the degree of intensity of the stance.

Automatically determining these types of relationships from a given text is a goal in both stance detection and argumentation mining research. Stance detection models seek to automatically determine a texts stance polarity (Favoring, Opposing, or Neutral) toward another text or topic based on its textual information [23]. Likewise, argumentation mining seeks to determine the stance relationship (Supporting, Attacking, or Neutral) between argumentation components in a text [33]. However, in both cases, attention is only payed to the stance polarity and very little attention has been payed to the intensity of the relationship. Some studies have tried to incorporate intensity into their predictions by expanding the number of classes to predict (Strongly For, For, Other, Against, and Strongly Against), however, this expansion lowered their classification performance considerably compared classification without intensity [29]. Thus, effective incorporation of stance intensity into stance classification remains an issue.

This is unfortunate, because research in Cyber Argumentation has shown that incorporating both stance polarity and intensity information into online discussions greatly improves the analysis of discussions and the various phenomena that arise during debate, including opinion polarization [28], and identifying outlier opinions [2], compared to using stance polarity alone. Thus, automatically identifying both the post's stance polarity and intensity, will allow these powerful analytical models to be used on unstructured debate data from platforms such as Twitter, Facebook, Wikipedia, comment threads, and online forums.

To that end, in this paper, we introduce a new research problem, stance polarity and intensity prediction in a responsive relationship between posts, which aims to predict a text's stance polarity and intensity which we combine into a single continuous agreement value. Given an online post A, which is replying to another online post B, we predict the stance polarity and intensity value of A towards B using A's (and sometimes B's) textual information. The stance polarity and intensity value is a continuous value, bounded from -1.0 to +1.0, where the value's sign (positive, negative, or zero) corresponds to the text's stance polarity (favoring, opposing, or neutral) and the value's magnitude (0 to 1.0) corresponds to the text's stance intensity.

Stance polarity and intensity prediction encapsulates stance detection within its problem definition and is thus a more difficult problem to address. While stance polarity can be identified through certain keywords (e.g. agree, disagree), intensity is a much more fuzzy concept. The difference between strong opposition and weak opposition is often expressed through subtle word choices and conversational behaviors. Thus, in order to accurately predict agreement intensity, a learned model must understand the nuances between word choices in the context of the discussion.

We explore five machine learning models for agreement prediction, adapted from the top performing models for stance detection: RidgeM regression, Ridge-S regression, SVR-RF-R, pkudblab-AP, and T-PAN-AP. These models were adapted from Mohammad et al. (2016) [23], Sobhani et al. (2016) [31], Mourad et al. (2018) [25], Wei et al. (2016) [38], and Dey et al. (2018) [7] respectively. We evaluated these models on a new dataset for stance polarity and intensity prediction, collected over three empirical studies using a cyber argumentation platform, the Intelligent Cyber Argumentation System (ICAS). This dataset contains over 22,000 online arguments from over 900 users discussing four important issues. In the dataset, each argument is manually annotated by their authoring user with an agreement value.

Results from our empirical analysis show that the SVR-RF-R ensemble model performed the best for agreement prediction, achieving an RMSE score of 0.596 for stance polarity and intensity prediction, and an accuracy of 70% for stance detection. Further analysis revealed that the models trained for stance polarity and intensity prediction often had better accuracy for stance classification (polarity only) compared to their stance polarity detection model counterparts. This result demonstrates that the added difficulty of detecting stance intensity does not come at the expense of detecting stance polarity. To our knowledge, this is the first time that learning models can be trained to predict an online posts stance polarity and intensity simultaneously. The contributions of our work are the following:

- We introduce a new research problem called stance polarity and intensity prediction, which seeks to predict a post's agreement value that contains both the stance polarity (value sign) and intensity (value magnitude), toward its parent post.
- We present a new empirical dataset for stance polarity and intensity prediction. The dataset, collected over three years of empirical studies, is large compared to similar datasets for stance detection, containing over 22,000 online arguments annotated by their original authors, collected using a cyber argumentation platform.
- We apply five machine learning models on our dataset for agreement prediction. Our empirical results reveal that an ensemble model with many hand-crafted features performed the best, with an RMSE of 0.595, and that models trained for stance polarity and intensity prediction do not lose significant performance for stance detection.

5.3 Related Work

5.3.1 Stance Detection

Stance detection research has a wide interest in a variety of different application areas including opinion mining [13], sentiment analysis [24], rumor veracity [6], and fake news detection [19]. Prior works have applied stance detection to many types of debate and discussion settings, including congressional floor debates [5], online forums [8],[13], persuasive essays [27], news articles [12],and on social media data like Twitter [23]. Approaches to stance detection depends on the type of text and relationship the stance is describing. For example, stance detection on Twitter often determines the author's stance (for/against/neutral) toward a proposition or target [23]. In this work we adapt the features sets and models used on the SemEval 2016 stance detection task Twitter dataset [23]. This dataset has many similarities to our data in terms of post length and topics addressed. Approaches to Twitter stance detection include SVMs [11, 23, 31], ensemble classifiers [25, 36], convolutional neural networks [14, 37, 38], recurrent neural networks [7, 39], and deep learning approaches [30, 34]. Due to the size of the dataset, the difference in domain, and time constraints, we did not test Sun et al. (2018)'s model [30] in this work, because we could not gather sufficient argument representation features.

5.3.2 Argumentation Mining

Argumentation mining is applied to argumentative text to identify the major argumentative components and their relationships to one another [33]. While stance detection identifies the relationship between an author's stance toward a concept or target, argumentation mining identifies relationships between arguments, similar to our task in agreement prediction. However, unlike our task, argumentation mining typically defines arguments based on argument components, instead

of treating an entire post as a single argument. In argumentation mining, a single text may contain many arguments.

The major tasks of argumentation mining include: 1) identify argumentative text from non-argumentative text, 2) classify argumentation components (e.g. Major Claim, Claims, Premise, etc.) in the text, 3) determine the relationships between the different components, and 4) classify the relationships as supporting, attacking, or neutral [20]. End-to-end argument mining seeks to solve all the argumentation mining tasks at once [10, 27], but most research focuses on one or two tasks at once. The most pertinent task to this work is the fourth task (though often times this task is combined with task 3). Approaches to this task include using textual entailment suites with syntactic features [4], or machine learning classifiers with different combinations of features including, structural and lexical features [27], sentiment features [32], and Topic modeling features [26]. We use many of these types of features in our Ridge-S and SVR-RF-R models.

5.3.3 Cyber Argumentation Systems

Cyber argumentation systems help facilitate and improve understanding of large-scale online discussions, compared to other platforms used for debate, such as social networking and media platforms, online forums, and chat rooms [15]. These systems typically employ argumentation frameworks, like IBIS [17] and Toulmin's structure of argumentation [35], to provide structure to discussions, making them easier to analyze. More specialized systems include features which improve the quality and understanding of discussions. Argumentation learning systems teach the users effective debating skills using argumentation scaffolding [3]. More complex systems, like ICAS and the Deliberatorium [15], provide several integrated analytical models which identify and measure various phenomena occurring in the discussions.

5.4 Background

5.4.1 ICAS Platform

In ICAS [21], arguments have two components, a textual component and an agreement value. The textual component is the written argument the user makes. ICAS does not limit the length of argument text, however, in practice the average argument length is about160 characters, similar to the length of a tweet. The agreement value is a numerical value which indicates the extent to which an argument agrees or disagrees with its parent. Unlike other argumentation systems, this system allows users to express partial agreement or disagreement with other posts. Users are allowed to select agreement values from a range of -1 to +1 at 0.2 increments that indicate different partial agreement values. Positive values indicate partial or complete agreement, negative values indicate partial or complete disagreement, and a value of 0 indicates indifference or neutrality. These agreement values represent each post's stance polarity (the sign) and intensity (the magnitude). These agreement values are distinctly different from other argumentation weighting schemes where argument weights represent the strength or veracity of an argument (see [1, 18]). Each agreement value is selected by the author of the argument and is a mandatory step when posting. Models for Stance Polarity and Intensity Prediction.

5.5 Models for Stance Polarity and Intensity Prediction

This section describes the models we applied to the stance polarity and intensity prediction problem. We applied five different models, adapted from top performing stance classification models based on their performance and approach on the SemEval 2016 stance classification Twitter dataset [23].

5.5.1 Ridge Regressions (Ridge-M and Ridge-S)

Our first two models use a linear ridge regression as the underlying model. We created two ridge regression models using two feature sets.

The first ridge model (Ridge-M) used the feature set described in Mohammad et al. (2016) [23] as their benchmark. They used word 1-3 grams and character 2-5 grams as features. We filtered out English stop words, tokens that existed in more than 95% of posts, and tokens that appear in less than 0.01% of posts for word N-grams and fewer than 10% for character N-grams. There were a total of 838 N-gram features for the Ridge-M model.

The second ridge model (Ridge-S) used the feature set described in Sobhani, Mohammad, and Kiritchenko's follow-up paper (2016) [31]. In that paper, they found the sum of trained word embeddings with 100 dimensions, in addition to the N-gram features outlined by Mohammad et al. (2016) [23], to be the best performing feature set. We trained a word-embedding (skip-gram word2vec) model on the dataset. For each post, and summed the embeddings for each token in the post were summed up and normalized by the total number of tokens of a post to generate the word embedding features. Ridge-S had 938 total features.

5.5.2 Ensemble of Regressions (SVR-RF-R)

This model (SRV-RF-R) consisted of an average voting ensemble containing three different regression models: an Epsilon-Support Vector Regression model, a Random Forest regressor, and a ridge regression model. This model is an adaption of the ensemble model presented by Mourad et al. (2018) [25] for stance detection. Their model used a large assortment of features including linguistic features, Topic features, Tweet-specific features, Labeled based features, Word-Embedding features, Similarity Features, Context features, and Sentiment Lexicon

features. They then used the feature selection technique reliefF [16] to select the top 50 features for usage. Due to the changes in context (Twitter vs Cyber Argumentation), we constructed a subset of their feature set, which included the following features:

- Linguistic Features: Word 1-3 grams as binary vectors, count vectors, and tf-idf weighted vectors. Character 1-6 grams as count vectors. POS tag 1-3 grams concatenated with their words (ex: word1 pos1 ...) and concatenated to the end of the post (ex: word1, word2, ..., POS1, POS2, ...).
- Topic Features: Topic membership of each post after LDA topic modeling (Blei et al., 2003) is run on the entire post corpus.
- Word Embedding Features: The 100 dimensional word embedding sums for each word in a post and the cosine similarity between the summed embedding vectors for the target post and its parent post.
- Lexical Features: Sentiment lexicon features outlined in Mourad et al. (2018), excluding the DAL and NRC Hashtag Lexicons.

We tested using the top 50 features selected using reliefF and reducing the feature size to 50 using Principal Component Analysis (PCA), as well as using the full feature set. We found that the full feature set (2855 total) performed significantly better than the reliefF and PCA feature sets. We used the full feature set in our final model.

5.5.3 pkudblab-PIP

The highest performing CNN model, pkudblab, applied to the SemEval 2016 benchmark dataset was submitted by Wei et al. (2016) [38]. Their model applied a convolutional neural network on the word embedding features of a tweet. We modified this model for agreement

prediction, the resulting model's (pkudblab-PIP) architecture is shown in Figure 5-1. We used pre-trained embeddings (300-dimension) published by the word2vec team [22]. Given an input of word embeddings of size d by |s|, where d is the size of the word embedding and |s| is the normalized post length, the input was fed into a convolution layer. The convolution layer contained filters with window size (m) 3, 4, and 5 words long with 100 filters (n) each. Then the layers were passed to a max-pooling layer, and finally passed through a fully-connected sigmoid layer to produce the final output value. We trained the model using a mean squared error loss function and used a 50% dropout layer after the max-pooling layer.

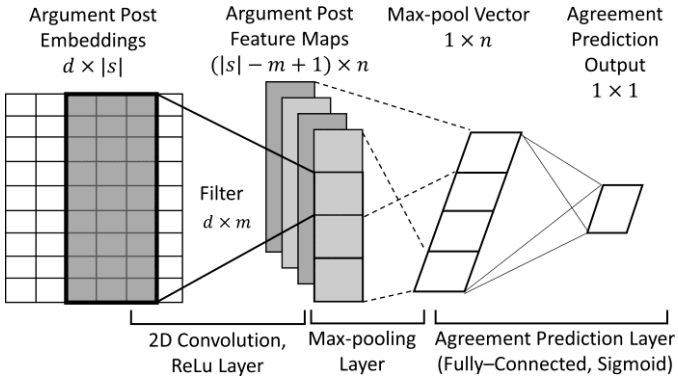


Figure 5-1 The architecture of pkudblab-PIP for stance polarity and intensity prediction.

5.5.4 T-PAN-PIP

The RNN model (T-PAN-PIP) is adapted from the T-PAN framework by Dey et al. (2018) [7], which was one of the highest performing neural network models on the SemEval 2016 benchmark dataset. The T-PAN framework uses a two-phase LSTM model with Attention, based on the architecture proposed by Du et al. (2017) [9]. We adapted this model for regression by making some modifications. Our adapted model (T-PAN-PIP) uses only a single phase architecture, resembling Du et al.'s original design (2017) [9], where the output is the predicted agreement value, instead of a categorical prediction.

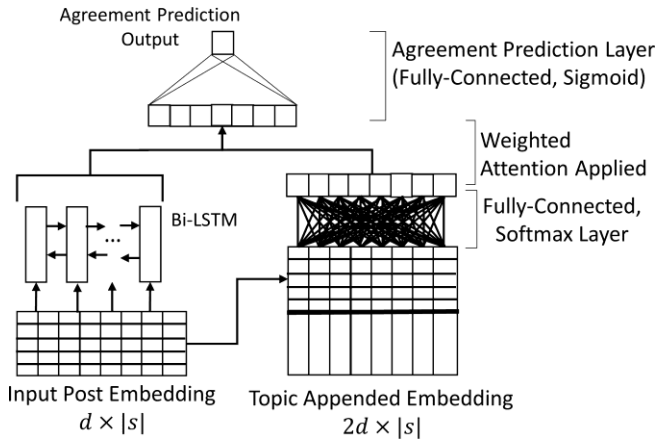


Figure 5-2 The architecture of T-PAN-PIP for stance polarity and intensity prediction.

Figure 5-2 illustrates the architecture of T-PANPIP. It uses word embedding features (with embedding size 300) as input to two network branches. The first branch feeds the word embeddings into a bi-directional LSTM (Bi-LSTM) with 256 hidden units, which outputs the hidden states for each direction (128 hidden units each) at every time step. The other branch appends the average topic embedding from the topic text (i.e. the text of the post that the input is responding) to the input embeddings and feeds that input into a fully-connected softmax layer to calculate what Dey et al. (2018) [7] called the "subjectivity attention signal". The subjectivity attention signals are a linear mapping of each input word's target augmented embedding to a scalar value that represents the importance of each word in the input relative to the target's text. These values serve as the attention weights that are used to scale the hidden state output of the Bi-LSTM. The weighted attention application layer combines the attention weighs to their corresponding hidden state output as shown in (1).

$$Q = \frac{1}{|s|} \sum_{s=0}^{|s|-1} a_s h_s \tag{1}$$

Where as is the attention signal for word s, hs is the hidden layer output of the Bi-LSTM for word s, |s| is the total number of words, and Q is the resulting attention weighted vector of size

256, the size of the hidden units output by the Bi-LISTM. The output Q feeds into a fully-connected sigmoid layer and outputs the predicted agreement value. We train the model using a mean absolute error loss function.

5.6 Empirical Dataset Description

The dataset was constructed from three separate empirical studies collected in Fall 2017, Spring 2018, and Spring 2019. In each study, a class of undergraduate students in an entry level sociology class was offered extra credit to participate in discussions in ICAS. Each student was asked to discuss four different issues relating to the content they were covering in class. Please refer to Chapter 2 for more details about the empirical studies.

The combined dataset contains 22,606 total arguments from 904 different users. Of those arguments, 11,802 are replying to a position and 10,804 are replying to another argument. The average depth of a reply thread tends to be shallow, with 52% of arguments on the first level (reply to position), 44% on the second level, 3% on the third level, and 1% on the remaining levels (deepest level was 5).

When a student posted an argument, they were required to annotate their argument with an agreement value. Overall, argument agreement values skew positive. Figure 5-3 displays a histogram of the agreement values for the arguments in the dataset.

The annotated labels in this data-set are self-labeled, meaning that the when a user replies to a post, they provide their own stance polarity and intensity label. The label is a reflection of the author's intended stance toward a post, where the post's text is a semantic description of that intention. While these label values are somewhat subjective, they are an accurate reflection of their author's agreement, which we need to capture to analyze opinions in the discussion.

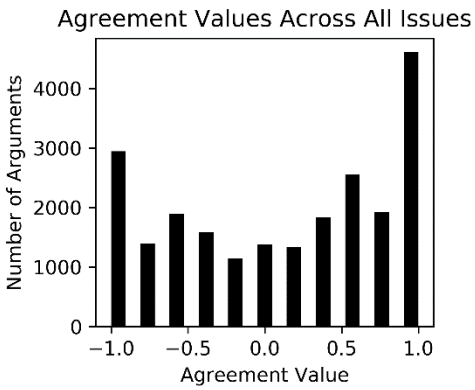


Figure 5-3 A histogram of the different agreement values across all of the issues in the cyber argumentation.

5.7 Empirical Study Evaluation

5.7.1 Agreement Prediction Problem

In this study we want to evaluate the models' performance on the stance polarity and intensity prediction problem. We separated the dataset into training and testing sets using a 75-25 split. For the neural network models (pkudblab-AP and TPAN-AP), we separated out 10% of the training set as a validation set to detect over-fitting. The split was performed randomly without consideration of the discussion issue. Each issue was represented proportionally in the training and testing data sets with a maximum discrepancy of less than 1%.

For evaluation, we want to see how well the regression models are able to predict the continuous agreement value for a post. We report the root mean-squared error (RMSE) for the predicted results.

5.7.2 Agreement Prediction Models for Stance Detection

We wanted to investigate whether training models for agreement prediction would degrade their performance for stance detection. Ideally, these models should learn to identify both stance intensity without impacting their ability to identify stance polarity.

To test this, we compared each model to their original stance classification models described in their source papers. Thus, ridge-H is compared with an SVM trained on the same feature set (SVM-H), ridge-S is compared to a Linear-SVM trained on the same feature set (SVM-S), SVR-RF-R is compared to a majority-voting ensemble of a linear SVM, Random Forest, and Naïve Bayes classifier using the same feature set (SVM-RF-NB), pkudblab-PIP is compared to the original pkudblab model trained using a softmax cross entropy loss function, and T-PAN-PIP is compared to the original T-PAN model trained using a softmax cross entropy loss function. We trained the classification models for stance detection by converting the continuous agreement values into categorical polarity values. When converted into categorical values, all of the positive agreement values are classified as Favoring, all negative values are classified as Opposing, and zero values are classified as Neutral. In the dataset, 12,258 arguments are Favoring (54%), 8962 arguments are Opposing (40%) and 1386 arguments are Neutral (6%). To assess the stance detection performance of the models trained for agreement prediction, we converted the predicted continuous agreement values output by the models into the categorical values using the same method.

For evaluation, we report both the accuracy value of the predictions and the macro-average F1-scores for the Favoring and Opposing classes on the testing set. This scoring scheme allows us to treat the Neutral category as a class that is not of interest [25].

5.8 Evaluation Results

5.8.1 Agreement Prediction Results

The results for agreement prediction are shown in Table 5.1. A mean prediction baseline model is shown in the table to demonstrate the difficulty associated with the problem. The neural

network models perform worse than both the ridge regression and ensemble models. Ridge-S performed slightly better than Ridge-M due to the sum word embedding features. The best performing model was the SVR-RF-R model with an RMSE of 0.596.

Table 5.1: The results of the regression models for the Agreement prediction task. The best result is bolded.

Model	**RMSE**
Baseline (Mean)	0.718
Ridge-M	0.620
Ridge-S	0.615
SVR-RF-R	**0.596**
pkudblab-PIP	0.657
T-PAN-PIP	0.623

5.8.2 Agreement Prediction Models for Stance Detection Results

We compare the models trained on the agreement prediction task to their classification model counterparts in terms of performance on the stance detection task. Tables 5.2 and 5.3 show the comparison between the models in terms of accuracy and (macro) F1-score.

SVR-RF-R has the best accuracy and F1-score for stance detection, which outperformed its classifier counterpart (SVM-RF-NB) by 2.12% in accuracy and +0.016 in F1-score. Three of the models trained for stance polarity and intensity prediction, SVR-RF-R, Ridge-S, and T-PAN-PIP, outperformed their classifier counterparts in accuracy by 1-2% and F1-score by +0.009 on average. Two of the models trained for stance polarity and intensity prediction, Ridge-H and pkudblab-PIP, slightly under-performed their classifier counterparts in accuracy by -0.38% and F1-score by -0.011 on average.

Table 5.2: The classification accuracy of the stance polarity prediction models and the stance polarity and intensity prediction models for Stance Detection (polarity only) classification.

Stance Polarity Prediction Model		**Polarity and Intensity Prediction Model**		
Model	**Accuracy**	**Model**	**Accuracy**	**Diff**
Baseline (Most Frequent)	54.36%	Baseline (Mean)	54.36%	0.00%
SVM-H	**68.48%**	Ridge-M	68.16%	-0.32%
SVM-S	67.63%	Ridge-S	68.84%	+1.21%
SVM-RF-NB	68.31%	SVR-RF-R	**70.43%**	**+2.12%**
pkudblab	67.28%	pkudblab-PIP	66.89%	-0.39%
T-PAN	65.55%	T-PAN-PIP	66.64%	+1.09%

Table 5.3: The classification F1-score of the stance polarity prediction models and the stance polarity and intensity prediction models for Stance Detection (polarity only) classification.

Stance Polarity Prediction Model		**Polarity and Intensity Prediction Model**		
Model	**F1-score**	**Model**	**F1-score**	**Diff**
Baseline (Most Frequent)	0.352	Baseline (Mean)	0.352	0.000
SVM-H	0.701	Ridge-M	0.695	-0.006
SVM-S	0.697	Ridge-S	0.703	+0.006
SVM-RF-NB	**0.705**	SVR-RF-R	**0.721**	**+0.016**
pkudblab	0.688	pkudblab-PIP	0.672	-0.016
T-PAN	0.673	T-PAN-PIP	0.678	+0.005

5.9 Discussion

The models behaved very similarly on the agreement prediction problem, with the difference between the best performing model and the worst performing model being only 0.061. Overall, the best model received an RMSE of 0.596 which is reasonably good but can be improved.

T-PAN-PIP had the worst performance, which is surprising, as it was the only model to include the parent post's information into its prediction, which should have helped improve its performance. It is possible that its architecture is unsuitable for agreement prediction; other architectures have been deployed that include a post's parent and ancestors into a stance prediction,

which might be more suitable for agreement prediction. More investigation should be paid to better incorporating a post's parent information.

The difference in performance between the agreement prediction models and the classification models on the stance detection task was small and sometimes better. This demonstrates that the models learning to identify stance intensity do so without significant loss of performance in identifying stance polarity.

Larger gains in performance will likely require information about the post's author. Some post authors will state strong levels of agreement in their statements, but annotate their argument with weaker agreement levels. For example, one author wrote:

"Agree completely. Government should stay out of healthcare."

Then annotated that argument with an agreement value of +0.6. The authors were instructed on how to annotate their posts, but the annotations themselves were left to the post's author's discretion. Thus, including author information into our models, would likely improve the stance polarity and intensity prediction results.

5.10 Conclusion

We introduce a new research problem called stance polarity and intensity prediction in a responsive relationship between posts, which predicts both an online post's stance polarity and intensity value toward another post. This problem encapsulates stance detection and adds the additional difficulty of detecting subtle differences in intensity found in text. We introduced a new large empirical dataset for agreement prediction, collected using a cyber argumentation platform. We implemented five models, adapted from top performing stance detection models, for evaluation on the new dataset for agreement prediction. Our empirical results demonstrate that the ensemble

model SVR-RF-R performed the best for agreement prediction and models trained for agreement prediction learn to differentiate between intensity values without degrading their performance for determining stance polarity. Research into this new problem of agreement prediction will allow for a more nuanced annotation and analysis of online debate.

5.11 References

[1] L. Amgoud and J. Ben-Naim, "Weighted Bipolar Argumentation Graphs: Axioms and Semantics," in IJCAI'18 Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018, pp. 5194–5198.

[2] R. S. Arvapally, X. F. Liu, F. F.-H. Nah, and W. Jiang, "Identifying outlier opinions in an online intelligent argumentation system," Concurrency and Computation: Practice and Experience, p. e4107, 2017, doi: 10.1002/cpe.4107.

[3] P. Bell and M. C. Linn, "Scientific arguments as learning artifacts: designing for learning from the web with KIE," International Journal of Science Education, vol. 22, no. 8, pp. 797–817, Aug. 2000, doi: 10.1080/095006900412284.

[4] F. Boltužić and J. Šnajder, "Back up your Stance: Recognizing Arguments in Online Discussions," in Proceedings of the First Workshop on Argumentation Mining, Baltimore, Maryland, 2014, pp. 49–58, doi: 10.3115/v1/W14-2107.

[5] C. Burfoot, S. Bird, and T. Baldwin, "Collective Classification of Congressional Floor-debate Transcripts," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Stroudsburg, PA, USA, 2011, pp. 1506–1515.

[6] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours," in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, 2017, pp. 69–76.

[7] K. Dey, R. Shrivastava, and S. Kaushik, "Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention," in Advances in Information Retrieval, 2018, pp. 529–536.

[8] R. Dong, Y. Sun, L. Wang, Y. Gu, and Y. Zhong, "Weakly-Guided User Stance Prediction via Joint Modeling of Content and Social Interaction," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, New York, NY, USA, 2017, pp. 1249–1258, doi: 10.1145/3132847.3133020.

[9] J. Du, R. Xu, Y. He, and L. Gui, “Stance classification with target-specific neural attention networks,” presented at the International Joint Conferences on Artificial Intelligence, 2017.

[10] S. Eger, J. Daxenberger, and I. Gurevych, “Neural End-to-End Learning for Computational Argumentation Mining,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancourver, Canada, 2017, vol. 1, pp. 11–22, doi: 10.18653/v1/P17-1002.

[11] H. Elfardy and M. Diab, “CU-GWU Perspective at SemEval-2016 Task 6: Ideological Stance Detection in Informal Text,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, 2016, pp. 434–439, doi: 10.18653/v1/S16-1070.

[12] A. Hanselowski et al., “A Retrospective Analysis of the Fake News Challenge Stance Detection Task,” arXiv:1806.05180 [cs], Jun. 2018.

[13] K. S. Hasan and V. Ng, “Stance Classification of Ideological Debates: Data, Models, Features, and Constraints,” in Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013, pp. 1348–1356.

[14] Y. Igarashi, H. Komatsu, S. Kobayashi, N. Okazaki, and K. Inui, “Tohoku at SemEval-2016 Task 6: Feature-based Model versus Convolutional Neural Network for Stance Detection,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, 2016, pp. 401–407, doi: 10.18653/v1/S16-1065.

[15] M. Klein, “How to Harvest Collective Wisdom on Complex Problems : An Introduction to the MIT Deliberatorium,” 2011.

[16] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, “Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF,” Applied Intelligence, vol. 7, no. 1, pp. 39–55, Jan. 1997, doi: 10.1023/A:1008280620621.

[17] W. Kunz and H. W. J. Rittel, “Issues as elements of information systems,” presented at the Working Paper No. 131, Berkeley, 1970, vol. 131.

[18] G.-A. Levow et al., “Recognition of stance strength and polarity in spontaneous speech,” in 2014 IEEE Spoken Language Technology Workshop (SLT), 2014, pp. 236–241, doi: 10.1109/SLT.2014.7078580.

[19] A. E. Lillie and E. R. Middelboe, “Fake News Detection using Stance Classification: A Survey,” arXiv:1907.00181 [cs], Jun. 2019.

[20] M. Lippi and P. Torroni, “Argumentation Mining: State of the Art and Emerging Trends,” ACM Trans. Internet Technol., vol. 16, no. 2, pp. 10:1–10:25, Mar. 2016, doi: 10.1145/2850417.

[21] X. (Frank) Liu, E. Khudkhudia, L. Wen, V. Sajja, and M. C. Leu, “An Intelligent Computational Argumentation System for Supporting Collaborative Software Development Decision Making,” in Artificial Intelligence Applications for Improved Software Engineering Development: New Prospects, F. Meziane and S. Vadera, Eds. IGI Global, 2010, pp. 167–180.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv:1301.3781 [cs], Jan. 2013.

[23] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, “SemEval-2016 Task 6: Detecting Stance in Tweets,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, 2016, pp. 31–41, doi: 10.18653/v1/S16-1003.

[24] S. M. Mohammad, “Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text,” in Emotion Measurement, H. L. Meiselman, Ed. Woodhead Publishing, 2016, pp. 201–237.

[25] S. S. Mourad, D. M. Shawky, H. A. Fayed, and A. H. Badawi, “Stance Detection in Tweets Using a Majority Vote Classifier,” in The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018), 2018, pp. 375–384.

[26] H. Nguyen and D. Litman, “Context-aware Argumentative Relation Mining,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 2016, pp. 1127–1137, doi: 10.18653/v1/P16-1107.

[27] I. Persing and V. Ng, “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, San Diego, California, 2016, pp. 1384–1394, doi: 10.18653/v1/N16-1164.

[28] J. Sirrianni, X. (Frank) Liu, and D. Adams, “Quantitative Modeling of Polarization in Online Intelligent Argumentation and Deliberation for Capturing Collective Intelligence,” in 2018 IEEE International Conference on Cognitive Computing (ICCC), 2018, pp. 57–64, doi: 10.1109/ICCC.2018.00015.

[29] P. Sobhani, D. Inkpen, and S. Matwin, “From Argumentation Mining to Stance Classification,” in Proceedings of the 2nd Workshop on Argumentation Mining, Denver, CO, 2015, pp. 67–77, doi: 10.3115/v1/W15-0509.

[30] P. Sobhani, D. Inkpen, and X. Zhu, “Exploring deep neural networks for multitarget stance detection,” Computational Intelligence, vol. 35, no. 1, pp. 82–97, 2019, doi: 10.1111/coin.12189.

[31] P. Sobhani, S. Mohammad, and S. Kiritchenko, “Detecting Stance in Tweets And Analyzing its Interaction with Sentiment,” in Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, Berlin, Germany, 2016, pp. 159–169, doi: 10.18653/v1/S16-2021.

[32] C. Stab and I. Gurevych, “Parsing Argumentation Structures in Persuasive Essays,” Computational Linguistics, vol. 43, no. 3, pp. 619–659, 2017, doi: 10.1162/COLI_a_00295.

[33] M. Stede and J. Schneider, Argumentation Mining, vol. 11. Morgan & Claypool Publishers, 2018.

[34] Q. Sun, Z. Wang, Q. Zhu, and G. Zhou, “Stance Detection with Hierarchical Attention Network,” in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2399–2409.

[35] S. E. Toulmin, The Uses of Argument. Cambridge: Cambridge University Press, 2003.

[36] M. Tutek et al., “TakeLab at SemEval-2016 Task 6: Stance Classification in Tweets Using a Genetic Algorithm Based Ensemble,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, 2016, pp. 464–468, doi: 10.18653/v1/S16-1075.

[37] P. Vijayaraghavan, I. Sysoev, S. Vosoughi, and D. Roy, “DeepStance at SemEval-2016 Task 6: Detecting Stance in Tweets Using Character and Word-Level CNNs,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, 2016, pp. 425–431.

[38] W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang, “pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, 2016, pp. 384–388, doi: 10.18653/v1/S16-1062.

[39] G. Zarrella and A. Marsh, “MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, 2016, pp. 458–463.

Chapter 6: Predicting Stance Polarity and Intensity in Cyber Argumentation with Deep Bi-directional Transformers

6.1 Abstract

In online deliberation, participants argue in support or opposition to one another's arguments and ideas to advocate their position. Often their stance expressed in their posts are implicit and must be derived from the post's text. Existing stance detection models predict the polarity of the user's stance from the text, but do not consider the stance's intensity. We introduce a new research problem, stance polarity and intensity prediction in response relationships between posts. This problem seeks to predict both the stance polarity and intensity of a replying post toward its parent post in online deliberation. Using our cyber argumentation platform, we have collected an empirical dataset with explicitly labeled stance polarity and intensity relationships. In this work, we create six models: five are adapted from top performing stance detection models and another novel model that fine-tunes the deep bi-directional transformer model BERT. We train and test these six models on our empirical dataset to compare their performance for stance polarity and intensity prediction and stance detection. Our results demonstrate that our method of encoding the stance polarity and intensity labels allows the models to predict stance polarity and intensity without compromising their accuracy for stance detection, making these models more versatile. Our results reveal that a novel split architecture for fine-tuning the BERT model outperforms the other models for stance polarity and intensity prediction by 5% accuracy. This work is the first to train models for predicting both the stance polarity and intensity in one combined task while maintaining good accuracy.

6.2 Introduction

Online platforms, such as Facebook, Twitter, and Wikipedia, have become the primary virtual public forums for people around the world to come together to discuss and debate issues of local, national, and international importance. With such massive participation, these online discussions contain a wealth of valuable information about public opinion on various topics. However, due to the limited structure of the discourse data produced in these platforms, analyzing the discussion information is an increasingly difficult task.

One crucial task in analyzing online discussions and debates is determining the different argumentative stance relationships between online posts in a discussion. Typically, in online debates, when a user replies to another user's post, they either argue for (supporting) or against (attacking) the entirety or some part of the original post. Thus, in terms of stance, the argumentative relationships between two posts include both the stance polarity (attacking/supporting/neutral) and intensity (degree of support/attack) from the child post (the replying post) toward the parent post.

Automatically identifying the stance relationships between posts has many potential research applications and is a goal in the fields of stance detection [1], [2] and argumentation mining research [3]. Stance detection research seeks to develop predictive models to classify the polarity (Supporting, Attacking, or Neutral) of a text's stance toward another text, topic, entity, or theme [1]. Stance detection has many application areas, including fake news detection [4] and rumor veracity detection [5]. Similarly, argumentation mining seeks to identify and classify the relationships between arguments and their components from a given text, including the stance polarity between arguments [3]. However, in both research areas, attention is paid primarily to the polarity of the stance relationships, while the intensity is often ignored.

Some stance detection research has tried to incorporate both stance polarity and intensity into a single predictive model by expanding the classification categories to include intensity information (e.g., Strongly For, For, Other, Against, Strongly Against) [6]; however, these expanded categories resulted in significantly lower model performance compared to classifying polarity alone. Thus, effective incorporation of stance intensity into stance prediction remains an issue. Including the stance intensity into stance polarity prediction has two main benefits. The first benefit is that including the intensity in stance prediction allows for the consideration of partial agreement. Often in discussions, users will express partial approval or disapproval of other's ideas and arguments, instead of simply fully supporting or opposing them. This partial agreement may not be captured by standard stance detection models, because they can only distinguish the polarity of the stance. This inability to capture partial agreement can make it difficult to accurately capture the rationale behind users' opinions on complex issues. Even in highly polarized discussions, such as the abortion debate in the U.S., users from opposite sides often still agree on underlying values and concepts related to the topic. By capturing the partial agreement of users in a discussion, researchers can gather a more nuanced and comprehensive analysis of the users' opinions on important, complex issues. Secondly, research in cyber argumentation has demonstrated that incorporating both stance polarity and intensity information into analytical models provides a more nuanced analysis of various deliberation phenomena, such as capturing users' rationale [7], collective opinion analysis [8], argument credibility [9], identify factions in discussions [10], argumentation polarization analysis [11], and opinion outlier detection [12], compared to using stance polarity only. Thus, by developing a model to predict both the stance polarity and intensity relationship between online posts in online deliberation, these powerful cyber argumentation

models can be applied to the online discussions from non-cyber argumentation platforms, such as Twitter, Facebook, and Reddit.

In this work, we address the issue of stance polarity and intensity prediction in a responsive relationship between posts. To enable a model to predict both the stance polarity and intensity of the stance relationship while still maintaining good accuracy, we encode the stance relationship as a single continuous value. This value represents the partial agreement between the posts, which we call the agreement value. Agreement values are bounded between -1.0 and +1.0, where the stance polarity is encoded in the argument value's sign (positive is supporting, negative is attacking, zero is neutral), and the stance intensity is encoded as the value's magnitude (0 to 1.0). This formulation allows for a model to predict the stance polarity and intensity without creating many separate categories.

By its nature, stance polarity and intensity is a difficult problem because it includes both stance detection and stance intensity recognition. In addition to the stance polarity information, models trained for this task must also associate stance intensity information to various words during training. This added burden placed on the models suggests that current state of-the-art stance detection models may not be most suitable for stance polarity and intensity detection if they are not able to capture the stance intensity information effectively. In this work, we explore six different stance polarity and intensity prediction machine learning models.

Five of the models, presented in Section 5.5, are adapted from the top-performing models for stance detection: Ridge M Regression, Ridge-S Regression, SVR-RF-R, pkudblab-PIP, and T-PAN-PIP, adapted from Mohammad et al. (2016) [1], Sobhani et al. (2016) [13], Mourad et al. (2018) [14], Wei et al. (2016) [15] and Dey et al. (2018) [16] respectively. The sixth model we explore is a new model that applies the pretrained deep bi-directional Transformers model BERT

[17] for stance polarity and intensity prediction. BERT is a pre-trained language model, whose purpose is to calculate representation of text that includes both word semantics and local context information. The BERT model has been used to generate language representations that have been applied effectively to many downstream natural language tasks. We test several different configurations for fine-tuning the pre-trained BERT model for stance polarity and intensity prediction, including using different fine-tuning architectures, using different sizes of the BERT model, and freezing or unfreezing the BERT weights during fine-tuning.

We train each of the six models on an empirical dataset of over 22,000 online arguments from over 900 users collected using a cyber argumentation platform, the Intelligent Cyber Argumentation System (ICAS). In this platform, when a poster creates a new argument in reply to another post, they must explicitly annotate their argument with an agreement value. Thus, every argument in the discussions in ICAS have an annotated agreement value associated with it. We train and evaluate the models on this empirical data.

The results of this research demonstrate that the five adapted stance detection models perform similarly in terms of accuracy when predicting stance polarity and intensity as they do when predicting only stance polarity. These results suggest that our method of encoding stance polarity and intensity as agreement values can be effectively used to incorporate stance intensity into the predictions, without penalizing the accuracy of the model, and, in the case of some models, can improve the accuracy of the stance prediction. Our results of comparing several different architectures and configurations for the BERT model show that using a novel Split architecture, where both the child argument and parent argument are fed into BERT separately, achieved much higher accuracy than using a standard Combined architecture, where the arguments are fed into the BERT model together. Lastly, a comparison of the six different models shows that the fine-

tuned BERT model using a Split architecture had the best performance for stance polarity and intensity prediction with a root mean squared error (RMSE) of 0.528. To our knowledge, this research is the first time that learning models have been trained to predict an online post's stance polarity and intensity simultaneously in cyber argumentation.

The contributions of our work are as follows:

- We introduce the research problem of stance polarity and intensity prediction. We offer and evaluate a method of encoding the stance polarity and intensity relationship as an agreement value. Our empirical results using this encoding method demonstrate that models trained for stance polarity and intensity maintain their accuracy for stance polarity detection, which is an improvement over prior methods of incorporating stance intensity.

- We investigate and develop a stance polarity and intensity prediction model that fine-tunes the pre-trained deep bidirectional transformer model BERT. We investigate several different fine-tuning architectures and configurations for BERT. Our results show that separately encoding each post using the Split architecture significantly increased the accuracy of the predictions compared to encoding both posts together. This architecture is novel and distinctly different from prior works using BERT for stance detection and other natural language understanding tasks.

- We compare the performances of the fine-tuned BERT model and the five adapted models on the stance polarity and intensity prediction task. Our empirical results show that the fine-tuned BERT model using the Split architecture outperformed the other models in terms of RMSE and regression accuracy.

6.3 Background

6.3.1 BERT

One effective approach for many NLP tasks is to develop pre-trained language models to learn representations of words in specific contexts. These pre-trained models can then be fine-tuned by adding a thin network or layer on the output of the generic language model to solve specific NLP tasks [19], [17]. One advantage of this method is that using a pre-trained language model reduces the number of training iterations necessary for fine tuning [17] because the language representations have already been learned during pre-training. Prior transfer learning approaches to dealing with text data focused mainly on using pre-trained word embeddings. However, these embeddings are static and do not consider the local context in which the words are appearing. More modern language models, such as OpenAI GPT [19] and BERT [17], address this issue by incorporating the local context into the initial word embeddings, using a variety of different techniques. The embeddings produced from these models have much more accurate word meaning and association information encoded within them, making them very useful for downstream tasks. This approach should be advantageous for tasks where acquiring large datasets is difficult, such as our task of predicting stance polarity and intensity.

Recently, Devlin et al. (2019) [17] published the Bidirectional Encoder Representations from Transformers, or BERT, model. BERT uses a bidirectional Transformer architecture [21]. Evaluations of BERT have demonstrated its effectiveness on a diverse set of natural language understanding tasks. By utilizing the pre-trained BERT model, a fine-tuned model for stance polarity and intensity prediction will contain the learned knowledge from the pre-trained model as well as learn the new associations relevant to the stance polarity and intensity task.

Prior work incorporating BERT into stance detection, and its related applications of Fake news detection and rumor veracity research, have shown that this strategy is effective [22], [20], [18], [25]. However, none of these works have addressed the issue of predicting both stance polarity and intensity simultaneously.

6.4 Fine-Tuning BERT Model

For implementing the Fine-Tuning of BERT, we used the Transformers library by Hugging Face for implementation [23]. We experimented with multiple different designs. First, we examined two architectures of the model in terms of inputs and outputs from the BERT model, shown in Figures 6-1 and 6-2.

Figure 6-1 has the architecture we label Combined. This architecture encodes both the input post and the parent post into a single output from the BERT model, which is then fed through the thin network layer. This setup allows the words from the parent and child posts to be embedded with respect to one another. This architecture matches the architecture for Sentence Pair Classification from the original BERT paper (see Figure 4a in [17]), and prior works using BERT for stance detection applications [22], [20], [18].

Figure 6-2 has the architecture which we label Split. This architecture encodes the input post and the parent post separately, through the same BERT model, producing one output for each post and then feeding the concatenated output into the thin network layer. This architecture does not encode a post relative toward one another and instead does so independently. The output of the Split model feeds each post into the thin layer, which is a shallow dense network on top of the output of the BERT model that learns to determine the relationship between the posts. This

approach contrasts the Combined model, where the thin layer learns the relationship based on one combined embedding.

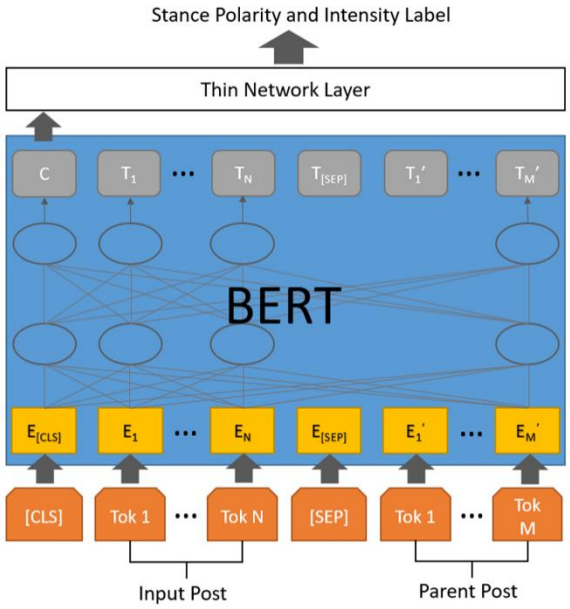


Figure 6-1 The architecture for the Combined BERT model.

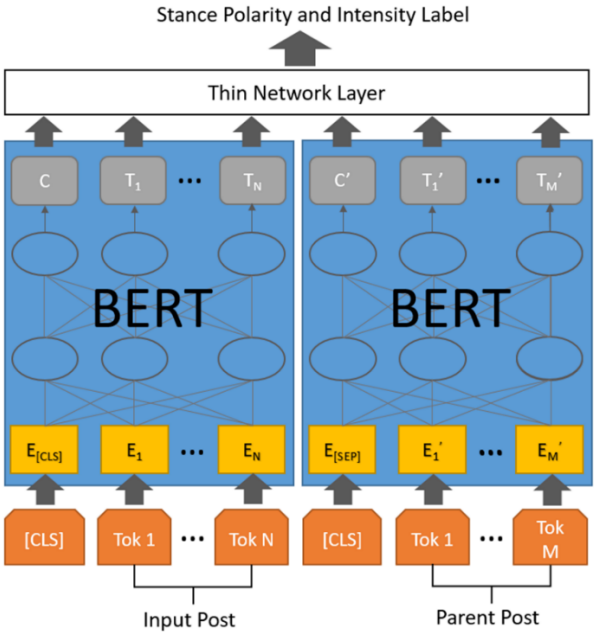


Figure 6-2 The architecture for the Split BERT model

Since both the input and parent posts are passed through the same BERT model, this does not significantly increase the number of trainable parameters in the model. To our knowledge, this architecture has not been explored in stance detection or stance detection adjacent research.

In addition to the model architecture configuration, other aspects were also examined including:

- Freezing/Unfreezing the BERT weights during training: Freezing the BERT weights meant that they were not further trained during the fine-tuning learning while unfreezing them did alter their values during training.
- BERT Model Size: The Transformers library used to implement the pre-trained BERT model had two instances: the BERT base model (12 layers, 768 Hidden state size, 12-head transformers, and 110M parameters) which we label small, and a large BERT model (24-layer, 1024 hidden state size, 16-head transformers, and 340M parameters).

The thin network layer is a linear layer followed by a Tanh layer. We experiment with several different thin network configurations (e.g., linear + tanh + linear, linear + tanh + linear + tanh, and linear + sigmoid), however using different thin network layers did not produce meaningfully different results. The output from the BERT model depended on the BERT pre-trained model size (768 for Small and 1024 for Large) and whether the architecture was Combined (1x BERT output) or Split (2x BERT output), and the output size of the thin network layer was one.

Each model was trained using the ADAM optimizer [24]. The input text was limited to 512 words. All the frozen models (BERT parameters not trained) used training batch size 64, and

learning rate 0.001, while unfrozen model (BERT parameters trained) used batch size two and learning rate $2 * 10^{-5}$. All models were trained using the MSE loss function.

6.5 Experimental Setup

6.5.1 Experiments

Our experiments had two primary objectives:

1. Determine which architectures and procedures yielded the best results for fine-tuning the BERT model for stance polarity and intensity prediction.
2. Compare our fine-tuned BERT model with the adapted stance detection models for the stance polarity and intensity prediction task.

For comparing with previously investigated stance polarity and intensity models, we compared the fine-tuned BERT model to 6 models investigated previously in Chapter 5. These models were adaptations of top performing models on the SemEval 2016 stance classification Twitter dataset [10].

All models were trained using the same dataset as described in Section 5.6. For training and testing the dataset was divided using a 75-25 split. For the neural network-based models (Fine Tuned BERT, pkudblab-PIP, and T-PAN-PIP) 10% of the training set was separated out as a validation set. The datasets were split randomly without consideration of the discussion issue. Each issue was represented proportionally in both training and testing datasets with a maximum discrepancy of less than one percent.

6.5.1.1 Comparing BERT Fine-Tuning Architectures

The second task compares various fine-tuning architectures and configurations for stance polarity and intensity prediction. In total, we tested six different configurations using the two types of architectures (Combined or Split), BERT model sizes (Small or Large), and either freezing or unfreezing the BERT weights during training (frozen or unfrozen). Each configuration was trained using the same training, testing, and validation datasets. The training was done using early stopping if the validation loss did not improve for five consecutive epochs, with a maximum of 20 training epochs. The models were trained on an NVIDIA Quadro P4000 video card using Python with the huggingface Transformer libraries [23]. The details for each of the trained models are in Table 6.1. Due to the memory limits of the graphics card, we were not able to test the configuration with a large BERT model that had unfrozen weights during training.

Table 6.1: The configurations tested for fine-tuning BERT.

Architecture	**BERT Size**	**Frozen Weights**	**Learning Rate**	**Total Training Epochs**	**Best Validation Epoch**
Combined	Small	Yes	0.001	20	15
Combined	Small	No	$2.0 * 10^{-5}$	7	2
Combined	Large	Yes	0.001	20	18
Split	Small	Yes	0.001	20	17
Split	Small	No	$2.0 * 10^{-5}$	7	2
Split	Large	Yes	0.001	12	6

6.5.1.2 Comparing Model Performance for Stance Polarity and Intensity Prediction

To evaluate the performance of the models for stance polarity and intensity prediction, we report both RMSE of the testing dataset and a weighted percentage we call Regression Accuracy (Reg Acc), which takes the testing RMSE as a percentage of the maximum RMSE possible. The

maximum possible RMSE is calculated by measuring the worst possible prediction on the testing data labels.

To calculate the worst possible predictions, we created a prediction model that takes in a label and outputs the prediction with the most distance from that labels, while still being within range of an agreement value (-1.0, +1.0). If the label is less than one, the model will predict one, and if the label is greater than or equal to zero, it will predict a negative one. This model ensures the worst possible outcome. For our testing dataset, the maximum RMSE was 1.6833. The regression accuracy is then calculated, as shown in (1).

$$\text{Regression Accuracy} = 1 - \frac{Instance\ RMSE}{Maximum\ RMSE} \quad (1)$$

This representation displays the error as an accuracy value, such that a 0.0 regression accuracy would indicate the worst possible RMSE value, and a value of 1.0 would indicate perfect accuracy.

6.6 Results

6.6.1 Fine-Tuning BERT Results

The results for the various architectures and configurations for fine-tuning the BERT model are shown in Table 6.2. In every configuration, the Split architecture outperformed the combined architecture by around 5.6% in regression accuracy and 0.1 RMSE. The smaller pre-trained BERT model tended to perform slightly better compared with the larger model by around 0.66% for both Combined and Split architectures. Unfreezing the BERT model weights while training also increased performance by 2.66% on the Split model and 1.54% on the Combined architecture. The

best performing configuration used the Split architecture, the small pre-trained BERT model, and unfrozen parameters during training, and had a regression accuracy of 68.58% and RMSE of 0.528.

Table 6.2: The performance for each architecture and configuration for the Fine-Tuned BERT model for stance polarity and intensity prediction.

Architecture	BERT Size	BERT Parameters	Testing RMSE	Regression Accuracy
Combined	Small	Frozen	0.6576	60.94%
Combined	Small	Unfrozen	0.6316	62.48%
Combined	Large	Frozen	0.6772	59.77%
Split	Small	Frozen	0.5737	65.92%
Split	Small	Unfrozen	**0.5288**	**68.58%**
Split	Large	Frozen	0.5761	65.77%

6.6.2 Stance Polarity and Intensity Results

The results for comparing both the fine-tuned BERT model and the adapted models for stance polarity and intensity prediction on the testing dataset are shown in Table 6.3. The best Split BERT model (Split/Small/Unfrozen) significantly outperformed the best adapted model, SVR-RF-R, by slightly less than four points of regression accuracy and 0.068 RMSE.

Table 6.3: The results for the different stance polarity and intensity prediction models on the testing set.

Model	Model Type	RMSE	Reg Acc
Baseline	Mean value prediction	0.718	57.35%
pkudblab-PIP	Convolutional Neutral Network	0.657	60.97%
Best Combined BERT	Combined/Small/Unfrozen BERT Fine-Tune	0.632	62.48%
T-PAN-PIP	RNN + Attention	0.623	62.99%
Ridge-M	Ridge Regression	0.620	63.17%
Ridge-S	Ridge Regression	0.615	63.47%
SVR-RF-R	Ensemble	0.596	64.60%
Best Split BERT	Split/Small/Unfrozen BERT Fine-Tune	**0.528**	**68.58%**

The best Combined BERT model (Combined/Small/Unfrozen) performed in the middle of the pack of the adapted models. The adapted models performed similarly relative to one another on the stance polarity and intensity prediction task as they did on the stance detection task, with SVR-RF-R being the best model out of the adapted models.

A breakdown of the testing set results from the Best Split BERT model reveals that the instances with stance intensity are the extremes (near -1 or +1) were a larger source for error than the instances with lower intensities. Figure 6-3 shows the testing set results for the Best Split BERT model broken down by the ground-truth label. Intensities between the range −0.4 and +0.4 had an RMSE of 0.4 or below while the instances at the extremes (less than -0.6 and greater than 0.6) had RMSE values of 0.49 or above.

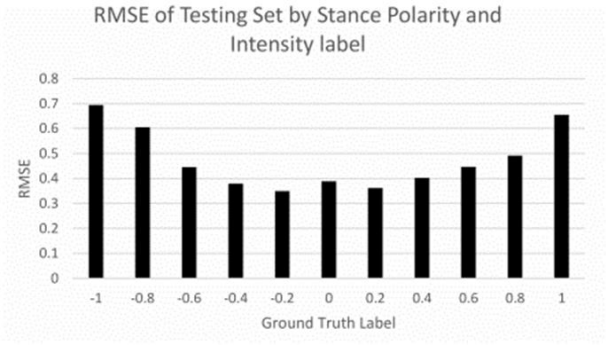


Figure 6-3 Breakdown of the testing set prediction RMSE of the Best Split BERT model by stance polarity and intensity label.

The input argument length and the topic issue of the instances had very little impact on the performance of the Best Split BERT model. The word count of the input argument had almost no relationship with prediction RMSE, with a correlation value of 0.0004. Likewise, the issue the argument originates from has very little impact on the error. Table 6.4 shows a breakdown of the Best Split BERT model's RMSE for testing data by the instance issue. The difference in RMSE between the best performing issue, Same Sex Adoption, and the worst performing issue, Guns on Campus, was only 0.0394.

Table 6.4: Breakdown of the testing set prediction RMSE of the best Split BERT model by issue.

Issue	RMSE
Same Sex Adoption	0.5101
Religion and Medicine	0.5204
Healthcare	0.5337
Guns on Campus	0.5495

Table 6.5: The classification accuracy and F1-scores of the stance polarity prediction models and the stance poalarity and tensity prediciton modes for stance dection (polarity only) classification.

Model	Accuracy	F1-score
Baseline (Mean)	54.36%	0.352
Ridge-M	68.16%	0.695
Ridge-S	68.84%	0.703
SVR-RF-R	70.43%	0.721
pkudblab-PIP	66.89%	0.672
T-PAN-PIP	66.64%	0.678
Best Split BERT	**76.02%**	**0.780**

The best Split BERT model also outperformed all the adapted models in the stance detection task (i.e. predicting the stance polarity only) as well, as shown Table 6.5, with an accuracy of 76.02% and F1-score of 0.780. This result is a 5.59% increase in accuracy over the best performing adapted model SVR-RF-R. Figure 6-4 shows a confusion matrix for the polarity predicted by the Best Split BERT model for the Favor (value greater than zero)and Oppose (value less than zero) categories. The neutral value (zero) was underrepresented in the testing set and omitted from the confusion matrix.

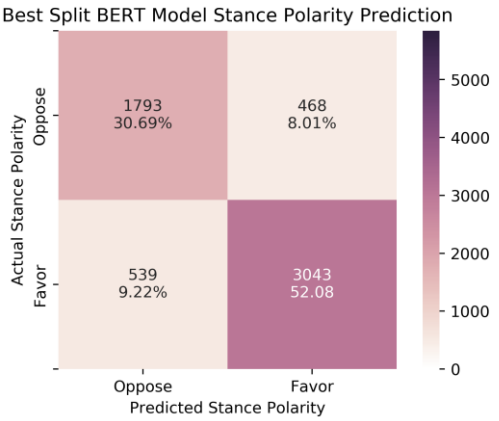


Figure 6-4 A confusion matrix for the Stance polarities of the testing dataset predicted by the Best Split BERT model.

These results suggest that the Best Split BERT model produces predictions that are consistent across the four different issues and across inputs of varying word counts and is very good at determining the polarity of the stance relationships with 76.02% accuracy. However, the model struggles to identify strong stance intensity in the relationships, with more error occurring when the actual stance intensity is closer to one.

6.7 Discussion

The experiments compared the overall performances of the fine-tuning BERT models with the adapted models reveals that the strategy of using pre-trained language models is beneficial for stance polarity and intensity prediction, but only when the Split BERT architecture was used. The Combined BERT architecture performed about the same as the other neural network models, T-PAN-PIP and pkudblabPIP, which were models that were trained from scratch and did not use a pre-trained model. Thus, a straightforward approach to incorporating the BERT model, such as the Combined architecture, does not provide any improvement in performance compared to the other models, while the Split architecture outperforms them in all the configurations. Overall, the adapted models' performance for stance polarity and intensity matched their relative performances

on stance detection, with SVR-RFR having the best performance, being only outperformed by the Split BERT model.

The Split architecture does have a larger output space since it has two outputs (one from each post), which could be causing the improved performance. However, we tested having multiple outputs with the Combined architecture (such as one output on the head [CLS] token and one on the middle [SEP] token that separates the parent and child posts). The results were still significantly worse than the Split architecture. Our results support the idea that encoding each post separately is more effective for a task that is identifying contrast between posts.

More broadly, this result suggests that when fine-tuning language models, finding the proper architecture for incorporating the pre-trained model is crucial for leveraging the benefits of transfer learning. The prior works using BERT did not explore various architectural setups, so it is not clear if the split architecture is advantageous for all stance detection applications or only our specific task of stance polarity and intensity prediction. However, in this case, it made a significant difference.

6.8 Conclusion

We continue exploring our new research problem, stance polarity and intensity prediction in response relationships between posts in online deliberation. This task encapsulates stance detection and includes the additional task of determining the intensity of the stance relationship. In this work, we developed a novel model that finetunes the pre-trained BERT language model for stance polarity and intensity prediction. We experimented using different architectures and configurations for fine-tuning the BERT model, including a novel Split architecture which encodes the parent and child posts separately through the BERT model and combines them in the output

layers. We trained and tested the models on an empirical dataset collected using a cyber argumentation platform. Our results demonstrate that the fine-tuned BERT model using the novel Split architecture was the best performing model on the dataset. To our knowledge, this finetuning architecture is new and has not been utilized in the stance detection literature prior. This Split architecture may prove useful in many other related tasks in stance detection and argumentation mining.

6.9 References

[1] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 task 6: Detecting stance in tweets," in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval2016). Association for Computational Linguistics, 2016, pp. 31–41. [Online]. Available: http://aclweb.org/anthology/S16-1003

[2] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in Emotion Measurement, H. L. Meiselman, Ed. Woodhead Publishing, 2016, pp. 201– 237. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780081005088000096

[3] M. Stede and J. Schneider, Argumentation Mining, ser. Synthesis Lecutres on Human Langauge Technologies. Morgan & Claypool Publishers, 2018, vol. 11. [Online]. Available: https://www. morganclaypool.com/doi/10.2200/S00883ED1V01Y201811HLT040

[4] A. E. Lillie and E. R. Middelboe, "Fake news detection using stance classification: A survey," arXiv:1907.00181 [cs], 2019. [Online]. Available: http://arxiv.org/abs/1907.00181

[5] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski, "RumourEval 2019: Determining rumour veracity and support for rumours," in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. pp 845 –854.

[6] P. Sobhani, D. Inkpen, and S. Matwin, "From argumentation mining to stance classification," in Proceedings of the 2nd Workshop on Argumentation Mining. Association for Computational Linguistics, 2015, pp. 67–77. [Online]. Available: http://aclweb.org/anthology/ W15-0509

[7] R. S. Arvapally and X. F. Liu, "Empirical evaluation of intellligent argumentation system for collaborative software project decision making," in 5th Annual ISC Research Symposium, 2011, p. 6.

[8] X. F. Liu, R. Wanchoo, and R. S. Arvapally, “Intelligent computational argumentation for evaluating performance scores in multi-criteria decision making,” in 2010 International Symposium on Collaborative Technologies and Systems, 2010, pp. 143–152.

[9] R. Arvapally and X. F. Liu, “Analyzing credibility of arguments in a web-based intelligent argumentation system for collective decision support based on k-means clustering algorithm: Knowledge management research & practice: Vol 10, no 4,” Knowledge Management Research & Preactice, vol. 10, no. 4, pp. 326–341, 2012. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1057/kmrp.2012.26

[10] R. S. Arvapally, X. F. Liu, and W. Jiang, “Identification of faction groups and leaders in web-based intelligent argumentation system for collaborative decision support,” in 2012 International Conference on Collaboration Technologies and Systems (CTS), 2012, pp. 509–516.

[11] J. W. Sirrianni, X. F. Liu, and D. Adams, “Quantitative modeling of polarization in online intelligent argumentation and deliberation for capturing collective intelligence,” in 2018 IEEE International Conference on Cognitive Computing (ICCC), 2018, pp. 57–64. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICCC.2018.00015

[12] R. S. Arvapally, X. F. Liu, F. F.-H. Nah, and W. Jiang, “Identifying outlier opinions in an online intelligent argumentation system,” Concurrency and Computation: Practice and Experience, p. e4107, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.4107

[13] P. Sobhani, S. Mohammad, and S. Kiritchenko, “Detecting stance in tweets and analyzing its interaction with sentiment,” in Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics, 2016, pp. 159–169. [Online]. Available: http://aclweb.org/anthology/S16-2021

[14] S. S. Mourad, D. M. Shawky, H. A. Fayed, and A. H. Badawi, “Stance detection in tweets using a majority vote classifier,” in The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018), ser. Advances in Intelligent Systems and Computing, A. E. Hassanien, M. F. Tolba, M. Elhoseny, and M. Mostafa, Eds. Springer International Publishing, 2018, pp. 375–384.

[15] W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang, “pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, 2016, pp. 384–388. [Online]. Available: http://aclweb.org/anthology/S16-1062

[16] K. Dey, R. Shrivastava, and S. Kaushik, “Topical stance detection for twitter: A two-phase LSTM model using attention,” in Advances in Information Retrieval, ser. Lecture Notes in

Computer Science, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Springer International Publishing, 2018, pp. 529–536.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[18] M. Fajcik, L. Burget, and P. Smrz, "BUT-FIT at SemEval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers," in Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2019, pp. 1097–1104. [Online]. Available: https://www.aclweb.org/anthology/ S19-2192

[19] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/ research-covers/language-unsupervised/languageunderstandingpaper.pdf

[20] C. Dulhanty, J. L. Deglint, I. B. Daya, and A. Wong, "Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection," in AI for Social Good workshop at NeurIPS, 2019. [Online]. Available: http://arxiv.org/abs/1911.11951

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in NIPS, 2017.

[22] W. Fang, M. Nadeem, M. Mohtarami, and J. Glass, "Neural multitask learning for stance prediction," in Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER). Association for Computational Linguistics, 2019, pp. 13–19. [Online]. Available: https://www.aclweb.org/anthology/D19-6603

[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "HuggingFace's transformers: State-of-the-art natural language processing," arXiv:1910.03771 [cs], 2019. [Online]. Available: http://arxiv.org/abs/1910.03771

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 [cs], 2014. [Online]. Available: http: //arxiv.org/abs/1412.6980

[25] R. Yang, W. Xie, C. Liu, and D. Yu, "BLCU nlp at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation," in Proceedings of the 13th International

Workshop on Semantic Evaluation. Association for Computational Linguistics, 2019, pp. 1090–1096. [Online]. Available: https://www.aclweb.org/anthology/S19-2191

Appendix

Appendix A: List of Published Papers

Chapter 3 through Chapter 6 are partial reproductions of papers that have been published or considered for publication at the following outlets:

Chapter 3:

> J. Sirrianni, X. (Frank) Liu, and D. Adams, "Quantitative Modeling and Analysis of Argumentation Polarization in Cyber Argumentation," Accepted for publication in IEEE Transactions on Computational Social Systems. Oct. 2020.
>
> The above journal article is an extension of the following published conference paper:
>
> > J. Sirrianni, X. (Frank) Liu, and D. Adams, "Quantitative Modeling of Polarization in Online Intelligent Argumentation and Deliberation for Capturing Collective Intelligence," in 2018 IEEE International Conference on Cognitive Computing (ICCC), San Francisco, California, USA, Jul. 2018, pp. 57–64, doi: 10.1109/ICCC.2018.00015.

Chapter 4:

> J. Sirrianni, X. F. Liu, M. M. Rahman, and D. Adams, "An Opinion Diversity Enhanced Social Connection Recommendation Re-ranking Method Based on Opinion Distance in Cyber Argumentation with Social Networking," in 2019 IEEE International Conference on Cognitive Computing (ICCC), Milan, Italy, Jul. 2019, pp. 106–113, doi: 10.1109/ICCC.2019.00029.

Chapter 5:

> J. Sirrianni, X. (Frank) Liu, and D. Adams, "Agreement Prediction of Arguments in Cyber Argumentation for Detecting Stance Polarity and Intensity," in The 58th Annual Meeting of the Association for Computational Linguistics, Seattle, Washington, USA, Jul. 2020, pp. 5746-5758, doi: 10.18653/v1/2020.acl-main.509.

Chapter 6:

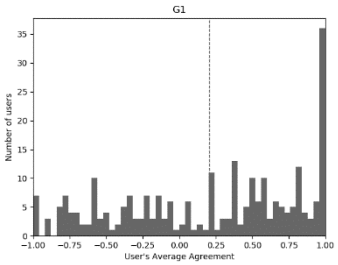
> J. Sirrianni, X. (Frank) Liu, and D. Adams, "Predicting Stance Polarity and Intensity in Cyber Argumentation with Deep Bi-directional Transformers," Submitted for review in IEEE Transactions on Computational Social Systems. Sept. 2020.

Appendix B: Full Polarization Model Results

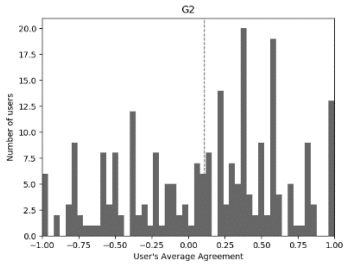
The full list of the polarization values for each of the positions in the dataset for each of the polarization models presented in Chapter 3 is shown in Table A.1. The labels for the positions are assigned as such: the first letter reflects the issue the position is under (G = Guns on Campus, H = Healthcare, S = Same Sex Adoption, R = Religion and Medicine), and the number represents the ideological tilt of the position (1 = Strong Conservative, 2 = Moderately Conservative, 3 = Moderately Liberal, 4 = Strong Liberal). The histograms of each of the user overall agreement distributions for each of the positions in the dataset are shown in Figure A-1.

Table A.1: The polarization value for all of the positions for each polarization model.

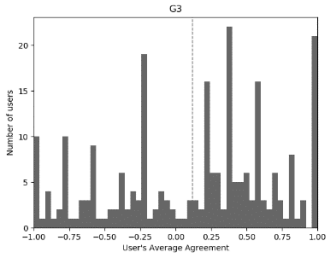
Position	**MAP**	**FM**	**MBLB**
G1	0	0.2649	0.4289
G2	0.0032	0.1998	0.3616
G3	0.0027	0.2350	0.4135
G4	0.0034	0.2328	0.2059
H1	0.0148	0.1808	0.3493
H2	0.0239	0.1819	0.2558
H3	0.0210	0.1629	0.2567
H4	0.0001	0.2259	0.4601
R1	0.0264	0.1781	0.3001
R2	0.0297	0.1652	0.1994
R3	0.0002	0.2183	0.4662
R4	0.0048	0.1741	0.1496
S1	0	0.2437	0.2269
S2	0.0029	0.2102	0.2674
S3	0.0059	0.2133	0.4633
S4	0	0.2914	0.2032



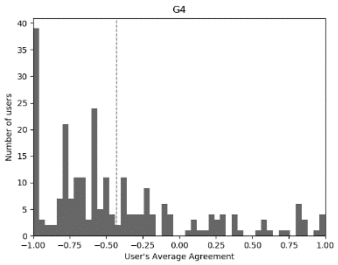
(a) Agreement Distribution for G1



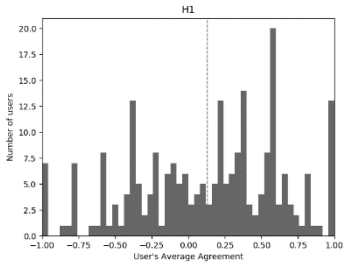
(b) Agreement Distribution for G2



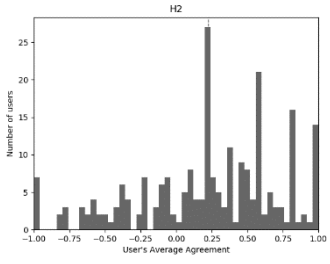
(c) Agreement Distribution for G3



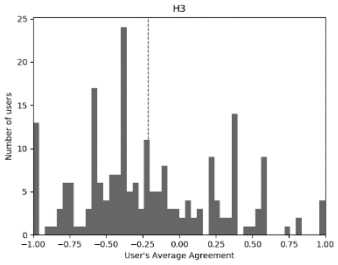
(d) Agreement Distribution for G4



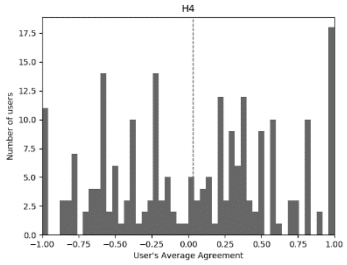
(e) Agreement Distribution for H1



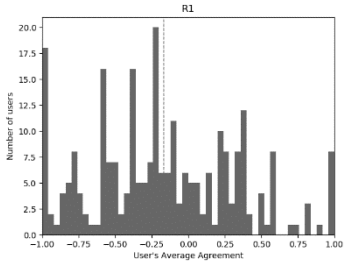
(f) Agreement Distribution for H2



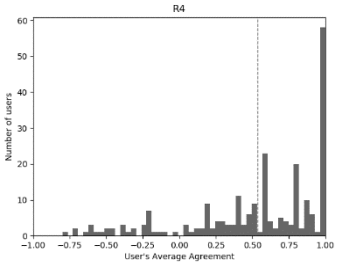
(g) Agreement Distribution for H3



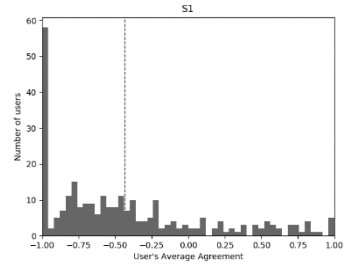
(h) Agreement Distribution for H4



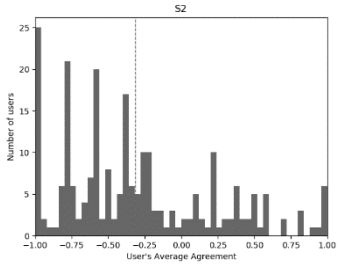
(i) Agreement Distribution for R1



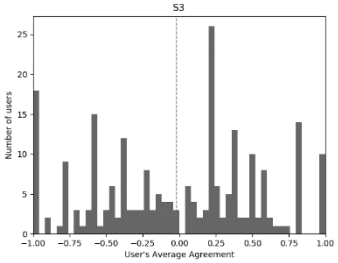
(j) Agreement Distribution for R4



(k) Agreement Distribution for S1



(l) Agreement Distribution for S2



(m) Agreement Distribution for S3

Figure A.1 Histogram of users by their overall average agreement for each remaining position in the spring 2018 dataset.

Appendix C: IRB Protocol Approval Letter

UNIVERSITY OF ARKANSAS

To:	Xiaoqing Liu JBHT 0504
From:	Douglas James Adams, Chair IRB Committee
Date:	03/17/2020
Action:	**Expedited Approval**
Action Date:	03/17/2020
Protocol #:	1710077940R002
Study Title:	Empirical Study of Structured Cyber Argumentation
Expiration Date:	11/12/2020
Last Approval Date:	03/17/2020

The above-referenced protocol has been approved following expedited review by the IRB Committee that oversees research with human subjects.

If the research involves collaboration with another institution then the research cannot commence until the Committee receives written notification of approval from the collaborating institution's IRB.

It is the Principal Investigator's responsibility to obtain review and continued approval before the expiration date.

Protocols are approved for a maximum period of one year. You may not continue any research activity beyond the expiration date without Committee approval. Please submit continuation requests early enough to allow sufficient time for review. Failure to receive approval for continuation before the expiration date will result in the automatic suspension of the approval of this protocol. Information collected following suspension is unapproved research and cannot be reported or published as research data. If you do not wish continued approval, please notify the Committee of the study closure.

Adverse Events: Any serious or unexpected adverse event must be reported to the IRB Committee within 48 hours. All other adverse events should be reported within 10 working days.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, study personnel, or number of participants, please submit an amendment to the IRB. All changes must be approved by the IRB Committee before they can be initiated.

You must maintain a research file for at least 3 years after completion of the study. This file should include all correspondence with the IRB Committee, original signed consent forms, and study data.

cc: Douglas James Adams, Investigator
Joseph W Sirrianni, Investigator
Md Mahfuzer Rahman, Investigator
Najla Althuniyan, Investigator
Zhang Hu, Investigator