Mining, Assessing, and Improving Arguments in NLP and the Social Sciences

Gabriella Lapesa[1], Eva Maria Vecchi[1], Serena Villata[2], Henning Wachsmuth[3]
[1]University of Stuttgart, Institute for Natural Language Processing
[2]Université Côte d'Azur, Inria, CNRS, I3S,
[3]Leibniz University Hannover, Institute of Artificial Intelligence
`name.surname@ims.uni-stuttgart.de`
`villata@i3S.unice.fr, h.wachsmuth@ai.uni-hannover.de`

Abstract

Computational argumentation is an interdisciplinary research field, connecting Natural Language Processing (NLP) to other disciplines such as the social sciences. This tutorial will focus on a task that recently got into the center of attention in the community: *argument quality assessment*, that is, what makes an argument good or bad? We structure the tutorial along three main coordinates: (1) the notions of argument quality across disciplines (how do we recognize good and bad arguments?), (2) the modeling of subjectivity (who argues to whom; what are their beliefs?), and (3) the generation of improved arguments (what makes an argument better?). The tutorial highlights interdisciplinary aspects of the field, ranging from the collaboration of theory and practice (e.g., in NLP and social sciences), to approaching different types of linguistic structures (e.g., social media versus parliamentary texts), and facing the ethical issues involved (e.g., how to build applications for the social good). A key feature of this tutorial is its interactive nature: We will involve the participants in two annotation studies on the assessment and the improvement of quality, and we will encourage them to reflect on the challenges and potential of these tasks.

1 Introduction

Computational argumentation is a field encompassing varying tasks on the automated analysis and synthesis of natural language arguments. Until recently, research in Natural Language Processing (NLP) mostly dealt with *Argument Mining* (AM), that is, the identification of argumentative claims that convey a stance towards some controversial issue, along with evidence given as reasons for the claims. AM has been studied for various genres (Mochales and Moens, 2011; Habernal and Gurevych, 2017; Dusmanu et al., 2017a) and argument models (Toulmin, 1958; Walton et al., 2008; Freeman, 2011). As Lawrence and Reed (2019) point out, the "reason giving" function of argumentation is what makes AM specific: "Although opinion mining and sentiment analysis provide techniques that are proving to be enormously successful [...] they can only tell us what opinions are being expressed and not why people hold the opinions they do". *Reason giving*, however, is only one of two main functions of argumentation, the other being *persuasion*. The dynamics of these "two souls" of argumentation are complex and the balance between reason giving and persuasion varies depending on the communication setting.

Mining Arguments In this tutorial, we start from the body of research on AM. Unlike recent NLP tutorials on argumentation (Budzynska and Reed, 2019; Bar-Haim et al., 2021), however, our focus is a task that recently got into the center of attention: *argument quality assessment*, that is, to rate or to compare how good arguments are with respect to one or more defined quality dimensions.

The NLP Perspective: Assessing Argument Quality Let us start with the concrete example of argument quality annotations in Figure 1, taken from Lauscher et al. (2020). The topic is "freedom of speech", and the stance is "against" (i.e., the government has the right to censorship). Quality is assessed here in four dimensions: *cogency* (is the conclusion adequately supported with relevant and sufficient premises?), *effectiveness* (how persuasive is the argument?), *reasonableness* (is the argument good in the context of the debate in which it is framed?), and *overall quality*.

The example illustrates the challenges which we take as coordinates of this tutorial. The first challenge is the identification and definition of appropriate *dimensions* for quality assessment: for example, in this case, the effectiveness label conflates several aspects. The second challenge in quality assessment is *subjectivity*. In our example, the three annotators (linguistics experts) clearly disagree in their assessment. Lauscher et al. (2020)

	Cogency	Effectiveness	Reasonableness	Overall
Title: Should 'blogging' be a capital crime? Iran is considering it... **Stance:** A government has the right to censor speech (...) **Text:** My government doesn't give me freedom of speech, so I have to argue for this side. Freedom of speech is bad because ... um ... then Our Leader's beliefs could be challenged. No one wants that. I mean, if everyone would just say and believe what Our Leader says to, we wouldn't need those firing squads altogether! Everyone wins.				
Annotator 1	4	1	1	2
Annotator 2	4	5	3	4
Annotator 3	2	2	2	2

Figure 1: Argument quality assessment from Lauscher et al. (2020): Example argument, annotated for four dimensions by three annotators, with partial agreement.

report that a crucial factor of disagreement of Annotators 1 and 2 was their perception of the ironic tone behind the text. Interestingly, for both of them, the text has a medium-high degree of cogency (so it is logically pretty "healthy"). A further challenge would be to improve the quality of this argument: How would we make this argument more effective? Do we need more irony, less irony, or a stronger statement of the stance?

To inform participants about argument quality, the tutorial will systematically review existing research on argument quality based on the literature (Wachsmuth et al., 2017), outlining the subjectiveness of quality dimensions as a key problem. In an interactive annotation session, participants will explore and discuss the assessment of quality on real-life arguments. They will be encouraged to take a critical standpoint to the annotation guidelines, learning in a concrete scenario how difficult it is to establish a trade-off between expressivity of the annotation schema and feasibility of the task.

The Social Science Perspective: Assessing Deliberative Quality   To demonstrate the impact of argument quality in practice, the tutorial will bridge research in NLP with the social sciences, looking at deliberative democracy in particular. Deliberative democracy is an approach to democratic processes which does not focus on the output of decision-making, but on the discourse exchange that precedes it (Bächtiger and Parkinson, 2019). Crucially, deliberative theory scholars have been asking the same question as computational argumentation: What makes a contribution to a discussion good? This has led to the development of a *discourse quality index* to assess the quality of a discourse contribution (Steenbergen et al., 2003; Gerber et al., 2016). While the dimensions of the index partially overlap with argument quality di-

mensions, some bring in a different perspective, for example, whether the discourse participants make reference to "common good" or whether they engage with other participants.

Modeling Subjectivity   Next, we will deal with subjectivity, modeling the parties involved in debates along with their values and beliefs. The connections of argument quality and deliberative quality highlight the subjective nature of argumentation, one of the three main coordinates of this tutorial. Subjectivity has been the trigger of an "affective turn" in both deliberative theory and computational argumentation. In the former, this has implied a switch from a purely rational perspective on deliberation to one which incorporates emotions, personal narratives, humor (Hoggett and Thompson, 2002; Black, 2020; Esau, 2018; Esau and Friess, 2022). In the latter, the affective turn has brought personal argumentation at center stage, highlighting the role played by human values (Kiesel et al., 2022), moral discourse (Alshomary et al., 2022), and narratives (Falk and Lapesa, 2022).

In the tutorial, we aim to foster participants to reflect on the two-fold role that subjectivity plays in quality assessment: subjective factors in quality assessment (e.g., interpretation of humor, as in the example above), and subjective factors in the production of an argument (e.g., all the "personal argumentation" ingredients listed before).

Improving Arguments   The subjectivity topic will lead to another interactive session where the goal is to improve the quality of arguments. Limitations will be discussed as well as first research on quality-related argument generation (Gurcke et al., 2021; Skitalinskaya et al., 2022), before the tutorial concludes with an outlook on future perspectives.

2 Diversity

We believe that exposing the students to the deliberative perspective of argumentation will be fruitful and enriching, as it might not be known to the typical *CL audience. It is our goal that participants leave our tutorial having learned the value of taking multiple disciplinary perspectives into account, even in a rather technical (logic- and NLP-oriented) subject such as computational argumentation. Besides, our focus on subjectivity and personal argumentation as positive features (and not bugs) brings individuals and their differences at center stage, contributing to inclusivity in the field.

3 Learning Outcomes

This introductory tutorial aims to offer an elaborate, systematic, and interdisciplinary understanding of the assessment and improvement of the quality of natural language arguments:

- Basics of argument mining, computational models of argumentation, argument quality assessment, and argument generation;

- Understanding of the NLP perspective: impact of assessing argument quality in practice;

- Understanding of the social sciences perspective: goals of deliberation (cooperative decision making) and real-world applications;

- Hands-on experience on the challenges of assessing and improving argument quality.

4 Tutorial Outline

Part I (60 min.) Mining Arguments

- Overview of computational argumentation
- Argument mining: Humans vs. computers
- Achieved results and open challenges

Part II (60 min.) The NLP Perspective: Assessing Argument Quality

- What makes an argument "good"?
- Logical, rhetorical, and dialectical dimensions of argument quality
- Subjectiveness as the key challenge for annotation and modeling
- Discussion of the notions of argument quality: Are they sufficient? Are they all necessary?

Part III (60 min.) Interactive Session 1

- Annotation: Assessment of sample arguments
- Consolidation: To what extent participants agree? Where not, and why?
- Discussion: What are alternative strategies to subjective quality annotation?

Part IV (45 min.) The Social Sciences Perspective: Assessing Deliberative Quality

- Direct democracy, deliberative theories, and e-deliberation
- Deliberative quality: Features and annotation
- Integration of deliberative features in computational architectures
- Application: Argument quality for social good

Part V (30 min.) Modeling Subjectivity

- Authors, audiences, and third parties
- Human values, moral foundations, narratives
- Issues with subjectivity

Part VI (60 min.) Interactive Session 2

- Annotation: Rewriting of sample arguments
- Consolidation: What was improved and how?
- Discussion: What can be improved, what not?

Part VII (45 min.) Improving Argument Quality

- Generation methods for improving arguments
- Challenges and lessons learned
- Conclusions and next steps for the field

5 Tutorial Breadth

The key objective of this tutorial is to provide a comprehensive overview of recent and current research on the assessment and improvement of quality in computational argumentation, in both NLP and the social sciences. We estimate that at most one quarter of the tutorial will cover our own work.

6 Presenters

Gabriella Lapesa leads the research group E-DELIB (*Powering-up E-DELIBeration: towards AI-supported moderation*) at the Institute for Natural Language Processing, University of Stuttgart. Her group works at the intersection between NLP (AM) and social science (Deliberative Theory) to develop methods and tools to support moderators in deliberative discussion. As a research associate in the project MARDY (*Modeling ARgumentation Dynamics in Political Discourse*, University of Stuttgart and Bremen), she works on NLP methods to scale-up the analysis policy debates in multiple textual sources (i.e., who claims what in the debate on immigration or Covid-19?). Gabriella has co-chaired the 9th Argument Mining workshop (2022). With Eva Maria Vecchi, she co-taught a course on interdisciplinary AM at ESSLLI 2022.[1]

Eva Maria Vecchi has a background in linguistics and mathematics and holds a Ph.D. degree in cognitive and neurosciences. She is a postdoctoral researcher at the Institute for Natural Language Processing at IMS Stuttgart, working on the E-DELIB

---

[1] https://sites.google.com/view/esslli2022-am-in-nlp-ss/

project. Her focus is on the interdisciplinary effort between NLP techniques for argument mining (AM) and theories in the social sciences with the goal of a more collaborative, productive, and ethical endeavor for e-Deliberation. She has taught courses and tutorials on AM and other topics, most recently with Gabriella Lapesa at ESSLLI 2022. Her current research aims at a better understanding of the role bias has in computational argumentation and e-Deliberation, particularly the impact it has on the models, implementation, and social aspects of computational argumentation.

Serena Villata is a research director in computer science at CNRS, and she pursues her research at the I3S laboratory in Sophia Antipolis (France). Her research area is computational argumentation, with a focus on legal and medical texts, political debates and social network harmful content (abusive language, disinformation). Her work conjugates argument-based reasoning frameworks with natural language arguments extracted from text. She is the author of over 150 scientific publications on the topic. She holds a Chair of the Interdisciplinary Institute for AI 3IA Côte d'Azur on "Artificial Argumentation for Humans". Serena has co-chaired the 7th Workshop on Argument Mining at COLING 2020. She has also given tutorials on Argument Mining at ESSLLI 2017[2] and IJCAI 2016[3].

Henning Wachsmuth is the head of the Natural Language Processing Group at Leibniz University Hannover. He is an internationally leading researcher on computational argumentation with more than 60 publications on the topic, many at major NLP and AI venues. Other interests include social bias mitigation, computational reframing, and explainable NLP. Henning has co-chaired the 6th Workshop on Argument Mining at ACL 2019, and has given tutorials on argumentation at ASIRF 2018 (Cole and Achilles, 2019), EuroCSS 2018,[4] KI 2019 (Benzmüller and Stuckenschmidt, 2019), and KI 2020 (Schmid et al., 2020). He is an initiator of the CLEF shared task series Touché on argument retrieval (Bondarenko et al., 2022), and co-chaired SemEval tasks on argument reasoning comprehension (Habernal et al., 2018), propaganda

technique detection (Da San Martino et al., 2020), and identifying human values in arguments.[5]

7 Target Audience / Prerequisites

The tutorial targets both participants who are new to the field of computational argumentation and those who need a comprehensive overview of techniques and applications. As the tutorial is interdisciplinary by design, it is also of interest to participants from a social sciences background who hope to integrate their knowledge within NLP. Finally, we expect the tutorial to attract attention from people interested in NLP techniques that currently impact the social and political world, in general. Basic knowledge of linguistics and computational linguistics is required. A general interest in the collaboration between NLP and social sciences is expected; relevant material, however, will be introduced without requiring prior knowledge.

8 Other Information

Tutorial Type: Introductory, 6 hours
Tutorial Materials: Tutorial materials and related information will be made available online.

9 Recommended Reading List

Survey Papers (Cabrio and Villata, 2018; Lawrence and Reed, 2019; Vecchi et al., 2021)
Mining Arguments (Habernal and Gurevych, 2017; Daxenberger et al., 2017; Dusmanu et al., 2017b)
Assessing Argument Quality (Wachsmuth et al., 2017; Lauscher et al., 2020; Marro et al., 2022)
Assessing Deliberative Quality (Steenbergen et al., 2003; Gerber et al., 2016)
Improving Arguments (Hua and Wang, 2018; Alshomary et al., 2020; Gurcke et al., 2021)
Challenges (Durmus et al., 2019; Toledo-Ronen et al., 2020; Spliethöver and Wachsmuth, 2020)

Acknowledgments

---

[2]https://www.irit.fr/esslli2017/courses/39.html
[3]https://ijcai-16.org/index.php/welcome/view/accepted_tutorials/
[4]http://symposium.computationalsocialscience.eu/2018/

[5]Ongoing task found here: https://touche.webis.de/semeval23/touche23-web

References

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online. Association for Computational Linguistics.

Andre Bächtiger and John Parkinson. 2019. *Towards a New Deliberative Quality*. Oxford University Press, Cambridge, MA, USA.

Roy Bar-Haim, Liat Ein-Dor, Matan Orbach, Elad Venezian, and Noam Slonim. 2021. Advances in debating technologies: Building AI that can debate humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Christoph Benzmüller and Heiner Stuckenschmidt, editors. 2019. *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, volume 11793 of *Lecture Notes in Computer Science*. Springer.

Laura Black. 2020. Framing democracy and conflict through storytelling in deliberative groups. *Regular Issue*, 9(1).

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of touché 2022: ArgumentÂ retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 311–336, Cham. Springer International Publishing.

Katarzyna Budzynska and Chris Reed. 2019. Advances in argument mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 39–42, Florence, Italy. Association for Computational Linguistics.

Elena Cabrio and S. Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 5427–5433.

Amelia W. Cole and Linda Achilles. 2019. Autumn school for information retrieval and foraging 2018. *SIGIR Forum*, 52(2):87?91.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *14th International Workshop on Semantic Evaluation (SemEval 2020)*, pages 1377–1414, Barcelona (online). Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017a. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017b. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322.

Katharina Esau. 2018. Capturing citizens' values: On the role of narratives and emotions in digital participation. *Analyse & Kritik*, 40(1):55–72.

Katharina Esau and Daniel Friess. 2022. What creates listening online? exploring reciprocity in online political discussions with relational content analysis. *Journal of Deliberative Democracy*, 18(1):1–16.

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553, Dublin, Ireland. Association for Computational Linguistics.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

Marlené Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2016. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1).

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval 2018 Task 12: The Argument Reasoning Comprehension Task. In *12th International Workshop on Semantic Evaluation (SemEval 2018)*. Association for Computational Linguistics.

Paul Hoggett and Simon Thompson. 2002. Toward a democracy of the emotions. *Constellations*, 9(1):106–126.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of EMNLP*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Ute Schmid, Franziska Klügl, , and Diedrich Wolter, editors. 2020. *KI 2020: Advances in Artificial Intelligence: 43rd German Conference on AI, Bamberg, Germany, September 21–25, 2020, Proceedings*. Springer-Verlag, Berlin, Heidelberg.

Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. 2022. Claim optimization in computational argumentation.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

M. Steenbergen, Andre Baechtiger, Markus Spörndli, and J. Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.