Learning to Improve Arguments:

Automated Claim Quality Assessment and Optimization

Dissertation by Gabriella Skitalinska

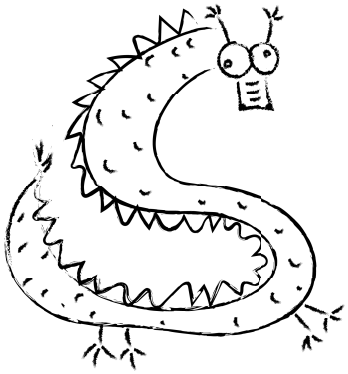
Dissertation by Gabriella Skitalinska submitted for the degree of Doctor of Engineering (Dr.-Ing.) to the faculty of Mathematics and Computer Science at the University of Bremen.

Bremen, December 13th, 2023

Supervisors
Prof. Dr. Andreas Breiter, University of Bremen
Prof. Dr. Henning Wachsmuth, Leibniz University Hannover

hic sunt dracones

* used by medieval cartographers to denote
unexplored or uncharted territories
typically accompanied by illustrations of legendary monsters



Abstract

Possessing strong argumentative writing skills is a crucial competency for academic and professional success. Such skills enable individuals to articulate their thoughts, beliefs, and opinions effectively while engaging in constructive discourse. Not only do they facilitate personal expression, but they also foster critical thinking and the ability to communicate persuasively. However, argumentation skills are challenging to acquire, especially for novice writers. This prompts the need to develop scalable *computational* solutions capable of guiding writers in improving their argumentative writing skills and assisting them in effectively communicating their ideas, regardless of their skill level.

Despite recent advancements in machine learning and natural language processing and extensive studies on argument quality in the past, the questions of automating argumentative writing support remain largely unexplored. In this thesis, we aim to address this gap and explore the following research question: *What makes a good argument and how can we computationally model this knowledge to develop tools supporting individuals in improving their arguments?* To do so, we suggest using *human revisions of argumentative texts* as a basis to understand and model quality characteristics of arguments. We argue that akin to how individuals learn through revisions to recognize gaps in their reasoning, organize ideas, and convey arguments in a clear and concise manner, computational models can be similarly conditioned to develop such competencies.

In this thesis, we make several contributions to the field of computational argumentation, specifically in automated argument assessment and generation. In particular, we introduce several new tasks focusing on identifying low quality content, characterizing the flaws within them, and suggesting types of improvement to increase their quality. The differences between the tasks and

their scope allow for a more nuanced and targeted assessment when capturing argument quality, making them applicable to a wide range of content quality control applications in online moderation or education processes.

To enable making such assessments with cutting-edge computational methods, we compile the first large-scale corpus of argumentative claim revisions from a popular online debate platform. With this data in hand, we investigate the important aspects, inter-dependencies, and attributes that shape the perceived quality of the argument and assess the impact of the revision processes on the various dimensions of argument quality. We find that working with revision-based data offers many opportunities and allows us to learn a more general notion of argument quality, which generalizes well across the topics, aspects, and stances covered in argumentative text. However, it also comes with several challenges related to the representativeness and reliability of data, topical bias in revision behaviors, appropriate model complexities and architectures, and the need for context when judging claims. In a detailed analysis, we outline the strengths and weaknesses of various approaches and strategies exploiting different types of knowledge specific to text and argument revisions to tackle said challenges. For example, we find that using revision distance-based sampling can improve performance when identifying claims that require improvement and incorporating contextual information allows to make more accurate quality assessments.

Finally, keeping in mind the lessons learned from quality assessment tasks, we address the problem of automatically generating improved versions of argumentative texts. Specifically, we propose a neural approach that first generates a diverse range of candidate claims and then selects the best candidate via a ranking process using several argument and text quality metrics. We empirically show that our approaches can perform a diverse range of improvement types and successfully revise argumentative texts. Moreover, the results show that the proposed solutions generalize well to other domains, such as instructional texts, news, scientific articles, and encyclopedia entries.

With this work, we take another step towards automatically assessing the quality of argumentative texts and generating their improved versions. We have done so by adopting a new perspective that looks at argument quality through the lens of revisions. By proposing a set of methods that can guide writers and help them improve their argumentative writing skills and produce more compelling and persuasive texts, we showcase that, with the right approach, the art of persuasion becomes an attainable endeavor.

Zusammenfassung

Starke argumentative Schreibfähigkeiten sind eine entscheidende Kompetenz für akademischen und beruflichen Erfolg. Solche Fähigkeiten ermöglichen es dem Einzelnen, seine Gedanken, Überzeugungen und Meinungen effektiv zu artikulieren und gleichzeitig einen konstruktiven Diskurs zu führen. Sie verbessern nicht nur die persönliche Ausdrucksweise, sondern fördern auch kritisches Denken und die Fähigkeit, überzeugend zu kommunizieren. Allerdings ist es gerade für unerfahrene Autoren eine Herausforderung, sich Argumentationskompetenzen anzueignen. Daraus ergibt sich die Notwendigkeit, skalierbare *maschinengestützte* Lösungen zu entwickeln, welche Autoren, unabhängig von ihrem Fähigkeitsniveau, dabei unterstützen, argumentative Schreibfähigkeiten zu verbessern und Ideen effektiv zu kommuninizieren.

Trotz jüngster Fortschritte beim maschinellen Lernen und der Verarbeitung natürlicher Sprache, sowie umfangreichen Studien zur Argumentqualität in der Vergangenheit, bleiben die Fragen der Automatisierung der Unterstützung argumentativen Schreibens weitgehend unerforscht. In dieser Arbeit wollen wir diese Lücke schließen und die folgende Forschungsfrage untersuchen: *Was macht ein gutes Argument aus und wie können wir dieses Wissen computergestützt modellieren, um Werkzeuge zu entwickeln, die Einzelpersonen bei der Verbesserung ihrer Argumente unterstützen?* Dazu schlagen wir vor, *menschliche Überarbeitungen argumentativer Texte* als Grundlage zu verwenden, um Qualitätsmerkmale von Argumenten zu verstehen und zu modellieren. Wir argumentieren, dass ähnlich wie Einzelpersonen durch Überarbeitungen lernen, Lücken in Ihrer Argumentation zu erkennen, Ideen zu ordnen und Argumente klar und prägnant zu vermitteln, auch auf maschinellem Lernen basierende Modelle ähnlich konditioniert werden können, um ebendiese Kompetenzen zu entwickeln.

In dieser Arbeit leisten wir mehrere Beiträge zum Bereich der computergestützten Argumentation, insbesondere zur automatisierten Argumentbewertung und -generierung. Konkret führen wir mehrere neue Aufgaben ein, die sich darauf konzentrieren, minderwertige Inhalte zu identifizieren, die darin enthaltenen Mängel zu charakterisieren und Arten von Verbesserungen zur Steigerung ihrer Qualität vorzuschlagen. Die Unterschiede zwischen den Aufgaben und ihrem Umfang ermöglichen eine differenziertere und gezieltere Beurteilung bei der Erfassung der Argumentqualität und machen sie für eine Vielzahl von Anwendungen zur Inhaltsqualitätskontrolle in Online-Moderationen oder Bildungsprozessen anwendbar.

Um solche Bewertungen mit modernsten Computermethoden zu ermöglichen, stellen wir den ersten groß angelegten Korpus von (mehrstufigen) Revisionen argumentativer Behauptungen aus einer beliebten Online-Debattenplattform zusammen. Mit diesen Daten untersuchen wir die wichtigen Aspekte, gegenseitigen Abhängigkeiten und Attribute, die die wahrgenommene Qualität des Arguments prägen, und bewerten die Auswirkungen der Revisionsprozesse auf die verschiedenen Dimensionen der Argumentqualität. Wir stellen fest, dass die Arbeit mit revisionsbasierten Daten viele Möglichkeiten bietet und es uns ermöglicht, einen allgemeineren Begriff der Argumentqualität zu erlernen, der sich gut auf die in argumentativen Texten behandelten Themen, Aspekte und Standpunkte übertragen lässt. Die Daten bringen jedoch auch mehrere Herausforderungen mit sich, die sich auf die Repräsentativität und Zuverlässigkeit der Daten, thematische Voreingenommenheit im Revisionsverhalten, geeignete Modellkomplexitäten und -architekturen sowie die Notwendigkeit des Kontexts bei der Beurteilung von Ansprüchen erstrecken. In einer detaillierten Analyse skizzieren wir die Stärken und Schwächen verschiedener Ansätze und Strategien, die sich unterschiedliche Arten von argumentations- und textanalysespezifischen Wissen und Methodiken zu Nutze machen, um diese Herausforderungen zu bewältigen. Wir stellen beispielsweise fest, dass die gesonderte Beachtung von Distanzen zwischen einzelnen Revisionen die Identifizierung von verbesserungswürdigen argumentativen Behauptungen verbessern kann, des weiteren stellen wir fest, dass die Einbeziehung von Kontextinformationen genauere Qualitätsbewertungen ermöglicht.

Abschließend beschäftigen wir uns, unter Berücksichtigung der erarbeiteten Erkenntnisse aus der vorangegangenen Forschung bezüglich der Bewertung der Argumentqualität, mit dem Problem der automatischen Generierung verbesserter Versionen argumentativer Texte. Konkret schlagen wir einen

neuronalen Ansatz vor, der zunächst eine Vielzahl von Kandidaten für eine gegebene argumentative Behauptung generiert und dann über einen Ranking-Prozess unter Verwendung mehrerer Argument- und Textqualitätsmetriken den besten Kandidaten auswählt. Wir zeigen empirisch, dass unsere Ansätze vielfältige Verbesserungen bewirken und argumentative Texte erfolgreich überarbeiten können. Darüber hinaus zeigen die Ergebnisse, dass sich die vorgeschlagenen Lösungen gut auf andere Bereiche wie Lehrtexte, Nachrichten, wissenschaftliche Artikel und Enzyklopädieeinträge übertragen lassen.

Mit dieser Arbeit gehen wir einen weiteren Schritt hin zur automatisierten Beurteilung und Verbesserung der Qualität argumentativer Texte. Dies ist uns gelungen, indem wir eine neue Perspektive eingenommen haben, die die Qualität der Argumente durch die Linse von Revisionen betrachtet. Indem wir eine Reihe von Methoden vorschlagen, die Autoren anleiten und helfen können, ihre argumentativen Schreibfähigkeiten zu verbessern und überzeugendere Texte zu produzieren, zeigen wir, dass die Kunst der Überzeugung mit dem richtigen Ansatz zu einem erreichbaren Unterfangen wird.

Acknowledgments

Completing this doctoral dissertation has been a roller coaster journey with its share of thrilling highs, stomach-churning lows, and some loops that had me wondering why I got on this ride in the first place. But, just like any amusement park adventure, I couldn't have done it alone, so I'd like to express my gratitude to the people and organizations who have supported me throughout this endeavor.

First, I would like to thank my advisors, Henning Wachsmuth and Andreas Breiter. Thank you for keeping me on track with your mentorship and guidance. Your expertise and insightful feedback have been instrumental in shaping my understanding of what it means to be a researcher. Stepping back, I would also like to thank Mikhail Alexandrov, who was the supervisor of my bachelor and master theses. Thank you for igniting my passion for academic exploration and ultimately inspiring me to pursue a Ph.D. In addition, I'd like to thank my employers, the University of Bremen, Leibniz University Hannover, and the Deutsche Forschungsgemeinschaft (DFG 1342, project number 374666841), who funded my research.

To my fellow grad students and collaborators, who are busy catching thrill rides of their own — I wish I met you sooner. Without you, this roller coaster would have been far less enriching and enjoyable.

Lastly, I would like to thank Jonas Klaff, who dedicated the past four years to supporting me through all the highs and lows of this journey ~~while subtly recalibrating my inner compass so that it always points to *beach*~~. Thank you for making sure I get enough papers ~~sun~~, stay motivated ~~hydrated~~, and surf not only google.scholar, but also proper waves in the Indian Ocean.

Contents

1. Introduction

Argumentation skills are considered one of the most critical competencies for academic and professional success. However, the ability to compose a coherent argument and effectively deliver it to an audience can be challenging to acquire. In this chapter, we first motivate the benefit of assisting individuals in argumentative writing, public discourse and deliberative processes with computational methods (Section 1.1). Then, we discuss the limitations of current approaches to computational argumentation, specifically, argument quality assessment and argument generation, that impede the development of effective and robust solutions (Section 1.2). Section 1.3 describes how we propose to tackle these problems by learning from human revisions of argumentative texts using natural language processing techniques. In Section 1.4, we present our main contributions, followed by the publication record in Section 1.5. Finally, we conclude with the outline of this dissertation (Section 1.6).

1.1 Motivation

Argumentation plays a major role in the advancement of society and is an integral part of everyday life, particularly in decision-making. As individuals, we routinely engage in argumentation when deciding whether or not to buy a house, switch to a healthier lifestyle, or adopt a pet. As such, we use arguments to form a personal opinion about something. On the other hand, arguments can be viewed as a form of communication in which individuals express their opinions, beliefs, and ideas to persuade others or to come to an agreement through deliberation (Atkinson et al., 2013). For example, as a society, we need to decide how to tackle issues related to human rights, environmental problems, population control, or global poverty. Making such choices is not

simple and involves reasoning about the causes and consequences of such issues, the advantages and disadvantages of proposed solutions, and their alternatives. Developing argumentation skills enables people to assess such problems of varying complexity, structuredness, and context and to make fair and informed choices even when confronted by problems with no definite solutions. However, not only the content of the argument and the reasoning abilities of the person are instrumental in such decision-making or persuasion (Oswald, 2011), but also their speaking and writing skills which enable them to present their arguments in a clear and compelling manner (Aristotle, 2007).

The rise of various discussion forums and online debate platforms, such as Idebate[1], CreateDebate[2], and Kialo[3] has given users many opportunities to share and express their opinions, views, and thoughts over a wide range of topics, making public discourse more accessible to people with any background. Prior, such opportunities to participate in debates were mainly offered in "*debate clubs*". Such clubs were particularly prominent in countries such as the UK and USA and were often associated with higher education institutions. However, while debate clubs have historically been valuable resources for individuals looking to develop their argumentation skills, they have often been exclusive and not available to everyone. Figure 1.1 shows examples of three debates found on Kialo related to abortion rights, vaccinations, and even fictional scenarios such as the immorality of killing vampires. While the first two debate topics discuss common concerns around human rights, which are practically relevant to our society, the third debate presents a fictional scenario to provoke debate about societal issues. Though such debates may not be directly related to real-world problems, they still can be seen as instrumental in fostering argumentation skills, as they enable one to imagine vividly different possibilities, present issues in an open-ended way, and are engaging.

As the accessibility of online platforms for discussions and debates increases, it becomes increasingly important to facilitate and support informed and rational discourse, particularly in political contexts. This includes ensuring effective communication, rational debate, and sensible collective behaviors. However, research has shown that people generally lack proficiency in argumentation, often mistaking opinions for fact-based claims and disregarding conflicting viewpoints instead of refuting them (Byrne, 1989; Wolfe et al., 2009) or simply possess poor argumentative writing skills rendering them unable

[1] https://idebate.net
[2] https://www.createdebate.com
[3] https://www.kialo.com

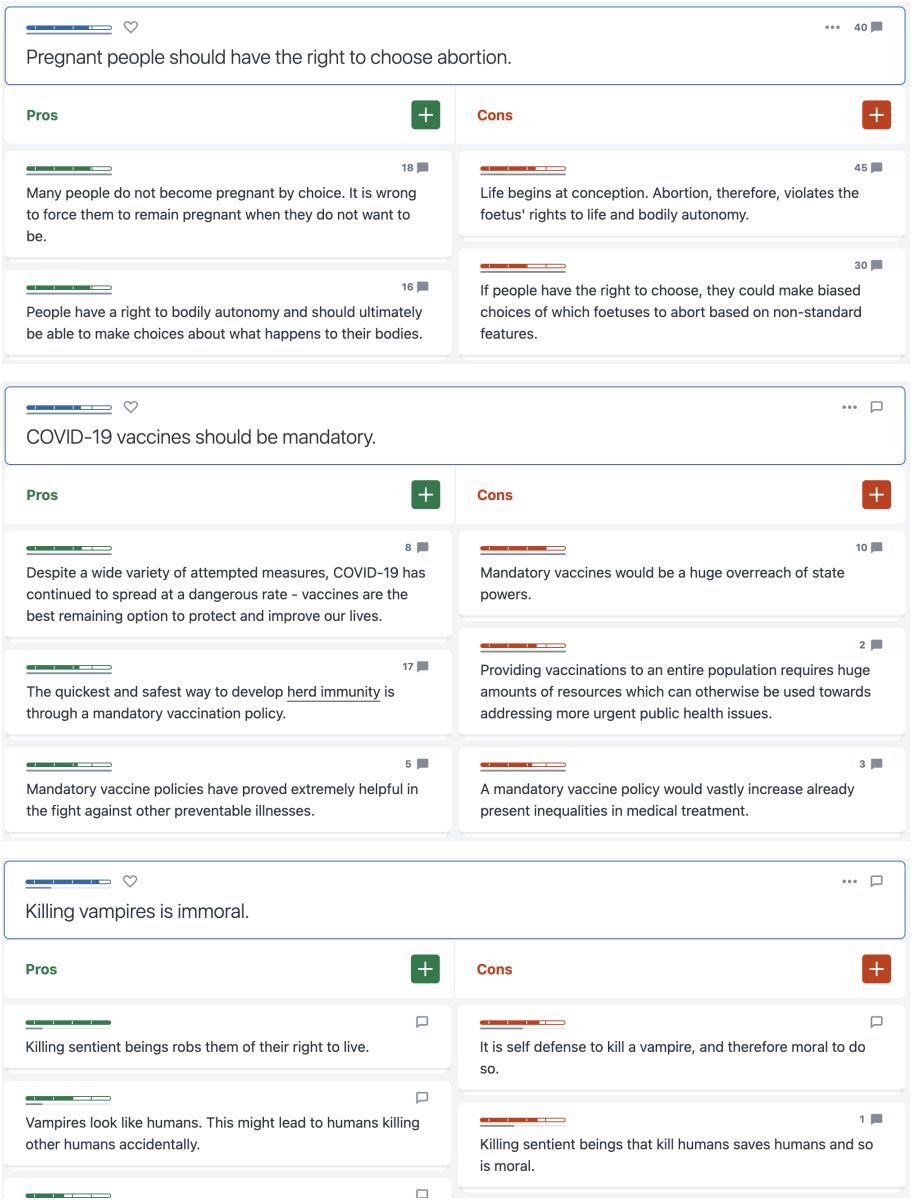


Figure 1.1: Examples of three debates found on Kialo covering various topics, such as COVID-19 vaccinations, abortion rights, and the immorality of killing vampires, consisting of the main thesis of the debate along with several supporting and opposing arguments. Each argument also includes an impact meter reflecting its impact on the thesis of the debate based on user votes, while the meter for the debate's thesis reflects its veracity. Finally, each argument has a comments section, where participants can "reply" to the argument with comments, questions, or revision requests, all of which would be visible to all participants in the debate.

to clearly express their viewpoint (Knudson, 1992; Butler and Britt, 2011). To overcome such issues, common teaching practices suggest using *text revision combined with personal feedback* as an effective tool for learning argumentation skills (Ferretti and Lewis, 2013).

By learning to perform text revisions, individuals learn how to identify gaps in their reasoning, to organize their ideas more coherently and to present their arguments in a clear and concise manner (Flower et al., 1986; Fitzgerald, 1987). However, learning when and how to successfully execute revisions can be challenging, especially for novice writers. While some revisions simply fix grammar issues, remove typos and biased language, others can reduce ambiguity, further clarify unclear points, or even change the structural layout of the original text, requiring expertise not only in the topic at hand but also in text editing. Thus, additional guidance, specifically instructional feedback, is required to allow learners to gain insights into their argumentation strengths and weaknesses and learn to identify areas for improvement (Ferretti and Lewis, 2013). Such feedback is typically given by teachers or peers (Afrin et al., 2021; Latifi et al., 2021; Beck et al., 2019); however, providing it can be time-consuming and impractical to deliver on a case-by-case basis both in educational and online public debate scenarios.

This prompts the need to develop scalable *computational* solutions capable of supporting participants in effectively communicating their ideas regardless of their skill level. Such technologies could inform future revisions and guide the development of argumentation skills while helping ensure that the participants in the discourse treat each other with respect, stay on topic, take up other opinions, and sufficiently justify their positions. As argumentative writing plays a critical role in professional development as well, we believe that research enabling automated *quality assessment of argumentative texts* and *generation of actionable feedback and suggestions* can be useful for assisting argumentative writing even outside of educational scenarios, for example, by supporting online participation in deliberative processes and public debates. Furthermore, such technologies could provide a foundation for a deeper understanding of the linguistic characteristics that distinguish high-quality arguments.

1.2 Computational Argumentation: Background and Challenges

In order to address the above-mentioned concerns and enable the creation of automated tools to support argumentative writing, this thesis focuses on understanding the factors of argument quality using methods of computational argumentation.

Computational argumentation is a sub-field of artificial intelligence that focuses on the application of computational methods for analyzing and synthesizing argumentation and human debate. It is based on techniques from natural language processing and computational linguistics, which are concerned with the development of methods to automatically interpret, synthesize, and comprehend human language in general. Computational argumentation covers a wide range of problems distributed across four main subareas (Lauscher et al., 2022):

- *Argument mining* considers the extraction of argumentative structures of varying granularity from natural language text, for example, whole argumentative sentences or argument components, such as claims or supporting evidence (e.g., Lippi and Torroni (2016); Stab and Gurevych (2017a); Cabrio and Villata (2018)).

- *Argument assessment* methods are used to determine particular properties of arguments in their context, for example, their stance towards some target (Bar-Haim et al., 2017), covered aspects (Ajjour et al., 2019a) or various dimensions of argument quality, such as clarity (Persing and Ng, 2013), strength (Habernal and Gurevych, 2016) or persuasiveness of the arguments (Toledo et al., 2019).

- *Argument reasoning* deals with understanding the reasoning process behind an argument and covers such tasks as predicting the entailment relationship between a premise and a hypothesis (Williams et al., 2017) or recognizing fallacies of certain reasoning types (Habernal et al., 2018; Delobelle et al., 2019).

- *Argument generation* tackles a variety of synthesis tasks, such as summarizing arguments (Wang and Ling, 2016; Syed et al., 2020), or generating potential counter-arguments (Hidey and McKeown, 2019; Alshomary et al., 2021c), conclusions based on premises (Alshomary et al., 2020), or additional supporting claims (Zukerman et al., 2000; Schiller et al., 2021; Al Khatib et al., 2021).

In this thesis, we will primarily focus on two subareas of computational argumentation: argument quality assessment and argument generation, as on the one hand, we are interested in understanding the characteristics of argumentative texts and the factors influencing the user's perception of such texts, e.g., what makes an argument persuasive, strong, acceptable, or sufficient (Wachsmuth et al., 2017a), to provide feedback on how to increase the text quality further. Yet, on the other hand, we also intend to use the learned insights to generate *optimized* versions of written text, which individuals could use as additional guidance in improving their content. Here, we define an "*optimized*" version as a text version that improves upon the original text in terms of overall argument quality while preserving the original meaning as far as possible (Skitalinskaya et al., 2023).

But what makes a good argument? Though different quality dimensions have been considered in argumentation, a common understanding of argument quality is still missing. For example, researchers studying argument quality from a theoretical viewpoint defined such quality dimensions as cogency (Johnson and Blair, 2006), fallaciousness (Hamblin, 1970), strength (Perelman, 1971), effectiveness (Perelman, 1971) and reasonableness (Van Eemeren, 2015). However, practitioners object that such quality dimensions are difficult to capture with computational methods in practice and focus on assessing more practical quality dimensions such as clarity (Persing and Ng, 2013), acceptability (Cabrio and Villata, 2018), or relevance (Wachsmuth et al., 2017c). To do so, typically, methods employing absolute or relative assessments to rate or compare various arguments are considered, with absolute assessments being less common, as necessary annotations are more difficult to obtain and model (Toledo et al., 2019).

Transferring such comparisons into actionable feedback to guide future improvements of the text is not very straightforward and, in certain cases, not possible without additional information. Let's consider the following scenario, Argument A is compared to Argument B in terms of relevance to the topic of discussion, and A is found to be more relevant than B. Does this mean that B is of low relevance and needs improvement? What if A and B are both of low relevance, with A being only marginally more relevant? Here, it is largely unclear how to interpret the result to create any feedback on improving either argument. Moreover, although extensive research has been done on various general text editing tasks, such as paraphrasing (Max and Wisniewski, 2010), sentence simplification (Botha et al., 2018), grammatical error correction

(Lichtarge et al., 2019) and bias neutralization (Pryzant et al., 2020), no work so far has studied how to actually improve *arguments* using text generation methods. Thus, we can formulate the first research gap as follows:

RESEARCH GAP 1:
While existing approaches based on natural language processing techniques can capture single quality dimensions and help characterize good arguments, they are not designed to identify issues within them or guide writers on how to improve the quality of their arguments, let alone generate improved versions of their texts automatically.

To further complicate matters, argument quality can be considered on different granularity levels, such as argument components, arguments, or full debates, and from various perspectives (Wachsmuth et al., 2017b). Many of these perspectives depend on personal beliefs, stance on an issue, and the weighting of different aspects of the topic, making them inherently subjective (Kock, 2007). Existing research largely ignores this limitation of inherent subjectivity and learns to predict argument quality based on subjective assessments of human annotators (Persing and Ng, 2013; Stab and Gurevych, 2017b; Toledo et al., 2019; Gretz et al., 2020). Some efforts have been made to control for *topic and stance* when comparing the convincingness of arguments (Habernal and Gurevych, 2016), for *personality and ideology* when assessing their effect on quality perception (Lukin et al., 2017; El Baff et al., 2020), or for the *argument* itself by abstracting from it and assessing its relevance only structurally (Wachsmuth et al., 2017c). However, *none of these approaches controls for the concrete aspects of a topic that the arguments claim and reason about*, which leads us to the following research gap:

RESEARCH GAP 2:
A general approach to assessing argument quality independently of the specific subject matter, beliefs and biases of the participants and audience, or context of the argument is still missing.

Another issue to consider is the lack of annotated corpora that deal with argumentative texts. Even though there are several datasets covering different

	Paper introducing Dataset	Quality dimension	Source	Size
Education	(Persing et al., 2010)	organization	student essays	34 topics, 1 003 essays
	(Persing and Ng, 2013)	thesis clarity	student essays	13 topics, 830 essays
	(Feng et al., 2014)	global coherence	student essays	34 topics, 1 003 essays
	(Rahimi et al., 2015)	evidence	student essays	11 topics, 2 392 essays
	(Persing and Ng, 2015)	strength	student essays	10 topics, 1 000 essays
	(Stab and Gurevych, 2017b)	sufficiency	student essays	402 essays
	(Zhang et al., 2017)	diverse	student essays	60 essays (3 drafts each)
	(Kashefi et al., 2022)	diverse	student essays	86 essays (3 drafts each)
Web-based media	(Cabrio and Villata, 2012)	acceptability	debate portals	200 argument pairs
	(Boltužić and Šnajder, 2015)	prominence	forum discussions	4 topics, 3 104 sentences
	(Habernal and Gurevych, 2016)	convincingness	debate portals	32 topics 16K arg. pairs
	(Wachsmuth et al., 2017c)	relevance	diverse	26 012 arguments
	(Gleize et al., 2019)	convincingness	Wikipedia	118 topics, 1 884 sentences
	(Toledo et al., 2019)	convincingness	debate portals	6K arguments
	(Durmus et al., 2019a)	impact	debate portals	471 topics, 47k arguments
	(Gretz et al., 2020)	overall quality	crowd-sourced	71 topics, 30K arguments
	(Ng et al., 2020)	diverse	diverse	3 domains, 5 284 arguments

Table 1.1: Available datasets for argument assessment tasks, grouped by their application domain.

topics and domains, the existing annotated data is still limited, especially in terms of size (typically less than 2000 arguments/argumentation units per dataset; and less than 50k arguments/argumentation units per dataset in "*big datasets*"). As no corpora enabling optimized argument generation has been introduced by the research community to date, in Table 1.1 we only give an overview of released datasets used in argumentation assessment grouped by their application domain. Here, most datasets consider annotations of single quality dimensions for a narrow selection of topics and domains, which limits their practical applicability. Though annotations provided in (Zhang et al., 2017; Kashefi et al., 2022) cover various quality dimensions and aspects, the limited amount of data provided (less than 90 essays) makes it unsuitable to use for computationally modeling generalizable argument quality features. To enable modeling argument quality more effectively, robustly and enable generalization across various domains of argumentative text, larger and more diverse datasets covering argument editing behaviors are required. Thus, we formulate the third research gap as follows:

RESEARCH GAP 3: Though there are various datasets covering different topics and domains, most of them are limited in size and diversity making them unsuitable for modeling generalizable aspects of argument quality, and none allow for modeling optimized argument generation.

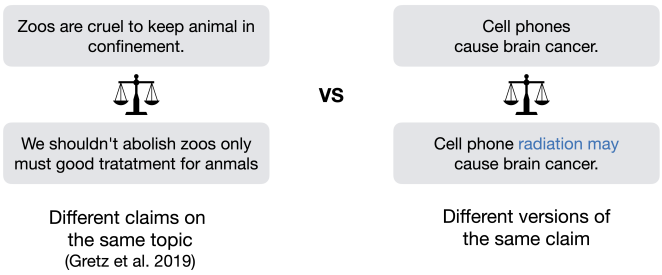


Figure 1.2: Example of an argument pair used to learn relative argument quality characteristics in Gretz et al. (2020) in contrast to the argument pair extracted from argument revision histories.

1.3 Revision-based Claim Quality Assessment and Optimization

As mentioned previously in Section 1.1, the composition and delivery of arguments are among the factors that determine their effectiveness. Individuals need to select, arrange and present their ideas taking into account the social and cultural context of the audience (Wachsmuth et al., 2017b). For written texts, such optimal phrasing is often found through cycles of revisions and text editing (Fitzgerald, 1987; Freeley and Steinberg, 2013). The dissertation at hand is largely based on this observation, specifically, we suggest using *revisions of argumentative texts* as a basis to understand and model *general* quality characteristics of arguments. Here, when we speak of revisions, we refer to any change that occurs during the writing process, including error corrections, rephrasing, and removing or replacing content while preserving the original meaning (Fitzgerald, 1987).

Specifically, to address Research Gaps 1 and 2, we suggest to consider argument quality from a novel perspective: instead of comparing *arguments with different content and meaning* as done in prior work, we suggest to compare *different versions of the same argument* (illustrated in Figure 1.2). By comparing the quality of different versions of the *same* argumentative text, we argue that we can learn *general* quality characteristics of such texts (Skitalinskaya et al., 2021) and, to a wide extent, abstract from prior perceptions and weightings (addresses Research Gap 2). Moreover, by taking into account not just the quality differences between revisions but also the reasoning and purpose of the performed edits, we can learn not only to identify existing flaws in texts but also to suggest types of changes required to tackle said flaws and even automatically generate optimized alternatives for the arguments (addresses Research Gap 1).

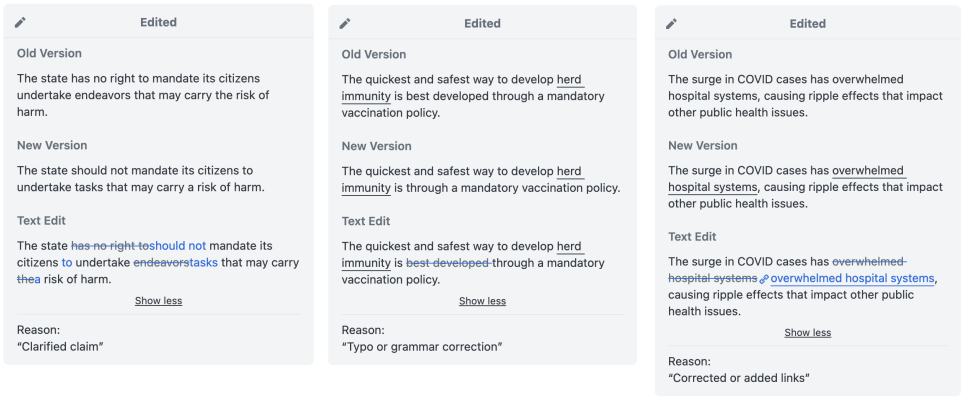


Figure 1.3: Examples of argument revisions found on Kialo along with the reasons provided by the users performing the revision.

To enable such assessments, we compile a new corpus of argumentative claims from an online debate platform, www.kialo.com. Kialo is a typical example of an online debate portal that encourages collaborative argumentative discussions, where participants jointly develop nuanced pro/con debates on a wide rage of topics. But what differentiates Kialo from other portals is that it supports *editing claims* and tracking the history of changes made in a discussion. This allows all users to collaborate and improve existing claims by suggesting edits, which are subject to review and approval by the debate's moderation team. Examples of such suggestions are presented in Figure 1.3 showcasing three distinct reasons for revision provided by the users suggesting it, such as *claim clarification*, *typo or grammar correction*, and *corrected or added links*. Through organized community discussions of each suggested change, this collaborative process should continually improve the quality of claims and lead to a topically diverse collection of claims. Such process of collaborative editing creates a high quality, diverse set of supporting and attacking claims for each controversial debate topic and helps gain deeper insights on various argument characteristics, such as impactfulness, persuasiveness, clarity, etc. Due to the popularity of the platform and the abundance of content, we were able to collect over 508K arguments, which is an order of magnitude larger than any existing corpus for studying argument quality in the field (addresses Research Gap 3).

1.3.1 Research Questions

By capturing implicit revision patterns, we aim to model argument quality irrespective of the discussed topics, aspects, and stances to enable the development of writing support systems where users are provided with access to automatically generated, structured feedback information, guiding them through the process of improving their arguments.

Thus, our first research question is dedicated to understanding the benefits and disadvantages of working with the compiled revision-based corpora, which we accompany with two additional sub-questions:

RQ1 What *quality-related phenomena* are typical of argument revisions on online debate platforms?

- **RQ1.1** What *quality flaws* and *revision types* are typical in online debates?
- **RQ1.2** What challenges does the *revision-based nature* of the corpora pose for computationally modeling argument quality?

As we are concerned with understanding the factors of argument quality and modeling them computationally, we formulate the next research question and relevant sub-questions as follows:

RQ2 How to approach the modeling of argument quality computationally to enable the *analysis of arguments in need of improvement*?

- **RQ2.1** How to model argument quality to enable *identification of low quality argumentative texts*?
- **RQ2.2** How to model argument quality to enable *identification of specific flaws* within an argumentative text?
- **RQ2.3** How to model argument quality to enable *comparison* of several versions of the same argumentative text?

Finally, we are also interested in generating optimized versions of argumentative texts automatically. In order to understand the benefits of using contextual information, such as the main thesis of the debate or related ar-

guments, and explore the potential of using the obtained generative models even outside the original domain of online debates, such as, for example news, encyclopedia entries, scientific papers, and instructional texts, we investigate the following research question and relevant sub-questions:

RQ3 How to approach the *generation of improved argumentative texts* using computational methods?

- **RQ3.1** How to model *argument quality* and specific *quality flaws* to enable automated generation of improved argumentative texts?
- **RQ3.2** How can including *contextual information* in the modeling process assist in further improvements of the generated argumentative texts?
- **RQ3.3** How can the obtained insights and computational models assist in automatically improving the argument quality of texts from *other domains and sources*?

1.3.2 Suggested Tasks

To answer the outlined research questions, we introduce and systematically address a set of four argument assessment and generation tasks that are specifically designed for working with revision-based data. In particular, we consider the following tasks:

- **T1 Suboptimal-Claim Detection** Given an argumentative claim, decide whether it is in need of further revision or can be considered to be phrased more or less optimally in terms of overall argument quality (Skitalinskaya and Wachsmuth, 2023).
- **T2 Claim Improvement Suggestion** Given an argumentative claim, select all types of quality issues from a defined set that should be improved when revising the claim (Skitalinskaya and Wachsmuth, 2023).
- **T3 Argumentative Claim Ranking** Given two or more versions of the same argumentative claim, determine which one is of higher argument quality (Skitalinskaya et al., 2021).
- **T4 Argument Quality Optimization** Given as input an argumentative claim, potentially along with context information on the debate, generate a revised version of the claim that (a) improves upon the original claim

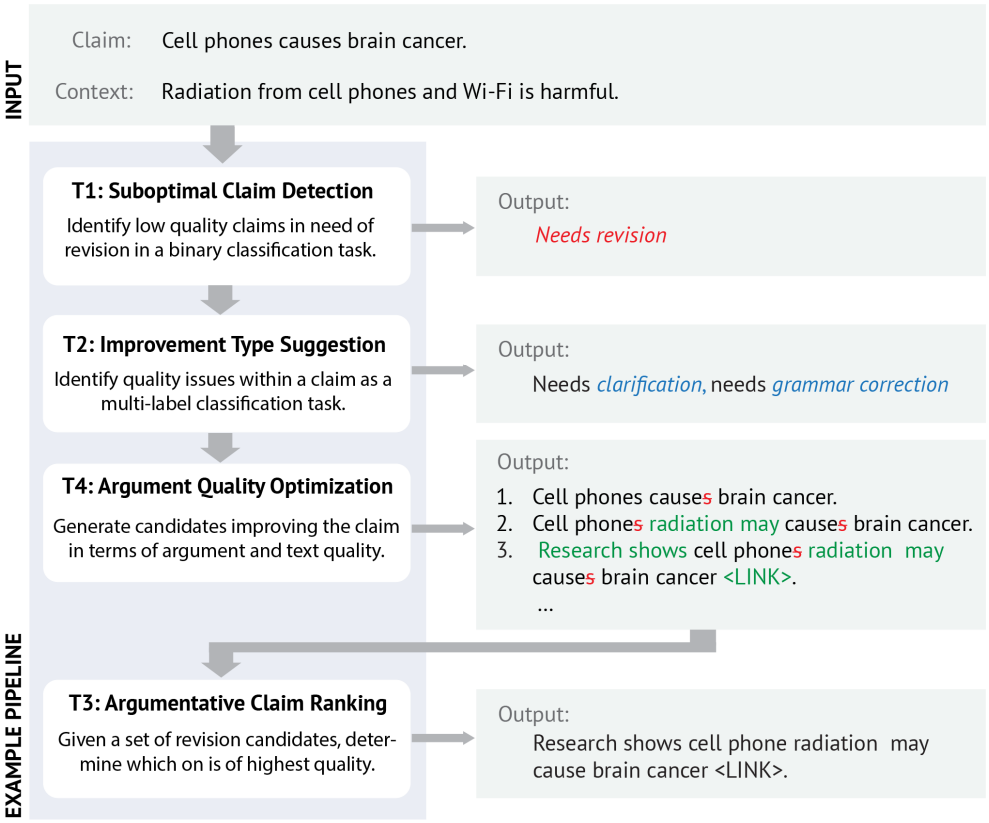


Figure 1.4: Example pipeline combining all four argument assessment and generation tasks explored in this thesis along with example outputs of user feedback based on the outcomes of the text classification and generation approaches.

in terms of text quality and/or argument quality, and (b) preserves the meaning of original claim as far as possible (Skitalinskaya et al., 2023).

Within the given context of tasks, we explore quantitative approaches, such as classical statistical machine-learning algorithms and deep-learning solutions adopted to facilitate automated language processing, primarily focusing on classification methods(SVM (Cortes and Vapnik, 1995), Glove (Pennington et al., 2014), FLAIR (Akbik et al., 2018), BERT (Devlin et al., 2019), SBERT (Reimers and Gurevych, 2019), DeBERTa (He et al., 2021), among others) and text generation models (e.g., Transformer (Vaswani et al., 2017), BART (Lewis et al., 2020)).

Figure 1.4 illustrates an example pipeline, which combines all four argument assessment and generation tasks explored in this thesis. With this figure, we aim to show how solutions to the considered tasks can be used to sup-

port individuals in assessing and writing argumentative text. Here, the input refers to an argumentative claim proposed by an individual, accompanied by contextual information. While providing such contextual information is not mandatory, it can include details such as the main thesis or topic of the debate, related claims, and other pertinent information. In the example illustrated in the figure, the user suggests an argumentative claim (*"Cell phones causes brain cancer."*) and specifies its parent claim that the argument is intended to support (*"Radiation from cell phones and Wi-Fi is harmful."*).

The provided input is then passed to the pipeline component, where solutions to the suggested text classification and generation tasks are used to identify various patterns and characteristics of argument quality and suggest improvements. Based on the outcomes of each task, an output with feedback is generated. The feedback message highlights any concerns found in the argumentative claims and offers suggestions on how to modify them or provides automatically generated alternative candidates. Specifically, the outcome of Task 1 indicates that the claim still requires further revision. To understand the issues the argumentative text is suffering from, we can further apply solutions to Task 2. As shown in the output to the task in Figure 1.4, the claim requires further clarification and contains a grammatical error.

Once the quality issues have been identified, solutions to Task 4 can automatically generate alternative versions of the same claim improving upon the original version in terms of text and argument quality. For example, in Figure 1.4 as feedback, the user is presented with several alternative suggestions on how to further improve the argumentative text, such as:

1. *"Cell phones cause~~s~~ brain cancer."*
2. *"Cell phone~~s~~ radiation may cause~~s~~ brain cancer."*
3. *"Research shows cell phone~~s~~ radiation may cause~~s~~ brain cancer <LINK>."*

Although all the suggested revisions can be considered plausible improvements, they differ in terms of the nature and extent of improvement. For instance, the first suggestion primarily focuses on fixing grammatical errors in the claim, while the second revision aims to clarify that it is the *radiation* emitted by cell phones that causes brain cancer. On the other hand, the third suggestion goes beyond grammar and clarification and provides relevant research evidence to support the claim. While individuals can manually choose the optimal suggestion from such a set of plausible revisions, an automatic approach can also be employed to evaluate and rank generated suggestions (Task 3). This approach can be applied not only to computer-generated argu-

ments but also to human-generated ones. In our example in Figure 1.4, the output of Task 3 displays the optimal suggestion (Candidate 3) among the set of automatically generated texts obtained from the outcome of Task 4.

Though the provided example showcases a scenario oriented at supporting individuals, the same solutions used in the outlined tasks can be used to support moderators and teachers, enabling them to efficiently assess and moderate large amounts of content while providing them with the feedback necessary to also guide other individuals in improving their arguments.

1.4 Contributions

In the following, we present contributions to the area of argument assessment and generation that can be attributed to the field of computational argumentation. Below, we give a compact overview of the most important ones grouped by the type of contribution.

Corpora. One of the central outcomes of this thesis is a collection of large-scale corpora consisting of argumentative claim revisions to analyze online editing behaviors and argument quality in online debate communities.

- **ClaimRev.** We create one of the largest corpora covering over 124,000 debate topics for studying argument quality within the computational argumentation field to date. A novel aspect of the corpus is its coverage of different versions of the same claim, with annotations regarding the claim quality, reasons, and types of revisions collected from an online debating website [4]. This is the first dataset to target the quality assessment and revision processes on a claim level.

- **Extension** As the original corpus only consists of revision histories of claims that have undergone a revision, we extend the original ClaimRev corpus with a collection of claims that remained unchanged throughout time. Adding such claims aims to fill the gap observed in the original corpus, where claims of good quality are only represented by final versions of revised arguments, enabling to distinguish between claims that contain flaws and get revised from ones that are already of high quality.

Models and Empirical Insights. Leveraging the resources introduced above, we conduct a series of analyses aimed at gaining a better understanding of the challenges of modeling argument quality based on human revision histories.

[4] www.kialo.com

- **Intention Taxonomy.** To understand the actions and intentions behind revisions in online debate platforms, we introduce a taxonomy for classifying argumentative claim edits into eight distinctive categories, capturing common revision intentions such as elaboration and disambiguation. This taxonomy allows a fine-grained analysis of revisions and serves as a basis for a systematic comparison of revision candidates generated both by humans in the original corpus and the suggested approaches.

- **Revision Types and Argument Quality.** To explore the relationship between the revision types covered in the collected corpus and argument quality in general, we assess whether each revision improved any of the 15 dimensions of argument quality defined by Wachsmuth et al. (2017b). We provide a detailed analysis of the correlations between the revision types and quality dimensions and summarize our findings along the logical, rhetorical, and dialectical quality aspects.

- **Challenges in Modeling Argument Quality** Finally, we delineate the main challenges of revision-based data, covering issues related to the representativeness and reliability of data, topical bias in revision behaviors, appropriate model complexities and architectures, and the need for context when judging claims. In a detailed analysis, we outline the strengths and weaknesses of strategies exploiting different types of knowledge specific to text and argument revisions to tackle said challenges.

Methods. Based on the obtained insights, we propose several new methods addressing the identified challenges using state-of-the-art language modeling techniques.

- **Suboptimal Claim Detection and Claim Improvement Suggestion.** We present a systematic comparison of approaches based on different contextualized representations and analyze their impact on various types of writing issues for the tasks of suboptimal claim detection and claim improvement suggestion. To tackle the noisy nature of revision-based data, we propose a new sampling strategy based on revision distance, which makes it generalizable to other domains without additional annotations and judgments.

- **Classification and Ranking of Argumentative Claims.** To explore the capabilities of existing approaches to capture the relative quality difference between various versions of the same argumentative text, we compare

traditional and transformer-based models in the tasks of claim quality classification and ranking. We demonstrate consistent performance even in cross-category scenarios, suggesting that learned features generalize well across debate topics.

- **Argument Quality Optimization.** We are the first to present an approach to argument quality optimization that first generates multiple candidate optimizations of an argumentative claim and then identifies the best one using quality based reranking. We demonstrate the generalization capabilities of the approach on out-of-domain datasets, showcase the benefits of using contextual information, and summarize the challenges concerning the automation of certain optimization types.

1.5 Publication record

Several parts of this thesis have been previously published in international peer-reviewed conference proceedings from major events in natural language processing, e.g., ACL, EACL, INLG. We list these publications below, along with author contribution statements, and indicate the chapters of this thesis which build upon them:

- Chapter 3: **Skitalinskaya G.**, Wachsmuth H. (2023) To Revise or Not To Revise: Learning to Detect Improvable Claims for Argumentative Writing Support, In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Main Volume*, pp. 15799–15816, Toronto, Canada

  Contributions: S.G. collected the data, designed and carried out the experiments. All authors analyzed and discussed the findings and contributed to writing the final manuscript. H.W. supervised the work.

- Chapter 4: **Skitalinskaya G.**, Klaff J., Wachsmut H. (2021) Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale, In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1718–1729, Kiev, Ukraine

  Contributions: S.G. collected the data and devised the main conceptual ideas. S.G. and K.J. annotated a subset of samples for data quality evaluations, and designed the experiments. S.G. carried out the experiments and analyzed the findings. All authors contributed to writing the final manuscript.

- Chapter 5: **Skitalinskaya G.**, Spliethoever M., Wachsmuth H. (2023) Generation of Optimized Claims in Computational Argumentation, In *Proceedings of the 16th International Conference on Natural Language Generation*, pp. 134–152, Prague, Czechia (Honorable Mention for Best Paper Award)

  Contributions: S.G. collected the data and devised the main conceptual ideas. S.G. designed and carried out the experiments. S.G. and S.M. annotated a subset of samples and established a taxonomy of revision intentions to enable deeper analysis of typical reviewing behaviors and model performance. All authors discussed the findings and contributed to writing the final manuscript. H.W. supervised the work.

1.6 Outline

This dissertation is structured in 6 chapters. We provide a brief overview of the organization and the content of each chapter:

Chapter 2: *Background*
In Chapter 2, we first introduce fundamental concepts pertaining to computational argumentation and writing research. Then we give an overview of recent developments, in particular, related to argument quality assessment, argument generation, and the process of text revision.

Chapter 3: *Suboptimal-Claim Detection and Claim Improvement Suggestion*
This chapter addresses Research questions 1 and 2 and is based on Skitalinskaya and Wachsmuth (2023). Specifically, we target Sub-questions 1.2, 2.1, and 2.2. As part of the paper, we summarize the main challenges of performing argument quality assessments covering issues related to the representativeness and reliability of data, topical bias in revision behaviors, appropriate model complexities and architectures, and the need for context when judging claims. In a series of experiments, we provide various solutions to the aforementioned challenges in relationship to each of the tasks separately.

Chapter 4: *Quality-based Ranking of Argumentative Text Revisions*
This chapter addresses our second Research question and is based on Skitalinskaya et al. (2021). Specifically, we target Sub-question 2.3. As part of the paper, we describe how revision histories of argumentative claims can be used to analyze and compare the quality of argumentative claims. Here, we introduce the details and statistics of the ClaimRev dataset. Applications of prominent approaches, detailed evaluations, and open challenges are presented as well.

Chapter 5: *Generation of Optimized Argumentative Texts*
This chapter addresses Research questions 1 and 3 and is based on Skitalinskaya et al. (2023). Here, we tackle Sub-questions 1.1, 3.1, 3.2 and 3.3. As part of the paper, we work towards not only being able to automatically *assess* but also to *optimize* argumentative text. We present an approach that generates multiple candidate optimizations of a claim and then identifies the best one using quality-based reranking. The quality-based assessments combine both text quality metrics and argument quality measurements, which are compared against human judgments. Furthermore, various conditioning techniques are adapted to the task, and their performances are compared on relevant revision-based corpora, covering not only argumentative claims but also formal and instructional texts.

It should be noted, that Chapters 3, 4 and 5 present the original content of published papers within this cumulative thesis as well as an additional section, titled "*Implications for the Thesis*". This section relates the findings presented in the papers to the research questions of this thesis.

Chapter 6: *Discussion and Conclusion*
Finally, this chapter explains the main conclusions of the papers published within this cumulative dissertation as well as their interrelationships. Here, we also discuss the limitations of our research, summarize our contributions, and provide directions for future work. Potential future research directions are proposed from both empirical and theoretical perspectives.

2. Background and Related Work

In this chapter, we introduce fundamental concepts relevant to the main areas of focus in this thesis, namely, computational argumentation, text revision, and natural language processing. First, we provide an overview of relevant research on argument quality and text revision from a non-computational perspective (Section 2.1 and Section 2.2, respectively). The aim is to examine the various approaches, theories, and methodologies that have been developed to understand and improve the process of argumentation. We then give a brief introduction to the machine learning methods for representing and modeling argumentative texts considered in this thesis (Section 2.3). Finally, in Section 2.4, we give an overview of the recent developments from a computational perspective within the two relevant sub-areas of computational argumentation: argumentation assessment and argumentation generation.

2.1 Argumentation

For millennia, the study of argumentation has been an important part of intellectual and philosophical inquiry. Different definitions of *argumentation* have been proposed in literature, often focusing on some particular aspects or perspectives of the phenomena. In this work we adopt the definition of Johnson (2012), which describes argumentation as follows:

Definition 1. (Argumentation). *The sociocultural activity of constructing, presenting, interpreting, criticizing, and revising arguments.*

Early theories of argument have been formed by cultures all around the world and can be traced to, for example, ancient Indian, Chinese, Persian,

Middle Eastern, and Greek traditions. In India, debating practices emerged as a means to address epistemological and religious questions, such as the meaning of life, the existence of the after-life, and the explication of the soul (Solomon, 1976). The concept of dharma, or right conduct, was a central theme in Indian philosophy, and debates often revolved around the proper way to live a moral and virtuous life (Matilal et al., 1999).

In ancient Chinese culture, individuals were seen as integral components of a closely-knit community, be it a family or a village. As such, the conduct of an individual was expected to be influenced by the expectations of the group. This is reflected in the philosophical ideologies and debating practices of the most prominent schools of thought at the time, Confucianism and Taoism. In their teachings, they emphasized the importance of respect, ethical behavior, and the cultivation of personal virtues (Munro, 1985).

However, it is the ancient Greek and Roman study of argument that has laid so many of the foundations for the way we think about argumentation today. In contrast to Chinese traditions and their value of reciprocal social obligation, the Greeks valued personal agency and freedom in their tradition of debate. As the concept of democracy began to emerge in Athens (fifth century B.C.E.), even ordinary citizens were given the opportunity to participate in the political process and engage in discussions regarding civic matters (Hansen, 1999). Public speaking and debate became highly valued skills, and the study of rhetoric emerged as a discipline of its own, with philosophers and educators developing theories and techniques for effective communication and persuasion, many of which continue to influence contemporary discourse (Hansen, 2005).

Aristotle Aristotle's On Rhetoric (Aristotle, 2007) is considered one of the most influential works on the subject. According to Aristotle, effective argumentation requires not only logical reasoning but also the ability to establish credibility and appeal to the emotions of the audience.

Plato Aristotle's teacher Plato, on the other hand, believed that argumentation should be used primarily for seeking truth and understanding rather than for persuasion. In his dialogues, such as the Phaedrus (Plato, 1961) and the Gorgias (Lamb, 1925), Plato presents a critical view of rhetoric, arguing that it can be used to deceive and manipulate people. He advocates for a form of dialectical inquiry, in which individuals engage in critical questioning and constructive debate to arrive at a shared understanding of truth.

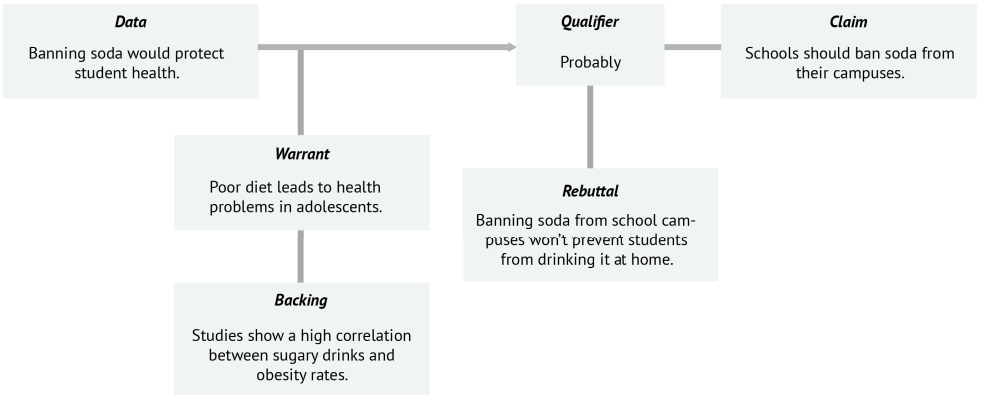


Figure 2.1: The Toulmin model of argumentation, illustrated with an argument example concerning the ban of soda from school campuses.

Cicero Cicero, a Roman statesman and orator, drew on the works of Aristotle and Plato to develop his own theory of argumentation. In his De Oratore (Cicero, 1903), Cicero emphasizes the importance of style, delivery, and adaptability in persuasive communication and argues that effective speakers must have a deep understanding of human nature and the ability to appeal to the values and interests of their audience.

Leveraging the historical foundations of persuasive discourse, logical analysis, rhetorical strategies, and the models of argument built upon them, this dissertation explores how these principles can be computationally modeled and used to help people improve their argumentation skills by revising and refining their arguments.

2.1.1 Models of Argument

There has been considerable effort in defining and conceptualizing arguments, as well as developing models of argumentation reflecting their internal (micro) or external (macro) structure. On the one hand, such models enable descriptive and explanatory analysis of how people typically engage in argumentation by breaking down complex arguments into smaller components, which can be assessed for their quality and inter-relations. On the other hand, such theories can also be used to guide and regulate the course of an argument helping distinguish valid reasoning from fallacious argumentation. One of the most notable models is Toulmin's Model (Toulmin, 1958), which influenced many works focusing on computationally modeling argumentation quality.

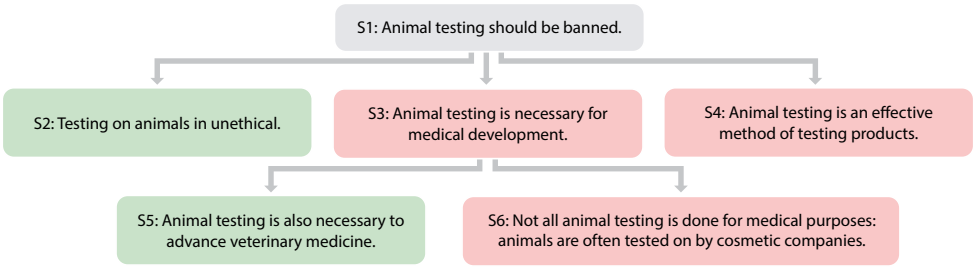


Figure 2.2: Sample of a debate found on Kialo on the topic of whether animal testing should be banned presented as a hierarchical tree structure. The topic of debate represents the root node (colored gray), whereas statements supporting (colored green) or challenging (colored red) other statements are connected with directed edges.

Toulmin's model A nuanced model of argumentation describing the different parts necessary in a well-formed argument and their interrelationships (Toulmin, 1958). Specifically, Toulmin introduces six key components (see Figure 2.1):

1. *Claim.* An argumentative statement whose merit the audience seeks to establish. This assertion reveals the individual's opinion on the matter.
2. *Data.* Evidence or facts which support the claim. Alternatively referred to as premise or ground.
3. *Warrant.* An assumption or reasoning of how the evidence logically and justifiably supports the claim.
4. *Backing.* Additional evidence or reasoning which provides support for the warrant.
5. *Rebuttal.* Conditions of exception in which the argumentative statement does not hold.
6. *Qualifier.* The degree of certainty and the strength of the justification, typically encoded in terms such as certainly, presumably, or probably.

However, real-world arguments do not usually appear in this format and typically lack some of these elements. Moreover, the component assignment, as defined by Toulmin, depends on the specific context and how the argumentative text is used within an argument. Consider the following example of a debate from Kialo illustrated in Figure 2.2 presented as a hierarchical tree structure. Each debate is a collection of statements expressing viewpoints or arguments related to the topic of debate (root node). The directed edges and block colors containing each statement reflect their interrelationships:

red nodes challenge parent statements, while green nodes support them. In this example, Statement 3 (S3 in the figure) can be seen as a *rebuttal* when considered together with its parent statement (S1). However, if its children statements (S5 and S6) are considered, it can be seen as a *claim*, according to Toulmin's model.

In this thesis, we do not differentiate between such fine-grained components as defined by Toulmin and adopt a more loose definition of a claim that unites these components under one concept:

Definition 2. (Argumentative claim). *An argumentative statement that asserts a particular viewpoint, facts, or reasoning and seeks to persuade others to accept or believe it.*

2.1.2 Argument Quality

Assessing the quality of arguments is a central theoretical and practical concern, yet, it has proven challenging to establish clear criteria for what constitutes a "*good*" argument.

In classical rhetoric, the quality of argumentation is often assessed according to three key elements: logos, ethos, and pathos (Aristotle, 2007). These elements are known as the *modes of persuasion*, and they refer to different ways of appealing to an audience in order to persuade them to accept an argument. While *logos* seeks to persuade the audience through sound reasoning, scientific evidence, and logical analysis, *pathos* focuses on persuasion by arousing certain emotions, which are beneficial for influencing their decision-making in a desired way. On the other hand, *ethos*, refers to the credibility and trustworthiness of the person making the argument, and includes the use of authority, expertise, and reputation of the speaker or writer.

In Wachsmuth et al. (2017a,b), the authors present a comprehensive overview of existing classical and modern theories to assessing argument quality, combining both theoretical and practical viewpoints. The authors suggest a unified framework (Figure 2.3) that defines overall argument quality as being composed of three main categories: Cogency, Effectiveness, and Reasonableness.

- **Cogency** covers the logical aspects of argument quality. According to Govier (2013), a cogent argument satisfies the following conditions: it has premises that are rationally acceptable and support the conclusion

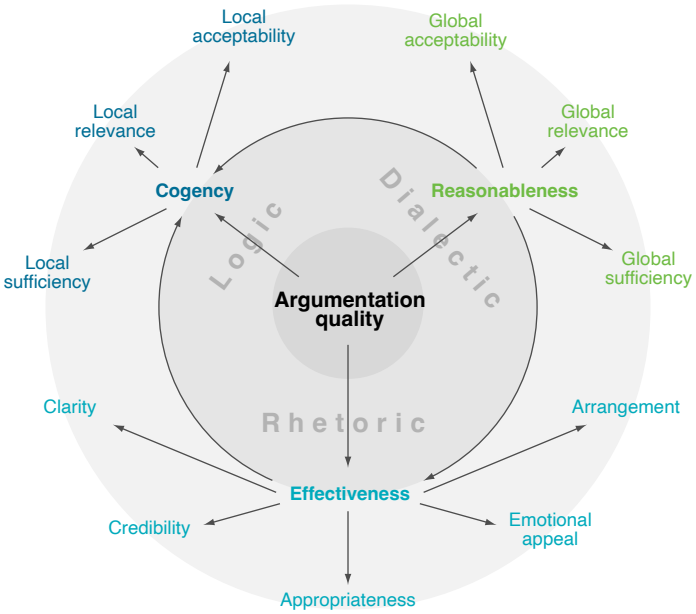


Figure 2.3: Taxonomy of argument quality proposed by Wachsmuth et al. (2017a).

in a way that is relevant and provides good grounds. Thus, as reflected in the taxonomy in Figure 2.3, within Cogency, one can assess: whether the premises are believable (*Local Acceptability*), the relevance of the premises to the conclusion (*Local Relevance*), and whether the premises give enough support to rationally draw the conclusion (*Local Sufficiency*).

- **Effectiveness** covers the rhetorical aspects of argument quality and reflects the persuasive power of how an argument is delivered. Here, Wachsmuth et al. (2017b) outline such quality dimensions as *Arrangement*, which assesses the structure of the argument, *Clarity*, which reflects how unambiguous and understandable the language used is, *Appropriateness* in a given context, the usage of *Emotional Appeal* in persuasion, and *Credibility*, which takes into account the rhetor's worthiness to be believed in.

- **Reasonableness** covers the dialectical aspects of argumentation and reflects the quality of an argument in the context of a debate. Here, the dimensions target quality aspects that ensure that the argument contributes to the issue's resolution (*Global Relevance*) in a sufficient way (*Global Sufficiency*) that is acceptable to the target audience (*Global Acceptability*). While the outlined quality dimensions resemble those

of Cogency, within this category, the arguments are judged specifically by their reasonableness for achieving agreement (Van Eemeren and Grootendorst, 2004).

Each of these categories is represented by several relevant quality dimensions. Although each dimension captures distinct aspects of a text's argumentative quality, they are interconnected and collectively contribute to the overall quality of the argument.

In this work, we use the taxonomy of Wachsmuth et al. (2017a) to disentangle which argument quality dimensions are typically addressed when revising arguments and how they relate to certain types of revision performed.

2.2 Text Revision

Revision is widely recognized as a crucial aspect of writing, contributing to the quality of the final work and enhancing writers' knowledge (Murray, 1978; Bamberg, 1978; Lowenthal, 1980; Sommers, 1980; Beck et al., 1991; Myhill and Jones, 2007; Hyland and Hyland, 2019). In the early 1980s, there was considerable debate among scholars regarding the exact definition of revision (Fitzgerald, 1987). It was discussed whether the term should solely refer to the product (changes made to a text) or the mental processes undertaken by authors, or both aspects. As a result, different terms like *reprocessing* (Bereiter and Scardamalia, 1987) and *reviewing* (Hayes and Flower, 1981) were introduced by the community to capture the mental aspects of revision, distinguishing them from the actual changes made to the text. In this thesis, we combine both aspects and not only explore the different textual changes but also try to capture the intentions of the authors that led to such modifications.

2.2.1 Theories of Revision

Different models and taxonomies have been proposed to describe the subprocesses involved in the revision of written text, each offering valuable insights into the process.

Problem-Solving Models Most work views text revision as a *problem-solving* task where writers actively engage in identifying and resolving issues to enhance the quality and effectiveness of their written work. For example, Flower and Hayes (1981), Beach (1984) and Scardamalia (1983) model revision as a goal-directed process, where, first, the writers infer their intentions or goals, then define the challenges in achieving those intentions, and, finally, make changes to the text that align with their intended meanings. In (Beach and

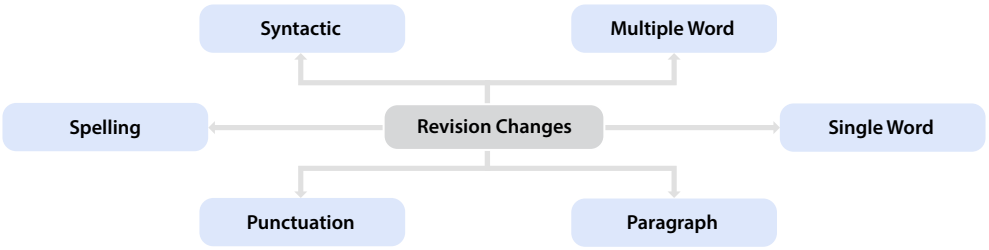


Figure 2.4: Taxonomy of revision types proposed by Stallard Jr (1972).

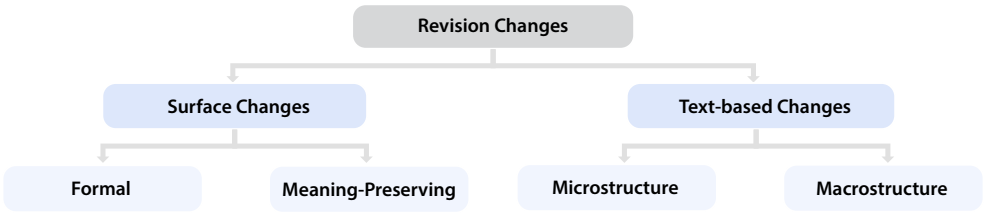


Figure 2.5: Taxonomy of revision types proposed by Faigley and Witte (1981).

Eaton, 1984), the authors found that college students who can clearly define their intentions are more adept at identifying problems and implementing appropriate revisions to enhance their drafts.

Revision Category Taxonomies To systematically analyze and understand the changes made during the revision process, several coding schemes and taxonomies were developed to categorize revisions in written texts. Early schemes, such as Stallard's study (Stallard Jr, 1972) introduced basic classifications that lacked theoretical grounding and contained non-mutually exclusive categories, such as spelling, syntax, punctuation, as well as multiple-word, paragraph, and single-word changes (see Figure 2.4). The study also did not distinguish between surface changes that do not affect the meaning of the text from meaning-altering changes.

Later taxonomies addressed some of these limitations, for example, Sommers (1980) introduced the distinction between revision operations and linguistic levels, as well as mutually exclusive revision categories. Building on discourse analysis research, Faigley and Witte (1981) introduced a taxonomy, which accounted for revisions related to the semantic structure of the text, considering surface and meaning changes, as well as micro-structure and macro-structure features (Figure 2.5). While Bronner and Monz (2012) suggested to classify revisions into fluency edits, aimed at improving style and readability, and factual edits, which alter the meaning.

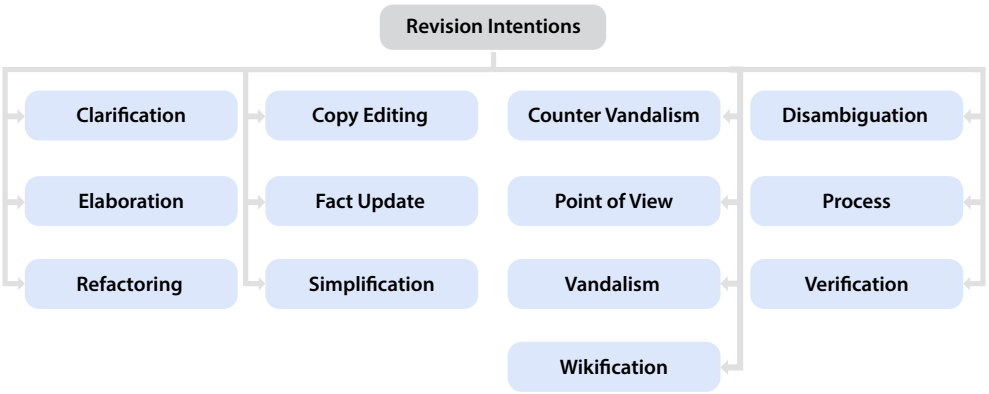


Figure 2.6: Taxonomy of revision intentions proposed by Yang et al. (2017).

As noted in previous studies (Bazerman, 2016), the process of text revision is influenced by the specific domains or contexts in which the writing takes place. Specifically, the text should meet the requirements and expectations of the intended audience or the conventions of a specific genre or discipline. Recognizing this, recent research has looked into developing domain-specific taxonomies for text revision. For instance, Daxenberger and Gurevych (2013) proposed a 21-category taxonomy for Wikipedia editing, which included not only surface edits but also domain-specific policy edits, such as vandalism and reversion of changes. Similarly, Yang et al. (2017) introduced a 13-category taxonomy that focuses on capturing the semantic intentions, providing insights into the reasons behind each revision (see Figure 2.6). While such taxonomies are more fine-grained and capture important parts of the revision process, they do not fully transfer to other domains limiting their practical applicability.

Given our focus on argumentative texts, it is relevant to provide an overview of the revision categories introduced specifically for this domain. While many of the above-mentioned works have focused on studying revisions of argumentative texts, such as student essays (Sommers, 1980; Faigley and Witte, 1981; Daxenberger and Gurevych, 2013), only Zhang and Litman (2015) incorporate argument theory concepts in their categorization. Specifically, they outline two main categories: surface changes and text-based changes (see Figure 2.7). Under the surface changes category, similar to previous work, they include sub-categories such as organization, conventions/grammar/spelling, and word usage/clarity. In the text-based changes category, they incorporate Toulmin's argument structure and outline such sub-categories as claims or ideas, warrant/reasoning/backing, rebuttal or reservation, general content, and evidence. It is worth noting that while the broader categories align with

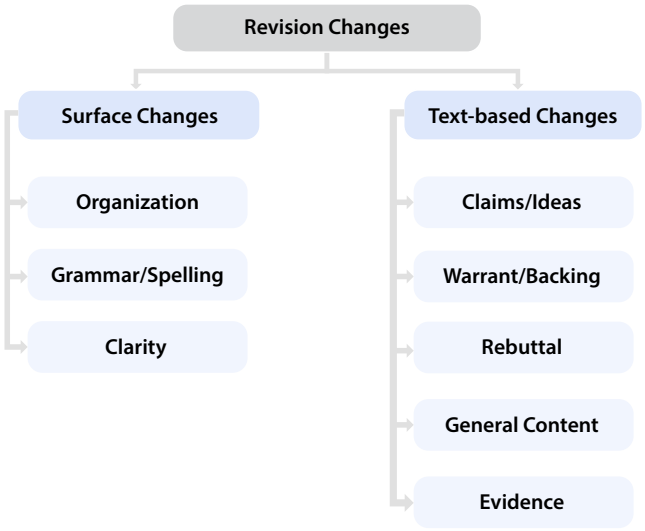


Figure 2.7: Taxonomy of revision types proposed by Zhang and Litman (2015).

the taxonomy of Faigley and Witte (1981), the specific sub-categories differ.

While the considered taxonomies include information regarding the argumentation structure (Zhang and Litman, 2015) and general linguistic and semantic text quality features (Faigley and Witte, 1981), such revision types do not reflect the *actions* and *intentions* of each revision, which is crucial information when providing feedback and guiding users on why and how to improve their texts. Though Yang et al. (2017) touch upon this in their work, their taxonomy is tailored to the needs and revisions of Wikipedia articles, making it hard to employ to other domains of text. This motivates us to further explore the relationships between revision actions and intentions and develop a new taxonomy tailored to the domain of argumentative texts building on the outlined work (see Section 5.7 for more details).

2.3 Natural Language Processing

As part of our work, we make use of several important concepts related to natural language processing. Below we summarize the basic concepts relevant to our work.

2.3.1 Types of Learning

Machine Learning Opposed to classical imperative programming, where each step of solving the task should be explicitly coded, in machine learning, the goal is to algorithmically derive a set of rules from data that captures mean-

ingful relationships and patterns (Bishop and Nasrabadi, 2006). Such learning is facilitated by computational algorithms that are designed to process and analyze input data, capture patterns of interest, and produce computational models to make predictions or decisions based on the learned knowledge. In (Mitchell, 1997), the author emphasizes the iterative nature of machine learning and further specifies that the performance of the computational model should improve with the number of examples it has seen in order to be called *learning*.

Literature typically distinguishes three types of learning computational models (Bishop and Nasrabadi, 2006; Goodfellow et al., 2016):

- **Supervised Learning** In a supervised learning setting, the aim is to extract a set of rules from *labeled* examples, where each example consists of input data paired with corresponding desired target values (Bishop and Nasrabadi, 2006). What distinguishes supervised learning from other learning settings is that the exact mapping between every input and target value is known.

- **Unsupervised Learning** In contrast, in an unsupervised learning setting, the goal is to identify hidden patterns or structures from *unlabeled* data, i.e., without prior knowledge of the correct target value (Goodfellow et al., 2016).

- **Semi-supervised Learning** Semi-supervised learning combines the benefits of both approaches mentioned above for scenarios when only a limited amount of labeled examples is available along with an abundant number of unlabeled samples (Russell, 2010). The idea is to leverage the unlabeled data to enhance the labeled samples by augmenting it using various techniques, such as, for example, self-training (Lee, 2013) and co-training (Blum and Mitchell, 1998). In this work, we consider a technique called *distant supervision* (Mintz et al., 2009), where labels are induced by leveraging existing knowledge bases, databases, or heuristics, resulting in an automated yet potentially noisy labeling process.

For more details on each type of learning, refer to the following main textbooks on pattern recognition, artificial intelligence, and deep learning (Bishop and Nasrabadi, 2006; Russell, 2010; Goodfellow et al., 2016).

The dataset we work with in this thesis consists of revision histories of argumentative claims, where each history defines a chain $(v_1, ..., v_m)$. Here, each

claim v_t is a revised version of the previous claim, v_{t-1} with $1 < t \leq m$. The only annotations available to us are the intention labels for consecutive claim version pairs (v_{t-1}, v_t) denoting the purpose of revision, such as clarification. Fortunately, we can use distant supervision methods to automatically infer certain labels. For instance, we can derive relative argument quality labels between different versions of the same claim (v_j, v_k) by assuming that each revision improves the quality of the claim. Thus, if $k > j$ (v_j appears before v_k in the revision history) the assigned label would be 1, denoting that the quality of v_k is greater than v_j or 0 if otherwise. Deriving such distantly supervised labels enables us to explore the applicability of *supervised* approaches in solving complex problems across different text classification and generation tasks in Chapters 3 to 5. While other types of learning could also be considered for these tasks, for the purposes of this thesis, we focused solely on supervised and semi-supervised approaches.

2.3.2 Text Representation

To be processed by computational models, textual input needs to be transformed into numerical representations in a way that captures the semantic and contextual information inherent in language. Various techniques and methods have been suggested to tackle the problem on the word level, including:

- **Bag-of-Words (BoW)**: BoW transforms text into a numerical representation by counting the occurrences of words without considering their order or context. The approach can be applied to a sentence, document, or even a collection of texts. As a result, the input is transformed into a vector, where each dimension represents a distinct word, and the value assigned to it indicates its frequency in the text.

- **Static Word Embeddings**: Static word embeddings are created by training on extensive datasets utilizing unsupervised learning methods. These methods aim to capture the word semantics and relations by analyzing the co-occurrence patterns present within the training data. As a result, each *word* of the input is transformed into a vector, and this representation does not change depending on the context in which the word appears. Word2Vec(Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) are examples of the most widely recognized word embedding methods.

- **Contextualized Word Embeddings**: Unlike static word embeddings, con-

textualized word embeddings, such as Flair (Akbik et al., 2018), ELMo (Embeddings from Language Models) (Peters et al., 2018), BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), and other transformer-based embeddings, generate word representations that are sensitive to context. These models consider the surrounding words and sentences to create embeddings that capture word meaning variations based on their usage within a specific context.

2.3.3 Models

Computational models in machine learning can be broadly classified into two main categories: *statistical* and *neural* models. In this thesis, we consider both of these categories. In the following, we provide an overview of the models used in our research and elaborate on the methods employed for their training.

Statistical Machine Learning

Linear Regression Linear regression is used to establish a linear relationship between one or more input variables, also known as independent variables, and a continuous output variable, known as the target (Bishop and Nasrabadi, 2006). The goal is to find the best-fitting line that minimizes the difference between the predicted values and the actual values of the target variable. The linear regression model is represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Here, y represents the target variable, $x_1, x_2, \ldots, x_n$ denote the independent variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients that represent the relationship between the variables. The best-fitting coefficients can be estimated, for example, using a technique called ordinary least squares (OLS) (Bishop and Nasrabadi, 2006), which minimizes the sum of the squared differences between the predicted and actual values. Once the coefficients are estimated, the linear regression model can be used to make predictions for new input values.

Support Vector Machines Support Vector Machines (SVMs) are a class of machine learning algorithms that aims to find the best possible decision boundary that separates different classes of data points (Joachims, 2006). For linearly separable data, the decision boundary is represented by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

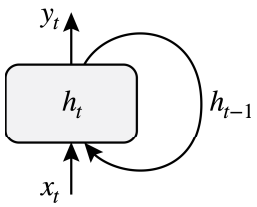


Figure 2.8: The basic structure of a recurrent neural network.

Here, y is the predicted class label, $x_1, x_2, \ldots, x_n$ are the input features, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients to be estimated. SVMs use the kernel trick to handle non-linearly separable data by mapping features into a higher-dimensional space. The optimization objective is to minimize

$$\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{m}\xi_i,$$

Here, $\frac{1}{2}\|\beta\|^2$ represents the regularization term that penalizes the magnitude of the coefficients β, and $C\sum_{i=1}^{m}\xi_i$ is the cost term that sums the slack variables ξ_i with a regularization parameter C. Slack variables are introduced to handle cases where finding a hyperplane that linearly separates the classes is infeasible by allowing some data points to be misclassified. Here, C controls the margin vs. classification error trade-off. SVMs can employ not only linear kernel functions as described above but also polynomial or Gaussian functions to handle complex decision boundaries in high-dimensional spaces (Joachims, 2006).

Neural Machine Learning

Recurrent Neural Networks Recurrent Neural Networks (RNNs) are a class of neural networks designed to process sequential data by maintaining a hidden state that carries information from previous time steps (Rumelhart et al., 1986). The key feature of RNNs is their ability to capture temporal dependencies by recursively applying the same set of weights across multiple time steps. The output of an RNN at each time step can be computed as:

$$h_t = \phi(W_{xh} \cdot x(t) + W_{hh} \cdot h_{t-1} + b_h)$$

Here, ϕ is the activation function, x_t is the input at time step t, h_{t-1} is the hidden state at the previous time step, W_{xh} and W_{hh} are weight matrices, and b_h is the bias vector.

The output of an RNN at each time step can be computed as:

$$y_t = W_o \cdot h_t + b_o(2)$$

Here, W_o and b_o are the weight matrix and bias vector for the output layer, respectively.

By recursively applying these formulas, an RNN can capture and propagate information through time, allowing it to model complex sequential patterns and dependencies. Figure 2.8 illustrates the basic structure of an RNN. Here the recurrent connections are depicted via cyclic edges. However, RNNs suffer from the vanishing or exploding gradient problem, which can make it difficult for them to learn long-term dependencies. This limitation led to the development of more advanced architectures like LSTMs that address these issues.

Long Short-Term Memory Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that addresses the vanishing gradient problem by incorporating memory cells and gating mechanisms. The basic building block of an LSTM is the memory cell, which consists of three main components: the input gate (i), the forget gate (f), and the output gate (o) (Hochreiter and Schmidhuber, 1997). These gates regulate the flow of information into and out of the cell. The input gate determines how much new information should be stored in the memory cell (c), the forget gate decides what information should be discarded, and the output gate controls the amount of information to be outputted.

Formally, an LSTM cell can be described as follows:

$$\begin{aligned}
i(t) &= \sigma(W_{ix}x(t) + U_{ih}h(t-1) + b_i) \\
f(t) &= \sigma(W_{fx}x(t) + U_{fh}h(t-1) + b_f) \\
o(t) &= \sigma(W_{ox}x(t) + U_{oh}h(t-1) + b_o) \\
g(t) &= \tanh(W_{gx}x(t) + U_{gh}h(t-1) + b_g)
\end{aligned}$$

Here, $x(t)$ represents the input at time step t, $h(t-1)$ denotes the hidden state at the previous time step, and σ is the sigmoid function. And updating the cell state is performed as follows:

$$c(t) = f(t) \odot c(t-1) + i(t) \odot g(t)$$

Here, $c(t-1)$ is the previous cell state, and $\odot$ denotes the Hadamard product

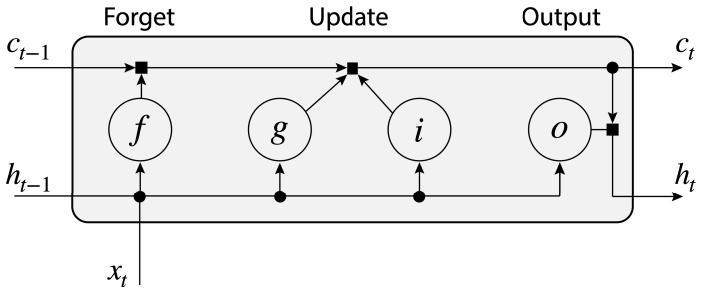


Figure 2.9: The basic structure of Long Short-Term Memory neural network.

(element-wise multiplication).

Finally, the output of the LSTM can be computed as:

$$h(t) = o(t) \odot \tanh(c(t))$$

Figure 2.9 presents a visual representation of the basic LSTM structure, incorporating all the aforementioned components.

Transformers Transformers are a type of multi-layer neural network architecture by Vaswani et al. (2017) that has made significant contributions to natural language processing tasks. Unlike LSTMs, which process sequences sequentially, transformers employ certain mechanisms that allow them to consider all words in the sequence simultaneously. Specifically, transformers utilize self-attention mechanisms, multi-head attention, position-wise feed-forward networks, and residual connections with layer normalization to process input sequences (illustrated in Fig. 2.10). These operations are performed within each transformer layer, enabling the model to capture long-range dependencies. We explain each of the introduced concepts below.

Self-Attention The core component of transformers is the self-attention mechanism, which allows the model to weigh the importance of different words in a sentence when processing the input sequence. The self-attention mechanism calculates attention weights for each word in the input sequence by comparing its relevance to other words. Given an input sequence of length n, the self-attention mechanism computes the attention weights using the following formula:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q represents the query matrix, K represents the key matrix, and V

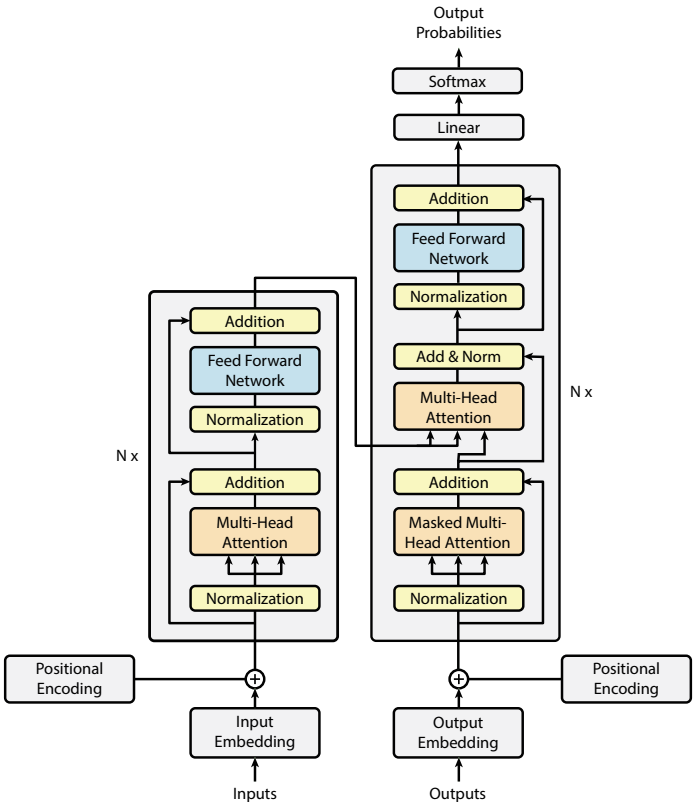


Figure 2.10: The encoder-decoder structure of the Transformer architecture based on Vaswani et al. (2017) with specifications from Xiong et al. (2020).

represents the value matrix. The query matrix corresponds to the representation of the current input element that the attention mechanism is focusing on. The key matrix represents the learned representation of the input sequence, based on which the relevance or similarity between the current input element (the query) and all other elements in the sequence is measured. The similarity scores are then transformed into attention weights using a softmax function. The division by $\sqrt{d_k}$ helps stabilize the gradients during training, and d_k represents the dimensionality of the query and key vectors. Finally, the value matrix represents the learned representation of the input sequence carrying the context information associated with each input element. Combining them with attention scores forms the final output of the attention mechanism.

Multi-head Attention To enable the model to attend to different parts of the input representation simultaneously, multi-head attention is used. In multi-head attention, the input embeddings are linearly projected into multiple subspaces, called *"heads"*. The outputs of these attention heads are concate-

nated and linearly transformed to produce the final output. Mathematically, the multi-head attention is computed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, head_2, ..., head_h)W_O$$

where each $head_i$ is computed as $\text{Attention}(QW_i^Q, KW_i{}^K, VW_i{}^V)$, and W^O is the output weight matrix.

Feedforward Network To further process the attention outputs independently at each position, a feedforward network is applied. This network consists of two linear transformations with a non-linear activation function in between.

The formula for the position-wise feed-forward network can be written as:

$$FFN(x) = max(0, x \cdot W_1 + b_1) \cdot W_2 + b_2$$

Here, x represents the input from the previous attention step, and W_1, b_1, W_2, and b_2 are learned weight matrices and biases for the feed-forward network.

Positional Encodings To provide the model with the necessary information about the relative or absolute positions of tokens, positional encodings are added to the word embeddings. The formula for generating the positional encoding vector for a given position pos and dimension i is as follows:

$$PE(pos, 2i) = sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$
$$PE(pos, 2i+1) = cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

Here, d is the dimensionality of the model. The value 10000 is chosen based on heuristics and can be adjusted according to the length of the input sequence.

Residual Connections and Layer Normalization To facilitate better gradient flow during training, residual connections, and layer normalization are applied after each sub-layer (attention and position-wise feed-forward network) within a transformer layer.

The formula for residual connection and layer normalization is given by:

$$LayerNorm(x + Sublayer(x))$$

Here, x represents the input to the sub-layer, and $Sublayer(x)$ denotes the

output from the sub-layer (either self-attention or position-wise feed-forward network).

By stacking multiple transformer layers, information can propagate through the network, allowing the model to capture complex dependencies within the input sequence.

Architecture Variations The original transformer architecture includes both an encoder and a decoder component, as illustrated in Figure 2.10. The encoder processes the input sequence, capturing its information and context. The decoder generates an output sequence based on the encoded input and incorporates attention mechanisms to attend to both the input and previously generated tokens.

However, variations of transformer models have been developed that feature either an encoder-only or a decoder-only setup, in addition to models that retain both components. While encoder-only models are suitable for tasks such as language understanding, sentiment analysis, or representation learning, models with a decoder component can be used in tasks oriented on sequence generation, such as text summarization or machine translation. Model architectures with both an encoder and decoder component are often referred to as *sequence-to-sequence* models, and can also be implemented using recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks, as done in early works on text generation (Sutskever et al., 2011; Salehinejad et al., 2017; Wu et al., 2016; Pawade et al., 2018).

Training Methods In this thesis (Chapters 3 to 5), we consider the following pretrained language models based on the encoder-only architecture: BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), DeBERTa (He et al., 2021), and the following sequence-to-sequence models: LSTM (Wiseman and Rush, 2016), Transformer (Vaswani et al., 2017), and BART (Lewis et al., 2020). While the models may share certain similarities in their architecture and objectives, there are some differences in their training methods.

- **BERT** is trained using a masked language modeling (MLM) objective. During training, a portion of the input text is randomly masked, and the model is tasked with predicting the masked tokens based on the surrounding context (Devlin et al., 2019).

- **ELECTRA** employs a training approach called *"discriminative masked language modeling"*. Instead of masking tokens and predicting them,

ELECTRA trains a generator model to replace certain tokens with plausible alternatives and then trains a discriminator model to distinguish between the original tokens and the replacements. (Clark et al., 2020)

- **DeBERTa** extends BERT by incorporating disentangled attention and enhanced decoding mechanisms. It introduces two additional training objectives: disentangled attention objective and sequence-level training objective. The disentangled attention objective encourages the model to attend to different aspects of the input independently, improving its interpretability. The sequence-level training objective focuses on predicting the next sentence given the previous one, promoting coherence in text generation (He et al., 2021).

- **BART** follows an encoder-decoder architecture and is trained using a combination of denoising auto-encoding and sequence-to-sequence learning. During training, BART corrupts the input text by randomly masking, shuffling, and deleting words. It then learns to reconstruct the original text from the corrupted version, serving as an auto-encoder (Lewis et al., 2020).

2.4 Related Work

After having introduced the fundamental concepts and theories underlying this thesis, we now give an overview of related work on computational approaches tackling argument quality assessment, argument generation, and processing revisions of argumentative texts.

2.4.1 Argument Quality Assessment

The assessment of argument quality can be approached from various perspectives as discussed in Sections 1.2 and 2.1.2. In the recent years, there has been a sharp increase of research and interest in learning how to make such quality assessments automatically, resulting in two main lines of work. The first is oriented more on the domain of education and mainly focuses on the assessment of *argumentative essays* of students. While the second works with less formal texts extracted from *web-based media*, such as online forums, debate portals, and digital encyclopedias. Assessing argument quality in such different contexts presents unique challenges and considerations, below we provide a comprehensive overview of studies for each line of work categorized according to the dimensions of argument quality, as introduced in Section 2.1.2.

Student Essays When assessing argument quality in student essays, the focus is typically on academic writing and structured argumentation. Evaluators often look for elements such as logical coherence, evidence-based reasoning, the use of counterarguments, and adherence to academic conventions. The evaluation criteria may also include factors like clarity of expression, organization, and adherence to the assigned prompt or guidelines.

Organization One of the first works on argument quality assessment is the work of Persing et al. (2010), who propose to predict the *organization* of student essays. To do so, they apply an SVM with a string kernel to sequences of paragraph and sentence labels, representing the structure of the written text. While paragraph labels denote the general structure of the essay, such as the introduction, conclusion, or rebuttal, the sentence labels provide more fine-grained information, denoting such aspects as the thesis, main idea, suggestion, or elaboration.

Clarity and Strength Similarly, in (Persing and Ng, 2013) and (Persing and Ng, 2015), the authors propose to model the *clarity of a thesis* and the *strength of an argument* by training a linear SVM incorporating both traditional features, such as word tokens, cue words, part of speech tag n-grams, and semantic frames or features reflecting the structure of the text, such as major claims, claims, and premises. Their results show that incorporating argumentative information can further improve the model's ability to predict argumentative quality aspects of the essays.

Sufficiency In another study, (Stab and Gurevych, 2017b) address the task of assessing the *sufficiency* of arguments in essays. They aim to determine whether the premises of an argument provide enough evidence to support its claim. To do so, they explore the effectiveness of feature-rich SVMs as well as convolutional neural networks (Collobert et al., 2011). Their findings indicate that arguments lacking sufficient exhibit display specific lexical indicators that can be reasonably captured with the above-mentioned approaches. Gurcke et al. (2021) reexamined the sufficiency quality criterion through a novel task setup involving the generation of conclusions from premises that may or may not be sufficient. Here, the objective is to assess premise sufficiency by evaluating the generated conclusion's quality. When approaching the task, the authors explore the benefits of assessing sufficiency based on the output of large-scale pre-trained language models. In the experiments, the authors fine-tuned

BART (Lewis et al., 2020) and then learned to distinguish sufficient from insufficient arguments using a modified RoBERTa model (Liu et al., 2019) based on the generated conclusion with the original argument.

Web-Based Media Assessing argument quality in online forums and debate platforms introduces additional challenges due to the interactive and dynamic nature of the discussions. Online platforms often involve multiple participants with diverse backgrounds and motivations, and the arguments can vary widely in terms of style, tone, and level of formality. Evaluating argument quality in this context requires considering factors such as relevance to the topic, the use of logical reasoning, the consideration of alternative viewpoints, and the ability to engage in productive and respectful discourse.

Acceptability Cabrio and Villata (2012) aim to automatically determine the logical relationships between arguments presented in online debates and predict the *acceptability* of arguments from the now defunct Debatepedia.org. They suggest to do so by building upon an argumentation framework proposed by Dung (1995), which models arguments as a graph structure. Specifically, they leverage textual entailment relations to construct the graph and only measure the acceptability of the main argument based on the obtained structure.

On the other hand, Yang et al. (2019a) and Yang et al. (2019b) focused on assessing *local* acceptability in news editorials, i.e., how much the given text is rationally worthy of being believed to be true (Wachsmuth et al., 2017b). In (Yang et al., 2019a) analyzed the reasons to see why given information is accepted or rejected. Their results show show a positive correlation between local acceptability and verifiability, and a negative correlation between local acceptability and disputability. Respectively, the authors of (Yang et al., 2019b) develop an approach aimed at identifying reliable annotations of local acceptability based on the provided reasons for each label.

Effectiveness A lot of attention has been given to understanding the characteristics of *persuasive* text, e.g., what distinguishes persuasive from non-persuasive text. In (Tan et al., 2016), the researchers examined the ChangeMyView community on Reddit to investigate how user interaction dynamics and language features influence persuasion. They found

that both user interaction and linguistic characteristics influence the success in persuasion. On the other hand, Hidey et al. (2017) studied whether specific semantic types or combinations of semantic types exist that are typical exclusively for persuasive essays. Their results suggest that emotional appeal is strongly correlated with persuasion and appears mainly in premises.

Habernal and Gurevych (2016) developed a crowdsourced corpus of argument pairs and asked annotators to determine which argument is more convincing. Based on such relative assessments, they modeled convincingness with feature-rich SVMs and LSTM neural networks with convolution and attention mechanisms to detect unconvincing arguments. Similarly, Gleize et al. (2019) and Toledo et al. (2019) took a relative approach to argument quality. They created crowsourced corpora by asking annotators whether they would recommend a particular argument to a friend while asking them to disregard their own opinion on a topic. Then they modeled the quality of the arguments using various neural solutions, including transformer-based approaches and bidirectional LSTMs.

While framing the task as a relative assessment allows to some extent to abstract from annotator bias and simplify the problem, some research also developed corpora to facilitate *absolute* assessments (Gretz et al., 2020; Fromm et al., 2023). For example, Gretz et al. (2020) introduced one of the largest to date crowdsourced corpora with pointwise quality annotations in the form of scores and binary judgments. In their study, the authors examined the correlation between these annotations and the 15 dimensions outlined in the framework proposed by Wachsmuth et al. (2017b) revealing that mainly two dimensions, namely *global relevance* and *effectiveness*, were captured by the annotations.

As previously mentioned, certain argument quality dimensions depend not only on the language employed but also on the attributes of the author and audience. In Durmus et al. (2019b), the authors tried to predict which argumentative claims will have the most *impact* on the audience. To do so, they modeled argument trees from an online debate portal, Kialo, using audience votes as indicators of impact. They found that incorporating the surrounding context of the claim, e.g., preceding claims, into the modeling process leads to significant performance improvements. In Lukin et al. (2017), the authors considered personality

traits in persuasion dialogues on social and political issues. They found that personality factors can affect belief change, with conscientious, open, and agreeable people being more convinced by emotional arguments. Similarly, El Baff et al. (2020) study the impact of argumentative texts on people depending on their political ideology and find that style has a significant influence on how (liberal) editorials affect (liberal) readers.

Appropriateness Although the concept of appropriateness is considered an important aspect of argument quality, it has been widely overlooked, with no systematic studies on how to define, assess, and model the appropriateness of arguments until very recently. In (Ziegenbein et al., 2023), the authors present a taxonomy of appropriateness derived from rhetoric and argumentation theory. To assess whether the suggested sub-dimensions of appropriateness can be modeled computationally, initial experiments have been conducted with a transformer-based model (DeBERTa, He et al. (2021)), achieving results close to human-level performance.

Relevance Wachsmuth et al. (2017c) propose a model for determining the *relevance* of arguments by abstracting from the content and ranking arguments solely based on structural relations. In this approach, the relevance of an argument's conclusion is decided by what other arguments reuse it as a premise. In contrast, Guo and Singh (2023) propose to use Siamese networks to determine whether an argumentative claim is relevant to another claim. As training data they suggest to extract pairs of claims from deeply nested tree-structured debates from online platforms, such as Kialo. They consider direct relations, i.e., claims, where one directly supports or opposes the other claim, ancestor-descendant relations, where the claims belong to the same branch of a tree are but not directly related, and randomly sampled pairs.

Opposed to only focusing on the assessment of single quality dimensions, Lauscher et al. (2020) suggest to model the three core argument quality dimensions (Wachsmuth et al., 2017b) – Cogency, Effectiveness, Reasonableness (see Section 2.1 for details) – in a multi-task learning setup. In such setup, for each quality dimension, a separate prediction layer with an individual task loss is employed. The total training loss is then defined as the sum of all losses. Doing so enables exploiting interactions between the dimensions to boost predictive performance. Specifically, the authors considered the BERT model architecture

and applied the modeling to a multi-domain dataset, which includes q&a forum posts, online debate content, and business reviews. Their results suggest that multi-task learning can lead to improvements for all considered quality dimensions.

Alternative Viewpoints on Argument Quality Assessment Though the aforementioned dimensions offer a comprehensive overview from the Argumentation Theory perspective, other fields and research communities have dedicated their attention to assessing argument quality as well. Notably, from a Social Science perspective the Deliberative Theory community has produced different theories of argument quality, where arguments are assessed not only by their persuasive power but also by their potential to contribute to a reasoned agreement among participants (Atkinson et al., 2013). From a computational perspective, Park et al. (2012) and Falk and Lapesa (2023) have approached the task of modeling the deliberative quality of arguments in an online public rulemaking platform. For instance, in (Park et al., 2012) the authors focus on facilitate moderation to help maximize individual contribution and promote quality discussions among the users. They do so by using an SVM to predict the type of action a moderator should perform given an argumentative text. Here, the type of actions considered were limited to two types: broadening the scope of the discussion and improving argument quality. More recently, (Falk and Lapesa, 2023) performed a comprehensive analysis to understand the relationships between argument quality and deliberative quality dimensions by combining several datasets and modeling them with transformer-based adapters (Pfeiffer et al., 2021) in multi-task setup. Their results suggest that automatically assessing certain individual quality dimensions can be improved by injecting knowledge about related dimensions (for example, dimensions that come from similar datasets or conceptually related dimensions).

Another way of assessing the quality of arguments is by identifying certain flaws or fallacies they may contain. A fallacy is an erroneous or deceptive argument that might seem persuasive but lacks logical validity (Hamblin, 1970). Fallacies can take various forms (over 150 types), each highlighting a different way in which an argument fails to provide valid or reliable reasoning (Van Eemeren and Grootendorst, 1987; Tindale, 2007). While most computational work on detecting fallacies mainly employed rule-based systems and theoretical frameworks (Gibson et al., 2007; Nakpih and Santini, 2020), some more recent work considered neural methods as part of their approaches (Habernal et al., 2018; Jin et al., 2022; Goffredo et al., 2022). For

instance, Habernal et al. (2018) employed self-attentive LSTM models to predict a certain type of fallacy - *ad hominem*, where the arguer attacks the person instead of the claim. In (Jin et al., 2022), the authors used transformer-based models to detect 13 types of logical fallacies exploiting coreference resolution and entity linking to abstract from the surface of the arguments and identify logical fallacies that are structurally fallacious in their arguments. Similarly, Goffredo et al. (2022) jointly finetuned four transformer-based models to detect fallacies in political debates. Each model processed a certain part of the argument, i.e., the dialogue context, argument components (premise and claim), fallacious argument snippet, argument relation (attack or support). They show that detecting argument components, relations, and context in debates is a necessary step when detecting fallacies.

We complement the existing body of work by introducing a novel perspective to modeling argument quality. Opposed to simply considering various claims with different content, we focus on modeling the differences between several versions of the same argumentative claim. By adopting this approach, we gain a comprehensive understanding of quality characteristics that are not limited to the specifics mentioned in the text. Moreover, this enables a more refined analysis of how individual claims can adapt and increase their quality through revisions. However, this perspective also presents its own set of challenges, including the need for careful data collection and methodological innovation. In Chapter 3 we provide a detailed overview of these challenges and propose solutions. Despite the dissimilarity in framing, we consider the same scope of argument quality dimensions and the same spectrum of computational approaches as in previous work covering both traditional (Persing et al., 2010; Stab and Gurevych, 2017b; Habernal and Gurevych, 2016) and neural approaches (Gurcke et al., 2021; Toledo et al., 2019; Ziegenbein et al., 2023). Guided by related work (Persing and Ng, 2015; Durmus et al., 2019b), we also consider the benefit of employing contextual information (Chapters 3 and 5) and assess the relationship between different types of context and argument quality dimensions (Chapter 5).

2.4.2 Argument Generation

While argument quality assessment has been a focal point in computational argumentation research, the exploration of argument generation has received comparatively less attention in the field. However, recent advancements in natural language generation and the emergence of large language models, such

as BART (Lewis et al., 2020), T5 (Raffel et al., 2020) have paved the way for significant progress in this area, opening new opportunities for developing systems that can engage in argumentative exchanges or support such exchanges between humans.

In traditional text generation systems, a lot of attention has been given to content selection and planning (Reiter and Dale, 1997). These processes involve determining the relevant information to be included in the generated text and organizing it in a coherent and logical manner. Early approaches mainly relied on hand-crafted features and rules based on discourse theory (Hovy, 1993) and expert knowledge (Reiter et al., 2000). For example, Elhadad (1995) and Zukerman et al. (2000) developed rules based on theories of argumentation, specifically, they explore the usage of common sense relations (*topos*) and discourse strategies to guide content selection and text planning.

Recent advancements in neural generation models have significantly reduced the need for human effort in system engineering by introducing an end-to-end trained conditional text generation framework. This framework enables learning the rules and patterns for content selection and text planning automatically, enabling a variety of argument generation tasks. We will give an overview of the most recent approaches to argument generation, focusing mainly on such tasks as argument synthesis, reframing, and summarization.

Argument Synthesis Argument synthesis can be defined as the task of generating new arguments that support or refute a given viewpoint or hypothesis. Typically, the task is approached using controlled text generation techniques, where the goal is to generate text that meets certain criteria, which can be specified through various forms of input, such as prompts, keywords, templates, or constraints. For example, Schiller et al. (2021) learn to generate argumentative text by conditioning a transformer-based model (Keskar et al., 2019) on topics, stances, and certain aspects the argument should address. To promote the generation of argumentative claims covering different aspects of a claim Park et al. (2019) introduce a diversity penalty as part of a sequence-to-sequence framework. Opposed to these works, in our task of argument optimization (Chapter 5), the goal is to improve the delivery of an argument without changing its meaning, i.e., the aspects of the claim should remain the same.

Alshomary et al. (2021a) emphasize the importance of the audience's characteristics and suggest generating claims on a given topic that also match the morals of a given user based on their stance towards "*big issues*". Although

they showed that it is possible to encode beliefs into claims, they did not investigate how effective these claims are in persuading an audience.

Other work focuses on the generation of counter-arguments that challenge or oppose a given statement on a controversial issue. Hua and Wang (2019) use a sequence-to-sequence model (Sutskever et al., 2014) to generate counter-arguments by using automatically extracted keyphrases from Wikipedia as input to the generation model. While Hidey and McKeown (2019) proposed a neural model that edits the original claim semantically to produce a claim with an opposing stance. On the other hand, Alshomary et al. (2021c) approached the task of counter-argument generation by identifying weak premises and undermining them.

Argument Reframing Another popular direction of research is argument reframing, i.e., changing the way an argument is presented or perceived. Often such changes include the polarity or sentiment of the argument, for example, Hu et al. (2017); Lai et al. (2019) trained variational auto-encoders (Mueller et al., 2017) and generative adversarial networks (Goodfellow et al., 2020), respectively, to generate sentences with a specific polarity. Later work utilized large transformer models to tackle the same task, as shown in (Sudhakar et al., 2019). Other work has employed neural models to *"flip"* the bias of news headlines based on political convictions (Chen et al., 2018b), or to reframe arguments to appear more trustworthy by combining controlled generation with entailment modeling (Chakrabarty et al., 2021). For example, Chakrabarty et al. (2021) first generated several candidate arguments, then measured the entailment of the generated candidates to the input and selected the one with the highest score as the optimal candidate. We take a similar approach in the task of claim optimization (Chapter 5) and develop our own post-processing approach that, based on the fluency, argument quality, and semantic similarity of the generated outputs, selects the best one.

Argument Summarization Argument summarization is the task of generating a concise summary of an argumentative text (e.g., essay, article) or collection of texts (e.g., debate, online discussion) that captures its main claims, evidence, and reasoning. Previous work on this task considered training attention-based neural networks for generating abstractive summaries of opinionated text by selecting the select the most representative opinionated sentences using sub-modular optimization (Wang and Ling, 2016). While the authors of (Egan et al., 2016) propose to generate summaries of debates by

extracting meaningful key points made in the debate, which are then grouped into discussion summaries. Similarly, Bar-Haim et al. (2020) and Alshomary et al. (2021b) learn to summarize debates based on key points. While the first paper does so by taking into account salience (the number of arguments a key point is matched to), the latter proposed a graph-based extractive summarization approach, which utilizes a PageRank variant to rank sentences in the input arguments by quality and outputs the top-ranked sentences as key points. Other domains of argumentative texts have also been considered, for instance, Syed et al. (2020) applied a graph-based extractive summarization approach to generate concise and fluent summaries of opinions in news editorials.

In contrast to approaches that solely concentrate on a specific narrow task, some researchers have considered a more holistic approach to modeling argumentation by developing full debating systems that seamlessly integrate multiple tasks, including argument mining, retrieval, and generation. One noteworthy advancement in this domain is the autonomous debating system proposed by (Slonim et al., 2021), known as Project Debater. This system has the ability to engage in competitive debates with humans and relies on a hybrid approach, which combines various retrieval, mining, clustering, and rephrasing components. However, the system does not offer writing assistance or support guiding users in improving the delivery of their arguments. Hence, one of the practical applications of our work could be the improvement of such systems by integrating methods proposed in this thesis as a post-processing mechanism or directly within the system.

2.4.3 Argument Revision Processing

Analyzing and modeling revisions of argumentative texts in terms of their quality is rather understudied. While so far no research tackled *rewriting* argumentative texts automatically to enhance their quality, some work on classifying and assessing argumentative writing revisions has been already carried out. In the following sections, we summarize the relevant studies on the topic highlighting the methodologies, findings, and limitations that served as a foundation for our research. Furthermore, we give an overview of recent revision generation approaches developed for other natural language processing tasks, such as simplification and grammar error correction, that inspired our work.

Revision Type Classification As mentioned in Section 2.2, various taxonomies have been proposed to categorize different types of revision as well as their

intentions and purposes. The most relevant work on computationally modeling such revisions is the work of Zhang and Litman (2015) and Zhang et al. (2017), where the authors propose to classify the reasons why writers perform revisions on a sentence level. Specifically, they suggest applying Random Forest Tree classifiers (Breiman, 2001) or SVMs in a supervised setup to various hand-crafted text features, including n-grams, position of sentence in the text, named entity features, discourse markers, and part-of-speech tags. The authors distinguish surface and text-based changes, and their results suggest that text-based changes are significantly correlated with writing improvement, while surface changes are not.

In a follow-up work (Zhang and Litman, 2016), the authors found that using features derived from the context around the revised sentence, such as coherence and cohesion, can further lead to classification performance improvements. Kashefi et al. (2022) extend the corpus of Zhang and Litman (2015) with more fine-grained annotations enabling a more granular distinction of content-level and surface-level revision changes. They propose to classify revision purposes by applying an XGBoost classifier (Chen and Guestrin, 2016) to various features representing textual (length and position), syntactic (part-of-speech), semantic (embedding), and discourse (transition words) aspects. Their results suggest that solely using embeddings from a pre-trained language model can produce competitive results even without the suggested textual, syntactic, and discourse features when predicting revision purposes.

Although neural solutions have not yet been widely considered in argument revision type classification, other domains have reported on their successful usage in similar tasks. For instance, in (Du et al., 2022), the authors fine-tune a transformer-based model to classify revision intentions in Wikipedia-style articles, academic essays, and news articles. Specifically, they consider clarity, fluency, coherence, style, and meaning-changing edits.

It should be noted that all of these works consider two versions as input and do not aim to identify issues within the texts but only to categorize the type of change introduced by the revision. Some work, such as (Bhat et al., 2020), consider detecting sentences requiring revision in other domains, such as instructional texts. However, due to the distinct objectives, varying perceptions of quality, and, consequently, different types of revisions performed across the domains, the results do not fully transfer to argumentative texts.

Assessment of Revision Quality Compared to characterizing the types of revisions typically found in argumentative texts, little attention has been given

to modeling and understanding the changes within any of the argument quality dimensions (e.g., argument strength, clarity, persuasiveness, etc.) that occur during the revisions.

The only works we are aware of that analyzes revision quality of argumentative texts are the studies of Afrin and Litman (2018), Liu et al. (2023), and Afrin and Litman (2023). In (Afrin and Litman, 2018), the authors propose to automatically assess whether a revision of an argumentative essay has been successful, i.e., improves the quality of the text or not, by comparing the two versions. To do so, they suggest using various hand-crafted features, such as n-grams and named entities, and feeding them into a Random Forest Tree classifier. To deal with class imbalance, they apply SMOTE (Chawla et al., 2002), which is an oversampling technique where new synthetic examples are generated for underrepresented classes. They experiment with expert and non-expert(student) revisions and find that using expert data to model quality is particularly important when predicting low-quality revisions. Similarly, Liu et al. (2023) explore how recent large language models, such as ChatGPT [1], can be incorporated in argument revision assessment. Specifically, they suggest predict successful revisions by using Chain-of-Thought prompting (Wei et al., 2022) to first extract claim, evidence, and reasoning sentences from input essays separately and then classify summaries of the extracted content with a transformer-based model.

On the other hand, in (Afrin and Litman, 2023), the authors introduce the concept of *revision desirability*. Under desirable revisions, the authors understand, revisions which have a hypothesized utility in improving an essay in response to certain feedback. The feedback is given to the students before they begin their revisions, e.g., “Explain how the evidence connects to the main idea and elaborate:" or “Explain how the evidence helps you make your point". In experiments, the authors suggest to feed BERT embeddings to a BiLSTM model to predict whether the performed revision was desirable. Their results suggest that contextual information, such as neighboring sentences or all revised sentences in the essay along with feedback provided before the revision, leads to performance improvements.

Though similar quality assessment tasks have been approached in other domains, such as news articles, online encyclopedia entries, and instructional texts (see Section 2.2), such results typically do not fully transfer to argumentative texts as they do not account for the domain-specific argument quality

[1] https://openai.com/blog/chatgpt

characteristics. This discrepancy arises from the different objectives, quality standards and definitions, and consequently, different types of revisions that are inherent to each domain.

As previously mentioned, when it comes to essay revisions, there is a possibility that the quality of the text may decrease, particularly when these revisions are carried out by individuals who may not be experts (students). In contrast, in online platforms that implement rigorous moderation procedures, the likelihood of such a decrease in quality due to revisions becomes very low and unlikely. In this thesis, we work with revision histories from the online debate platform, Kialo, which not only employs moderators to overlook the quality of the created content but also allows the users themselves to suggest improvements and revisions to existing arguments. In an annotation study (see Chapter 4 for details), we explore the impact of these revisions on the quality of claims and find that in 93% of cases, the revisions improved their quality.

Revision Generation Despite being largely disregarded in the field of computational argumentation, automated revision generation has recently garnered attention from practitioners in the broader field of natural language processing. Primarily, these efforts have targeted specific narrow tasks such as sentence simplification (Botha et al., 2018), grammatical error correction (Lichtarge et al., 2019), bias neutralization (Pryzant et al., 2020) and used Wikipedia revision histories (Faruqui et al., 2018) as training data.

For instance, in the case of sentence simplification, Botha et al. (2018) approach the problem as a *split-and-rephrase*, where larger sentences are broken down into shorter ones that together convey the same meaning. To do so, they suggest employing the *COPY512* sequence-to-sequence architecture of Aharoni and Goldberg (2018), which is a one-layer, bi-directional LSTM with attention and a copying mechanism that dynamically interpolates the standard word distribution with a distribution over the words in the input sentence.

In (Lichtarge et al., 2019), for the task of grammar error correction, the authors proposed an iterative decoding algorithm that allows a transformer-based sequence-to-sequence model to make multiple incremental corrections until no further edits are required. The stopping criteria are determined by the model's confidence, i.e., the model would rewrite the text only if the confidence surpasses a prespecified threshold, which is found empirically.

In the case of subjective bias neutralization, Pryzant et al. (2020) used a collection of original and debiased sentence pairs mined from Wikipedia edits

to train an encoder-decoder neural network. While the encoder is represented by a BERT-based classifier trained to identify problematic words. The decoder is a bidirectional LSTM with attention, copy, and coverage mechanisms. Due to the specifics of the proposed architecture, the model is limited to single-word edits and can only handle the simplest instances of bias.

More recently, the development of larger language models, such as BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and PEGASUS (Zhang et al., 2020) has allowed researchers to expand their conceptualization of revision tasks. In particular, Faltings et al. (2021) explored natural language conditioned editing as a means for controllable text generation using a T5-based sequence-to-sequence model. The suggested approach required a command in natural language describing the nature of the revision that needs to be performed and generating the revised text based on a grounding consisting of snippets of web page results. Similarly, Du et al. (2022) and (Rajagopal et al., 2022) considered conditioning sequence-to-sequence models, such as BART and PEGASUS, by using phrases denoting the intent of the edit the model should perform. While Rajagopal et al. (2022) work only with Wikipedia edits, Du et al. (2022) model revisions as an iterative process and combine Wikipedia edits with revisions of scientific texts and news articles during modeling. Inspired by these works, in Chapter 5, we explore the benefits of using controllable text generation to condition the inputs on additional contextual information to support the generation of relevant and grounded argumentative claims.

Having introduced the fundamental concepts underlying this thesis and discussing relevant previous research, we now move forward and introduce the publications that this thesis builds upon as separate chapters.

3. Suboptimal Claim Detection and Claim Improvement Suggestion

This chapter presents the original content of the following paper (Skitalinskaya and Wachsmuth, 2023) by Gabriella Skitalinskaya and Henning Wachsmuth: *"To Revise or Not to Revise: Learning to Detect Improvable Claims for Argumentative Writing Support"* in the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics 2023 (ACL'23).

Abstract

Optimizing the phrasing of argumentative text is crucial in higher education and professional development. However, assessing whether and how the different claims in a text should be revised is a hard task, especially for novice writers. In this work, we explore the main challenges to identifying argumentative claims in need of specific revisions. By learning from collaborative editing behaviors in online debates, we seek to capture implicit revision patterns in order to develop approaches aimed at guiding writers in how to further improve their arguments. We systematically compare the ability of common word embedding models to capture the differences between different versions of the same text, and we analyze their impact on various types of writing issues. To deal with the noisy nature of revision-based corpora, we propose a new sampling strategy based on revision distance. Opposed to approaches from prior work, such sampling can be done without employing additional annotations and judgments. Moreover, we provide evidence that using contextual information and domain knowledge can further improve prediction results. How useful a certain type of context is, depends on the issue the claim

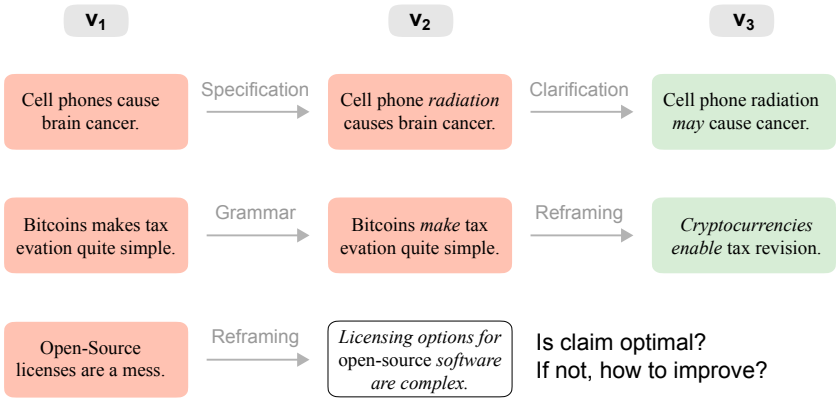


Figure 3.1: Examples of revision histories of argumentative claims from the online debate platform Kialo. Colors denote claims considered optimal (light green) and claims requiring revisions (medium red). We seek to identify if and how a new claim should be revised.

is suffering from, though.

3.1 Introduction

Text revision is an essential part of professional writing and is typically a recursive process until a somehow *optimal* phrasing is achieved from the author's point of view. Aside from proofreading and copyediting, text revision subsumes substantive and rhetorical changes not only at the lexical, syntactic, and semantic levels, but also some that may require knowledge about the topic of discussion or about conventions of the domain or genre. An optimal phrasing is especially important in argumentative writing, where it is considered a key component in academic and professional success: An argument's style directly affects its persuasive effect on the audience (El Baff et al., 2020).

But how to know whether an argument is phrased well enough and no more revisions are needed? Most existing approaches to argument quality assessment score arguments on different aspects of a topic or compare one to another, rather than detecting issues within arguments to highlight potential improvements (see Section 3.2 for details). Beyond those, Zhang and Litman (2015) analyze the nature of revisions in argumentative writing. They annotate revisions at various levels to learn to classify changes that occur. Others compare revisions in terms of quality on essay level (Afrin and Litman, 2018) or claim level (Skitalinskaya et al., 2021). Still, the question of whether a given argumentative text should be revised remains unexplored.

Figure 3.1 illustrates the underlying learning problem. What makes re-

search on detecting the need for revision challenging is the noisy and biased nature of revision-based corpora in general and respective argument corpora specifically. Not only is it uncertain whether a text will be revised in the future and how, but also the inherent subjectivity and context dependency of certain argument quality dimensions (Wachsmuth et al., 2017b) pose challenges.

In this work, we investigate how to best develop approaches that identify argumentative claims in need of further revision, and that decide what type of revision is most pressing. We delineate the main challenges originating from the nature of revision-based corpora and from the notion of argument quality. To tackle these challenges, we exploit different types of knowledge specific to text and argument revisions: (a) the number of revisions performed, in the past and the available future, (b) the types of revision performed, such as typo correction vs. clarification, (c) contextual information, such as the main thesis or parent claim in the given debate, (d) topic knowledge, such as debates belonging to the same topical categories, and (e) the nature of revisions and their concordance with training processes of embedding representations.

In systematic experiments on a claim revision corpus, we provide evidence that some explored approaches can detect claims in need of revision well even in low-resource scenarios, if appropriate sampling strategies are used. While employing contextual information leads to further improvements in cases where linguistic cues may be too subtle, we find that it may also be harmful when detecting certain types of issues within the claim.

We argue that technologies that identify texts in need of revision can highly benefit writing assistance systems, supporting users in formulating arguments in a better way in order to optimize their impact. The main contributions of this paper are:

1. An overview of the main challenges in assessing whether a claim needs revision;
2. a detailed analysis of the strengths and weaknesses of strategies to tackle said challenges, guiding future research on revisions in argumentation and other domains;
3. a systematic comparison of approaches based on different contextualized representations for the tasks of suboptimal-claim detection and claim improvement suggestion.[1]

[1]Data, code, and models from our experiments are found at `https://github.com/webis-de/ACL-23`.

3.2 Related Work

Foundational studies of writing specify two main revision sub-processes: evaluating (reading the produced text and detecting any problems) and editing (finding an optimal solution and executing the changes) (Flower et al., 1986). In this work, we focus on the former in the domain of *argumentative texts*. Although considerable attention has been given to the computational assessment of the quality of such texts, very few works consider the effects of revision behaviors on quality.

Existing research largely focuses on the absolute and relative assessment of single quality dimensions, such as cogency and reasonableness (Marro et al., 2022). Wachsmuth et al. (2017b) propose a unifying taxonomy that combines 15 quality dimensions. They clarify that some dimensions, such as acceptability and rhetorical effectiveness, depend on the social and cultural context of the writer and/or audience. A number of approaches exist that control for topic and stance (Habernal and Gurevych, 2016), model the audience (Al Khatib et al., 2020), or their beliefs (El Baff et al., 2020), and similar. However, they all compare texts with *different* content and meaning in terms of the aspects of topics being discussed. While such comparisons help characterize good arguments, they are not geared towards identifying issues within them, let alone towards guiding writers on how to improve the quality of their arguments.

The only works we are aware of that study revisions of argumentative texts are those of Afrin and Litman (2018), Skitalinskaya et al. (2021), and Kashefi et al. (2022). The first two suggest approaches that enable automatic assessments of whether a revision can be considered successful, that is, it improves the quality of the argumentative essay or claim. The third extends the corpus of Zhang and Litman (2015) to complement 86 essays with more fine-grained annotations, enabling the distinction of content-level from surface-level revision changes at different levels of granularity. All these approaches to characterizing the type of revision and its quality require two versions as input. In contrast, we seek to identify whether an argumentative text needs to be revised at all and, if so, what type of improvement should be undertaken. In such framing, the solutions to the tasks can also be used to support argument generation approaches, for example, by helping identify weak arguments for counter-argument generation (Alshomary et al., 2021c), as well as automated revision approaches, for example, by providing required revision types or weak points as prompts (Hua and Wang, 2020; Skitalinskaya et al., 2023).

Due to the lack of corpora where revisions are performed by the authors

of texts themselves, researchers utilize collaborative online platforms. Such platforms encourage users to revise and improve existing content, such as encyclopedias (Faruqui et al., 2018), how-to instructions (Anthonio et al., 2020), Q&A sites (Li et al., 2015), and debate portals (Skitalinskaya et al., 2021). Studies have explored ways to automate content regulation, namely text simplification (Botha et al., 2018), detection of grammar errors (Lichtarge et al., 2019), lack of citations (Redi et al., 2019), biased language (De Kock and Vlachos, 2022), and vagueness (Debnath and Roth, 2021). While Bhat et al. (2020) consider a task similar to ours – detecting sentences in need of revision in the domain of instructional texts – their findings do not fully transfer to argumentative texts, as different domains have different goals, different notions of quality, and, subsequently, different revision types performed.

Revision histories of peer-reviewed content help alleviate the shortcomings typical for self-revisions, where a writer may fail to improve a text for lack of expertise or skills (Fitzgerald, 1987). Yet, they also introduce new challenges stemming from sociocultural aspects, such as opinion bias (Garimella et al., 2018; Pryzant et al., 2020) and exposure bias (Westerwick et al., 2017). Approaches to filtering out true positive and negative samples have been suggested to tackle such issues. These include community quality assessments, where high quality content is determined based on editor or user ratings and upvotes (Redi et al., 2019; Chen et al., 2018a), timestamp-based heuristics, where high-quality labels are assigned to content that has not been revised for a certain time period (Anthonio et al., 2020), and complementary crowdsourced annotation (Asthana et al., 2021). However, all this requires domain-specific information which may not be available in general. In our experiments, we analyze the potential of sampling data solely based on revision characteristics, namely revision distance (the number of revisions between a certain claim version and its final version).

Moreover, studies have shown that writing expertise is domain-dependent, revealing commonalities within various professional and academic writing domains (Bazerman, 2016). Although certain quality aspects can be defined and evaluated using explicit rules, norms, and guidelines typical for a domain, not all quality aspects can be encoded in such rules. This raises the need to develop approaches capable of capturing implicit revision behaviors and incorporating additional context relevant to the decision-making process (Flower et al., 1986; Boltužić and Šnajder, 2016). Below, we outline the main challenges stemming from the noisy and biased nature of revision-based corpora as well

as from the context dependence of certain argument quality aspects. We then establish potential data filtering and sampling strategies targeting said issues.

3.3 Tasks and Challenges

Revision-based data provides many opportunities; yet, it also comes with several challenges that arise at different stages of the experiment design process. In the following, we define the tasks we deal with in this paper, summarize the main challenges, and outline our approaches to these challenges.

3.3.1 Tasks

Previous work has studied how to identify the better of two revisions. However, this does not suffice to support humans in optimizing their arguments, as it remains unclear when a claim is phrased optimally. We close this gap by studying two new tasks:

Suboptimal-Claim Detection Given an argumentative claim, decide whether it is in need of further revision or can be considered to be phrased more or less optimally (binary classification).

Claim Improvement Suggestion Given an argumentative claim, select all types of quality issues from a defined set that should be improved when revising the claim (multi-class classification).

Reasons for revision can vary strongly, from the correction of grammatical errors to the clarification of ambiguity or the addition of evidence supporting the claim. In our experiments, we select quality issues sufficiently represented in the given data.

3.3.2 Challenges

To tackle the given tasks on revision-based data, the following main challenges need to be faced:

- *Data.* Compiling a dataset that is (a) representative and reliable and (b) free of topical bias.
- *Model.* Selecting a model for the task whose complexity and architecture fit the data.
- *Context.* Incorporating complementary contextual knowledge useful for the tasks at hand.

We detail each challenge below and discuss how we approach it in our experiments.

Representativity and Reliability Compiling a reliable dataset from claim revision histories that represents argumentative claim quality well is not straightforward. While examples of suboptimal quality are rather easy to find, since each revision signals that something is wrong with the previous version, identifying examples of high (ideally, optimal) quality text remains a challenge. The main reason is that such texts remain unchanged throughout time in collaborative systems, yet the same holds for low-quality texts that may have been overlooked and never revised.

Prior work largely employs external information and additional quality assessments to sample representative examples (see Section 3.2), limiting scalability. In this paper, we complement existing efforts by suggesting a scalable sampling strategy solely based on revision history characteristics, namely revision distance, which denotes the number of revisions that occurred until an optimal (final) state of the claim was reached. The proposed strategy as illustrated in Figure 3.2 only considers claim histories with 4 or more revisions (chosen empirically). At each revision distance i from 1 to 4, a dataset Di is compiled, where all final versions of claims are considered as positive examples not needing a revision, and claim versions at revision distance i are considered as negative ones.

Another problem is identifying flaws that need to be tackled in a revision. Although a claim may suffer from multiple flaws at the same time, not all of them may be eliminated in the same revision. In the dataset introduced in Section 3.4, revisions may be accompanied by a label describing the type of improvement performed. Still, such labels are skewed towards improvements addressed by the community and do not account for other flaws in the text.

To address these issues, we explore three ways of extracting quality labels from revision histories:

- We consider the revision distance between positive and negative examples when identifying claims in need of revision (Section 3.6.2).
- We extend the given dataset with examples of claims that were never revised (Section 3.4).
- We frame the improvement suggestion task as a multi-class classification task, where only the next most probable improvement type is predicted. This better reflects the iterative nature of revision processes and accounts for the lack of explicit quality labels (Section 3.6.5).

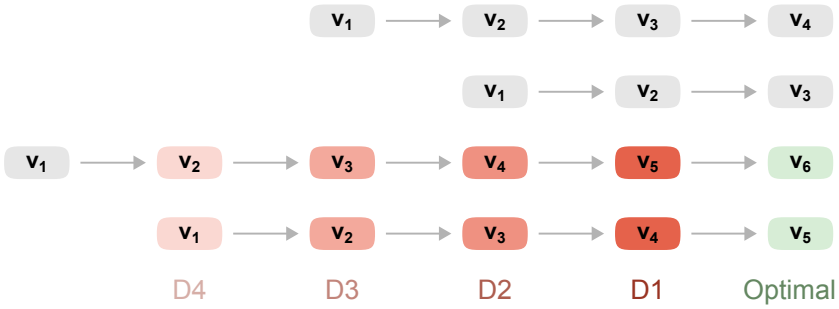


Figure 3.2: Illustration of sampling approach. Shades of red denote claims requiring revisions at different revision distances from 1 to 4, while final versions are green and represent optimally phrased claims.

Topical Bias Despite the best efforts, histories of collaborative online revisions may contain noise, be it due to accidental mistakes or biases of users or moderators. Different users may have conflicting opinions on what should or should not be edited, and certain debate topics may be seen as controversial, making it even more difficult to assess the quality of claims and suggest further improvements.

Accounting for such bias is inherently difficult and also depends on the prominence of such behaviors in the given community. We do not solve this issue here, but we explore it:

- We determine the extent to which bias differs across topical debate categories by assessing performance differences when including claims on specific topics or not (Section 3.6.3).

Model Complexity and Architecture Learning quality differences across several versions of the same argumentative claim likely requires a model whose architecture aligns with the idea of revisions. To determine the best model, we carry out two complementary steps:

- We train several types of models of varying complexity, including statistical and neural approaches to both obtaining claim representations and classification (Section 3.5).
- We disentangle how pre-training, fine-tuning, and the final classification affect the performance of claim assessment (Section 3.6.1).

Contextuality As mentioned in Section 3.2, some quality aspects require domain knowledge. However, determining what kind of information should be included when deciding whether a text needs a revision remains an open

question. Some revisions may be typical for debate as a whole, for example, related to a desired structure, layout and style of citations, or choice of words for main concepts in the debate. In such cases, conditioning models on the debate thesis may be beneficial. Others may depend on the parent claim, which is supported or opposed by the claim in question, and affects whether clarifications or edits improving the relevance of the claim are needed, and potentially general domain knowledge as well (Gretz et al., 2020).

Therefore, we explore contextuality as follows:

- We compare benefits of using contextual debate information of varying specificity when detecting suboptimal claims and recommending revision types (Sections 3.6.4–3.6.5).

3.4 Data

In our experiments, we use ClaimRev (Skitalinskaya et al., 2021): a corpus of over 124,312 claim revision histories, collected from the online debating platform Kialo,[2] which uses the structure of dialectical reasoning. Each debate has a main thesis, which is elaborated through pro and con claims, which allows to consider each comment as a self-contained, relevant argument. Each revision history is a set of claims in chronological order, where each claim represents a revised version of the preceding claim meant to improve a claim's quality, which holds in 93% of all cases according to an annotation study (Skitalinskaya et al., 2021).

We extend the corpus by extracting 86,482 unrevised claims from the same set of debates as in ClaimRev, which have been introduced *before* the reported date of data collection (June 26, 2020). Since claims that have been introduced shortly before this date are still likely to receive revisions, we additionally filter out claims that have undergone a revision within six months after the initial data collection (December 22, 2020). We remove all revision histories, where claim versions have been reverted to exclude potential cases of edit wars.

Our final corpus is formed by 410,204 claims with 207,986 instances representing optimally phrased claims (positive class) and 202,218 instances requiring revisions (negative class). All claims in need of further refinement are also provided with labels indicating the specific type of improvement the claim could benefit from. In this work, we limit ourselves to the three most common types, covering 95% of all labels revisions in the ClaimRev dataset: clarification,

[2]Kialo, `https://www.kialo.com`

Subset	Type	# Instances
Positive	Final in history	121 504
	Unrevised	86 482
Negative	Clarification	61 142
	Typo/Grammar	57 219
	Links	17 467
	Other/Unlabeled	66 390
Overall		410 204

Table 3.1: Number of instances in the extended corpus. Positive examples represent claims considered as optimally phrased. Negative examples require revisions.

typo/grammar correction, and adding/correcting links. Specifically, *clarification* means to adjust/rephrase a claim to make it more clear, *typo/grammar correction* simply indicates linguistic edits, and *adding/correcting links* refers to the insertion or editing of evidence in the form of links that provide supporting information or external resources to the claim. Statistics of the final dataset are shown in Table 3.1. Ensuring that all versions of the same claim appear in the same split, we assign 70% of the histories to the training set and the remaining 30% are evenly split between the dev and test sets.

3.5 Methods

To study the two proposed tasks, we consider two experimental settings: (i) extracting claim representations by using embeddings as input to an SVM (Joachims, 2006), (ii) adding a classifier layer on top of pre-trained transformer models with further fine-tuning (FT).

In our experiments, we consider the following approaches to generating claim representations:

- *Glove* (Pennington et al., 2014). A static word embedding method
- *Flair* (Akbik et al., 2018). A contextual character-level embedding method
- *BERT* (Devlin et al., 2019). A standard baseline pre-trained transformer
- *ELECTRA* (Clark et al., 2020). A transformer with adversarial pre-training fitting our tasks
- *DeBERTa* (He et al., 2021). A transformer that achieved state-of-the-art performance on the SuperGLUE benchmark (Wang et al., 2019).

Approach	Model	Accuracy	Ma. F_1	P	R	F_1
Embed.	Glove	54.9	54.9	54.9	50.0	52.1
+ SVM	Flair	60.1	60.1	60.2	**56.9**	58.5
	BERT	62.1	61.8	63.5	54.7	58.8
	ELECTRA	63.2	62.9	65.1	55.0	59.6
	DeBERTa	61.5	61.2	63.2	52.9	57.6
Fine-tuned	FT-BERT	63.1	61.7	70.1	44.2	54.2
	FT-ELECTRA	63.8	62.9	68.8	49.0	57.2
	FT-DeBERTa	**67.1**	**66.6**	**71.3**	55.9	**62.6**
Random baseline		50.0	50.0	50.0	50.0	50.0

Table 3.2: Suboptimal-claim detection: Accuracy, macro F_1, and precision/recall/F_1 of the suboptimal class for all tested models, averaged over five runs. Per approach, all gains from one row to another are significant at $p < .001$ according to a two-sided student's t-test.

3.6 Experiments

Based on the data from Section 3.4 and the methods from Section 3.5, we now present a series of experiments aimed at understanding the effects and possible solutions to the four challenges from Section 3.3: (1) the right model complexity and architecture to capture differences between claim revisions; (2) representative and reliable examples of high and low quality; (3) the impact of topical bias in the data; (4) contextuality, where the quality of a claim may depend on its surrounding claims.

3.6.1 Model Complexity and Architecture

First, we explore the ability of the methods to predict whether a claim can be considered as optimal or needs to be revised. We treat all unrevised claims and final versions of claims as not needing revisions and all preceding versions as requiring revisions.

Table 3.2 presents the results of integrating several embeddings with a linear SVM classifier and fine-tuned transformer-based language models. Although we see gradual substantial improvements as we increase the complexity of the models used to generate the word embeddings, the best results (accuracy 67.1, macro F_1 66.6) indicate the difficult nature of the task itself. Low results of *Glove* (both 54.9) indicate that general word co-occurrence statistics are insufficient for the task at hand. And while switching to contextualized word embeddings, such as *Flair*, leads to significant improvements, pre-trained transformers perform best.

The difference between the transformer-based models suggests that the pre-training task and attention mechanism of models impact the results notably. Unlike BERT, *ELECTRA* is pretrained on the replaced-token detection task, mimicking certain revision behaviors of human editors (e.g., replacing some input tokens with alternatives). Using ELECTRA boosts accuracy from 62.1 to 63.2 for non-finetuned models and from 63.1 to 63.8 for fine-tuned ones. *FT-DeBERTa* further improves the accuracy to 67.2, suggesting that also separately encoding content and position information, along with relative positional encodings, make the model more accurate on the given tasks. We point out that, apart from considering alternative pre-training strategies, other sentence embeddings and/or pooling strategies may further improve results.

Error Analysis Inspecting false predictions revealed that detecting claims in need of revisions concerning corroboration (i.e., *links*) is the most challenging (52% of such cases have been misclassified). This may be due to the fact that corroboration examples are underrepresented in the data (only 13% of the negative labeled samples). Accordingly, increasing the number of training samples could lead to improvement. In the appendix, we provide examples of false negative and false positive predictions. They demonstrate different cases where claims are missing necessary punctuation, clarification, and links to supporting evidence.

3.6.2 Representativeness and Reliability

Next, we explore the relationship between revision distance and data reliability by using the sampling strategy proposed in Section 3.3. We limit our experiments to revision histories with more than four revisions and compile four datasets, each representing a certain revision distance. We use the same data split as for the full corpus, resulting in 11,462 examples for training, 2554 for validation, and 2700 for testing for each of the sampled datasets.

Table 3.3 shows the accuracy scores obtained by *FT-ELECTRA*, when trained and tested on each sampled subset Di. Not only does the accuracy increase when training on a subset with a higher revision distance (results per column), but also the same model achieves higher accuracy when classifying more distant examples (results per row). On one hand, this indicates that, the closer the claim is to its optimal version, the more difficult it is to identify flaws. On the other hand, when considering claims of higher revision distance, the model seems capable of distinguishing optimal claims from other improved but suboptimal versions.

Training Subset	D1	D2	D3	D4	Average
D1	53.8	56.3	58.1	60.5	57.2
D2	55.1	58.4	60.1	63.6	59.4
D3	55.2	58.4	61.1	64.2	59.7
D4	**55.5**	**59.0**	**61.8**	**65.6**	**60.5**
Full training set	56.8	61.2	64.3	65.8	62.0

Table 3.3: Accuracy of FT-ELECTRA averaged over five runs, depending on sample training subset and test subset Di; i denotes the revision distance from 1 to 4.

Comparing the results of training on D4 with those for the full training set, we see that the D4 classifier is almost competitive, despite the much smaller amount of data. For example, the accuracy values on the D4 test set are 65.6 and 65.8, respectively. We conclude that the task at hand can be tackled even in lower-resource scenarios, if a representative sample of high quality can be obtained. This may be particularly important when considering languages other than English, where argument corpora of large scale may not be available.

3.6.3 Topical Bias

To measure the effects of topical bias, in a first experiment we compare the accuracy per topic category of *FT-DeBERTa* and *FT-ELECTRA* to detect whether identifying suboptimal claims is more difficult for certain topics. In Table 3.4, we report the accuracy for the 20 topic categories from Skitalinskaya et al. (2021). We find that the task is somewhat more challenging for debates related to topics, such as *justice*, *science*, and *democracy* (best accuracy 63.6–65.2) than for *europe* (69.1) or *education* (68.9). We analyzed the relationship between the size of the categories and the models' accuracy, but found no general correlation between the variables indicating that the performance difference does not stem from how represented each topic is in terms of sample size.[3]

In a second experiment, we evaluate how well the models generalize to unseen topics. To do so, we use a leave-one-category-out paradigm, ensuring that all claims from the same category are confined to a single split. We observe performance drops for both *FT-DeBERTa* and *FT-ELECTRA* in Table 3.4. The differences in the scores indicate that the models seem to learn topic-independent features that are applicable across the diverse set of categories, however, depending on the approach certain topics may pose more challenges than others, such as *religion* and *europe* for *FT-DeBERTa*.

[3] A scatter plot of size vs. accuracy is found in the appendix.

	FT-ELECTRA		FT-DeBERTa	
Category	**Full**	**Across**	**Full**	**Across**
Education	67.0	66.2	**68.9**	68.6
Technology	65.7	64.3	66.9	**67.0**
Philosophy	65.0	65.2	**67.3**	67.1
Europe	65.3	64.6	**69.1**	66.5
Economics	64.8	65.1	**68.0**	66.2
Government	65.2	64.7	**67.7**	66.8
Law	64.5	62.0	**67.7**	65.9
Ethics	64.7	64.4	**67.4**	66.1
Children	64.2	62.0	**67.2**	66.0
Society	64.5	63.6	**67.1**	66.2
Health	65.0	64.7	**68.7**	66.5
Religion	64.2	63.9	**67.5**	63.4
Gender	63.4	62.9	**66.8**	65.0
ClimateChange	63.2	62.8	**66.0**	63.8
Politics	62.6	62.2	**66.5**	64.7
USA	62.0	62.2	**65.4**	64.0
Science	61.9	61.0	**65.2**	62.8
Justice	60.2	58.6	**63.6**	61.2
Equality	62.9	61.2	**67.5**	65.5
Democracy	61.3	60.3	**65.2**	63.4

Table 3.4: Topical bias: Accuracy of FT-ELECTRA an FT-DeBERTa across 20 topic categories, when trained on the full dataset (*full*) and in a cross-category setting using a leave-one-out strategy (*across*).

Overall, the experiments indicate that the considered approaches generalize fairly well to unseen topics, however, further evaluations are necessary to assess whether the identified topical bias is due to the inherent difficulty of certain debate topics, or the lack of expertise of participants on the subject resulting in low quality revisions, requiring the collection of additional data annotations.

3.6.4 Contextuality

In our fourth experiment, we explore the benefits of incorporating contextual information. We restrict our view to the consideration of the main thesis and the parent claim, each representing context of different levels of broadness. We do so by concatenating the context and claim vector representations in SVM-based models, and by prepending the context separated by a delimiter token when fine-tuning transformer-based methods.

Table 3.5 reveals that, overall, adding context leads to improvements regardless of the method used. Across all approaches, including the narrow context of the parent claim seems more important for identifying suboptimal claims, with the best result obtained by *FT-DeBERTa* (accuracy of 68.6).

Model	Accuracy	Ma. F_1	P	R	F_1
Glove+SVM	54.9	54.9	54.9	50.0	52.1
+ thesis	55.9	55.8	55.6	53.1	54.3
+ parent	56.9	56.9	56.3	57.3	56.8
Flair+SVM	60.1	60.1	60.2	56.9	58.5
+ thesis	62.4	62.4	62.0	61.4	61.7
+ parent	62.8	62.8	61.9	**64.4**	63.1
BERT+SVM	62.1	61.8	63.5	54.7	58.8
+ thesis	63.5	63.4	64.2	59.0	61.5
+ parent	63.8	63.8	64.0	61.0	62.5
ELECTRA+SVM	63.2	62.9	65.1	55.0	59.6
+ thesis	65.0	64.9	66.1	60.0	62.9
+ parent	65.2	65.1	65.4	62.6	64.0
DeBERTa+SVM	61.5	61.2	63.2	52.9	57.6
+ thesis	62.5	62.2	63.9	55.1	59.2
+ parent	63.3	63.2	64.0	59.0	61.4
FT-BERT	63.1	61.7	70.1	44.2	54.2
+ thesis	64.1	63.0	70.1	47.6	56.7
+ parent	65.7	65.4	67.5	58.8	62.8
FT-ELECTRA	63.8	62.9	68.8	49.0	57.2
+ thesis	64.4	63.5	69.2	50.4	58.2
+ parent	64.8	64.6	66.0	59.3	62.4
FT-DeBERTa	67.1	66.6	71.3	55.9	62.6
+ thesis	67.3	67.0	70.1	59.5	64.2
+ parent	**68.6**	**68.4**	**71.4**	60.8	**65.7**
Random baseline	0.50	0.50	0.50	0.50	0.50

Table 3.5: Contextuality: Results of all evaluated models when including the *thesis* or *parent* as contextual information, averaged over five runs. For each approach, all gains from one row to another are significant at $p < .001$ according to a two-sided student's t-test.

The results also suggest that classification approaches employing non-finetuned transformer embeddings and contextual information can achieve results comparable to fine-tuned models, specifically models with a high similarity of the pretraining and target tasks (Peters et al., 2019). However, some quality aspects may require more general world knowledge and reasoning capabilities, which could be incorporated by using external knowledge bases. We leave this for future work.

Setup	Accuracy	Ma. F_1	F_1-Score		
			Clarif.	Typo	Links
FT-ELECTRA	56.0	49.0	62.4	52.4	34.5
+ parent	56.2	50.3	62.0	53.6	35.3
+ thesis	57.5	52.0	63.4	54.4	38.3
FT-DeBERTa	59.9	55.4	63.7	60.2	42.5
+ parent	60.3	56.0	63.6	61.2	43.0
+ thesis	**62.0**	**57.3**	**65.2**	**63.1**	**43.4**
Random baseline	33.4	31.4	38.5	33.4	45.3

Table 3.6: Claim improvement suggestion: Accuracy, macro F_1-score, and the F_1-score per revision type for ELECTRA+SVM and FT-DeBERTa with and without considering context, averaged over five runs.

3.6.5 Claim Improvement Suggestion

While previous experiments have shown the difficulty of distinguishing between claims in need of improvements and acceptable ones, the aim of this experiment is to provide benchmark models for predicting the type of improvement that a certain claim could benefit from. Here, we limit ourselves to the three most common types of revision: *clarification*, *typo and grammar correction* (includes style and formatting edits), and *adding/correcting links* to evidence in the form of external sources. Additional experiments covering an end-to-end setup by extending the classes to include claims that do not need revisions can be found in the appendix. We compare two best performing models from previous experiments, FT-ELECTRA and FT-DeBERTa, on a subset of 135,828 claims, where editors reported any of the three types.

Table 3.6 emphasizes the general benefit of utilizing contextual information for claim improvement suggestion. Though, depending on the specific revision type, the addition of contextual information can both raise and decrease performance. For example, despite the slightly improved overall accuracy of considering the parent claim as context, the F_1-score for *clarification* edits drops for all considered approaches (from 63.7 to 63.6 for FT-DeBERTa and from 62.4 to 62.0 for FT-ELECTRA). On the other hand, in the case of *links*, both types of contextual information lead to increased F_1-scores. Generally, we notice that opposed to the task of suboptimal-claim detection, providing the main thesis of the debate leads to higher score improvements overall. When comparing the approaches directly, we observe that *FT-DeBERTa* consistently outperforms *ELECTRA+SVM* in accuracy, achieving 62.0 when considering the main thesis.

Overall, our experiments indicate that to identify whether certain ap-

proaches to generating text representations are more suitable than others, one needs to consider the relationships between improvement type and context as well. In future work, we plan to focus on the problem of further defining and disentangling revision types to enable a deeper analysis of their relationships with contextual information.

Error Analysis Inspecting false predictions of the best performing model (FT-DeBERTa) revealed that the typo/grammar correction class seems to be confused frequently with both the clarification class and the links class (see the appendix for a confusion matrix). Our manual analysis suggests that editors frequently tackle more than one quality aspect of a claim in the same revision, for example, specifying a claim and fixing grammatical errors, or, removing typos from a link snippet. In the collected dataset, however, the revision type label in such cases would reflect only one class, such as *clarification* or *adding/correcting links*, respectively. These not fully accurate labels reduce the models ability to properly distinguish said classes. We provide examples of misclassifications obtained by FT-DeBERTa in the appendix, illustrating cases where both the true label and the predicted label represent plausible revision type suggestions.

3.7 Limitations

A limitation of our work is that we cannot directly apply our methods to the few existing revision-based corpora from other domains (Yang et al., 2017; Afrin and Litman, 2018; Anthonio et al., 2020) for multiple reasons: On the one hand, those corpora do not contain histories with more than one revision but only before-after sentence pairs). Some also consist of less than 1000 sentence pairs, rendering the quantitative experiments considered in this paper pointless. On the other hand, additional metadata useful for our analysis (e.g., revision types and contextual information) is either not available at all or only for a limited number of instances that is insufficient for training models.

Furthermore, the methods we evaluated utilize distantly supervised labels based on the assumption that each revision improves the quality of the claim and additional annotations provided by human editors. These annotations suffer from being coarse-grained, consisting of mainly three classes. However, each of the improvement types can be represented by several more fine-grained revision intentions. A point that we did not consider as part of this work is whether certain revisions can affect or inform future revisions within the same debate, for example, rephrasing of arguments to avoid repetition or ensuring

that all claims use the same wording for the main concepts. Often, such relationships are implicit and cannot be derived without additional information provided by the user performing the revision. We believe that collecting datasets and developing approaches, which enable distinguishing more fine-grained types of edits and implicit relationships, could not only enable deeper analysis and training more fine-grained improvement suggestion models, but also allow for better explanations to end users.

However, it should be noted that some of the considered methods rely on deep learning and have certain limitations when it comes to underrepresented classes, where the number of available training instances is very low. This is especially important when considering the task of claim improvement suggestion. We also point out in this regard that we only use the base versions of the BERT, ELECTRA, and DeBERTa models due to resource constraints. The results may vary, if larger models are used.

While common types of improvements likely differ across other domains and communities, we stress that our approaches are entirely data-driven, and are not tied to any specific quality definition. Therefore, we expect our data processing and filtering methods as well as the considered approaches to be applicable to other domains, where historical collaborative editing data similar to ours is available. When it comes to practice, several issues require further investigation, such as how to integrate recommendations in collaborative editing and educational environments, whether the recommended improvements will be accepted by users, and how they may impact the users' behavior. We leave these questions for future work.

3.8 Ethical Considerations

Online collaborative platforms face challenging ethical problems in maintaining content quality. On the one hand, they need to preserve a certain level of free speech to stimulate high quality discussions, while implementing regulations to identify editing behaviors defined as inappropriate. On the other hand, distinguishing such legitimate forms of regulation from illegitimate censorship, where particular opinions and individuals are suppressed, is a challenge of its own.

Our work is intended to support humans in different scenarios, including the creation or moderation of content on online debate platforms or in educational settings. In particular, the presented approaches are meant to help users by identifying argumentative claims in need of further improvements

and suggesting potential types of improvements, so they can deliver their messages effectively and honing their writing skills. However, the presented technology might also be subject to intentional misuse, such as the above-mentioned illegitimate censorship. While it is hard to prevent such misuse, we think that the described scenarios are fairly unlikely, as such changes tend to be noticed by the online community quickly. Moreover, the source of the used data (online debate platform Kialo) employs thorough content policies and user guidelines aimed at dealing with toxic behaviors, censorship, and discrimination. However, we suggest that follow-up studies stay alert for such behaviors and carefully choose training data.

3.9 Conclusion

Most approaches to argument quality assessment rate or compare argumentative texts that cover different aspects of a topic. While a few works studied which of two revisions of the same argumentative text is better, this does not suffice to decide whether a text actually needs revisions.

In this paper, we have presented two tasks to learn *when* and *how* to improve a given argumentative claim. We have delineated the main challenges of revision-based data, covering issues related to the representativeness and reliability of data, topical bias in revision behaviors, appropriate model complexities and architectures, and the need for context when judging claims. In experiments, we have compared several methods based on their ability to capture quality differences between different versions of the same text. Despite a number of limitations (discussed below), our results indicate that, in general, revision-based data can be employed effectively for the given tasks, contributing towards solutions for each of the considered challenges. Specifically, our suggested sampling strategy revealed that training on claim versions with a higher revision distance between them improves the performance when detecting claims in need of improvement. Moreover, we found that the impact of the available types of contextual information is not only task-dependent but also depends on the quality issue that a claim suffers from.

We argue that the developed approaches can help assist automated argument analysis and guide writers in improving their argumentative texts. With our work, we seek to encourage further research on improving writing support not only in debate communities but in educational settings as well.

Acknowledgments

We thank Andreas Breiter for his valuable feedback on early drafts, and the anonymous reviewers for their helpful comments. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 374666841, SFB 1342.

3.10 Implications for the Thesis

This chapter explores the first research question, namely, *What quality-related phenomena are typical of argument revisions on online debate platforms?* Specifically, we looked into what quality flaws and revision types are most common in online debates and revision based-corpora, as well as the challenges they pose for computationally modeling argument quality. The chapter has limited itself to considering two tasks: identifying claims in need of revisions and determining the types of quality issues that should be improved when revising the claim. Through a series of experiments, we have provided various solutions to the identified challenges in relation to each of the tasks separately. Hence, partially answering the second research as well: *How to approach the modeling of argument quality computationally to enable the analysis of arguments in need of improvement?* To do so, we exploit such information as the number of revisions performed in the past and the available future, the types of revision performed, contextual information, topic knowledge, etc. In systematic experiments, we show that the suggested quality assessments can be done even in low-resource scenarios, which can be especially appealing when dealing with languages and domains that have limited available resources. Overall, the chapter provides valuable insights into how to approach *absolute* quality assessments solely based on the claim's characteristics, without direct comparison to other claims.

In contrast, the following chapter will explore *relative* quality assessments, thus tackling the remaining sub-question of the second research question, namely, *How to model argument quality to enable comparison of several versions of the same argumentative text?*

4. Quality-based Ranking of Argumentative Text Revisions

This chapter presents the original content of the following paper (Skitalinskaya et al., 2023) by Gabriella Skitalinskaya, Jonas Klaff and Henning Wachsmuth: *"Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale"* in the Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics 2021 (EACL'21).

Abstract

Assessing the quality of arguments and of the claims the arguments are composed of has become a key task in computational argumentation. However, even if different claims share the same stance on the same topic, their assessment depends on the prior perception and weighting of the different aspects of the topic being discussed. This renders it difficult to learn topic-independent quality indicators. In this paper, we study claim quality assessment irrespective of discussed aspects by comparing different *revisions of the same claim*. We compile a large-scale corpus with over 377k claim revision pairs of various types from *kialo.com*, covering diverse topics from politics, ethics, entertainment, and others. We then propose two tasks: (a) assessing which claim of a revision pair is better, and (b) ranking all versions of a claim by quality. Our first experiments with embedding-based logistic regression and transformer-based neural networks show promising results, suggesting that learned indicators generalize well across topics. In a detailed error analysis, we give insights into what quality dimensions of claims can be assessed reliably. We provide the

data and scripts needed to reproduce all results.[1]

4.1 Introduction

Assessing argument quality is as important as it is questionable in nature. On the one hand, identifying the good and the bad claims and reasons for arguing on a given topic is key to convincingly support or attack a stance in debating technologies (Rinott et al., 2015), argument search (Ajjour et al., 2019b), and similar. On the other hand, argument quality can be considered on different granularity levels and from diverse perspectives, many of which are inherently subjective (Wachsmuth et al., 2017b); they depend on the prior beliefs and stance on a topic as well as on the personal weighting of different aspects of the topic (Kock, 2007).

Existing research largely ignores this limitation, by focusing on learning to predict argument quality based on subjective assessments of human annotators (see Section 4.2 for examples). In contrast, Habernal and Gurevych (2016) control for topic and stance to compare the convincingness of arguments. Wachsmuth et al. (2017c) abstract from an argument's text, assessing its relevance only structurally. Lukin et al. (2017) and El Baff et al. (2020) focus on personality-specific and ideology-specific quality perception, respectively, whereas Toledo et al. (2019) asked annotators to disregard their own stance in judging length-restricted arguments. However, none of these approaches controls for the concrete aspects of a topic that the arguments claim and reason about. This renders it difficult to learn what makes an argument and its building blocks good or bad in general.

In this paper, we study quality in argumentation irrespective of the discussed topics, aspects, and stances by assessing different revisions of the basic building blocks of arguments, i.e., claims. Such revisions are found in large quantities on online debate platforms such as *kialo.com*, where users post claims, other users suggest revisions to improve claim quality (in terms of clarity, grammaticality, grounding, etc.), and moderators approve or disapprove them. By comparing the quality of different revisions of the same instance, we argue that we can learn general quality characteristics of argumentative text and, to a wide extent, abstract from prior perceptions and weightings.

To address the proposed problem, we present a new large-scale corpus, consisting of 124k unique claims from kialo.com spanning a diverse range of topics related to politics, ethics, and several others (Section 4.3). Using distant

[1] Data and code: `https://github.com/GabriellaSky/claimrev`

Claim before Revision	Claim after Revision	Type
Dogs can help disabled people function better.	Dogs can help disabled people to navigate the world better.	Claim Clarification
African American soldiers joined unionists to fight for their freedom.	Black soldiers joined unionists to fight for their freedom.	Typo / Grammar Correction
Elections insure the independence of the judiciary.	Elections ensure the independence of the judiciary.	Typo / Grammar Correction
Israel has a track record of selling US arms to third countries without authorization.	Israel has a track record of selling US arms to third countries without authorization (*https://www.jstor.org/stable/1149008*).	Corrected / Added links

Table 4.1: Four examples of claims from Kialo before and after revision, along with the type of revision performed.

supervision, we derive a total number of 377k claim revision pairs from the platform, each reflecting a quality improvement, often, with a specified revision type. Four examples are shown in Table 4.1. To the best of our knowledge, this is the first corpus to target quality assessment based on claim revisions. In a manual annotation study, we provide support for our underlying hypothesis that a revision improves a claim in most cases, and we test how much the revision types correlate with known argument quality dimensions.

Given the corpus, we study two tasks: (a) how to compare revisions of a claim by quality and (b) how to rank a set of claim revisions. As initial approaches to the first task, we select in Section 4.4 a "traditional" logistic regression model based on word embeddings as well as transformer-based neural networks (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and SBERT (Reimers and Gurevych, 2019). For the ranking task, we consider the Bradley-Terry-Luce model (Bradley and Terry, 1952; Luce, 2012) and SVMRank (Joachims, 2006). They achieve promising results, indicating that the compiled corpus allows learning topic-independent characteristics associated with the quality of claims (Section 3.6). To understand what claim quality improvements can be assessed reliably, we then carry out a detailed error analysis for different revision types and numbers of revisions.

The main contributions of our work are: (1) A new corpus for topic-independent claim quality assessment, with distantly supervised quality improvement labels of claim revision pairs, (2) initial promising approaches to the tasks of claim quality classification and ranking, and (3) insights into what works well in claim quality assessment and what remains to be solved.

4.2 Related Work

In the recent years, there has been an increase of research on the quality of arguments and the claims and reasoning they are composed of. Wachsmuth et al. (2017b) describe argumentation quality as a multidimensional concept that can be considered from a logical, rhetorical, and dialectical perspectives. To achieve a common understanding, the authors suggest a unified framework with 15 quality dimensions, which together give a holistic quality evaluation at a certain abstraction level. They point out, that several dimensions may be perceived differently depending on the target audience. In recent follow-up work, Wachsmuth and Werner (2020) examined how well each dimension can be assessed only based on plain text only.

Most existing quality assessment approaches target a single dimension. On mixed-topic student essays, Persing and Ng (2013) learn to score the clarity of an argument's thesis, Persing and Ng (2015) do the same for argument strength, and Stab and Gurevych (2017b) classify whether an argument's premises sufficiently support its conclusion. All these are trained on pointwise quality annotations in the form of scores or binary judgments. Gretz et al. (2020) provide a corpus with crowdsourced quality annotations for 30,497 arguments, the largest to date for pointwise argument quality. The authors studied how their annotations correlate with the 15 dimensions from the framework of Wachsmuth et al. (2017b), finding that only *global relevance* and *effectiveness* are captured. Similarly, Lauscher et al. (2020) built a new corpus based on the framework to then exploit interactions between the dimensions in a neural approach. We present a small related annotation study for our dataset below. However, we follow Habernal and Gurevych (2016) in that we cast argument quality assessment as a relation classification problem, where the goal is to identify the better among a pair of instances.

In particular, Habernal and Gurevych (2016) created a dataset with argument convincingness pairs on 32 topics. To mitigate annotator bias, the arguments in a pair always have the same stance on the same topic. The more convincing argument is then predicted using a feature-rich SVM and a simple bidirectional LSTM. Other approaches to the same task map passage representations to real-valued scores using Gaussian Process Preference Learning (Simpson and Gurevych, 2018) or represent arguments by the sum of their token embeddings (Potash et al., 2017), later extended by a Feed Forward Neural Network (Potash et al., 2019). Recently, Gleize et al. (2019) employed a Siamese neural network to rank arguments by the convincingness of evidence.

In our experiments below, we take on some of these ideas, but also explore the impact of transformer-based methods such as BERT (Devlin et al., 2019), which have been shown to predict argument quality well (Gretz et al., 2020).

Potash et al. (2017) observed that longer arguments tend to be judged better in existing corpora, a phenomenon we will also check for below. Toledo et al. (2019) prevent such bias in their corpora for both pointwise and pairwise quality, by restricting the length of arguments to 8–36 words. The authors define quality as the level of preference for an argument over other arguments with the same stance, asking annotators to disregard their own stance. For a more objective assessment of argument relevance, Wachsmuth et al. (2017c) abstract from content, ranking arguments only based on structural relations, but they employ majority human assessments for evaluation. Lukin et al. (2017) take a different approach, including knowledge about the personality of the reader into the assessment, and El Baff et al. (2020) study the impact of argumentative texts on people depending on their political ideology.

As can be seen, several approaches aim to control for length, stance, audience, or similar. However, all of them still compare argumentative texts with different content and meaning in terms of the aspects of topics being discussed. In this work, we assess quality based on different revisions of the same text. In this setting, the quality is primarily focused on how a text is formulated, which will help to better understand what influences argument quality in general, irrespective of the topic. To be able to do so, we refer to online debate portals.

Debate portals give users the opportunity to discuss their views on a wide range of topics. Existing research has used the rich argumentative content and structure of different portals for argument mining, including *createdebate.com* (Habernal and Gurevych, 2015), *idebate.org* (Al-Khatib et al., 2016), and others. Also, large-scale debate portal datasets form the basis of applications such as argument search engines (Ajjour et al., 2019b). Unlike these works, we exploit debate portals for studying *quality*. Tan et al. (2016) predicted argument persuasiveness in the discussion forum *ChangeMyView* from ground-truth labels given by opinion posters, and Wei et al. (2016) used user upvotes and downvotes for the same purpose. Here, we resort to *kialo.com*, where users cannot only state argumentative claims and vote on the impact of claims submitted by others, but they can also help improve claims by suggesting revisions, which are approved or disapproved by moderators. While Durmus et al. (2019b) assessed quality based on the impact value of claims from kialo.com, we derive information on quality from the revision history of claims.

The only work we are aware of that analyzes revision quality of argumentative texts is the study of Afrin and Litman (2018). From the corpus of Zhang et al. (2017) containing 60 student essays with three draft versions each, 940 sentence writing revision pairs were annotated for whether the revision improves essay quality or not. The authors then trained a random forest classifier for automatic revision quality classification. In contrast, instead of sentences, we shift our focus to claims. Moreover, our dataset is orders of magnitude larger and includes notably longer revision chains, which enables deeper analyses and more reliable prediction of revision quality using data-intensive methods.

4.3 Data

Here, we present our corpus created based on claim revision histories collected from *kialo.com*.

4.3.1 A New Corpus based on Kialo

Kialo is a typical example of an online debate portal for collaborative argumentative discussions, where participants jointly develop complex pro/con debates on a variety of topics. The scope ranges from general topics (religion, fair trade, etc.) to very specific ones, for instance, on particular policy-making (e.g., whether wealthy countries should provide citizens with a universal basic income). Each debate consists of a set of claims and is associated with a list of related pre-defined generic categories, such as politics, ethics, education, and entertainment.

What differentiates Kialo from other portals is that it allows editing claims and tracking changes made in a discussion. All users can help improve existing claims by suggesting edits, which are then accepted or rejected by the moderator team of the debate. As every suggested change is discussed by the community, this collaborative process should lead to a continuous improvement of claim quality and a diverse set of claims for each topic.

As a result of the editing process, claims in a debate have a version history in the format of claim pairs, forming a chain where one claim is the successor of another and is considered to be of higher quality (examples found in Table 4.1). In addition, claim pairs may have a revision type label assigned to them via a non-mandatory free form text field, where moderators explain the reason of revision.

Base Corpus To compile the corpus, we scraped all 1628 debates found on Kialo until June 26th, 2020, related to over 1120 categories. They contain

Corpus	Type of Instances	Instances
$\text{ClaimRev}_{\text{BASE}}$	**Total claim pairs**	**210 222**
	Claim Clarification	63 729
	Typo/Grammar Correction	59 690
	Corrected/Added Links	17 882
	Changed Meaning of Claim	1 178
	Misc	10 464
	None	57 279
$\text{ClaimRev}_{\text{EXT}}$	**Total claim pairs**	**377 659**
	Revision distance 1	77 217
	Revision distance 2	27 819
	Revision distance 3	10 753
	Revision distance 4	4 460
	Revision distance 5	2 055
	Revision distance 6+	2 008
Both Corpora	**Claim revision chains**	**124 312**

Table 4.2: Statistics of the two provided corpus versions. $\text{ClaimRev}_{\text{BASE}}$: Number of claim pairs in total and of each revision type. $\text{ClaimRev}_{\text{EXT}}$: Number of claim pairs in total and of each revision distance. The bottom line shows the number of unique revision chains in the corpora.

124,312 unique claims along with their revision histories, which comprise of 210,222 pairwise relations. The average number of revisions per claim is 1.7 and the maximum length of a revision chain is 36. 74% of all pairs have a revision type. Overall, there are 8105 unique revision type labels in the corpus. 92% of labeled claim pairs refer to three types only: *Claim Clarification*, *Typo/Grammar Correction*, and *Corrected/Added Links*. An overview of the distribution of revision labels is given in Table 4.2. We refer to the resulting corpus as $\textit{ClaimRev}_{\textit{BASE}}$.

Data pre-processing included removing all claim pairs from debates carried out in languages other than English. Also, we considered claims with less than four characters as uninformative and left them out. As we seek to compare different versions of the *same* claim, claim version pairs with a general change of meaning do not satisfy this description. Thus, we removed such pairs from the corpus, too (inspecting the data revealed that such pairs were mostly generated due to debate restructuring). For this, we assessed the cosine similarity of a given claim pair using *spacy.io* and remove a pair if the score is lower than the threshold of 0.8.

Extended Corpus To increase the diversity of data available for training models, without actually collecting new data, we applied data augmentation.

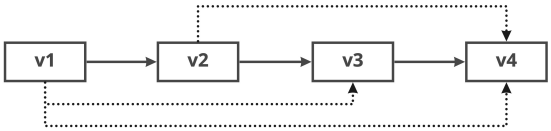


Figure 4.1: Visual representation of relations between revisions. Solid and dashed lines denote original and inferred non-consecutive relations respectively.

ClaimRev$_{\text{BASE}}$ consists of consecutive claim version pairs, i.e., if a claim v has four versions, it will be represented by three three pairs: (v_1, v_2), (v_2, v_3), and (v_3, v_4), where v_1 is the original claim and v_4 is the latest version. We extend this data by adding all pairs between non-consecutive versions that are inferrable transitively. Considering the previous example, this means we add (v_1, v_3), (v_1, v_4), and $(v2, v4)$. This is based on our hypothesis that every argument version is of higher quality than its predecessors, which we come back to below. Figure 4.1 illustrates the data augmentation. We call the augmented corpus *ClaimRev$_{EXT}$*.

For this corpus, we introduce the concept of *revision distance*, by which we mean the number of revisions between two versions. For example, the distance between v_1 and v_2 would be 1, whereas the distance between v_1 and v_3 would be 2. The distribution of the revision distances across ClaimRev$_{\text{EXT}}$ is summarized in Table 4.2.

The number of claim pairs of the 20 most frequent categories in both corpus versions are presented in Figure 4.2. We will restrict our view to the topics in these categories in our experiments.

4.3.2 Data Consistency on Kialo

While collaborative content creation enables leveraging the wisdom of large groups of individuals toward solving problems, it also poses challenges in terms of quality control, because it relies on varying perceptions of quality, backgrounds, expertise, and personal objectives of the moderators. To assess the consistency of the distantly-supervised corpus annotations, we carried out two annotation studies on samples of our corpus.

Consistency of Relative Quality In this study, we aimed to capture the general perception of claim quality on a meta-level, by deriving a data-driven quality assessment based on the revision histories. This was based on our hypothesis that every claim version is better than its predecessor. To test the

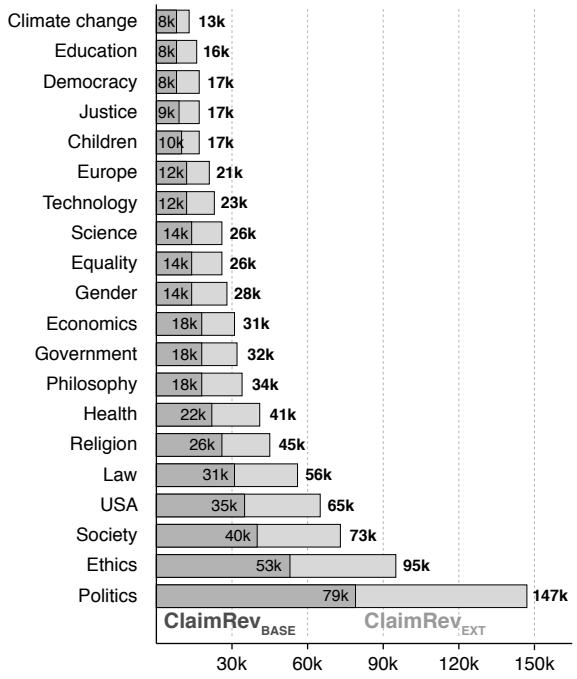


Figure 4.2: Number of claim revision pairs in each debate category of the two provided versions of our corpus (ClaimRev$_{\text{BASE}}$, ClaimRev$_{\text{EXT}}$).

validity of this hypothesis, two authors of this paper annotated whether a revision increases, decreases, or does not affect the overall claim quality. For this purpose, we randomly sampled 315 claim revision pairs, found in the supplementary material.

The results clearly support our hypothesis, showing an increase in quality in 292 (93%) of the annotated cases at a Cohen's κ agreement of 0.75, while 8 (3%) of the revisions had no effect on quality and only 6 (2%) led to a decrease. On the remaining 2%, the annotators did not reach an agreement.

Consistency of Revision Type Labels Our second annotation study focused on the reliability of the revision type labels. We restricted our view to the top three revision labels, which cover 96% of all revisions. We randomly sampled 140–150 claim pairs per each revision type, 440 in total. For each claim pair, the same annotators as above provided a label for the revision type from the following set: *Claim Clarification*, *Typo/Grammar Correction*, *Corrected/Added Links*, and *Other*.

Comparing the results to the original labels in the corpus revealed that the annotators strongly agreed with the labels, namely, with Cohen's κ of 0.82 and 0.76 respectively. The level of agreement between the annotators was

	Clarification	Grammar	Links
Cogency	**-0.31**	**-0.31**	**0.65**
Local Acceptability	**0.38**	-0.20	-0.19
Local Relevance	**0.44**	-0.25	-0.22
Local Sufficiency	-0.28	**-0.33**	**0.62**
Effectiveness	0.02	**-0.35**	**0.34**
Credibility	0.06	-0.16	0.10
Emotional Appeal	0.00	0.00	0.00
Clarity	-0.16	**0.35**	-0.18
Appropriateness	0.01	0.02	-0.04
Arrangement	0.00	0.00	0.00
Reasonableness	0.07	-0.04	-0.04
Global Acceptability	**0.37**	**0.42**	**-0.82**
Global Relevance	0.02	**-0.43**	**0.42**
Global Sufficiency	0.00	0.00	0.00
Overall	-0.05	0.00	0.05
Pairs with revision type	120	100	95

Table 4.3: Pearson's r correlation in our annotation study between increases in the 15 quality dimensions of Wachsmuth et al. (2017b) and the main revision types: Claim *Clarification*, Typo/*Grammar* Correction, Corrected/Added *Links*. Moderate and high correlations are shown in bold ($r \geq 0.3$).

even higher ($\kappa = 0.84$). In further analysis, we observed that most confusion happened between the revision types *Typo/Grammar correction* and *Claim Clarification*. This may be due to the non-strict nature of the revision type labels, which leaves space for different interpretations on a case-to-case basis. Still, we conclude that the revision type labels seem reliable in general.

4.3.3 Quality Dimensions on Kialo

To explore the relationship between the revision types on Kialo and argument quality in general, we conducted a third annotation study. In particular, for each of the 315 claim pairs from Section 4.3.2, one of the authors of this paper provided a label indicating whether the revision improved for each of the 15 quality dimensions defined by Wachsmuth et al. (2017b) or not. It should be noted that the annotators reached an agreement on the revision type for all these pairs.

Table 4.3 shows Pearson's r rank correlation for each quality dimension for the three main revision types. We observe a strong correlation between the revision type *Corrected/Added Links* and the logical quality dimensions *Cogency* (0.65) and *Local Sufficiency* (0.62), which matches the main purpose of such revisions: to add supporting information to a claim. The high negative

correlation of this revision type with *Global Acceptability* (-0.82) indicates that improvements regarding the dimension in question are more prominent in other types. Complementarily, *Claim Clarification* mainly improves the other logical dimensions (*Local Acceptability* 0.38, *Local Relevance* 0.44), matching the intuition that a clarification helps to ensure a correct understanding of the meaning. *Typo/Grammar corrections*, finally, rather seem to support an acceptable linguistic shape, improving *Clarity* (0.35) and *Global Acceptability* (0.42).

Finding only low correlations for many rhetorical dimensions (credibility, emotional appeal, etc.) as well as for overall quality, we conclude that the revisions on Kialo seem to target primarily the general form a well-phrased claim should have.

4.4 Approaches

To study the two proposed tasks, claim quality classification and claim quality ranking, on the given corpus, we consider the following approaches.

4.4.1 Claim Quality Classification

We cast this task as a pairwise classification task, where the objective is to compare two versions of the same claim and determine which one is better. To solve this task, we compare four methods:

Length To check whether there is a bias towards longer claims in the data, we use a trivial method which assumes that claims with more characters are better.

S-BOW As a "traditional" method, we employ the siamese bag-of-words embedding (S-BOW) as described by Potash et al. (2017). We concatenate two bag-of-words matrices, each representing a claim version from a pair, and input the concatenated matrix to a logistic regression. We also test whether information on length improves S-BOW.

BERT We select the BERT model, as it has become the standard neural baseline. BERT is a pre-trained deep bidirectional transformer language model (Devlin et al., 2019). For our experiments we use the pre-trained version *bert-base-cased*, as implemented in the *huggingface* library.[2] We fine-tune the model for two epochs using the Adam optimizer with learning rate 1e-5. [3]

[2]Huggingface library, `https://huggingface.co/transformers/pretrained_models.html`

[3]We chose the number of epochs empirically, picking the best learning rate out of {5e-7, 5e-6,1e-5,2e-5,3e-5}.

	v_1	v_2	v_3
v_1	0	0.018	0.002
v_2	0.982	0	0.428
v_3	0.998	0.572	0

Table 4.4: Example of a pairwise score matrix for ranking of three claim revisions, v_1–v_3, given the following pairwise scores: $(v_1, v_2) = (0.018, 0.982)$, $(v_2, v_3) = (0.428, 0.572)$, and $(v_1, v_3) = (0.002, 0.998)$.

SBERT We also use Sentence-BERT (SBERT) to learn to represent each claim version as a sentence embedding (Reimers and Gurevych, 2019), opposed to the token-level embeddings of standard BERT models. We fine-tune SBERT based on *bert-base-cased* using a siamese network structure, as implemented in the *sentence-transformers* library.[4] We set the numbers of epochs to one which is recommended by the authors (Reimers and Gurevych, 2019), and we use a batch-size of 16, Adam optimizer with learning rate 1e-5, and a linear learning rate warm-up over 10% of the training data. Our default pooling strategy is MEAN.

4.4.2 Claim Quality Ranking

In contrast to the previous task, we cast this problem as a sequence-pair regression task. After obtaining all pairwise scores using S-BOW, BERT, and SBERT respectively, we map the pairwise labels to real-valued scores and rank them using the following models, once for each method.

BTL For mapping, we use the well-established Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 2012), in which items are ranked according to the probability that a given item beats an item chosen randomly. We feed the BTL model a pairwise-comparison matrix for all revisions related to a claim, generated as follows: Each row represents the probability of the revision being better than other revisions. All diagonal values are set to zero. Table 4.4 illustrates an example for a set of three argument revisions.

SVMRank Additionally, we employ SVMRank (Joachims, 2006), which views the ranking problem as a pairwise classification task. First, we change the input data, provided as a ranked list, into a set of ordered pairs, where the (binary) class label for every pair is the order in which the elements of the pair should be ranked. Then, SVMRank learns by minimizing the error of the order relation when comparing all possible combinations of candidate

[4] Sentence-transformers library, `https://www.sbert.net/`

pairs. Given the nature of the algorithm we cannot work with token embeddings obtained from BERT directly. Thus, we utilize one of most commonly used approaches to transform token embeddings to a sentence embedding: extracting the special [CLS] token vector (Reimers and Gurevych, 2019; May et al., 2019). In our experiments we select a linear kernel for the SVM and use PySVMRank,[5] a python API to the SVM^{rank} library written in C.[6]

4.5 Experiments and Discussion

We now present empirical experiments with the approaches from Section 4.4. The goal is to evaluate how hard it is to compare and rank the claim revisions in our corpus from Section 4.3 by quality.

4.5.1 Experimental Setup

We carry out experiments in two settings. The first considers creating *random splits* over revision histories, ensuring that all versions of the same claim are in a single split in order to avoid data leakage. We assign 80% of the revision histories to the training set and the remaining 20% to the test set. A drawback of this setup is that it is not clear how well models generalize to unseen debate categories. In the second setting, we therefore evaluate the methods also in a *cross-category* setup using a leave-one-category-out paradigm, which ensures that all claims from the same debate category are confined to a single split. We split the data in this way to evaluate if our models learn independent features that are applicable across the diverse set of categories. To assess the effect of adding augmented data, we evaluate all models on both ClaimRev$_{\text{BASE}}$ and ClaimRev$_{\text{EXT}}$.

For quality *classification*, we report accuracy and the Matthews correlation coefficient (Matthews, 1975). We report the mean results over five runs in the random setting and the mean results across all test categories in the cross-category setting. To ensure balanced class labels, we create one false claim pair for each true claim pair by shuffling the order of the claims: $(v_1, v_2, true) \rightarrow (v_2, v_1, false)$, where the label denotes whether the second claim in the pair is of higher quality. We report results obtained by models trained on ClaimRev$_{\text{BASE}}$ and ClaimRev$_{\text{EXT}}$ as score pairs in Table 4.5.

To measure *ranking* performance, we calculate Pearson's r and Spearman's ρ correlation, as well as NDCG and MRR. We also compute the Top-1 accuracy, i.e. the proportion of claim sets, where the latest version has been ranked best.

[5] PySVMRank, `https://github.com/ds4dm/PySVMRank`

[6] SVM^{rank}, `www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

We average the results on each claim set across the test set for each metric. Afterwards we average the results across five runs or across all categories, depending on the chosen setting.

4.5.2 Claim Quality Classification

The results in Table 4.5 show that a claim's *length* is a weak indicator of quality (up to 61.3 accuracy). An intuitive explanation is that, even though claims with more information may be better, it is also important to keep them readable and concise.

Despite *SBOW*'s good performance on predicting convincingness (Potash et al., 2017), the claim quality in our corpus cannot be captured by a model of such simplicity (maximum accuracy of 65.4). We point out that adding other linguistic features (for example, part-of-speech tags or sentiment scores) may further improve SBOW. Exemplarily, we equip SBOW with length features and observe a significant improvement (up to 67.5).

As for the transformer-based methods, we see that *BERT* and *SBERT* consistently outperform SBOW in all settings on both corpus versions, with SBERT's accuracy of up to 77.7 being best.[7]

A comparison of the performance of the methods depending on the corpus used for training in Table 4.5 shows the effect of augmenting the original Kialo data. In most cases, the results obtained by models trained on $\text{ClaimRev}_{\text{EXT}}$ are comparable (slightly higher/lower) than results obtained by models trained on $\text{ClaimRev}_{\text{BASE}}$. This means that adding relations between non-consecutive claim versions does not improve the reliability of methods. Given that the performance scores obtained on the $\text{ClaimRev}_{\text{EXT}}$ test set are evidently higher than on the $\text{ClaimRev}_{\text{BASE}}$ test set, we can conclude that the augmented cases are easier to classify and the cumulative difference in quality is more evident.

We can also see in Table 4.5 that the trained models are able to generalize across categories; the accuracy and MCC scores in the random split and cross-category settings for each method are very similar, with only a slight drop in the cross-category setting. This indicates that the nature of the revisions is relatively consistent among all categories, yet reveals the existence of some category-dependent features.

To find out whether BERT really captures the relative revision quality and not only lexical features present in the original claim, we introduced a *Single*

[7] Additionally, we have experimented with an adversarial training algorithm, ELECTRA (Clark et al., 2020), and obtained results slightly better than BERT, yet inferior to SBERT. We omit to report these results here, since they did not provide any further notable insights.

	Test set: ClaimRev$_{BASE}$				Test set: ClaimRev$_{EXT}$			
	Random-Split		Cross-Category		Random-Split		Cross-Category	
Model	**Accuracy**	**MCC**	**Accuracy**	**MCC**	**Accuracy**	**MCC**	**Accuracy**	**MCC**
Length	61.3 / 61.3	0.23 / 0.23	60.7 / 60.7	0.21 / 0.21	60.8 / 60.8	0.22 / 0.22	60.0 / 60.0	0.20 / 0.20
SBOW	62.0 / 62.6	0.24 / 0.25	61.4 / 61.4	0.23 / 0.23	64.9 / 65.4	0.30 / 0.31	63.9 / 64.1	0.28 / 0.28
SBOW + Length	65.1 / 65.5	0.30 / 0.31	64.8 / 64.4	0.29 / 0.29	67.1 / 67.5	0.34 / 0.35	66.1 / 66.2	0.32 / 0.32
BERT	75.5 / 75.2	0.51 / 0.51	75.1 / 74.1	**0.51** / 0.49	76.4 / 76.5	0.53 / 0.53	76.2 / 75.4	0.53 / 0.51
SBERT	**76.2** / **76.2**	**0.53** / **0.52**	**75.5** / **75.4**	**0.51** / **0.51**	**77.4** / **77.7**	**0.55** / **0.55**	**76.8** / **76.8**	**0.54** / **0.54**
Random baseline	50.0 / 50.0	0.00 / 0.00	50.0 / 50.0	0.00 / 0.00	50.0 / 50.0	0.00 / 0.00	50.0 / 50.0	0.00 / 0.00
Single claim baseline	57.7 / 58.1	0.17 / 0.17	57.7 / 57.3	0.17 / 0.16	58.8 / 59.8	0.20 / 0.20	58.9 / 58.9	0.20 / 0.20

Table 4.5: Claim quality classification results: Accuracy and Matthew Correlation Coefficient (MCC) for all tested approaches in the random-split and the cross-category setting on the two corpus versions. The first value in each value pair is obtained by a model trained on ClaimRev$_{BASE}$, the second by a model trained on ClaimRev$_{EXT}$. All improvements from one row to the next are significant at $p < 0.001$ according to a two-sided Student's t-test.

claim baseline, analogous to the *hypothesis-only* baseline in natural language inference(Poliak et al., 2018). It can be seen that the accuracy and MCC scores are low across all settings (maximum accuracy of 59.8), which indicates that BERT indeed captures relative revision quality mostly.

4.5.3 Claim Quality Ranking

Table 4.6 lists the results of our ranking experiments, which show patterns similar to the results achieved in the classification task.

We can observe similar patterns in both of the selected ranking approaches: SBERT consistently outperforms all other considered approaches across all settings (up to 0.73 and 0.72 in Pearson's r and Spearman's ρ accordingly). BERT and SBERT outperform SBOW, indicating that transformer-based methods are more capable of capturing the relative quality of revisions. While BTL + BERT obtains results comparable to BTL + SBERT, we find that using the CLS-vector as a sentence embedding representation leads to lower results. We point out, though, that using other sentence embeddings and/or pooling strategies (for example, averaged BERT embeddings) may further improve results.

Similar to the results of the classification task, we observe only a slight performance drop in the cross-category setting when using BTL for ranking, yet an increase when using SVMRank, again emphasizing the topic-independent nature of claim quality in our corpus.

4.5.4 Error Analysis

To further explore the capabilities and limitations of the best model, SBERT, we analyzed its performance on each revision type and distance.

As the upper part of Table 4.7 shows, SBERT is highly capable of assessing revisions related to the correction and addition of links and supporting information. This revision type also obtained the highest correlations between quality dimensions and type of revision (see Table 4.3), which indicates that the patterns of changes performed within this type are more consistent. In contrast, we observe that the model fails to address revisions related to the changed meaning of a claim. On the one hand, this may be due to the fact that such examples are underrepresented in the data. On the other hand, the consideration of such examples in the selected tasks is questionable, since changing the meaning of claim is usually considered as the creation of a *new claim* and not a *new version* of a claim.

An insight from the lower part of Table 4.7 is that the accuracy of predictions increases from revision distance 1 to 4. We obtain better results when

	Random-Split					Cross-Category				
Model	r	ρ	**Top-1**	NDCG	MRR	r	ρ	**Top-1**	NDCG	MRR
BTL + SBOW+L	0.38	0.37	0.62	0.94	0.79	0.36	0.35	0.60	0.94	0.78
BTL + BERT	0.60	0.59	0.74	0.96	0.86	0.58	0.57	0.72	0.96	0.85
BTL + SBERT	0.63	0.62	0.77	**0.97**	0.87	0.62	0.61	0.75	0.97	0.86
SVMRank + SBOW+L	0.18	0.18	0.50	0.93	0.73	0.24	0.23	0.52	0.93	0.75
SVMRank + BERT CLS	0.50	0.49	0.67	0.95	0.84	0.51	0.51	0.67	0.96	0.84
SVMRank + SBERT	**0.70**	**0.70**	**0.79**	**0.97**	**0.90**	**0.73**	**0.72**	**0.80**	**0.98**	**0.91**
Random baseline	0.00	0.00	0.42	0.91	0.68	0.00	0.00	0.42	0.91	0.67

Table 4.6: Claim quality ranking results: Pearson's r and Spearman's ρ correlation as well as top-1 accuracy for all tested approaches in the random-split and the cross-category setting on $\text{ClaimRev}_{\text{EXT}}$.In all cases, SVMRank + SBERT is significantly better than all others at $p < 0.001$ according to a two-sided Student's t-test.

Task	Label	Accuracy	Instances
Type	Claim Clarification	69.7	12 856
	Typo/Grammar Correction	83.6	12 125
	Corrected/Added Links	89.3	3 660
	Changed Meaning of Claim	57.3	232
	Misc	67.2	2 130
	None	78.3	45 842
Distance	Revision distance 1	76.2	42 341
	Revision distance 2	79.6	17 478
	Revision distance 3	80.6	8 023
	Revision distance 4	81.0	3 979
	Revision distance 5	79.5	2 103
	Revision distance 6+	74.9	2 921
	All	**77.7**	**76 845**

Table 4.7: Accuracy of the best model, SBERT, on each single revision type and distance in $\text{ClaimRev}_{\text{EXT}}$, along with the number of instances per each case.

comparing non-consecutive claims than when comparing claim pairs with distance of 1. An intuitive explanation is that, since each single revision should ideally improve the quality of a claim, the more revisions a claim undergoes, the more evident the quality improvement should be. For distances > 5, the accuracy starts to decrease again, but this may be due to the limited number of cases given.

4.6 Conclusion and Future Work

In this paper, we have proposed a new way of assessing quality in argumentation by considering different revisions of the same claim. This allows us to focus on characteristics of quality regardless of the discussed topics, aspects, and stances in argumentation. We provide a new corpus of web claims, which is the first large-scale corpus to target quality assessment and revision processes on a claim level. We have carried out initial experiments on this corpus using traditional and transformer-based models, yielding promising results but also pointing to limitations. In a detailed analysis we have studied different kinds of claim revisions and provided insights into the aspects of a claim that influence the users' perception of quality. Such insights could help improve writing support in educational settings, or identify the best claims for debating technologies and argument search.

We seek to encourage further research on how to help online debate platforms automate the process of quality control and design automatic quality assessment systems. Such systems can be used to indicate if the suggested

revisions increase the quality of an argument or recommend the type of revision needed. We leave it for future work to investigate whether the learned concepts of quality are transferable to content from other collaborative online platforms (such as idebate.org or Wikipedia), or to data from other domains, such as student essays and forum discussions.

Acknowledgments

We thank Andreas Breiter for feedback on early drafts, and the anonymous reviewers for their helpful comments. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 374666841, SFB 1342.

4.7 Implications for the Thesis

The findings presented in this chapter answer the final sub-question of the second research question: *How to model argument quality to enable comparison of several versions of the same argumentative text?* Such an assessment allows for immediate quality control of the revision process, enabling users to get feedback on the effectiveness of their revisions and identify further areas of improvement. Within the chapter, we demonstrated different approaches to modeling argument quality that enable such comparison and ranking of text variations, taking into account several types of revision and their order in the revision history.

Considering the overall thesis, the chapter provides valuable insights into how revisions affect the quality of claims across various argument quality dimensions. Specifically, in Section 4.3, we analyzed various quality aspects, such as clarity, relevance, and coherence, to identify which revision types trigger improvements across each dimension and have the most impact on the overall persuasiveness of the argumentative text. These findings not only provide a deeper understanding of how revisions impact argument quality but also offer practical guidance for writers.

Using these insights, the next chapter will explore the third research question, which focuses on *How to automatically generate improved versions of argumentative texts using computational methods?*

5. Generation of Optimized Argumentative Texts

This chapter presents the original content of the following paper (Skitalinskaya et al., 2023) by Gabriella Skitalinskaya, Maximilian Spliethöver and Henning Wachsmuth: *"Claim Optimization in Computational Argumentation"* in the Proceedings of the 16th International Natural Language Generation Conference 2023 (INLG'23).

Abstract

An optimal delivery of arguments is key to persuasion in any debate, both for humans and for AI systems. This requires the use of clear and fluent claims relevant to the given debate. Prior work has studied the automatic assessment of argument quality extensively. Yet, no approach actually improves the quality so far. To fill this gap, this paper proposes the task of *claim optimization*: to rewrite argumentative claims in order to optimize their delivery. As multiple types of optimization are possible, we approach this task by first generating a diverse set of candidate claims using a large language model, such as BART, taking into account contextual information. Then, the best candidate is selected using various quality metrics. In automatic and human evaluation on an English-language corpus, our quality-based candidate selection outperforms several baselines, improving 60% of all claims (worsening 16% only). Follow-up analyses reveal that, beyond copy editing, our approach often specifies claims with details, whereas it adds less evidence than humans do. Moreover, its capabilities generalize well to other domains, such as instructional texts.

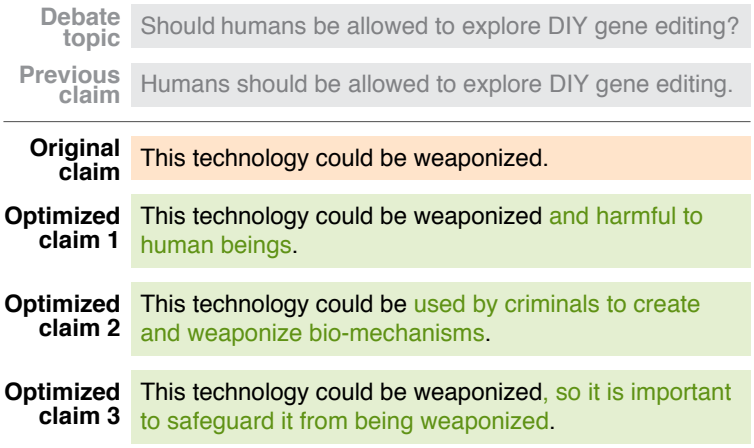


Figure 5.1: Examples of different optimized versions of an *original claim* found on the debate platform Kialo. All optimizations were generated by the approach proposed in this paper, using the *debate topic* as context.

5.1 Introduction

The delivery of arguments in clear and appropriate language is a decisive factor in achieving persuasion in any debating situation, known as *elocutio* in Aristotle's rhetoric (El Baff et al., 2019). Accordingly, the claims composed in an argument should not only be grammatically fluent and relevant to the given debate topic, but also unambiguous, self-contained, and more. Written arguments therefore often undergo multiple revisions in which various aspects are optimized (Zhang and Litman, 2015).

Extensive research has been done on the automatic assessment of argument quality and the use of large language models on various text editing tasks. Yet, no work so far has studied how to actually improve argumentative texts. However, developing respective approaches is a critical step towards building effective writing assistants, which could help learners write better argumentative texts (Wambsganss et al., 2021) or rephrase arguments made by an AI debater (Slonim et al., 2021). In this work, we close the outlined gap by studying how to employ language models for rewriting argumentative text to optimize its delivery.

We start by defining the task of *claim optimization* in Section 5.3, and adjust the English-language claim revision dataset of Skitalinskaya et al. (2021) for evaluation. The new task requires complementary abilities: On the one hand, different types of quality issues inside a claim must be detected, from grammatical errors to missing details. If not all quality aspects can be improved

simultaneously, specific ones must be targeted. On the other hand, improved claim parts need to be integrated with the context of the surrounding discussion, while preserving the original meaning as far as possible. Figure 5.1 shows three exemplary optimizations of a claim from the debate platform *Kialo*. The first elaborates what the consequence of weaponization is, whereas the second rephrases the claim to clarify what weaponizing means, employing knowledge about the debate topic. The third renders the stance of the claim explicit. We observe that different ways to optimize a claim exist, yet the level of improvement differs as well.

To account for the multiplicity of claim optimization, we propose a controlled generation approach that combines the capabilities of large language models with quality assessment (Section 5.4). First, a fine-tuned generation model produces several candidate optimizations of a given claim. To optimize claims, we condition the model on discourse context, namely the debate topic and the previous claim in the debate. The key to selecting the best optimization is to then score candidates using three quality metrics: *grammatical fluency*, *meaning preservation*, and *argument quality*. Such candidate selection remains understudied in many generative tasks, particularly within computational argumentation.

In automatic and manual evaluation (Section 5.5), we demonstrate the effectiveness of our approach, employing fine-tuned BART (Lewis et al., 2020) for candidate generation. Our results stress the benefits of quality assessment (Section 5.6). Incorporating context turns out especially helpful for making shorter claims—where the topic of the debate is difficult to infer—more self-contained. According to human annotators, our approach improves 60% of all claims and harms only 16%, clearly outperforming standard fine-tuned generation.

To gain further insights, we carry out a manual annotation of 600 claim optimizations and identify eight types typically found in online debate communities, such as *elaboration* and *disambiguation* (Section 5.7). Intriguingly, our approach covers similar optimization types as in human revisions, but we also observe limitations (Section 5.7). To explore to what extent it generalizes to other revision domains, we also carry out experiments on instructional texts (Anthonio and Roth, 2020) and formal texts (Du et al., 2022), finding that it outperforms strong baselines and state-of-the-art approaches.

In summary, the contributions of this paper are:

1. *a new task*, claim optimization, along with a manual analysis of typical

optimization types;

2. *a computational approach* that selects the best generated candidate claim in terms of quality;
3. *empirical insights* into the impact and challenges of optimizing claims computationally.[1]

5.2 Related Work

Quality assessment has become a key topic in computational argumentation research (Lapesa et al., 2023). Various quality dimensions exist in argumentation theory, as surveyed by Wachsmuth et al. (2017b) and assessed computationally in various works (Lauscher et al., 2020; Marro et al., 2022). Many of them relate to quality aspects we consider in this work, from clarity and organization (Wachsmuth et al., 2016) to the general evaluability of arguments (Park and Cardie, 2018), potential fallacies in their reasoning (Goffredo et al., 2022), and the appropriateness of the language used (Ziegenbein et al., 2023). Recently, (Skitalinskaya and Wachsmuth, 2023) tackled the question whether an argumentative claim is in need of revision, whereas Jundi et al. (2023) investigated where to best elaborate a discussion. While Gurcke et al. (2021) leverage claim generation for a refined assessment of argument quality, we are not aware of any prior work that actually optimizes arguments or their components in order to improve quality.

As shown in Figure 5.1, there can be several ways to optimize a given text. Our key idea is to select the best optimization among diverse candidates generated by a language model. Prior generation work on candidate selection hints at the potential benefits of such setup, albeit in other tasks and domains. In early work on rule-based conversational systems, Walker et al. (2001) introduced dialogue quality metrics to optimize template-based systems towards user satisfaction. Kondadadi et al. (2013) and Cao et al. (2018) chose the best templates for generation, and Mizumoto and Matsumoto (2016) used syntactic features to rank candidates in grammar correction. Recently, Yoshimura et al. (2020) proposed a reference-less metric trained on manual evaluations of grammar correction system outputs to assess generated candidates, while Suzgun et al. (2022) utilize pre-trained language models to select the best candidate in textual style transfer tasks.

In generation research on computational argumentation, candidate selec-

[1]Data, code, and models from our experiments are found at https://github.com/GabriellaSky/claim_optimization

tion remains largely understudied. Most relevant in this regard is the approach of Chakrabarty et al. (2021) which reframes arguments to be more trustworthy (e.g., less partisan). It generates multiple candidates and selects one based on the entailment relation scores to the input. Extending this idea, we select candidates based on various properties, including argument quality.

Understanding the editing process of arguments is crucial, as it reveals what quality dimensions are considered important. For Wikipedia, Daxenberger and Gurevych (2013) proposed a fine-grained taxonomy as a result of their multi-label edit categorization of revisions (Daxenberger and Gurevych, 2012). The taxonomy focuses solely on the editing actions performed, such as inserting, deleting, and paraphrasing. In contrast, Yang et al. (2017) identified various semantic intentions behind Wikipedia revisions, from *copy editing* to *content clarifications* and *fact updates*. Their taxonomy defines a starting point for our research. Not all covered intentions generalize beyond Wiki scenarios, though.

Wikipedia-based corpora have often been used in the study of editing and rewriting, including paraphrasing (Max and Wisniewski, 2010), grammar correction (Lichtarge et al., 2019), bias neutralization (Pryzant et al., 2020), and controllable text editing (Faltings et al., 2021; Du et al., 2022). Similarly, WikiHow enabled summarization (Koupaee and Wang, 2018) and knowledge acquisition (Zhou et al., 2019). However, neither of these includes *argumentative* texts. Instead, we thus rely on the corpus of Skitalinskaya et al. (2021), which consists of revision histories of argumentative claims from online debates. Whereas the authors *compare* claims in terms of quality, we propose and study the new task of automatically *optimizing* claim quality. Moreoever, we see the revision types they distinguish (clarification, grammar correction, linking to external sources) as too coarse-grained to represent the diversity of claim optimizations. We refine them manually into eight optimization types, allowing for a more systematic analysis. Skitalinskaya et al. (2021) also found low correlations between the revision types and 15 common argument quality dimensions (Wachsmuth et al., 2017b), suggesting that they are rather complementary. Primarily, they target the general form a well-phrased claim should have and its relevance to the debate.

For the analysis of argumentative text rewriting, Zhang and Litman (2015) incorporated both argumentative writing features and surface changes. To explore the classification of essay revisions, they defined a two-dimensional schema, combining the revision operation (e.g., modify, add, or delete) with the component being revised (e.g., reasoning or evidence). Moreover, Afrin

and Litman (2018) created a small corpus of between-draft revisions of 60 student essays to study whether revision improves quality. However, these works do not uncover the reasoning behind a revision operation and are more geared towards analysis at the essay level.

5.3 Task and Data

This section introduces the proposed task and pre-sents the data used for development and evaluation.

5.3.1 Claim Optimization

We define the claim optimization task as follows:

Task Given as input an argumentative claim c, potentially along with context information on the debate, rewrite c into an output claim $\tilde{c}$ such that

(a) $\tilde{c}$ improves upon c in terms of text quality and/or argument quality, and

(b) $\tilde{c}$ preserves the meaning of c as far as possible.

While we conceptually assume that c consists of one or more sentences and has at least one quality flaw, our approaches do not model this explicitly. Moreover, note that c might have multiple flaws, resulting in $n \geq 2$ candidate optimizations $\tilde{C} = \{\tilde{c}_1, \ldots, \tilde{c}_n\}$. In this case, the goal is to identify the candidate $c^* \in \tilde{C}$ that maximizes overall quality.

5.3.2 Data for Development and Evaluation

We start from the ClaimRev dataset (Skitalinskaya et al., 2021), consisting of 124,312 claim revision histories from the debate platform *Kialo*. Each history defines a chain $(c_1, ..., c_m)$, in which claim c_i is a revised version of the previous claim, c_{i-1} with $1 < i \leq m$, improving upon its quality. According to the authors, this holds in 93% of all cases.

From each revision chain, we derived all possible optimization pairs $(c, \tilde{c}) := (c_{i-1}, c_i)$, in total 210,222. Most revisions are labeled with their intention by the users who performed them, rendering them suitable for learning to optimize claims automatically.[2] Overall, 95% of all pairs refer to three intention labels: *clarification*, *typo/grammar correction*, and *corrected/added links*. To avoid noise from the few remaining labels, we condensed the data to 198,089 instances of the three main labels.[3]

[2] As 26% of all pairs were unlabeled, we trained a BERT model to assign such pairs one of the 6 most prominent labels.

[3] The labels of the removed instances denote changes to the meaning of c and statements from which no action or intention can be derived (e.g., "see comments", "moved as pro").

For the final task dataset, we associated each remaining pair $(c, \tilde{c})$ to its context: the *debate topic* τ (i.e., the thesis on Kialo) as well as the *previous claim* $\hat{c}$ (the parent on Kialo), which is supported or opposed by c (see Figure 5.1). We sampled 600 revision pairs pseudo-randomly as a test set (200 per intention label), and split remaining pairs into training (90%) and validation set (10%). As the given labels are rather coarse-grained, we look into the optimizations in more detail in Section 5.7.

5.4 Approach

We now present the first approach to automatic claim optimization. To account for the variety of possible optimizations, multiple candidate claims are generated that are pertinent to the context given and preserve the claim's meaning. Then, the best candidate is selected based on quality metrics. Both steps are detailed below and illustrated in Figure 5.2.

5.4.1 Seq2Seq-based Candidate Generation

To generate candidates, we fine-tune a Seq2Seq model on pairs $(c, \tilde{c})$, by treating the original claim c as encoder source and revised claim $\tilde{c}$ as the decoder target. In a separate experiment, we condition the model on context information, the debate topic τ and the previous claim $\hat{c}$, during fine-tuning to further optimize the relevance of generated candidates. The context is separated from c by delimiter tokens (Keskar et al., 2019; Schiller et al., 2021).

Multiple ways to improve c exist, especially if it suffers from multiple flaws, since not all flaws may be fixed in a single revision. Therefore, we first generate n suitable candidates, $\tilde{c}_1, \ldots, \tilde{c}_n$, among which the best one is to be found later (n is set to 10 in Section 3.6). However, the top candidates created by language models often tend to be very similar. To increase the diversity of candidates, we perform top-k sampling (Fan et al., 2018), where we first generate the most probable claim (top-1) and then vary k with in steps of 5 (e.g. top-5, top-10, etc).

5.4.2 Quality-based Candidate Reranking

Among the n candidates, we aim to find the optimal claim, c^*, that most improves the delivery of c in terms of text and argument quality. Similar to Yoshimura et al. (2020), we tackle this task as a candidate selection problem. In our proposed strategy, *AutoScore*, we integrate three metrics: (1) grammatical fluency, (2) meaning preservation, and (3) argument quality. This way, we can *explicitly* favor specific quality dimensions via respective models:

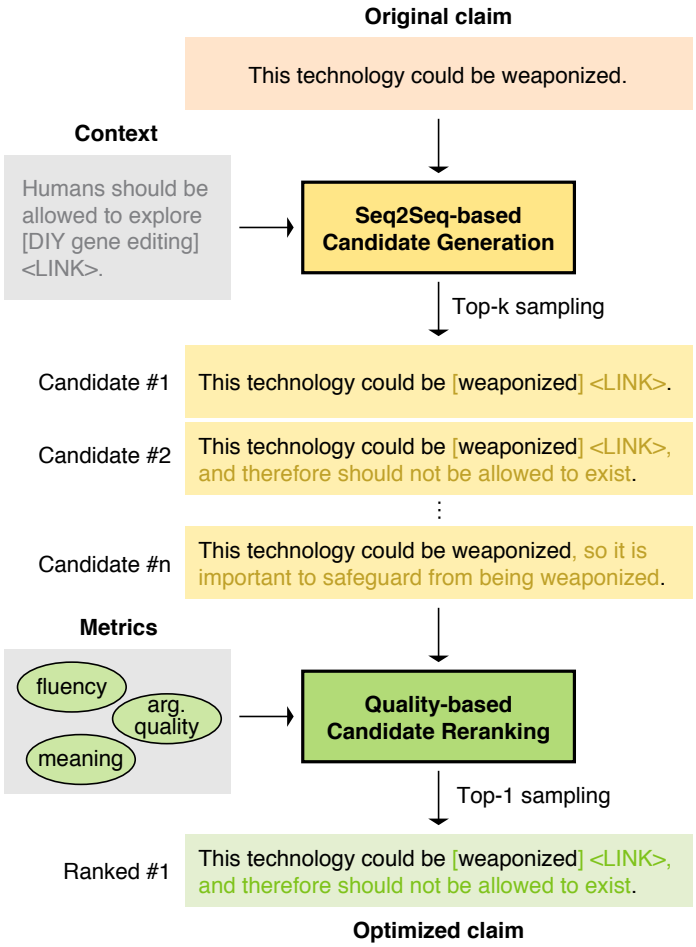


Figure 5.2: Proposed claim optimization approach: First, we generate n candidates from the *original claim*, possibly conditioned on context information. Then, the *optimized claim* is selected using three quality metrics.

Grammatical Fluency We learn to assess fluency on the MSR corpus (Toutanova et al., 2016) where the grammaticality of abstractive compressions is scored by 3–5 annotators from 1 (disfluent) to 3 (fluent). We chose this corpus, since multiple compressions per input make a trained model sensitive to the differences in variants of a text. For training, we average all annotator scores and make the task binary, namely, a text is seen as disfluent unless all annotators gave score 3. Then, we train BERT on the data to obtain fluency probabilities (details found in Appendix B.1). The accuracy of our model on the suggested data split is 77.4.

Meaning Preservation To quantify to what extent a generated candidate maintains the meaning of the original claim, we compute their semantic sim-

ilarity as the cosine similarity of the SBERT sentence embeddings (Reimers and Gurevych, 2019).

Argument Quality Finally, to examine whether the generated candidates are better than the original claim from an argumentation perspective, we fine-tune a BERT model on the task of pairwise argument classification using the ClaimRev dataset. Since this corpus is also used to fine-tune the Seq2Seq model, we apply the same training and validation split as described in Section 5.3.2 to avoid data leakage, and obtain 75.5 accuracy. We then use its probability scores to determine relative quality improvement (for more details see Appendix B.1). Given the three quality metrics, we calculate the final evaluation score, $AutoScore$, as the weighted linear sum of all three individual scores as

$$\alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument,$$

where $fluency$, $meaning$, and $argument$ are the normalized scores for the three outlined quality metrics. The three non-negative weights satisfy $\alpha + \beta + \gamma = 1$.

It should be noted that depending on the domain or writing skills of the users, there may be other more suitable datasets or approaches to capturing the outlined quality aspects, which could potentially lead to further performance improvements. While we do explore how well the suggested approaches transfer to certain other domains of text (see Section 5.7.3), identifying the optimal model for each quality dimension falls beyond the scope of this paper.

5.5 Experiments

This section describes our experimental setup to study how well the claims from Section 5.3 can be improved using our approach from Section 5.4. We focus on the impact of candidate selection.

5.5.1 Seq2Seq-based Candidate Generation

For candidate generation, we employ the pre-trained conditional language model BART (Lewis et al., 2020), using the *bart-large* checkpoint. However, other Seq2Seq architectures can also be considered within our approach (see Appendices B.1, B.2).

5.5.2 Quality-based Candidate Reranking

We evaluate our candidate selection approach in comparison to three ablations and four baselines:

Approach To utilize AutoScore for choosing candidates, the optimal weighting of its metrics must be determined. We follow Yoshimura et al. (2020), performing a grid search in increments of 0.01 in the range of 0.01 to 0.98 for each weight to maximize the Pearson's correlation coefficient between AutoScore and the original order of the revisions from revision histories in the validation set. Similar has been done for counterargument retrieval by Wachsmuth et al. (2018). The best weights found are $\alpha = 0.43$, $\beta = 0.01$, and $\gamma = 0.56$, suggesting that meaning preservation is of low importance and potentially may be omitted. We suppose this is due to the general similarity of the generated candidates, so a strong meaning deviation is unlikely.

Ablations To assess the impact of each considered quality metric used in AutoScore, we perform an ablation study, where optimal candidates are chosen based on the individual metric scores:

- *Max Fluency.* Highest grammatical fluency
- *Max Argument.* Highest argument quality
- *Max Meaning.* Highest semantic similarity

Baselines We test four selection strategies for 10 candidates generated via top-k sampling:

- *Unedited.* Return the original input as output.
- *Top-1.* Return the most likely candidate (obtained by appending the most probable token generated by the model at each time step).
- *Random.* Return candidate pseudo-randomly.
- *SVMRank.* Rerank candidates with SVMRank (Joachims, 2006). Using sentence embeddings we decide which of the claim versions is better, by fine-tuning SBERT (*bert-base-cased*) on the corpus of Skitalinskaya et al. (2021).

5.5.3 Evaluation

We explore claim optimization on all 600 test cases, both automatically and manually:

Automatic Evaluation We compare all content selection strategies against the reference revisions using the precision-oriented *BLEU* (Papineni et al., 2002), recall-oriented *Rouge-L* (Lin, 2004), *SARI* (Xu et al., 2016), which computes the average F_1-scores of the added, kept, and deleted n-grams in comparison to the ground truth revision output, and the *exact match accuracy*. We also compute

the semantic similarity of the optimized claim and the context information to capture whether conditioning claims on context affects their topic relevance.

Manual Evaluation As we fine-tune existing generation models rather than proposing new ones, we focus on the *candidate selection* in two manual annotation studies. For each instance, we acquired five independent crowdworkers via *MTurk*.

In the first study, the annotators scored all candidates with respect to the three considered quality metrics. We used the following Likert scales:

- *Fluency.* 1 (major errors, disfluent), 2 (minor errors), and 3 (fluent)
- *Meaning Preservation.* 1 (entirely different), 2 (substantial differences), 3 (moderate differences), 4 (minor differences), and 5 (identical)
- *Argument Quality.* 1 (notably worse than original), 2 (slightly worse), 3 (same as original), 4 (slightly improved), and 5 (notably improved)

A challenge of crowdsourcing is to ensure good results (Sabou et al., 2014). To account for this, we obtained the fina fluency, argument quality and meaning preservation scores using MACE (Hovy et al., 2013), a Bayesian model that gives more weight to reliable workers. In the given case, 39% of the 46 annotators had a MACE competence value > 0.3, which can be seen as reasonable in MTurk studies.

In the second study, we asked annotators to rank four candidates, returned by the content selection strategies, by perceived overall quality. If multiple candidates were identical, we showed each only once. While Krippendorff's α agreement was only 0.20 and percent agreement was 0.36% (majority voting), such values are common in subjective tasks (Wachsmuth et al., 2017b; Alshomary et al., 2021a).

5.6 Results and Discussion

Apart from evaluating the applicability of large generative language models to the task of argumentative claim optimization in general, our experiments focus on two questions: (1) Does the use of explicit knowledge about text and argument quality lead to the selection of better candidates? (2) Does the use of contextual information make the generated candidates more accurate and relevant to the debate?

Approach	BLEU	RouL	SARI	NoEd↓	ExM
Baselines					
Unedited	**69.4**	**0.87**	27.9	1.00	0.0%
BART + Top-1	64.0	0.83	39.7	0.31	7.8%
BART + Random	62.6	0.83	38.7	0.28	6.8%
BART + SVMRank	55.7	0.76	38.8	0.03	4.5%
Approach					
BART + AutoScore	59.4	0.80	**43.7**	**0.02**	**8.3%**
Ablation					
BART + Max Fluency	57.6	0.78	41.5	0.09	5.8%
BART + Max Argument	60.9	0.81	43.6	0.02	8.0%
BART + Max Meaning	69.0	0.87	33.8	0.72	5.2%

Table 5.1: Automatic evaluation: Performance of each candidate selection strategy on 600 test cases in terms of BLEU, Rouge-L, SARI, ratio of unedited cases, and ratio of exact matches to target reference.

5.6.1 Overall Claim Optimization Performance

Automatic Evaluation Table 5.1 shows the automatic scores of all considered candidate selection strategies. The high scores of the baseline *Unedited* on metrics such as BLEU and ROUGE-L indicate that many claim revisions change little only. In contrast, *Unedited* is worst on SARI, a measure taking into account words that are added, deleted, and kept in changes, making it more suitable for evaluation. Here, *BART+AutoScore* performs best on SARI (43.7) and exact match accuracy (8.3%).

The *BART+Max Meaning* ablation supports the intuition that the candidates with highest meaning preservation scores are those with minimal changes, if any (72% of the candidates remain identical to the input). Such identical outputs are undesirable, as the claims are not optimized successfully, which is also corroborated by the low weight parameter ($\beta = 0.01$) found for the meaning preservation metric when optimizing AutoScore (see Section 3.6).

Manual Evaluation Table 5.2 shows that human annotators prefer optimized candidates selected by *AutoScore*, with an average rank of 1.92. The difference to *Top-1* and *Random* is statistically significant ($p < .05$ in both cases) according to a Wilcoxon signed-rank test, whereas the gain over the second-best algorithm, *SVMRank*, is limited. Also, candidates of AutoScore and SVMRank are deemed more fluent than those of Top-1 and Random (2.33 vs. 2.29 and 2.26). In terms of argument quality, the results deviate from the automatic evaluation (Table 5.1), showing marginally higher scores for SVMRank and Top-1. Further analysis revealed that AutoScore and SVMRank agreed on the

Model	Strategy	Fluency	Argument	Meaning	Rank
BART	Top-1	2.29	3.61	3.65	2.16
	Random	2.26	3.50	3.53	2.06
	SVMRank	**2.33**	**3.69**	**3.66**	1.95
	AutoScore	**2.33**	3.61	3.57	**1.92**

Table 5.2: Manual evaluation: Scores on the 600 test cases generated by BART using our candidate selection strategy *AutoScore* or the baselines: fluency (1–3), argument quality and meaning (1–5), mean rank (1–4, lower better). AutoScore ranks significantly better than *Top-1* ($p < .005$), *Random* ($p < .05$), and *SVMRank* ($p < .1$).

Context	BLEU	Original	Previous	Topic
Claim only	59.4	0.95	0.55	0.55
+ Previous Claim	**60.3**	0.95	**0.57**	**0.57**
+ Debate Topic	60.0	0.95	0.55	0.55
Human-Baseline	100.0	0.94	0.55	0.55

Table 5.3: BLEU and semantic similarity score with respect to the *original* claim, the debate's *previous* claim, and its *topic* of BART+AutoScore, depending on the context given for the 600 test samples.

optimal candidate in 35% of the cases, partially explaining their close scores. Although SVMRank achieved high scores across the three quality metrics, we note that the annotators preferred candidates scores generated by AutoScore, highlighting the importance of more diverse revision changes reflected by lower meaning preservation scores.

Overall, our findings suggest that using candidate selection approaches that incorporate quality assessments (i.e., AutoScore and SVMRank) leads to candidates of higher fluency and argument quality while preserving the meaning of the original claim. In addition to Figure 5.1, examples of automatically-generated optimized claims can be found in the appendix.

5.6.2 Performance with Context Integration

General Assessment Table 5.3 shows the semantic similarity of claims optimized by our approach and context information, depending on the context given. The results reveal slight improvements when conditioning the model on the previous claim (e.g., 60.3 vs. 59.4 BLEU). To check whether this led to improved claims, two authors of the paper compared 600 claims generated with and without the use of the previous claim in terms of (a) which claim seems better overall and (b) which seems more grounded. We found that

using the previous claim as context improved quality in 12% of the cases and lowered it in 1% only, while leading to more grounded claims in 36%.

Qualitative Analysis Our manual inspection of a claim sample revealed the following insights:

First, conditioning on context reduces the number of erroneous specifications, particularly for very short claims with up to 10 words. This seems intuitive, as such claims often convey little information about the topic of the debate, making inaccurate changes without additional context likely.

Next, Kialo revisions often adhere to the following form: A claim introduces a statement and/or supporting facts, followed by a conclusion. This pattern was frequently mimicked by our approach. Yet, in some cases, it added a follow-up sentence repeating the original claim in different wording or generated conclusions containing fallacious or unsound phrases contradicting the original claim in others. Modeling context mitigated this issue.

Finally, we found that models conditioned on different contexts sometimes generated candidates optimized in different regards, whereas a truly optimal candidate would be a fusion of both suggestions.

5.7 Analysis

To explore the nature of claim optimization and the capabilities of our approach, this section reports on (a) what types of optimizations exist, (b) how well our approach can operationalize these, and (c) how well it generalizes to non-argumentative domains.

5.7.1 Taxonomy of Optimization Types

To understand the relationship between optimizations found in the data and the underlying revision intentions, two authors of this paper inspected 600 claim revisions of the test set. Opposed to actions, intentions describe the goal of an edit (e.g., making a text easier to read) rather than referring to specific changes(e.g., paraphrasing or adding punctuation). We build on ideas of Yang et al. (2017) who provide a taxonomy of revision intentions in Wikipedia texts. Claims usually do not come from encyclopedias, but from debate types or from monological arguments, as in essays (Persing and Ng, 2015). Therefore, we adapt the terminology of Yang et al. (2017) to gear it more towards argumentative texts.

As a result of a joint discussion of various sample pairs, we decided to distinguish eight optimization types, as presented in Table 5.4. Both authors then

#	Optimization	Description of the Type	Clarification	Grammar	Links
1	Specification	Specifying or explaining a given fact or meaning (of the argument) by adding an example or discussion without adding new information.	58	1	–
2	Simplification	Removing information or simplifying the sentence structure, e.g., with the intent to reduce the complexity or breadth of the claim.	43	–	–
3	Reframing	Paraphrasing or rephrasing a claim, e.g., with the intent to specify or generalize the claim, or to add clarity.	29	–	–
4	Elaboration	Extending the claim by more information or adding a fact with the intent to make the claim more self-contained, sound, or stronger.	23	–	–
5	Corroboration	Adding, editing, or removing evidence in the form of links that provide supporting information or external resources to the claim.	8	–	153
6	Neutralization	Rewriting a claim using a more encyclopedic or neutral tone, e.g., with the intent to remove bias or biased language.	7	–	–
7	Disambiguation	Reducing ambiguity, e.g., replacing pronouns by concepts mentioned before in the debate, or replacing acronyms with what they stand for.	7	–	1
8	Copy editing	Improving the grammar, spelling, tone, or punctuation of a claim, without changing the main point or meaning.	41	200	52

Table 5.4: Descriptions of the eight claim optimization types identified in the 600 test pairs. The right columns show the count of claims per type for each of the three intention labels from Skitalinskaya et al. (2021): *clarification*, typo/*grammar* correction, and correcting/adding *links*. Note, that a revision may be assigned to multiple categories.

Type	Human	Approach	Better	Same	Worse
Specification	59	152	65%	19%	16%
Simplification	43	18	61%	28%	11%
Reframing	29	21	62%	33%	5%
Elaboration	23	55	62%	18%	20%
Corroboration	161	38	53%	23%	24%
Neutralization	7	0	–	–	–
Disambiguation	8	8	63%	25%	12%
Copy editing	293	301	59%	26%	15%
Overall	623	593	60%	24%	16%

Table 5.5: Manual analysis: Comparison of the human-optimized claims of all 600 test cases (some have multiple) and of the claims optimized by BART+AutoScore (15 claims were unchanged). The three right columns show the ratio of optimized claims judged *better*, *same*, or *worse* than the original in terms of overall quality.

annotated all 600 test pairs for these types, which led to only 29 disagreement cases, meaning a high agreement of 0.89 in terms of Cohen's κ. These cases were resolved by both annotators together.[4]

Table 5.4 also shows cooccurrences of the types and intention labels. *Typo/grammar correction* and *correcting/adding links* align well with *copy editing* and *corroboration* respectively. In contrast, clarification is broken into more fine-grained types, where *specification* seems most common with 58 cases, followed by *simplification* and *reframing*. Examples of each type are found in the appendix.

We point out that the eight types are not exhaustive for all possible claim quality optimizations, but rather provide insights into the semantic and discourse-related phenomena observed in the data. We see them as complementary to the argument quality taxonomy of Wachsmuth et al. (2017b) as ways to improve the delivery-related quality dimensions: *clarity*, *appropriateness*, and *arrangement*.

5.7.2 Performance across Optimization Types

To enable comparison between the human optimizations and automatically generated outputs, two authors of the paper labeled 600 optimized claims with the types defined in Table 5.4. Due to resource constrains only the best performing approach, BART+AutoScore, was considered. Overall, our approach generates better claims in 60% of the cases, while 84% remain at least of similar

[4]We acknowledge that there is potential bias inherent in self-annotation. However, we would like to point out that no knowledge about the test set was used to develop the approach presented in Section 5.4.

Approach	BLEU	RouL	SARI	NoEd↓	ExM
WikiHow Dataset					
Unedited	65.7	0.85	28.4	1.00	0.00%
BART + Top-1	**64.7**	**0.83**	41.3	0.50	13.0%
BART + AutoScore	61.8	0.80	**48.5**	**0.08**	**16.0%**
IteraTeR Dataset					
Unedited	74.0	0.86	28.6	1.00	0.00%
BART + Top-1	**68.9**	**0.83**	37.0	0.07	0.00%
BART + AutoScore	64.8	0.80	**38.6**	**0.02**	0.00%

Table 5.6: Automatic evaluation: Performance of candidate selection strategies on data from other domains, in terms of BLEU, Rouge-L, SARI, ratio of unedited samples, and ratio of exact matches to target reference.

quality.

Most noteworthily, we observe that our approach performs optimizations of the type *specification* 2.5 times as often as humans, and more than double as many *elaboration* revisions (55 vs. 23). In contrast, it adds, edits, or removes evidence in the form of links (*corroboration*) four times less often than humans. The model also made fewer *simplifications* (18 vs. 43) and no *neutralization* edits, which may be due to data imbalance regarding such types.

In terms of average quality, *specification* (65%) and *disambiguation* edits (63%) most often lead to improvements, but the eight types appear rather balanced in this regard. The Jaccard similarity score between optimizations performed by humans and our approach is 0.37, mostly agreeing on copy edits (178 cases) and corroboration (22 cases). Given such low overlap, future work should consider conditioning models to generate specific optimizations.

5.7.3 Performance across Revision Domains

Lastly, we examine whether our approach, along with the chosen text quality metrics, applies to texts from other domains. We consider two datasets: *WikiHow* (Anthonio and Roth, 2020), containing revisions of instructional texts, and *IteraTeR* (Du et al., 2022), containing revisions of various formal texts, such as encyclopedia entries, news, and scientific papers. For our experiments, we use the provided document-level splits, and sample 1000 revision pairs pseudo-randomly as a final test set.

Table 5.6 shows automatic evaluation results. In both cases, *BART+Autoscore* leads to higher SARI scores (48.5 vs. 41.3 for WikiHow, 38.6 vs. 37.0 for IteraTeR), and notably reduces the number of cases where the models failed to revise the input (0.08 vs. 0.50 for WikiHow). The reported *BART+Top1* model

represents the approach of Du et al. (2022), indicating that our approach and its text quality metrics achieve state-of-the-art performance with systematic improvements across domains, when generating optimized content. However, as different domains of text have different goals, different notions of quality, and, subsequently, different revision types performed, integrating domain-specific quality metrics may further improve performance. We leave this for future work.

5.8 Limitations

This work contributes to the task of argumentative text editing, namely we explore how to revise claims automatically in order to optimize their quality. While our work may also improve downstream task performance on other tasks, it is mainly intended to support humans in scenarios, such as the creation and moderation of content on online debate platforms as well as the improvement of arguments generated or retrieved by other systems. In particular, the presented approach is meant to help users by showing examples of how to further optimize their claims in relation to a certain debate topic, so they can deliver their messages effectively and hone their writing skills.

However, our generation approach still comes with limitations and may favor revision patterns over others in unpredictable ways, both of which might raise ethical concerns. For example, it may occasionally produce false claims based on untrue or non-existent facts. We think, humans should be able to identify such cases in light of the available context though, as long as the improvements remain suggestions and do not happen fully automatically, as intended.

The presented technology might further be subject to intentional misuse. A word processing software, for example, could be conditioned to automatically detect and adapt claims made by the user in subtle ways that favors political or social views of the software provider. Such misuse might then not only change the intended message of the text, but also influence or even change the views of the user (Jakesch et al., 2023).

In a different scenario, online services, such as social media platforms or review portals, might change posted claims (e.g. social media posts, online reviews) to personalize them and increase user engagement or revenue. These changes might not only negatively affect the posting, but also the visiting user.

While it is hard to prevent such misuse, we think that the described scenarios are fairly unlikely, as such changes tend to be noticed by the online

community quickly. Furthermore, the presented architecture and training procedure would require notable adaptations to produce such high-quality revisions.

An aspect that remains unexplored in this work is the ability of the presented approaches to work with variations of the English language, such as African-American English, mainly due to the lack of available data. In this regard, the approach might unfairly disadvantage or favor particular language varieties and dialects, potentially inducing social bias and harm if applied in public scenarios. We encourage researchers and practitioners to stay alert for such cases and to choose training data with care for various social groups.

Finally, our work included the labeling of generated candidate claims on a crowdsourcing platform. As detailed in Section 5.5, we compensated MTurk workers $13 per hour, complying with minimum wage standards in most countries at the time of conducting the experiment.

5.9 Conclusion

With this paper, we work towards the next level of computational argument quality research, namely, to not only *assess* but also to *optimize* argumentative text. Applications include suggesting improvements in writing support and automatic phrasing in debating systems. We presented an approach that generates multiple candidate claim optimizations and then selects the best one using various quality metrics. In experiments, combining fine-tuned BART with such candidate selection improved 60% of the claims from online debates, outperforming several baseline models and candidate selection strategies. We showcased generalization capabilities on two out-of-domain datasets, but we also found some claim optimization types hard to automate.

In future work, we seek to examine whether recent large language models (e.g., Alpaca) and end-to-end models (where generation and candidate selection are learned jointly) can further optimize the quality of claims. As our approach so far relies on the availability of large claim revision corpora and language models, techniques for low-resource scenarios and languages should be explored to make claim optimization more widely applicable.

Acknowledgments

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 374666841, SFB 1342.

5.10 Implications for the Thesis

The findings presented in this chapter address the third research question: *How to approach the generation of improved argumentative texts using computational methods?* Specifically, we showcased how combining the capabilities of large language models with quality-based reranking can be used to automatically rewrite argumentative texts. Through a series of experiments, we have showcased the benefits of leveraging contextual information when addressing certain flaws in the text, such as the reduction of erroneous specifications, repetitions, and fallacious or unsound phrases contradicting the original claim. Additionally, we have demonstrated the generalization capabilities of the suggested approaches on out-of-domain datasets, such as scientific articles, news, and instructional texts, thereby shedding light on the adaptability of the suggested methods beyond their original context.

Overall, the chapter provides valuable insights into how to approach the modeling of argument quality for such optimizations of argumentative texts, thereby contributing to a deeper understanding of the intricate dynamics between linguistic proficiency and persuasive discourse. This research holds significance not only in advancing academic knowledge but also in its practical implications for enhancing writing assistance tools, ultimately aiding users in generating more impactful and effective written communication.

In conclusion, the following chapter summarizes our main findings from the individual chapters and their contributions while outlining their implications and limitations.

6. Discussion and Conclusion

As discussed in Chapter 1, argumentative writing is an essential skill that cultivates critical thinking, elevates public discourse, bolsters communication abilities, and empowers people to engage thoughtfully with information. In this thesis, through the use of advanced computational techniques, we have aimed to develop a set of methods that can guide writers and help them in improving their argumentative writing skills and in producing more compelling and persuasive texts. In this final chapter, we will reflect on the progress made within our work and consider the challenges that lie ahead for the future of computational argument assessment and text improvement generation. We start with summarizing the key findings and contributions of our work, as well as their potential impact on the field of computational argumentation and argumentative writing support technologies (Section 6.1). In Section 6.2, we then discuss the limitations of our work and areas we did not focus on while highlighting how future research could potentially address these shortcomings and unexplored directions. Finally, in Section 6.3, we conclude the thesis with a closing remark.

6.1 Summary

Our research tackles a broad question: *What makes a good argument and how can we computationally model this knowledge to develop methods supporting individuals in improving their arguments?* Despite extensive studies on argument quality in the past, the questions of supporting and automating argument improvement remain largely unexplored. This thesis investigated the question in-depth and assessed the potential benefits of computationally modeling argument quality based on human revision and text editing behaviors. Through this exploration,

we aimed to gain not only a thorough understanding of how such behaviors influence the quality of arguments but also how these insights can be leveraged to develop automated approaches that could guide users through the process of critically assessing and improving their written texts.

To this extent, this thesis focused on three main research questions. We start with the first research question, which targets the specifics of the domain of argumentative writing and revision-based data:

RQ1 What *quality-related phenomena* are typical of argument revisions on online debate platforms?

Previous work on argument quality assessment introduced various datasets covering different topics and domains, however, most of them are limited in size and diversity, making them unsuitable for modeling generalizable aspects of argument quality. While a significant amount of data can be obtained from online debate platforms that support collaborative revision processes, it is challenging to utilize this revision-based data to its full potential. On the one hand, certain challenges arise from the notion of argument quality, as certain quality dimensions are inherently subjective and depend on the surrounding context. On the other hand, the nature of revision-based corpora, in general, is known to be noisy and biased, making it difficult to compile reliable examples of high and low quality content.

In Chapter 3, we outlined the main challenges of dealing with revision-based corpora when modeling text quality. We addressed concerns related to the representativeness and reliability of data, potential bias in revision behaviors due to topic preferences, model complexities and architectures suitable for capturing differences between claim revisions, and the need for context when evaluating argumentative claims.

To confirm whether such revision-based data could be used to computationally model argument quality, we conducted a detailed analysis of the interplay between various types of revisions and their impact on the quality of the argumentative text. Specifically, in Chapter 4, we explored the potential of working with argumentative writing found in online debate platforms, such as Kialo, where vast amounts of data are readily available. Our analysis yielded compelling results, revealing that within the collected dataset, 93% of annotated revisions were associated with an evident enhancement in argument

quality. To better understand the differences between the types of revision performed and their effect on argument quality, we conducted a second annotation study. Here, the goal was to determine whether the revision improved for each of the 15 argument quality dimensions defined by Wachsmuth et al. (2017b) or not. We found that the collected revisions primarily target the general form a well-phrased claim should have, mainly focusing on logical and dialectical quality dimensions as opposed to rhetoric dimensions, such as credibility, emotional appeal, etc.

Furthermore, in Chapter 5, we explored the interplay between various types of improvement found in the data and the underlying revision intentions of the authors. As a result, we have developed a fine-grained taxonomy that categorizes revisions based on the actions (optimizations) taken to improve a certain text. This taxonomy provided a framework that was not only used to understand the effectiveness and significance of human revisions in enhancing the overall quality and persuasiveness of texts but also to evaluate automatically generated revision outputs.

The second research question addressed the problem of modeling such revision-based data to enable quality assessments of argumentative texts:

RQ2 How to approach the modeling of argument quality computationally to enable the *analysis of arguments in need of improvement*?

When dealing with a high volume of text-based contributions, whether in mass collaboration processes (online debate moderation, encyclopedia content curation) or educational scenarios (essay feedback generation and grading), manual analysis proves to be challenging, costly, and time-consuming. Thus, automating the process of quality control can be beneficial not only to the participants creating the content to learn how to write better texts but also to the moderators and teachers, who monitor and evaluate the quality of the generated data.

To this end, we considered several conceptualizations of argument quality assessment. Specifically, in Chapter 3, we begin with introducing the task of *Suboptimal Claim Detection*, where the goal is to identify low quality argumentative texts in need of revision. Once problematic texts are identified, as a next step, we propose the task of *Claim Improvement Suggestion*, where the goal is to predict what type of revision a given text would benefit from most.

Finally, in Chapter 4, we introduce the task of *Claim Quality Ranking*, where several versions of the same text are compared in order to find the best one. The differences in problem framing allow for a more nuanced and targeted assessment when capturing argument quality, making them applicable to a wide range of applications focusing on content quality control in online moderation processes or education.

For each of these tasks, we proposed and evaluated various solutions covering traditional and neural approaches. In experiments, we compared the suggested methods to determine their effectiveness in capturing quality differences between different versions of the same text. Although there were some limitations, we found that using revision-based data can be useful for these tasks and can help assess argument quality from a yet unexplored revision-oriented perspective. Overall, our results demonstrated that utilizing revision-based data allows the learned quality assessment models to generalize well across topics and make judgments regardless of the aspects and stances covered in the text. Our proposed sampling strategy showed that training on claim versions with a greater revision distance between them can improve performance when identifying claims that require improvement. Moreover, we have shown that incorporating contextual information in the modeling process is beneficial when making any argument quality assessments. Specifically, we have provided empirical evidence of the effectiveness of various context types depending on the task and quality issues that a text is suffering from.

While the first two research questions help us understand, characterize, and computationally model the quality of argumentative texts, they also offer guidance on how to approach the automated improvement of such texts by emphasizing important aspects, inter-dependencies, and attributes that shape the perceived quality of the argument. Consequently, the third research question addresses the problem of automatically generating improved versions of argumentative texts while keeping in mind the lessons learned from quality assessment tasks:

RQ3 How to approach the *generation of improved argumentative texts* using computational methods?

Automated systems that can generate improved argumentative texts have the potential to significantly enhance the efficiency and effectiveness of content

creation and moderation processes, saving time and resources. For example, such technologies could provide students with immediate feedback on their written assignments, offering corrections and suggestions for improvement while helping them learn from their mistakes and make revisions to their work. On the other hand, teachers could use automated tools to streamline the grading process, allowing them to focus more on providing personalized feedback on content rather than spending excessive time on basic language, formatting, structure, or argumentation issues. Similarly, content platforms, such as online debates, can use such automated tools to ensure that user-generated content meets specific quality standards before publication, reducing the need for manual moderation. However, despite the wide attention devoted by the research community to automated argument assessment, no work has considered improving the quality of said arguments.

In Chapter 5, we address this gap and introduce the task of *Claim Optimization*, where the goal is to rewrite an argumentative text without changing its original meaning in order to improve its delivery. To this end, we proposed various neural approaches that incorporate relevant contextual information and argument quality assessments. In particular, we proposed an approach that first generates a diverse range of candidate claims using a large language model like BART and then selects the best candidate via a ranking process using several argument and text quality metrics. We argue that decomposing the approach into two stages allows for finer control over the claim generation process, facilitating more nuanced and higher quality improvements in the final argumentative text. Using the proposed taxonomy of revisions, we demonstrated in both automatic and manual evaluations that the suggested approach outperforms several baselines covering nearly all types of revisions typically performed by humans. Moreover, the proposed solutions also generalize well to other domains, such as instructional texts, news, scientific articles, and encyclopedia entries.

Together, these contributions address several limitations and research gaps derived from previous work in computational argumentation and natural language processing (see Section 1.2), such as the absence of approaches capable of making general argument quality assessments independently of the subject matter, beliefs and biases of the participants or audience, the lack of computational solutions geared towards guiding writers on how and when to improve the quality of their argumentative texts, and the scarcity of large-scale corpora necessary to facilitate the tasks mentioned above.

6.2 Limitations and Future Work

Though the conducted analysis and presented approaches bring advances in how we can assess argument quality and assist users in improving their argumentative texts, the work also comes with limitations imposed by the research design and chosen methods.

Low-resource Scenarios and Multilinguality

A notable aspect left unexplored in this work is the adaptability of the proposed methodologies beyond the English language. This is primarily due to the lack of available data in the form of revision histories of argumentative texts for other languages. This leads us to another connected issue: the ability of the suggested approaches to successfully perform the considered tasks in cases where limited annotated data is available. Although, in Chapter 3, we explore low-resource scenarios for low argument quality detection, our approaches for text generation (Chapter 5) require large scale parallel corpora for training. Addressing these constraints is essential to enhancing the applicability and inclusivity of proposed solutions to generative language modeling and quality assessment across different languages and domains.

During the final writing stages of this thesis, large language models with hundreds of billions of parameters, such as GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023), have been introduced and made available to researchers and practitioners. While such models require significant computational resources for training, they have made it possible to directly process text generation tasks via in-context learning (Brown et al., 2020) without any fine-tuning. Recent research suggests that such large language models can successfully perform cross-lingual text classification and generation tasks, i.e., perform tasks in a target language although they were not implicitly trained on that language (Tanwar et al., 2023; Holmström et al., 2023). In future work, it would be interesting to explore how well such large language models can perform the argument assessment and generation tasks introduced in this thesis and understand their limitations when it comes to alleviating problems related to limited data availability and cross-lingual transfer.

Bias, Factuality, and Ethical Considerations

Generative language models, while capable of producing coherent and contextually relevant text, are known to suffer from several limitations pertaining to the level of bias and factuality of the generated content. In our work, to generate argumentative text improvements, we relied on pre-trained language

models, such as BART(Lewis et al., 2020) in Chapter 5, which occasionally may inadvertently generate inaccurate or misleading content and perpetuate or even exacerbate biases present in its training data, leading to biased and unfair generated outputs. Despite ongoing efforts to mitigate bias (Liu et al., 2021; Stahl et al., 2022) and ensure faithfulness and factuality (Guo and Singh, 2023), achieving a completely bias-free generation remains a challenging endeavor. In this regard, future work could consider integrating additional bias detection and fact verification mechanisms within the considered approaches, for example, as part of the reranking strategy proposed in Chapter 5.

Moreover, it is worth mentioning that the development of automated text improvement technologies comes with challenging ethical problems. On the one hand, such technologies need to preserve a certain level of free speech to stimulate high-quality discussions while implementing regulations to identify editing behaviors defined as inappropriate. On the other hand, distinguishing such legitimate forms of regulation from illegitimate censorship, where particular opinions and individuals are suppressed, is a challenge of its own. Without solutions to mitigate these negative effects, the technologies presented in the paper might be subject to intentional misuse. For example, one could condition the proposed generative approaches to automatically detect and adapt claims made by users to favor political or social views. Such misuse might then not only change the intended message of the text but also influence or even change the views of the user (Jakesch et al., 2023). Addressing these limitations is crucial for harnessing the full potential of generative language models while safeguarding against their unintended consequences on society.

The Role of Audience Characteristics and Emotions in Revisions

Another important aspect is that our suggested approaches, while capable of automatically refining and enhancing written arguments, do not take into account audience or participant characteristics, which have been shown to be integral to effective communication (Lukin et al., 2017; Durmus et al., 2019b; Alshomary et al., 2021a; Alshomary and Wachsmuth, 2021). Such characteristics could include prior beliefs, personalities, gender, and social background, among others. By failing to tailor their revisions to the unique perspectives and predispositions of readers, these models risk producing arguments that are less compelling or even counterproductive.

Furthermore, a significant gap exists in the exploration of emotional engagement within the context of argumentative text revisions. Emotions play a pivotal role in influencing the receptiveness and persuasiveness of an argu-

ment, yet due to the specificity of the considered corpora, where it was against platform policy to contribute arguments that are typically emotionally loaded, such as testimonies and anecdotical evidence. As such, we haven't explored the interplay of emotions and argument quality in the revision process. Addressing these limitations is essential for ensuring that generative language models can truly enhance the effectiveness of argumentative texts and foster more persuasive and emotionally resonant communication.

Personalization of Automated Argument Assessment and Revision

While the previous paragraph focuses on the audience characteristics, it is also important to take into account the domain specifics and characteristics of the user interacting with such argumentative writing support technologies. While in our work, we do not integrate such information in the argument assessment and revision approaches; we believe that adding such personalization into the process could lead to further improvements and enable a more user-focused and targeted support. For example, in educational scenarios, by considering factors such as a user's age, educational level, and prior argumentation skills, one could tailor argument assessment and revision support approaches to adjust the complexity and depth of their guidance to support the individual needs of a student, while ensuring that the content remains relevant and comprehensible. Thus offering a more effective and engaging learning experience. On the other hand, in online moderation scenarios, personalized revision support can help users understand and rectify their violations of community guidelines while ensuring high-quality content, thus promoting a more constructive and respectful online environment. However, dealing with such user-specific characteristics comes with its own challenges (Weidinger et al., 2022; Brown et al., 2022). Collecting and processing user-specific data to provide personalized moderation assistance can potentially encroach upon user privacy. Safeguarding sensitive information and ensuring that data handling practices align with legal and ethical standards, such as the General Data Protection Regulation (GDPR), would be crucial to address such concerns.

Beyond Text Quality Improvement in Argumentation Contexts

As previously mentioned, the development of large language models, such as GPT-4 and LLama, has led to a paradigm shift in the field of natural language processing by moving away from typical finetuning processes to in-context learning. Although in-context learning represents a cost-effective way to harness the power of such models, it is highly sensitive to prompt engineering

(Wei et al., 2022), i.e., designing and crafting input instructions to elicit desired responses or behaviors from the model to boost its performance. Consequently, various prompt methods have been suggested in recent work, such as few-shot prompting, i.e., prepending high-quality input-output demonstrations before task prompt (Vilar et al., 2023), automatic prompt learning, i.e., enabling the model itself to augment and curate high quality training examples to improve its own performance (Li et al., 2023), chain-of-thought prompting, i.e., allowing models to decompose multi-step problems into intermediate steps (Wei et al., 2022), etc.

Leaving out the computational argumentation context of our work, we would like to suggest a more broad perspective on how the considered concepts of iterative revision and text refinement can benefit the large language models and existing prompting techniques. Specifically, in future work, we are interested in exploring whether conditioning large language models to iteratively generate feedback and refine their own outputs through novel prompting techniques to provide a high-quality response in a single inference. We believe that such conditioning could lead to performance improvements beyond the considered tasks of argumentative text assessment and generation, but on other natural language processing tasks, such as text summarization (Zhang et al., 2023) and commonsense reasoning (Qiao et al., 2023). We want to explore whether such refinement is possible to achieve solely through prompting without requiring additional training data or reinforcement learning (Sutton and Barto, 2018).

6.3 Closing Remark

With this work, we took a step forward towards automatically assessing the quality of argumentative texts and generating their improved versions. We have done so by adopting a new perspective that looks at argument quality through the lens of revisions. This approach has allowed us to gain insights into the essential aspects of the revision process and how it correlates with the broader theory of argument quality. We believe our work will be a valuable resource for researchers, moderators, debate enthusiasts, education specialists, and practitioners alike. As Mark Twain once noted, "*Writing is easy. All you have to do is cross out the wrong words.*", we hope our work simplifies the process of refining arguments and showcases that, with the right approach, the art of persuasion becomes an attainable endeavor.

Appendices

A. Experimental Details for Chapter 3

A.1 Implementation and Training Details

A.1.1 Generating Embeddings

All claim embeddings were generated using the flair library,[1] via DocumentPoolEmbeddings for non-transformer-based models, such as Glove and Flair, or TransformerDocumentEmbeddings for BERT and ELECTRA embeddings.

Glove + SVM We derived claim representations by averaging the obtained word representations and feed them as input to a linear SVM (Joachims, 2006). We initialized the 100-dimensional word embeddings pretrained on Wikipedia data ("glove-wiki-gigaword-100").

Flair + SVM We used the 2,048-dimension "news-forward" embeddings, produced by a forward bi-LSTM, trained on the One Billion Word Benchmark (Chelba et al., 2013) and feed the obtained embeddings to a linear SVM classifier.

BERT We use the case-sensitive pre-trained version (bert-base-cased).

A.1.2 Training SVM models

For faster convergence when dealing with a large number of samples, we use a SVM with a linear kernel, specifically, LinearSVC, as implemented in the sklearn library.[2] We set maximum iterations to 1000 and choose the regularization parameter out of {0.001, 0.01, 0.1, 1, 10}.

[1] flair, `https://github.com/flairNLP/flair`

[2] sklearn SVM, `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC`

False Positives	False Negatives
The HPV virus is harmful. (Clarif)	can be dangerous for bikers
Vertically farming is healthier for people. (Clarif)	Women are healthier than men
There would be disputed over the leaders (Typo/Grammar)	I can't support this. The math is way off. We have 15X the population and 55X the homicide rate.
The world is becoming too populated anyway. (Style)	People are likely to forget distressing memories.
The Czech Republic is funding travel TV shows in Korea. (Links)	The police of every country have abused their authority systemically at some point in history
A number of recreational drugs may have health benefits. (Links)	Podcasts cannot include music due to copyright issues, so they cannot replace radio entirely

Table A.1: Examples of False Positive and False Negative predictions obtained by FT-DeBERTa (without considering context). The true class for False Positives is reflected in the brackets at the end of each claim.

A.1.3 Fine-tuning Transformer-based models

We used the *bert-base-cased* pre-trained BERT version (110M parameters), the *electra-base-discriminator* pre-trained ELECTRA version (110M parameters), and the *deberta-base* pretrained DeBERTA version (140M parameters) as implemented in the huggingface library.[3] We set the maximum sequence length to 128 and 256 tokens, depending whether contextual information was used or not. We trained for a maximum of five epochs using the Adam optimizer with a warmup of 10000 steps and linear learning rate scheduler. We chose the learning rate out of {5e-7, 1e-6, 5e-6, 1e-5, 5e-4} and found that 1e-5 works best for BERT and DeBERTa, and 1e-6 – for ELECTRA. In all experiments, the batch size was set to 8. The training time on one RTX 2080Ti GPU was 80–160 minutes, depending on the chosen setup (with or without context information).

A.1.4 Data and Models

All dataset extensions and trained models are available under the CC-BY-NC license.

[3]Huggingface transformers, `https://huggingface.co/transformers/pretrained_models.html`

	Predicted			
	Clarification	**Typo**	**Links**	
Clarification	5884 (.64)	2593 (.28)	709 (.08)	
Typo	2788 (.33)	5214 (.61)	483 (.06)	*True*
Links	1020 (.39)	544 (.21)	1067 (.41)	

Table A.2: Claim improvement suggestion: Confusion matrix obtained by FT-DeBERTa without using context.

A.2 Prediction Outputs

A.2.1 Suboptimal Claim Detection

Table A.1 provides examples of false negative and false positive predictions obtained by FT-DeBERTa (without considering context) illustrating common patterns found in the results.

A.2.2 Claim Improvement Suggestion

Table A.2 presents the confusion matrix of predictions made by FT-DeBERTa (without considering context) illustrating misclassification patterns found in the results.

Table A.3 provides examples of misclassifications obtained by the best performing model (FT-DeBERTa), illustrating cases where both the true class label and the predicted class label represent plausible revision type suggestions.

A.2.3 End-to-end Setup

Table A.4 provides extended performance results obtained by approaches using ELECTRA and DeBERTA in an end-to-end setup, where both optimal claim detection and improvement suggestion tasks are combined into one multiclass classification task with four classes: *optimal* (claim does not need revisions), needs *clarification*, needs *typo* and/or grammar correction, needs editing of *links*.

The results suggest that in such setup it is highly difficult to detect claims requiring clarification edits (F1-scores of 15.3 (*FT-DeBERTa* with parent) and 1.5 (*FT-ELECTRA* with parent). Such low scores can be partially explained by (a) the high diversity of changes included in the class compared to *typo* and *links* classes, (b) the high imbalance of the data (percentage of samples per class: clarification (18%), typo (17%), links (5%), and optimal (60%)).

Table 3.6 emphasizes the general benefit of utilizing contextual information, however, similar to the results obtained in the task of claim improvement

Claim	True Label	Predicted Label
Freedom of speech is exceptionally good in the US, despite a recent decline in its acceptance	clarif	links
Muslim women must remove their burkas for their driver's license.	clarif	links
Voluntary help is beneficial to Germany	clarif	gram
indecent exposure violated the right of free expression, and is therefore an illegal law.	clarif	gram
Public restrooms should be gender neutral.	clarif	gram
Not all platforms aid terrorists' cause. Those who do not will not be censored or shut down.	typo	clarif
The use of nuclear weapons was required in order to end the Pacific War between the US and Japan.	typo	clarif
Nuclear weapons have spread to politically unstable states, for example Pakistan which experienced stagflation during the 1990s, a military coup in 1999 as well as a unsuccessful coup attempt in 1995.	typo	links
Many of the animals are now extinct, such as mammoths, mastodons, aurochs, cave bears ect.	typo	links
For example, the one who will have more than one wife, should equally treat all his wives.[Link](http://islamqa.info/en/14022)	links	clarif
Before the nuclear bombs were dropped 70% of suitable targets had already been completely destroyed by conventional bombing.	links	typo
For the Spanish bullfighting is a way to reconnect to old, traditional and great Spain and therefore a major source of identity.	links	gram
DDOS attacks are the online equivalent of a sit-in.	links	clarif

Table A.3: Examples of misclassifications obtained by TF-DeBERTa (without considering context).

Setup	Accuracy	Ma. F_1	F_1-Score Clarif.	Typo	Links	Optimal
FT-ELECTRA	62.7	32.4	0.0	33.6	19.1	76.8
+ parent	62.9	33.0	0.0	33.5	21.2	77.1
+ thesis	63.3	34.1	1.5	36.8	20.7	77.3
FT-DeBERTa	64.2	39.8	9.4	43.0	28.9	78.0
+ parent	64.8	40.3	9.1	45.4	28.1	78.5
+ thesis	**65.5**	**42.7**	**15.3**	47.0	**29.6**	78.8
Random baseline	25.0	21.1	20.8	19.8	8.4	35.5

Table A.4: Combining Improvement Suggestion and Optimal Claim Detection: Accuracy, macro F_1-score, and the F_1-score per revision type for ELECTRA+SVM and FT-DeBERTa with and without considering context, averaged over five runs.

suggestion, depending on the specific revision type, the addition of contextual information can both raise and decrease performance. Particularly, we observe decreased performance in *FT-DeBERTa* when detecting *clarifications* and *link* corrections while considering the parent claim as context. On the other hand, in the case of *typo/grammar* and *optimal* claims, both types of contextual information lead to increased F_1-scores. Generally, we notice that similar to the task of claim improvement suggestion, providing the main thesis of the debate leads to higher score improvements overall.

As indicated previously, further defining and disentangling revision types along with their relationships to contextual information could further benefit not only our understanding of revision processes in argumentative texts and their relationship to quality, but also help overcome modeling limitations identified in this paper.

A.3 Figures

A.3.1 Topical Categories

Figure A.1 depicts the relationship between how represented the topical category is in the corpus and the achieved prediction accuracy by FT-ELECTRA in the cross-category setting using a leave-one-out-strategy.

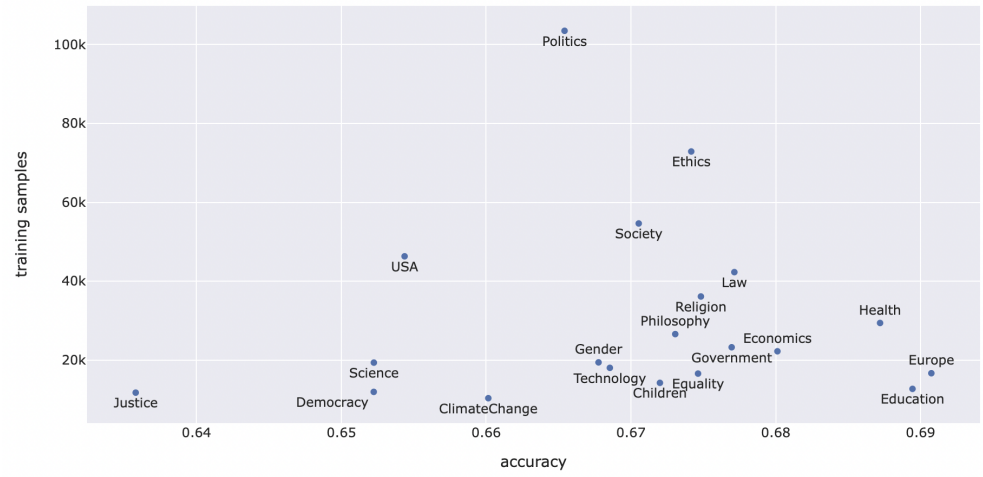


Figure A.1: Scatter plot of training sample size vs. accuracy for 20 topical categories of the extended ClaimRev corpus achieved by FT-DeBERTa in the cross-category setting.

B. Experimental Details for Chapter 5

B.1 Implementation and Training Details

B.1.1 BART-based models

For generation, we use the pre-trained BART model implemented in the fairseq library. The library and pre-trained models are BSD-licensed. We use the BART-large checkpoint (400M parameters) and further finetune the model for 10 epochs on 2 RTX 2080Ti GPUs. We use the same parameters as suggested in the fine-tuning of BART for the CNN-DM summarization task by fairseq and set MAX-TOKENS to 1024. The training time is 100-140 minutes, depending on the chosen setup (with or without context information).

During inference, we generate candidates using a top-k random sampling scheme (Fan et al., 2018) with the following parameters: length penalty is set to 1.0, n-grams of size 3 can only be repeated once, temperature is set to 0.7, while the minimum and maximum length of the sequence to be generated are 7 and 256 accordingly.

B.1.2 BERT-based models

For the automatic assessment of fluency and argument quality, we use the bert-base-cased pre-trained BERT version, as implemented in the huggingface library. The library and pre-trained models have the Apache License 2.0. We finetune the model for two epochs and use the parameters suggested in Skitalinskaya et al. (2021). The accuracy of the trained model for fluency obtained on the train/dev/test split suggested by the authors (Toutanova et al., 2016) is 77.4 and 75.5 for argument quality.

For labeling the missing or unassigned revision types, we use the same bert-base-cased pre-trained BERT model, but in a multi-label setup, where

Model	Strategy	BLEU	RouL	SARI	NoEd↓	ExM
BART	Top-1	64.0	0.83	39.7	0.31	7.8%
	Random	62.6	0.83	38.7	0.28	6.8%
	SVMRank	55.7	0.76	38.8	0.03	4.5%
	AutoScore	59.4	0.80	43.7	0.02	8.3%
Trans-	Top-1	43.6	0.64	0.30	0.12	0.8%
former	Random	42.4	0.63	0.30	0.13	1.0%
	SVMRank	41.8	0.63	0.31	0.10	1.2%
	AutoScore	40.5	0.62	0.30	0.10	1.3%
LSTM	Top-1	36.2	0.56	0.28	0.10	0.3%
	Random	36.0	0.56	0.28	0.10	0.3%
	SVMRank	36.2	0.56	0.29	0.10	1.0%
	AutoScore	34.1	0.52	0.28	0.10	1.0%

Table B.1: Automatic evaluation: Results for each combination of generation model and candidate selection strategy on the 600 test samples, in comparison to the human revisions: BLEU (0-100), ROUGE-L (RouL), SARI, ratio of unedited samples (NoEd), % of exact matches to target reference (ExM).

we consider the following 6 classes: claim clarification, typo or grammar correction, correcting or adding links, changing the meaning of the claim, splitting the claim, and merging claims. We fine-tune the model for two epochs using the Adam optimizer with a learning rate of 1e-5 and achieve a weighted F1-score of 0.81.

B.2 Alternative Generation Models

For comparison, we provide two additional baseline Seq2Seq model architectures, which help identify the complexity of the model needed for the task:

LSTM. Our first baseline is a popular LSTM variant introduced by Wiseman and Rush (2016). We use the *lstm_wiseman_iwslt_de_e* architecture, which is a two-layer encoder and decoder LSTM, each with 256 hidden units, and dropout with a rate of 0.1 between LSTM layers.

Transformer. The second model is based on the work of Vaswani et al. (2017). We use the *transformer_iwslt_de_en* architecture, a 6-layer encoder and decoder with 512-dimensional embeddings, 1024 for inner-layers, and four self-attention heads.

Tables B.1 and B.2 compare the automatic evaluation scores of all generation-content selection combinations.

Model	Strategy	Fluency	Meaning	Argument	Average
BART	Top-1	0.73	0.97	0.65	0.78
	Random	0.72	0.97	0.68	0.79
	SVMRank	0.72	0.94	0.76	0.81
	AutoScore	0.83	0.95	0.86	**0.88**
Trans-	Top-1	0.44	0.76	0.40	0.53
former	Random	0.41	0.76	0.38	0.52
	SVMRank	0.50	0.76	0.45	0.57
	AutoScore	0.68	0.75	0.61	**0.68**
LSTM	Top-1	0.27	0.68	0.31	0.42
	Random	0.27	0.68	0.31	0.42
	SVMRank	0.29	0.69	0.31	0.43
	AutoScore	0.52	0.65	0.53	**0.57**
Human		0.72	0.94	0.74	0.80

Table B.2: Results for each combination of generation model and candidate selection strategy on the 600 test samples, in comparison to the human revisions based on three quality metrics: fluency, meaning preservation and argument quality.

B.3 Automatic Evaluation

We use the following python packages and scripts to perform automatic evaluations: nltk (BLEU (Papineni et al., 2002)), rouge-score (ROUGE (Lin, 2004)), https://github.com/cocoxu/simplification/ SARI.py (SARI (Xu et al., 2016))

B.4 Claim Optimization Examples

For all eight optimization categories, we provide one or more examples illustrating each action in Table B.3.

B.5 Manual Quality Assessment Guidelines

Figure B.1 shows the annotation guidelines for the Amazon Mechanical Turk study.

B.6 System Outputs

Table B.4 provides examples of candidates selected by different content selection strategies along with human references illustrating common patterns found in the results. Table B.5 provides examples of candidates generated with and without utilizing context knowledge with insertions and deletions being highlighted in green and red fonts accordingly.

Type	Examples
Specification	Nipples are the openings of female-only exocrene glands that can have abnormal [secretions] <LINK> during any time of life, get erected by cold stimulation or sexual excitement (much more visibly than in men), get lumps or bumps and change color and size of areola during the menstrual cycle or pregnancy, so their display can break [personal space] <LINK> and privacy (which is stressful), affect public sensibilities and also be a [window] <LINK> for infections, allergies, and irritation. The idea behind laws, such as limiting the amount of guns, is to reduce the need to defend yourself from a gun or rapist. It is very common for governments to actively make certain forms of healthcare [harder for minority groups to access] <LINK>. They could also, therefore, make cloning technology hard to access.
Simplification	Very complex~~, cognitively meaningful behavior such as~~ behaviours like creating art are evidence of free will~~, because they exhibit the same lack of predictability as stochastic systems, but are intelligible and articulate clearly via recognizable vehicles~~.
Reframing	It reduces the oversight of the BaFin and thus increases ~~the risk of financial crisis~~ market failures.
Elaboration	It takes 2-4 weeks for HIV to present any symptom. The incubation period risk ~~can't be ruled out for~~ is higher for a member of high risk group, ~~effectively and timely~~ even though member of a low risk group is not completely safe. The decision is based on the overall risk, not on individual level.
Corroboration	[Person-based predictive policing technologies] <LINK> - that focus on predicting who is likely to commit crime rather than where is it likely to occur - violate the [presumption of innocence.] <LINK>.
Neutralization	Biden ~~does not~~ lacks the support ~~or agree with several key issues that are important to liberal voters.~~ of many liberal voting groups due to his stance on key issues concerning them.
Disambiguation	The USSR had [passed legislation] <LINK> to gradually eliminate religious belief within its borders. However the death penalty was more used in USSR than in Russia. ~~It~~ USSR had 2000 [death penalties] <LINK> per year in the 1980s whereas pre USSR Russia had [banned the death penalty] <LINK> in 1917 and almost never carried it out in the decades before that. ~~SRM~~ Solar geoengineering merely serves as a "technological fix" (Weinberg).[harvard.edu] <LINK>
Copy Editing	Women are experiencing record ~~level~~ levels of success in primaries.

Table B.3: Illustrative examples of optimization types identified in the paper. The green font denotes additions and the striked out red font denotes the removal of text snippets.

Instructions

In this task, your goal is to identify whether a claim has been successfully improved, without changing the overall meaning of the text.
Each task contains a set of pairs, where one claim is the "original claim," and the other an optimized candidate. Each of these pairs have the same original text, but different candidate optimizations.
Please rate each candidate along the following three perspectives: argument quality, fluency and semantic similarity. And, finally, please, rank all candidates relative to each other in terms of overall quality.

Argument Quality

Scale (1-5): 1 (notably worse than original), 2 (slightly worse), 3 (same as original), 4 (slightly improved), 5 (notably improved)

Does the optimized claim improve the argument quality compared to the original claim? Relevant changes include, but are not limited to:

- further specifying or explaining an existing fact or meaning
- removing information or simplifying the sentence structure with the intent to reduce the complexity or breadth of the claim
- rephrasing a claim with the intent to specify or generalize the claim, or to add clarity
- adding (substantive) new content or information to the claim or inserting an additional fact with the intent of making it more self-contained, more sound or stronger
- adding, editing or removing evidence in the form of links that provide supporting information or external resources to the claim
- removal of bias or biased language
- removal uncertainty. e.g. by replacing pronouns referring to concepts that have been mentioned in other claims of the debate, or by replacing acronyms with what they stand for
- improving the grammar, spelling, tone, or punctuation of a claim

Meaning

Scale (1-5): 1 (entirely different), 2 (substantial differences), 3 (moderate differences), 4 (minor differences), 5 (identical)

Does the transformed claim still have the same overall meaning as the original? It is OK if extra information is added, as long as it doesn't change the underlying people, events, and objects described in the sentence. You should also not strongly penalize for meaning transformations which aim to generalize or specify some aspects of the claim.

Fluency

Scale (1-3): 1 (major errors, disfluent), 2 (minor errors), 3 (fluent)

Is this sentence fluent English and does it make sense?

Examples with explanations can be found here.

Figure B.1: Guidelines provided to the crowdworkers to annotate automatically generated claim optimizations.

Strategy	Candidate
Original	Easy to cheat, and procastinate
Top-1	It is easy to cheat and procastinate .
SVMRank	The easy to cheat, and procastinate
AutoScore	The UBI is easy to cheat and easy to manipulate.
Human	Flexible schedules can easily increase procrastination
Original	Many games often attempt to depict the norms and attitudes of the real world.
Top-1	Many games often attempt to depict the norms and attitudes of the real world.
SVMRank	Many games often attempt to depict the norms and attitudes of the real world, and thus are likely to be accurate and relevant to what people are thinking and feeling in the real-world.
AutoScore	Many games often attempt to depict the norms and attitudes of the real world, in order to give players a greater understanding of the world around them.
Human	Many games often attempt to depict the norms and attitudes of the real world; gamers are likely to see social attitudes espoused in these games on issues of gender equality as indicative of progress in the real world.
Original	Owners of PWA take home [more profit] <LINK> from their sales than mobile apps.
Top-1	Owners of PWA take home [more profit] <LINK> from their sales than mobile apps.
SVMRank	Owners of PWA take home [more profit] <LINK> from their sales than mobile apps. This allows them to reinvest in their business and increase their profits.
AutoScore	Owners of PWA take home [more profit] <LINK> from their sales than those of mobile apps.
Human	Owners of PWAs take home [more profit] <LINK> from their sales than mobile apps.
Original	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that.
Top-1	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that.
SVMRank	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that. Therefore, people are not responsible for their actions, and are therefore not morally responsible.
AutoScore	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that, as it is not a conscious decision but rather a result of an involuntary reaction.
Human	The coercion to vote eliminates the idea of responsibility, which requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that.
Original	There was much progress, advancement and forward looking thinking in the 20th century in Europe across all disciplines.
Top-1	There was much progress, advancement and forward looking thinking in the 20th century in Europe across all disciplines. Therefore, it is logical to conclude that the world has changed in the last 150 years.
SVMRank	There was much progress, advancement and forward looking thinking in the 20th century in Europe across all disciplines, which shows the progress that was made during the century.
AutoScore	There was much progress, advancement, and forward looking thinking in the 20th century in Europe across all disciplines.
Human	There was much progress, advancement and forward ~~looking~~ thinking in the 20th century in Europe across all disciplines.
Original	A [catch-22] <LINK> situation currently exists in regards to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>).
SVMRank (Top-1)	A [catch-22] <LINK> situation currently exists in regards to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>). This prevents researchers from knowing how to fully assess the medicinal potential of drugs in a manner that is best suited for the patients who need it.
AutoScore	A [catch-22] <LINK> situation currently exists in regards to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>). This prevents researchers from researching possible medicinal uses for illegal drugs.
Human	A [catch-22] <LINK> situation currently exists in regard to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>).

Table B.4: Examples of optimizations selected by different candidate selection strategies along with human references. The green font denotes additions and the striked out red font denotes the removal of text snippets.

Control	Candidate
Topic	Should Men Be Involved And Trusted In Childcare?
Previous	Perpetrators of child abuse are [way more likely] <LINK> to be the child's parents.
Original	Mostly the mothers, too.
Claim only	The mothers are the most vulnerable.
+ Previous	The majority of perpetrators are the mothers, too.
+ Topic	It is mostly mothers who are responsible for childcare.
Topic	Would Trump or Biden be a better President?
Previous	The [Healthy Indiana Plan 2.0] <LINK>, initiated by Mike Pence, has provided access to healthcare to many people who could not previously afford it.
Original	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK>.
Claim only	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK>. This means that people with low income could be penalized for failure to pay.
+ Previous	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK> in exchange for providing healthcare.
+ Topic	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK>,which were imposed in 2015.
Topic	Artificial Intelligence (AI): Should an artificial general intelligence be created?
Previous	There is the chance an AGI might be able to think for itself and turn on its creators.
Original	AGI are susceptable.
Claim only	Many AGI are susceptable to the danger of death.
+ Previos	AGI are susceptible. They may be influenced by humans and by the biases of their creators.
+ Topic	AGI are susceptibleto being hacked.
Topic	Who should the Conservative Party choose to be the next UK Prime Minister?
Previous	[Jeremy Hunt] <LINK>
Original	Hunt is the best placed candidate to unify Parliament in order to get a Brexit deal passed.
Claim only	Hunt is the best placed candidate to unify Parliament and the public in order to get a Brexit deal passed.
+ Previous	Hunt could use his position as the Prime Minister to unify Parliament in order to get a Brexit deal passed.
+ Topic	Hunt is the best placed candidate to [unify Parliament] <LINK> in order to get a Brexit deal passed
Topic	Should high-income countries take in refugees?
Previous	Taking in refugees will increase criminality within host countries.
Original	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities.
Claim only	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities, making them more likely to get involved in crime.
+ Previous	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities. This will not increase criminality.
+ Topic	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities, which make it harder to find a job.
Topic	Mark Twain used the N-word in The Adventures of Huckleberry Finn. Should it be censored?
Previous	Changing the N-word would skip a piece of the *linguistic* past and thus everyday life. As a result, people could start to forget this part of history.
Original	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>".
Claim only	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>".This is a common trait of the American English language.
+ Previous	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>"by using the N-word in everyday conversation.
+ Topic	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>", which is a language that is often used by people who do not share his values.

Table B.5: Examples of different candidates generated by BART + AutoScore with and without context information. The green font denotes additions of text snippets.

References

Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective interfaces for student-driven revision sessions for argumentative writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246, New Orleans, Louisiana. Association for Computational Linguistics.

Tazin Afrin and Diane Litman. 2023. Predicting desirable revisions of evidence and reasoning in argumentative writing. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2550–2561, Dubrovnik, Croatia. Association for Computational Linguistics.

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 719–724, Melbourne, Australia. Association for Computational Linguistics.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019a. Modeling frames in argumentation. 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 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019b. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, pages 48–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument generation. 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 4744–4754, Online. Association for Computational Linguistics.

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021a. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics.

Milad Alshomary, Timon Gurcke, Shahbaz Syed, Philipp Heinisch, Maximilian Spliethöver, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. 2021b. Key point analysis via contrastive learning and extractive argument summarization. In *Proceedings of the 8th Workshop on Argument Mining*, pages 184–189, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021c. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, Online. 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.

Milad Alshomary and Henning Wachsmuth. 2021. Toward audience-aware argument generation. *Patterns*, 2(6):100253.

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Talita Anthonio and Michael Roth. 2020. What can we learn from noun substitutions in revision histories? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1359–1370, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse*, 2 edition. Oxford University Press, New York and Oxford.

Sumit Asthana, Sabrina Tobar Thommel, Aaron Lee Halfaker, and Nikola Banovic. 2021. Automatically labeling low quality content on wikipedia by leveraging patterns in editing behaviors. 5(CSCW2).

Katie Atkinson, Trevor Bench-Capon, and Douglas Walton. 2013. Distinctive features of persuasion and deliberation dialogues. *Argument & Computation*, 4(2):105–127.

Betty Bamberg. 1978. Composition instruction does make a difference: A comparison of the high school preparation of college freshmen in regular and remedial english classes. *Research in the Teaching of English*, 12(1):47–59.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Charles Bazerman. 2016. What do sociocultural studies of writing tell us about learning to write.

Richard Beach. 1984. The effect of reading ability on seventh graders' narrative writing.

Richard Beach and Sara Eaton. 1984. Factors influencing self-assessing and revising by college freshmen. *New directions in composition research*, 1:149–189.

Isabel L Beck, Margaret G McKeown, Gale M Sinatra, and Jane A Loxterman. 1991. Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading research quarterly*, pages 251–276.

Jordan Beck, Bikalpa Neupane, and John M Carroll. 2019. Managing conflict in online debate communities. *First Monday*.

Carl Bereiter and Marlene Scardamalia. 1987. The psychology of written composition. the psychology of education and instruction series.

Irshad Bhat, Talita Anthonio, and Michael Roth. 2020. Towards modeling revision requirements in wikiHow instructions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8407–8414, Online. Association for Computational Linguistics.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Denver, CO. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany. Association for Computational Linguistics.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366, Avignon, France. Association for Computational Linguistics.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jodie A Butler and M Anne Britt. 2011. Investigating instruction for improving revision of argumentative essays. *Written Communication*, 28(1):70–96.

Ruth M. J. Byrne. 1989. Everyday reasoning with conditional sequences. *The Quarterly Journal of Experimental Psychology Section A*, 41(1):141–166.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. 2021. ENTRUST: Argument reframing with language models and entailment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4958–4971, Online. Association for Computational Linguistics.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Chunyang Chen, Xi Chen, Jiamou Sun, Zhenchang Xing, and Guoqiang Li. 2018a. Data-driven proactive policy assurance of post quality in community q&a sites. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018b. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88, Tilburg University, The Netherlands. Association for Computational Linguistics.

MT Cicero. 1903. De oratore (js watson, ed. and trans.). *London: George Bell.(Original work published 55 BC)*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2461–2505.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India. The COLING 2012 Organizing Committee.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, Washington, USA. Association for Computational Linguistics.

Christine De Kock and Andreas Vlachos. 2022. Leveraging Wikipedia article evolution for promotional tone detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5601–5613, Dublin, Ireland. Association for Computational Linguistics.

Alok Debnath and Michael Roth. 2021. A computational analysis of vagueness in revisions of instructional texts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 30–35, Online. Association for Computational Linguistics.

Pieter Delobelle, Murilo Cunha, Eric Massip Cano, Jeroen Peperkamp, and Bettina Berendt. 2019. Computational ad hominem detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 203–209, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019a. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019b. 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.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Michael Elhadad. 1995. Using argumentation in text generation. *Journal of pragmatics*, 24(1-2):189–220.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, 32(4):400–414.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text editing by command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5274, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ralph P Ferretti and William E Lewis. 2013. Best practices in teaching argumentative writing. *Best practices in writing instruction*, 2:113–140.

Jill Fitzgerald. 1987. Research on revision in writing. *Review of educational research*, 57(4):481–506.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

Linda Flower, John R Hayes, Linda Carey, Karen Schriver, and James Stratman. 1986. Detection, diagnosis, and the strategies of revision. *College composition and Communication*, 37(1):16–55.

Austin J Freeley and David L Steinberg. 2013. *Argumentation and debate*. Cengage Learning.

Michael Fromm, Max Berrendorf, Evgeniy Faerman, and Thomas Seidl. 2023. Cross-domain argument quality estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13435–13448, Toronto, Canada. Association for Computational Linguistics.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1).

A Gibson, Glenn Rowe, and Chris Reed. 2007. A computational approach to identifying formal fallacy. *CMNA VII-Computational Models of Natural Argument*.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Trudy Govier. 2013. *A practical study of argument*. Cengage Learning.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Zhen Guo and Munindar P Singh. 2023. Representing and determining argumentative relevance in online discussions: A general approach. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 292–302.

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. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Charles L Hamblin. 1970. Fallacies. *Tijdschrift Voor Filosofie*, 33(1).

Mogens Herman Hansen. 1999. *The Athenian democracy in the age of Demosthenes: structure, principles, and ideology*. University of Oklahoma Press.

Mogens Herman Hansen. 2005. *The tradition of ancient Greek democracy and its importance for modern democracy*, volume 93. Kgl. Danske Videnskabernes Selskab.

John R Hayes and Linda Flower. 1981. *Uncovering cognitive processes in writing: An introduction to protocol analysis*. ERIC Clearinghouse.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Christopher Hidey and Kathleen McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. 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)*, pages 1756–1767.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Oskar Holmström, Jenny Kunz, and Marco Kuhlmann. 2023. Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 92–110, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Eduard H Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial intelligence*, 63(1-2):341–385.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. 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 591–602, Hong Kong, China. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Ken Hyland and Fiona Hyland. 2019. Contexts and issues in feedback on l2 writing. *Feedback in second language writing: Contexts and issues*, pages 1–22.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA. Association for Computing Machinery.

Ralph H Johnson. 2012. *Manifest rationality: A pragmatic theory of argument*. Routledge.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. Idea.

Iman Jundi, Neele Falk, Eva Maria Vecchi, and Gabriella Lapesa. 2023. Node placement in argument maps: Modeling unidirectional relations in high & low-resource scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5854–5876, Toronto, Canada. Association for Computational Linguistics.

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. 2022. Argrewrite v. 2: an annotated argumentative revisions corpus. *Language Resources and Evaluation*, pages 1–35.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Ruth E Knudson. 1992. The development of written argumentation: An analysis and comparison of argumentative writing at four grade levels. *Child study journal*.

Christian Kock. 2007. Dialectical obligations in political debate. *Informal Logic*, 27(3):233–247.

Ravi Kondadadi, Blake Howald, and Frank Schilder. 2013. A statistical NLG framework for aggregated planning and realization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1406–1415, Sofia, Bulgaria. Association for Computational Linguistics.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Chih-Te Lai, Yi-Te Hong, Hong-You Chen, Chi-Jen Lu, and Shou-De Lin. 2019. Multiple text style transfer by using word-level conditional generative

adversarial network with two-phase training. 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 3579–3584, Hong Kong, China. Association for Computational Linguistics.

Walter Rangeley Maitland Lamb. 1925. Plato: Lysis, symposium, gorgias. *Trans.) Loeb Classical Library. Harvard University Press*.

Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. 2023. Mining, assessing, and improving arguments in NLP and the social sciences. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dubrovnik, Croatia. Association for Computational Linguistics.

Saeed Latifi, Omid Noroozi, Javad Hatami, and Harm J.A. Biemans. 2021. How does online peer feedback improve argumentative essay writing and learning? *Innovations in Education and Teaching International*, 58(2):195–206.

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.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Guo Li, Haiyi Zhu, Tun Lu, Xianghua Ding, and Ning Gu. 2015. Is it good to be like wikipedia? exploring the trade-offs of introducing collaborative editing model to q&a sites. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1080–1091, New York, NY, USA. Association for Computing Machinery.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. 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)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021. A transformer-based framework for neutralizing and reversing the political polarity of news articles. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura, and Richard Correnti. 2023. Predicting the quality of revisions in argumentative writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 275–287, Toronto, Canada. Association for Computational Linguistics.

David Lowenthal. 1980. Mixing levels of revision. *Visible language*, 14(4):383–87.

R Duncan Luce. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bimal Krishna Matilal, Jonardon Ganeri, and Heeraman Tiwari. 1999. *Character of Logic in India, The*. State University of New York Press.

Bryan M. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Thomas M. Mitchell. 1997. *Machine Learning*, 1 edition. McGraw-Hill, Inc., USA.

Tomoya Mizumoto and Yuji Matsumoto. 2016. Discriminative reranking for grammatical error correction with statistical machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1133–1138, San Diego, California. Association for Computational Linguistics.

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544. PMLR.

Donald J Munro. 1985. Individualism and holism: Studies in confucian and taoist values.

Donald M Murray. 1978. Internal revision: A process of discovery. *Research on composing: Points of departure*, pages 85–103.

Debra Myhill and Susan Jones. 2007. More than just error correction: Students' perspectives on their revision processes during writing. *Written communication*, 24(4):323–343.

Callistus Ireneous Nakpih and Simone Santini. 2020. Automated discovery of logical fallacies in legal argumentation. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 11.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Steve Oswald. 2011. From interpretation to consent: Arguments, beliefs and meaning. *Discourse Studies*, 13(6):806–814.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

ChaeHun Park, Wonsuk Yang, and Jong Park. 2019. Generating sentential arguments from diverse perspectives on controversial topic. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 56–65, Hong Kong, China. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, dg.o '12, pages 173–182, New York, NY, USA. Association for Computing Machinery.

Dipti Pawade, A Sakhapara, Mansi Jain, Neha Jain, and Krushi Gada. 2018. Story scrambler-automatic text generation using word level rnn-lstm. *International Journal of Information Technology and Computer Science (IJITCS)*, 10(6):44–53.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Chaim Perelman. 1971. The new rhetoric. In *Pragmatics of natural languages*, pages 145–149. Springer.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long Papers*), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP* (*RepL4NLP-2019*), pages 7–14, Florence, Italy. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

HG Plato. 1961. Phaedrus. In *The Collected Dialogues of Plato*, pages 475–525. Princeton University Press.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017. Length, interchangeability, and external knowledge: Observations from predicting

argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Peter Potash, Adam Ferguson, and Timothy J. Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*, pages 146–155, Florence, Italy. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. 2015. Incorporating coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 20–30, Denver, Colorado. Association for Computational Linguistics.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: A dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524, Marseille, France. European Language Resources Association.

Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *The World Wide Web Conference*, WWW '19, pages 1567–1578, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ehud Reiter, Roma Robertson, and Liesl Osman. 2000. Knowledge acquisition for natural language generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 217–224.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Stuart J Russell. 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. 2017. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*.

Marlene Scardamalia. 1983. The development of evaluative, diagnostic, and remedial capabilities in children's composing. *The psychology of written language: A developmental approach*, pages 67–95.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.

Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. 2023. Claim optimization in computational argumentation. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 134–152, Prague, Czechia. Association for Computational Linguistics.

Gabriella Skitalinskaya and Henning Wachsmuth. 2023. To revise or not to revise: Learning to detect improvable claims for argumentative writing support. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15799–15816, Toronto, Canada. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Esther Abraham Solomon. 1976. Indian dialectics: Methods of philosophical discussion.

Nancy Sommers. 1980. Revision strategies of student writers and experienced adult writers. *College composition and communication*, 31(4):378–388.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Maja Stahl, Maximilian Spliethöver, and Henning Wachsmuth. 2022. To prefer or to choose? generating agency and power counterfactuals jointly for gender bias mitigation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 39–51, Abu Dhabi, UAE. Association for Computational Linguistics.

Charles Kelson Stallard Jr. 1972. *An analysis of the writing behavior of good student writers*. University of Virginia.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. 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 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models.

Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. 2020. News editorials: Towards summarizing long argumentative texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5384–5396, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Christopher W Tindale. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. 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 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Stephen E Toulmin. 1958. The uses of argument. *Philosophy*, 34(130).

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Frans H Van Eemeren. 2015. Reasonableness and effectiveness in argumentative discourse. *Dordrecht: Springer*, 10:978–3.

Frans H Van Eemeren and Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation*, 1:283–301.

Frans H Van Eemeren and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. 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.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 241–251. Association for Computational Linguistics.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 515–522, Toulouse, France. Association for Computational Linguistics.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. 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 4063–4077, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pages 195–200, Berlin, Germany. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa

Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Axel Westerwick, Benjamin K. Johnson, and Silvia Knobloch-Westerwick. 2017. Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs*, 84(3):343–364.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.

Christopher R Wolfe, M Anne Britt, and Jodie A Butler. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Wonsuk Yang, Jung-Ho Kim, Seungwon Yoon, ChaeHun Park, and Jong C Park. 2019a. A corpus of sentence-level annotations of local acceptability with reasons. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 288–297. Waseda Institute for the Study of Language and Information.

Wonsuk Yang, Seungwon Yoon, Ada Carpenter, and Jong Park. 2019b. Nonsense!: Quality control via two-step reason selection for annotating local

acceptability and related attributes in news editorials. 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 2954–2963, Hong Kong, China. Association for Computational Linguistics.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430, San Diego, California. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Summit: Iterative text summarization via chatgpt.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Yilun Zhou, Julie Shah, and Steven Schockaert. 2019. Learning household task knowledge from WikiHow descriptions. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 50–56, Macau, China. Association for Computational Linguistics.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 55–62.