Institut für
Maschinelle
Sprachverarbeitung

Universität Stuttgart
Pfaffenwaldring 5b
70569 Stuttgart

Master Thesis

What Makes a Good Argument?

Investigating subjective factors of argument strength

Author:	Carlotta Quensel[1]
Examiners:	Dr. Gabriella Lapesa
	Dr. Roman Klinger
Supervisor	Neele Falk

Start date:	01.05.2023
End date:	01.11.2023

[1] st174674@stud.uni-stuttgart.de

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.

Statement of Authorship

This thesis is the result of my own independent work, and any material from work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

(Carlotta Quensel)

Abstract

Argument quality assessment is a field of computational argument mining, in which the quality or strength of persuasive texts is rated automatically. The notions of what makes a good argument are manifold. Historically, argument quality pertained mostly to objective markers like clarity, logical soundness or coherence. As the field shifts to address subjectivity and persuasion, the definition of argument strength also broadens to include persuasiveness and the subjective complexities this shift brings with it. While many small studies on subjective features of arguments exist, there are no large-scale analyses of the relation between these features and argument strength. To address this gap, I model the influence of three subjective features on argument quality data from differently focused domains. My contribution is twofold: first, I conduct a regression analysis on argument strength with the features of storytelling, emotions and hedging, which argument research either approached onesidedly or only recently. Secondly, as there are no datasets available with annotations for all four dimensions, I compare different methods for automatically annotating argument data with labels for storytelling, emotions and hedging. My analysis shows a link between the features and argument strength as well as systematic differences between the two argument corpora. In evaluating different automatic annotation methods, I find advantages of modified training setups but also see some limitations in how far automatic methods reach for complex tasks like cross-domain emotion classification.

Contents

1 Introduction

We encounter arguments and persuasive rhetoric in many facets of everyday life, be it in discussions on social media or under news articles, in political speeches and citizen participation programs, in research articles arguing their contributions, in professional debates and talk shows or even in our private discussions and advertisements. In its core, argumentation has two sides, *reason-giving*, and *persuasion* and the analysis of rhetoric strategies goes back to antiquity (Aristotle, 2007). Since then, research on arguments is conducted in various social sciences and in Natural Language Processing (NLP) and with this, the notion of what makes a good argument is thus as varied as there are domains of argumentation and argumentation research.

In NLP, the investigations of natural language arguments are traditionally engaged in *argument mining*, the field of detecting arguments and their components, i.e., claims and their premises, and analyzing relationships like support and attack between those (Lawrence and Reed, 2019). For a long time, this concept of arguments as purely logical reasoning, and the preoccupation with domains like academic writing, student essays or professional debate, meant, that argument quality is conceived only in dimensions where it relates to reason-giving. Quality assessment as emerged from argument mining observes features such as clarity, evidence support, soundness or logical organization. Only recently, as political participation grows more popular with the possibility of digital deliberation, the NLP research on arguments opens itself to notions of persuasiveness and less objective markers of quality adopted from deliberative research and social sciences.

In deliberative politics, ideal political decision-making includes deliberation, i.e., the mutual exchange of opinions and thoughts to weigh multiple options in an open dialogue (Steenbergen et al., 2003). In this vein, digital deliberation projects, like the public dialogue on how to shape the future uses of the old Berlin Tempelhof airport (Liebeck et al., 2016), are less focused on persuasion in either direction but rather the process of gathering opinions. The strength of an argument in deliberative theory is therefore related more to its contribution to a constructive discourse than

any standalone quality score. As such, the discourse quality index (DQI, Steenbergen et al., 2003) measures dimensions like respect for other participants, valuing counter-arguments or uninterrupted discourse contributions. In deliberative research, many more complex factors that might influence the discourse quality have been studied, like storytelling, prior beliefs or socio-economic biases (cf. e.g., Black, 2008; Gerber et al., 2018).

In a recent shift towards subjectivity in both deliberative research and argument mining, more and more alternative ways of argumentation are analyzed for their effect on the discourse overall and persuasiveness or argument strength in particular (e.g., emotions and personality, Benlamine et al., 2015; 2017; Villata et al., 2017). With this diversification of the research space however comes a complication of definitions, as argument strength is defined pertaining to objective quality, discourse contribution or persuasiveness in different works. Only recently were efforts made to consolidate this diversity of concepts (Wachsmuth et al., 2017). This leads to a big number of analyses either accompanied by only a small dataset annotated for their own newly investigated feature(s) or without any accompanying data at all, reporting only a case study of a particular issue forum. This leaves a gap between the subjective and alternative argument features and the big datasets available for NLP research, as the latter, if at all annotated for argument strength, often include no other annotations, hindering large-scale research.

Even though these large corpora all belong in the NLP area of argumentation research, their argument strength annotations are diverse: on the one hand, argument quality scores are aggregated from multiple, crowdsourced judgements on arguments especially generated (e.g. IBM ArgQ, Toledo et al., 2019) or collected from debate portals (e.g., Habernal and Gurevych, 2016). On the other hand, some corpora come with persuasiveness labels intrinsic to the data source, like the *delta*-point an initiator of a debate on the Reddit forum *ChangeMyView* can award to posts that persuaded them (Tan et al., 2016) or the change in audience vote numbers for either side of a debate from start to finish on the online platform *debate.org* (Durmus and Cardie, 2018). Fromm et al. (2022) showed that argument quality notions are not contradictory between corpora, i.e., one classifier can reliably predict

argument quality in multiple corpora. But, while not contradictory, the different notions are also not equal, with cross-domain experiments displaying a considerable performance drop. With the current state of the argument mining field however, there is no way to reliably test if the reason for this behavior is related to domain differences in argument strength concepts. Many small case studies of one or more subjective features on a specific domain are contrasted by big corpora annotated for only argument strength, often with no explicit description of the annotation's conceptualization. With this background, I aim to address two problems.

RQ1 Firstly, there is currently no large-scale study of the influence of subjective features on argument strength or any comparative study on the on similarities and differences of how subjective factors interact with argument strength in different domains. To address this gap, I am investigating three features that were neglected in computational argument mining until recently: storytelling (Falk and Lapesa, 2022), emotions (Maia and Hauber, 2020) and hedging (Chatterjee et al., 2014). This investigation is carried out on two corpora annotated with different measures of argument strength, IBM ArgQ (Toledo et al., 2019) with debate-oriented argument collection and an aggregated quality annotation, and Cornell CMV (Tan et al., 2016) with posts from the Subreddit forum *ChangeMyView* annotated for persuasiveness by tracking the *delta*-points given to posts by the single initiator of a debate. Employing regression analysis on the argument strength annotations, I aim to answer:

> *Do storytelling, emotions and hedging influence argument strength both individually and in combination and is this influence contingent on the argument domain?*

RQ2 Secondly, the leap from small case studies to such large-scale investigations previously suffered from the bottleneck of missing annotations. There is no argument dataset annotated for all of these features and manual annotation processes are complex, resource-intensive endeavors. Thus, I am relying on established methods for automatic annotation, which I adopt from the original domain (e.g., emotion

analysis) to automatically generate annotations for the two argument corpora. As automatic annotation is error-prone especially in such cross-domain settings, I am including different approaches for each feature to answer:

> *Is it possible to employ automatic annotation and obtain meaningful results usable in the downstream regression analysis and which choice in annotation methods generalizes better to the argument domain?*

To address these questions, I am first giving an overview of related works on argument strength (section 2.1) to motivate my choice of features, to then explain the theoretical and methodological background of the subjective features (section 2.2). I am then describing my methods in aggregating the data for the analysis and the different variants I employ to ease the cross-domain annotation (section 3). This allows me to analyze (section 4) the impact of the subjective features by regressing all features individually and in combined models on the argument strength measure of both IBM ArgQ and Cornell CMV. Finally, I am discussing the results of both the automatic annotation and the regression with regard to the research questions (section 5) and drawing conclusions on the study and possible avenues for future works (section 6).

2 Related Works

2.1 Argument strength

In argument mining, arguments are classified as any persuasive text and split into components often called elementary units (EU) consisting of a claim, that is, an assertion taking stance on a certain topic, and a premise with supporting evidence (Toulmin, 2003). Different works however disagree on what constitutes a claim and what counts as a premise, and focus on a number of diverse argument components and questions. Leading from this, what makes a good argument is a complicated question without a unified answer in different works on argument quality, which necessitates a comprehensive review of diverging definitions.

Following the Aristotelian theory (Aristotle, 2007), persuasive arguments can be classified based on their rhetoric strategy: an argument may rely on logical reasoning (*logos*), personal credibility (*ethos*), or emotions (*pathos*). Wachsmuth et al. (2017) summarize notions of argument quality into three different dimensions, namely logic (cogency), dialectic (reasonableness) and rhetoric (effectiveness), which include components like linguistic clarity, relevance, emotional appeal or logical soundness (figure 1). In both computational argument mining and deliberation research in the social sciences, argument quality has been traditionally understood in terms of the *logos* strategy (a brief sketch of argument conceptualization in deliberative research is included in section 2.2.1). This is partly due to the initial understanding of argumentation as reason-giving, which regards objectivity as a feature of good argumentation. Further, argument mining as a data-driven computational field evolved from tasks in other domains. While works on deliberative theory involved manual transcriptions and analysis of face-to-face discussions like political debates, this only allowed for smaller case studies in which argument quality is not at the forefront of the analysis.

Instead, the growing number of argumentative text in online forums starting around twenty years ago called for automatic identification, extraction and analysis of argument segments (Lawrence and Reed, 2019) in text. Early works adopted conflict detection in the collaborative production of Wikipedia articles (Kittur et al., 2007), classification of citations' argumentative function in scientific articles (Teufel et al., 2006) and argumentative zoning of articles (Teufel et al., 2009) into categories resembling claims (hypotheses, goals and conclusions) and evidence (support and findings). The early argument corpus AraucariaDB (Reed, 2006) similarly contains text from domains from newspaper editorials and advertising to parliamentary records and judicial summaries. A big part of argument mining research is concerned with automatic essay scoring, an active field of research since 1960 (Shermis and Burstein, 2003), which naturally evolved into a central domain in argument mining as student essays include argumentation on a hypothesis organized into supporting and opposing evidence. With these predominantly formal, impersonal textual domains and the assumption of argumentation as persuasive reason-giving, it

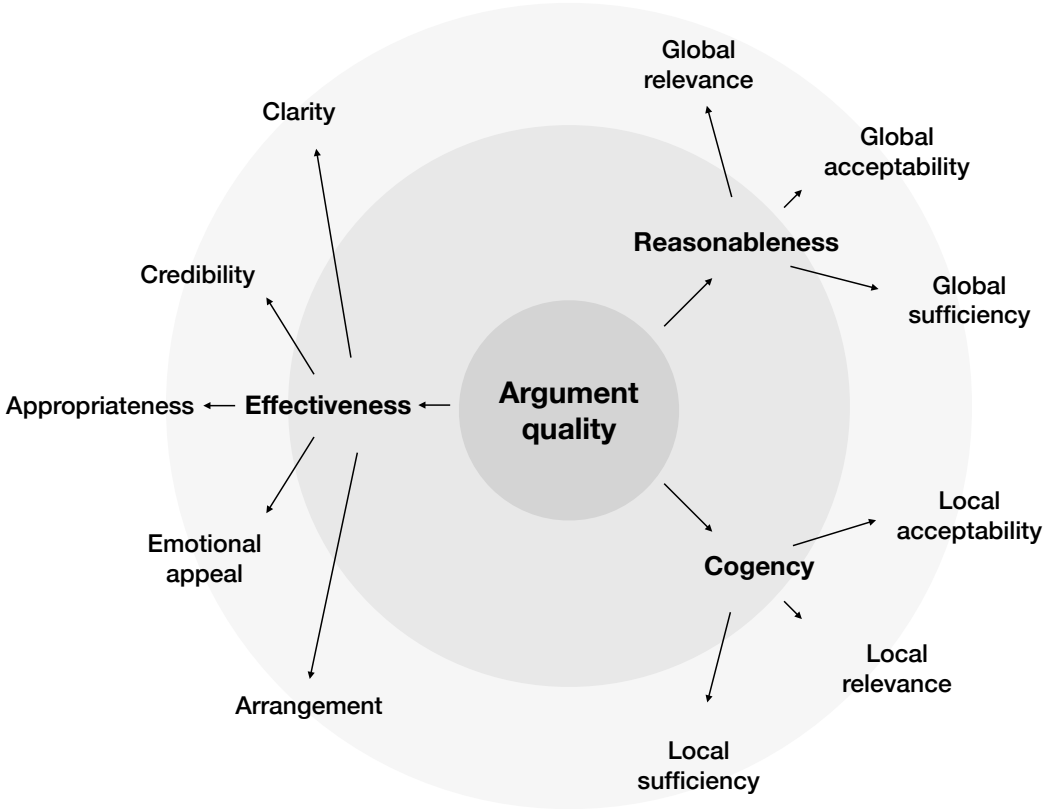


Figure 1: 15 hierarchically structured dimensions of argument quality in argument mining research adapted from Wachsmuth et al. (2017).

is evident how quality in early argument mining is contingent on the *logos* strategy.

As such, the quality definition and the argument domains are mutually dependent: impersonal argumentation will necessarily rely mostly on objective evidence and thus good arguments are those successfully using the *logos* strategy, and the automation of grading and feedback on student essays (Persing and Ng, 2013; 2014; Ong et al., 2014) necessitates a quality definition that complies with grading criteria. Thus, argument quality is equated with many of its components, with Persing et al. (2010) modeling argumentative structure and logical organization, Persing and Ng (2013) looking at linguistic clarity, and Rahimi et al. (2014) detecting the use of evidence. Ong et al. (2014) combine multiple features like grammar, flow and organization, the logic reasoning to arrive at hypotheses and the inclusion of opposing

claims into an automatic grading system. All these measures are certainly part of argument quality as they address intuitive markers of well-formed text in general – clear and correct language, sound reasoning and logical structure to follow. These markers however do not necessarily coincide with argument quality in terms of persuasiveness as Benlamine et al. (2017) show that the *logos* strategy is less effective than *pathos* based arguments in persuasive efforts during synchronous online discussions. When explicitly modeling persuasiveness or quality, corpora suffer from a similar bias. Assuming quality to be an objective, universal property renders a comprehensive definition unnecessary. Instead, letting multiple people judge argument quality is assumed to yield a universal result. When addressing argument quality feature-independent, such a universality of aggregated judgements makes detailed annotation guidelines unnecessary. In the creation of their large crowdsourced argument dataset IBM ArgQ, (Toledo et al., 2019; p. 5627) only include one sentence in their annotation guide: "Disregarding your own opinion on the topic, would you recommend a friend preparing a speech supporting/contesting the topic to use this argument as is in the speech?", relying on annotators' shared understanding and hierarchy of argument quality in debate speeches. They aggregate the binary judgements into scores denoting universal acceptability, and an argument with score .5 is interpreted as one of medium quality, though it could easily also be said to be a very divisive argument. Other corpora that address a particular marker of argument quality explicitly (rather than addressing the marker by itself and leaving the equation with quality implicit), do not distinguish between the quality marker and quality as a whole, e.g., while Swanson et al. (2015) explicitly state that they obtain argument quality annotation, their annotation guidelines explicitly ask crowdworkers to position a slider on a scale of how easily interpretable an argument is or if much context is needed. Even while addressing that "an exact definition of argument quality is potentially elusive" (Gretz et al., 2020; p. 7805), embracing the complexity and subjectivity of the issue is seemingly avoided by asking crowdworkers to rank two arguments in terms of convincingness instead of assigning point-wise scores (Swanson et al., 2015; Habernal and Gurevych, 2016; Toledo et al., 2019; Gretz et al., 2020), which improves inter-annotator agreement.

One holistic approach is the taxonomy of Wachsmuth et al. (2017) in figure 1. Trying to unify theoretical work on argument quality with the existing works in the NLP domain, the taxonomy includes 15 quality annotations, with overall quality split into three main components which all have three to five even finer quality dimensions. Together with the taxonomy, Wachsmuth et al. (2017) provided a small sample of 320 online debate portal arguments annotated for all 15 dimensions (Dagstuhl-15512-ArgQuality). When comparing the above datasets and assumptions to this approach, it becomes apparent that the objective focus results in disregarding a large area of what goes into argument quality.

Only recently did more subjective quality markers gain the attention of the argument mining field. The expansion of the available argument domains is one reason for this trend. With much more data from informal discussions in online forums, an insistence on objective, rigorously logical rhetoric as the sole dimension of argument quality becomes less and less viable. Corpora include forums dedicated to open debate like *RegulationRoom* or the Subreddit *ChangeMyView* and digital deliberation forums on local issues. While some works described below annotate argument quality similar to the above approaches, quality is often much more explicitly equated to persuasiveness. This is feasible because of built-in functions of discussion forums allowing for labels intrinsic to the data source, e.g., in *ChangeMyView*, debaters can award a *delta*-point to posts that changed their opinion, and extracting the points awarded by the original poster who started the discussion thread as done for Cornell CMV (Tan et al., 2016), depicts a genuine instance of a successful persuasion effort. A second reason for the shift is found in the adoption of methods and findings from deliberative research, which by that time upheld the significance of subjectivity in non-traditional arguments appealing to empathy and emotions through personal experiences (cf. Black and Lubensky, 2013; Maia and Hauber, 2020). This reason also brings with it a shift in the argument quality definition, as deliberative theory is often more concerned with how arguments contribute to a high-quality discussion marked by mutual, civil exchange and constructive contributions (Lapesa et al., 2023).

As *ethos* relates to personal credibility, argument quality dimensions associated

with *ethos* are harder to aggregate than those for *logos*, since they may need to include demographic information on the authors and annotators of the arguments, or, in the case of Wachsmuth et al. (2017) are time-intensive due to the fine granularity of annotation. Duthie et al. (2016) provide a scheme and dataset of ethos related characteristics in UK parliamentary debate, and other works include characteristics of argument author and audience. Lukin et al. (2017) tested subjects for personality traits and prior beliefs on socio-political issues to then provide them with either a curated monologue, a factual or an emotional argument and measured their belief-change. They showed that people open to experience were swayed by emotional arguments while agreeable personalities were convinced by factual arguments. In a similar vein, Durmus and Cardie (2018) explored the link between persuasiveness and voluntarily shared demographic factors of participants in an online debate forum, including age, gender, ethnicity, income level and education, political and religious ideology, and the president and political party they supported. They show that in religious debates, users change their opinion more readily, if the opinion comes from someone who shares their religious identity, which is mirrored in political debates for debaters and audiences with matching political ideology. Following the same line of research, Kiesel et al. (2022) developed a taxonomy and classification method for human values in arguments though did not interact these with argument quality in their small corresponding dataset. Wei et al. (2016) predict persuasiveness scores on the Reddit discussion forum *ChangeMyView* and uncover their correlation with metadata on the time and posting history revealing how argumentative features are more predictive of persuasiveness in the early discussion stages, while social interaction features, i.e., the attention a comment gets from other users is more informative late in the discussion. These works constitute a poignant example of the power of *ethos* in persuasion, with a shared identity and a debater's acquired reputation lending credibility to an argument. A linguistic feature of *ethos*-based arguments that has not received much attention in argument mining is the admission of uncertainty or hesitancy including hedge terms (e.g., *probably, I think*). Works include Chatterjee et al. (2014) who show that behaviors like hesitation are able to predict persuasiveness, though they do not mention the direction of this

influence and do not use textual features but rather paraverbal hesitation cues like filler words, silences or stuttering in video transcripts of movie reviews that they annotated for persuasiveness. In their corpus of mixed-domain discussions of education controversies, Habernal and Gurevych (2017) include hedging as one dimension they analyze in arguments categorized as making persuasive efforts or not. They show a biased distribution of hedges not occurring in heated discussions where each side is firm in their opinion, but rather in less black-and-white dialogues where participants are more ready to empathize and collaborate. Showing uncertainty might influence the *ethos* of an argument, either making the debater seem more approachable, self-reflected and thus trustworthy or conversely showing a lack of expertise and confidence. As verbal uncertainty may also indicate emotional uncertainty, hedging as an argument component bridges the gab between the two strategies of *ethos* and *pathos*.

A feature similarly connecting *pathos* and *ethos* is the inclusion of personal anecdotes and stories. Maia et al. (2020) assess how personal narratives influence the quality of the overall discussion using data from legislative public hearings and face-to-face discussions about the criminal responsibility age in Brazil. They argue that in natural discussions, logical reasoning is mixed with storytelling in complex ways, which mirrors Esau (2018) who assumes that argumentative language using *logos* is intuitively embedded in personal narratives to convey emotions and values. There are multiple investigations into the impact of emotions on individual argument strength and overall discourse quality. In the deliberative field, Maia and Hauber (2020) observe *anger, fear, indignation* and *compassion* in political discussions, showing how these emotions are distributed unevenly between different argument directions. The argument quality taxonomy and dataset by Wachsmuth et al. (2017) introduced above (figure 1) also includes emotional appeal in its 15 labels. Similarly, Fromm et al. (2022) conclude their work unifying argument quality corpora by modeling the impact of emotionality on argument quality. They find that emotions have either no significant impact or a slight negative effect in the case of the previously mentioned IBM ArgQ, though their annotation consist only of an emotion/no emotion label. There are multiple works combining sentiment analysis, i.e., detecting positivity

and negativity, with argument mining (Grosse et al., 2015; Stede, 2020; e.g.,), which however do not focus on argument quality. Furthermore, while these approaches investigate how emotion in general interacts with arguments, they view emotion one-dimensionally, distinguishing only positive and negative sentiment or even emotionality as a general concept. Conversely, after previously (Benlamine et al., 2015) showing the link between emotions and argumentation behavior, Benlamine et al. (2017) measure participants' mental workload, engagement and discrete emotions elicited during an online debate session with another participant using either a *logos*, *pathos* or *ethos* strategy. They show that while *logos* arguments require a higher mental load, the *pathos* strategy is most persuasive.

As shown above, the dimensions investigated in arguments are diverse, ranging in complexity and granularity between personal anecdotes, emotional appeal (Maia and Hauber, 2020; Benlamine et al., 2015), participant characteristics (Al Khatib et al., 2020; Lukin et al., 2017; Durmus and Cardie, 2018) and discourse context (Durmus et al., 2019; Luu et al., 2019). This diversity is a testimony to the complexity of modeling argument quality, but also hinders systematic comparisons between different proposed markers of argument quality.

Corpora with theory-based argument quality annotations like Wachsmuth et al. (2017) are costly to create because of the high complexity and subjectivity of the task, especially when trying to create resources for large-scale analyses unlike Wachsmuth et al.'s sample of 320 arguments. Only recently, (Ng et al., 2020) set out to close this gap with the grammarly Argument Corpus (GAC), simplifying the taxonomy to only overall argument quality and its three direct subcategories and annotating arguments from two online debate forums (*ConvinceMe* and *ChangeMyView*) and a community question and answer forum. This constitutes an important first step toward a unified approach to argument strength, though the diverging definitions of objective quality, well-formedness or persuasive power in the large number of existing corpora remain, as does the unbalanced nature of which quality markers are investigated on corpora with which quality definition. As each work focuses on one or two aspects of argumentation, there are no datasets available which facilitate comparisons between aspects. Many investigations into subjective features

either emerge from, or adopt the methods of, deliberative theory, resulting in small, non-generalizable case-studies or relying on (sensitive) demographic metadata that is not obtainable for existing corpora and large-scale assessments of argument data. The subjective features investigated on online forums and small samples may behave differently on the traditional corpora like IBM ArgQ.

Thus, a systematic, broad data-driven review of multiple subjective features associated with *ethos* and *pathos* strategies is lacking from argument mining research. The approaches described above already show an avenue for feature choice: Storytelling has long been addressed in deliberative theory but only recently came into focus of computational argument mining. As Falk and Lapesa (2022) argue, personal narratives are vital for inclusivity, because argument mining approaches only focused on the *logos* dimension exclude less educated groups. As an addendum on their main contribution, they showed how storytelling interacts positively with certain Wachsmuth et al. (2017) dimensions on Dagstuhl-15512-ArgQuality and GAC, though they do not include general argument quality corpora without the taxonomy, and as such do not report effects of storytelling on overall quality. Using their work as a starting point, investigating storytelling in relation to argument quality is thus a logical next step. Similarly, when addressing the *pathos* strategy, the most intuitive argument feature is emotionality, though apart from deliberative works (Maia and Hauber, 2020) and small studies analyzing facial expressions (Benlamine et al., 2015; 2017), there is no detailed exploration on individual emotions' influence argument quality and persuasion. This lack of resources necessitates the adoption of emotion analysis methods and resources for a first approach to closing this gap in argument mining. Lastly, the lack of hedging analyses in argument mining is especially inexplicable as it constitutes an easy surface feature that shows much about debater's confidence. Similar to emotion, the missing resources in the argument domain also necessitate an adaptation of hedging research from other domains. These three features cover a large span of subjectivity in arguments while allowing for post-annotation of large corpora. This means that a costly manual annotation process or aggregation of sensitive metadata can be avoided by deploying automatic methods from the respective research field in a cross-domain setting. In this vein,

the following sections elaborate on storytelling, emotions and hedging divorced from argument mining in order to gain the understanding necessary to successfully model them in the argument domain.

2.2 Subjective Features

2.2.1 Storytelling

Due to the narrow focus on objective forms of argumentation that only expanded in the last ten years, computational argument mining missed the deliberative potential of personal narratives. Personal narratives or storytelling in this case describes a series of events told sequentially to make a point (Ryfe, 2006; Polletta and Lee, 2006). While only recently addressed in computational argument mining, narratives have long been a part of political science research.

Early deliberation research also overlooked personal stories in favor of more abstract reason-giving that was declared to be a marker of high quality, successful deliberation (Polletta and Lee, 2006). As such, storytelling was seen as a personal variant of reason-giving that should be avoided in deliberation because of its subjective nature: it focuses on the individual rather than community and presents evidence that cannot be proven and only contended with a personal attack discrediting the storyteller (Black, 2008). The first investigations of narrative thus largely served to contend this view. Ryfe (2006) analyzed video recordings of five small deliberative groups hosted by the National Issue Forum and consisting of a facilitator and a diverse group of people deliberating face-to-face on public issues like school reform, internet privacy or campaign finance reform. These groups consisted of strangers initiated only through an information package on the issue at hand, who during the course of the deliberative process had to find disagreements and common ground to work towards. In this setting, (Ryfe, 2006; p. 73) found that a majority of contributions to the discussion are frames as narratives: “They tell stories about themselves, their family, and their friends. They tell stories about events in the news, people at work, and casual acquaintances. Sometimes, they use other

modes of talk: they argue, debate, or lecture. But the clear pattern is that they prefer to tell stories." These findings in typical deliberative settings directly contradict the then prevalent view of good deliberation as an exchange of reasons meant to persuade other participants (Polletta and Lee, 2006). Ryfe (2006) finds the reason participants rely on stories to convey their points in the unique interpersonal character of stories. Saving face in a group of strangers is thus difficult when plainly stating reasons about complex topics on which almost none of the participants are experts, and stories bridge these difficulties. Ryfe recognizes multiple different functions by which stories do this. They construct a person's identity in the discourse and lend credibility to their claim by disclosing personal information that makes them appear more sincere, friendly and trustworthy. Similarly upholding friendliness, participants often agreed with a previous point to then indirectly oppose it with a story that includes their disagreement. Furthermore, stories facilitate the perception of accountability, stakes and cognitive diversity in a deliberative group, which are main motivators for a deliberative frame of mind (Ryfe, 2006). Stories allow participants to construct their own opinion and stake in a discussion and disagree, persuade and be persuaded without direct confrontation and thus lower the barriers inherent to the deliberative setting. At the same time, Polletta and Lee (2006) also argued for storytelling to be legitimized as a component of good deliberation, as insisting on more abstract reason-giving would further disadvantage minority-voices. Observing groups in a deliberative online forum on the future of the World Trade Center site after the 9/11 terrorist attack, Polletta and Lee conducted a systematic comparison of reason-giving and storytelling claims. They observed that participants with lower income, without a college degree, who were non-white or women were more likely to use narrative in their claims than were people on the other end of these axes. Participants used stories especially when perceiving their own opinion or experiences as a minority-perspective. In these cases, telling stories provided writers with an otherwise missing expertise to challenge the universality of the majority opinion and to either present new issues and talking points or serve as a starting point determine available avenues and perspectives in a collaborative process. Seemingly in contrast, participants also used stories to argue against their

own position, which allowed them to signal empathy to the opponent and invalidate arguments aiming at a lack of contrary experience. Thus, Polletta and Lee (2006) see the deliberative opportunity of storytelling as related to its openness to interpretation, inviting empathy and exchange instead of insisting on abstract reasons. Black (2008) also based her investigation of storytelling and dialogic moments on this forum and came to similar conclusions: personal stories with their potential for identity negotiation and perspective taking engender a space for mutual dialogic collaboration in an otherwise persuasion-focused deliberative mode of communication. These early accounts of narratives in deliberation all point toward its role as less a personal, and thus subpar, way of reason-giving, but rather an alternative, more open approach to deliberation.

As measures of deliberative public engagement grew more popular, the possibilities of storytelling purported by the early works described above were accepted into deliberative theory and research in political science was able to focus on narrower questions than the validity of narratives as a whole. Black and Lubensky (2013) base their study on the link between storytelling and different modes of deliberation on transcripts from the Australian Citizen Parliament, a four-day deliberation on strengthening the political system with 150 randomly selected participants split into 24 discussion groups. The study showed the lack of storytelling in fast-paced discussions structured around listing, summarizing, ranking and voting on issues, while the World Café that had people mingle away from their assigned group and discuss value-based questions engendered storytelling. Other aspects of storytelling investigated during this time includes the use in argument conflict (Black, 2013) and in airing feelings of injustice (Maia and Garcêz, 2014). Black (2013) again looks at the forum on reconstructing the site of the World Trade Center and compares two discussions started by an adversarial story. Such adversarial stories frame an issue as two-sided with the storyteller trying to demonstrate their expertise and persuade other participants. In contrast, unitary frames of conflict emphasize shared values, openness to other perspectives and the potential for compromise. To reframe the conflict from adversarial to unitary, a facilitator redirects the conflict with questions about the participants' underlying values in a direct, process-oriented way. In

a similar situation without a facilitator, participants answer the adversarial story by reframing the conflict in a unitary way with stories of their own that push the conflict frame towards shared values and compromise without an explicit redirection away from the conflict. This natural redirection shows how storytelling might be more suited towards mediating deliberative conflicts than the more direct, process-oriented approach of facilitators (Black, 2013). Maia and Garcêz (2014) transfer earlier insights from deliberation to the broader field of political and social inclusion with a case study on feelings of injustice aired in Brazilian forums for Deaf people. This study confirms the potential of storytelling to help minorities challenge false universal and gain the empathy needed to introduce their struggle as a public issue.

A commonality between these studies is their reliance on small case studies, which is usual for deliberative research but hinders extrapolation from the study results in two ways. The narrow domain of, e.g., reconstructing after 9/11, the future of Australian democracy or the societal struggles of Deaf people, firstly forbids broader generalizations extending to different topics or deliberation forms. Secondly, the importance of those case studies amounts to a conflation of theoretical findings divorced from the topic with an analysis of the domain itself – common arguments, idiosyncracies of the deliberation form, extrapolations from just one line of argumentation. This is where computational argument mining with its data-driven approach to argumentation may further storytelling research. The state of such research in the argument mining field is however still sparse. At the same time as deliberative research accepted storytelling as a component of deliberation, first works in computational argument mining (Park and Cardie, 2018) adopted an approach similar to early deliberative theory, with storytelling classified as a personal type of evidence. In an investigation of different types of evidence, Park and Cardie (2014) annotated 9,476 individual sentences and clauses from 1,047 comments taken from the deliberation platform *RegulationRoom*. Apart from non-arguments and non-verifiable sentences, the annotation distinguishes verifiable non-experiential and experiential sentences, with the latter category of personal experience overlapping with definitions of storytelling. Later on, Park and Cardie (2018) extended the work to an annotation of elementary units as facts, experience, values, policy or references with

a relational annotation categorizing pairs of such elementary units into reason or evidence. The experience or *testimony* annotation is defined as "an objective proposition about the author's personal state or experience" (Park and Cardie, 2018; p. 1625), which lacks the sequential pattern of the common definition in deliberative theory. Experiential propositions can thus have narrative character (1) but also statements about the author themselves (2).

(1) *We receive repeated calls trying to get contact information, even though we request to be taken off their list.* (Park and Cardie, 2018)

(2) *My son has hypolycemia.* (Park and Cardie, 2014)

A similar category is that of anecdotes, which are used by Al-Khatib et al. (2017) as one of three types of evidence, defined as "personal experience of the author, a concrete example, an instance, a specific event, or similar," and by Song et al. (2016), who extract narratives and their protagonists from argumentative essays to recommend anecdotes as evidence structured as <*person, story, implication*>. These works however largely suffer similar constraints as the case studies of deliberative theory, namely a lack of generalizability due to different domains and different working definitions of narrative.

This changed only recently with Falk and Lapesa (2022) both stating and addressing the need for storytelling resources. They aggregate a corpus from three domains annotated for experiential evidence or testimony and storytelling and test different classification approaches on the data. This is posited as the first step towards exploring storytelling in argument mining to the same extent as in deliberative theory's two decades of case studies. With this thesis, I contribute to this exploration, linking storytelling with argument strength by harnessing large automatically annotated corpora.

2.2.2 Emotion

As already observed in section 2.1, when NLP argument research includes emotion, this feature is simplified to mean something completely different, i.e., sentiment,

referring to an overall negative or positive polarity in the text. Exceptions are rare and restricted in their approach; while Benlamine et al. (2015; 2017) use multiple emotion classes, they study the webcam footage of argument participants instead of textual emotions, and Maia and Hauber (2020) limit their investigation to *anger, fear, compassion* and *indignation*, which constitute a small and unusual set for general emotion analysis. Concerning emotion classification in text, available resources use diverse underlying emotion theories. The most relevant theories can be divided into the categorical approach of basic emotion theory and the continuous approaches of constructivist and evaluative theory.

Basic emotion (BE) theories model emotions as intuitive discrete categories like happiness, sadness or anger, that are in some feature universal. The choice and number of basic emotions vary by theory, as does the conceptualization, with emotions usually being explained as inherited (and thus universal) programs that evolved as they provide motivation, help in problem-solving or social interactions through distinctive bodily reactions, facial expressions and thoughts (Scarantino and de Sousa, 2021). Ekman (1992) identifies six fundamental emotions produced by facial muscles that are universally recognized[1]: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. This model serves as the foundation for many BE theories and is built upon by Plutchik (2001) by modeling emotion intensity and relations, and incorporating the notion that emotions can co-occur to create mixed emotions. The resulting emotion model (figure 2) thus explains how *apprehension*, *fear* and *terror* are different degrees of intensity of the same emotion, or how *optimism* emerges from the co-occurrence of *joy* and *anticipation*. These two widely used BE theories show the variability in approaching emotion as discrete categories with 6, 8, or, including intensities and mixed emotions, 24 or 32 distinct classes.

Constructivist theories divest from explaining emotion as independent classes without overlap. They instead assume a composite view of different emotions by observing continuous affect dimensions that each represent a different component of the emotional experience. Affect in this case means a neurophysiological representation of the most basic feelings that construct emotions. The VAD-model by Russell

[1]Recent studies challenge this universality beyond basic polarity, cf. (Gendron et al., 2014).

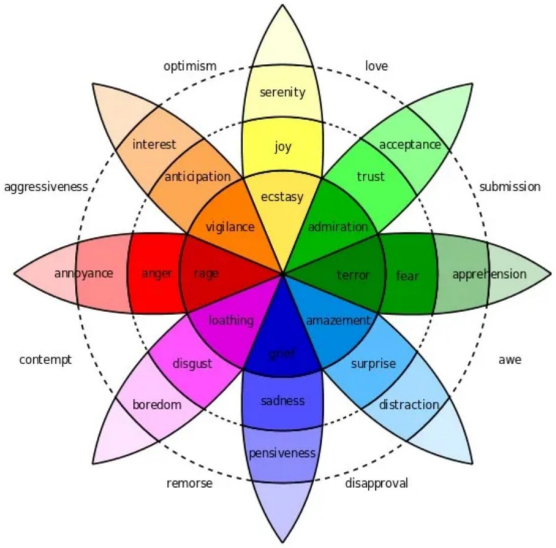


Figure 2: Emotion scheme by Plutchik (2001) with color saturation corresponding to emotion intensity.

and Mehrabian (1977) is used most commonly in NLP and models emotion along the affect dimensions of *valence* (degree of pleasantness), *arousal* (degree of calmness or excitement) and *dominance* (degree of control over a situation). Emotional states are thus constructed as vectors and discrete emotion labels can be mapped to these vectors as done by Buechel and Hahn (2016) for Ekman's (1992) fundamental emotions (cf. figure 3) according to ratings obtained by Russell and Mehrabian (1977).

Evaluative theories of emotions finally assume emotions to emerge from the cognitive appraisal of a stimulus. Thus, while resembling constructivist approaches in rejecting the distinct stand-alone categorization of basic emotion theories, the inclusion of cognitive processes differs from the constructivist assumption of spontaneous affect. In evaluative theory, cognitive appraisals are one of five components that constitute an emotion as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of a [...] stimulus-event" (Scherer, 2005). Other subsystems include a motivational component, physical symptoms, facial (and vocal) expressions and lastly the feeling itself. According to Scherer (2001), appraisals follow from the inciting stimulus-event

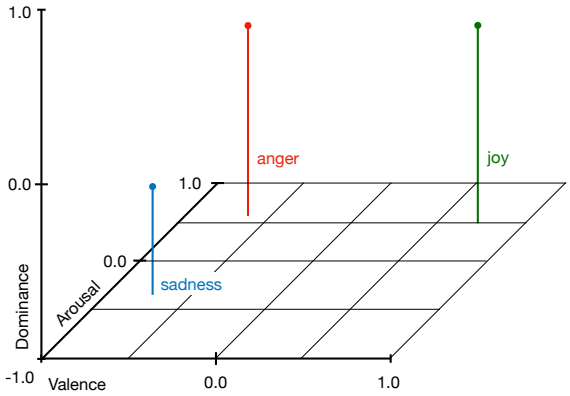


Figure 3: Positions of Ekman's basic emotions along the VAD-dimensions adapted from Buechel and Hahn (2016).

in four stages of increasing complexity; first evaluating the relevance of the stimulus, then its implication and the ability to cope with it and lastly its normative significance. Each of these stages includes multiple appraisals, such as checking for novelty in the first stage, finding causality and predicting the outcome in the second step, checking if this outcome can be controlled or adjusted to in the third step, and finally in the last step, all information is held up to personal morals and standards of self-image and shared cultural values. The individual appraisal dimensions can thus vary in number, Scherer's four stages contain 20 dimensions, but other theories include only six (Smith and Ellsworth, 1985) and the OCC model arrives at emotions through a flow-chart-like series of appraisals (Clore and Ortony, 2013).

All above theories are used in the creation of NLP emotion analysis resources, and have their own challenges associated with them. The biggest challenge for all theories is the implicit nature of emotions in text. Lexicons like the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2010; 2013) or WordNet Affect (Strapparava and Valitutti, 2004) allow for emotion or affect analysis through explicit mention of feelings and emotionally salient concepts (e.g., *delightful* indicating *happiness* or *shout* indicating negativity and high arousal), these analyses stay at a surface level since emotions are rarely explicit in text and context-dependent (e.g., *shout* could indicate *fear*, *anger* or even intense *joy*). While lexicons can be con-

structed by crowdsourcing emotion associations or aggregating synonyms of emotionally salient words, any method more complex than word matching necessitates annotating text instances with emotions. There are multiple annotation methods, which all suffer from the indirectness of text – unless an author stated their emotional state, expert annotators have to infer the writer emotions or approximate a universal reader reaction from their own feelings. This is a non-trivial problem, as a headline like *Italy defeats France in World Cup Final* written without emotion is perceived very differently if readers are fans of either team (Katz et al., 2007). Inter-annotator agreement in emotion annotation is typically low compared to other tasks, varying by annotation class and domain (cf. Troiano et al., 2023). When relying not on expert annotation but rather crowdworkers, Mohammad and Turney (2013) showed how changes to the annotation guidelines change the annotation quality and the inter-annotator agreement, which then Buechel and Hahn (2017b) showed to be domain-dependent. Some corpora avoid this perspective problem by relying on self-labelling on social media, e.g., the use of emotion hashtags on Twitter (Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017) or the reaction feature on Facebook posts (Krebs et al., 2018; Pool and Nissim, 2016). While these corpora arguably include true emotion labels for writers and/or readers, the collection from social media makes for noisy data. A similar approach to noisy self-labelling are corpora aggregated from self-reports, where subjects are asked for event descriptions of situations they associate with particular emotions, which are seen as an expert-annotation by the writer themselves (cf. Bostan and Klinger, 2018). As diverse as the aggregation and annotation choices are the annotation schemes and data domains used in the process. GoodNewsEveryone (Bostan et al., 2020) includes headlines span-annotated for experiencer, cause, target, and clue for emotions extended from Plutchik, and EmoBank (Buechel and Hahn, 2017a) includes individual sentences annotated for valence, arousal and dominance. Appraisal corpora include isear (Scherer and Wallbott, 1994) with self-reports for Ekman's six basic emotions and *shame* and a 25-part questionnaire that classed subjects' appraisal evaluation according to 15 appraisal dimensions in a four-step scheme as above (Scherer, 2005). Crowd-enVENT (Troiano et al., 2023) was aggregated sim-

ilarly but with self-reports for twelve different emotions (adding *boredom, pride, relief, surprise* and *trust* to the ISEAR set) and neutral events and evaluations for 21 appraisal dimensions and an additional annotation study to validate self-reported emotions and appraisals.

Thus, when approaching an emotion classification problem as in this thesis, the most important decision concerns the resource used. Methods using emotion lexicons have largely been replaced by neural models (Bostan and Klinger, 2018), and more recently, by transformers with pre-trained embeddings, such as BERT or RoBERTa. Thus, training data must be chosen from the multitude of options, all of which have advantages and disadvantages in terms of underlying emotion theory and domain. While discrete emotion labels following basic emotion theories are most easily understood, the inherent fuzziness and subjectivity also influences the training of models on data with lower inter-annotator agreement. VAD- and appraisal-based corpora however add a layer of abstraction to any interpretation. Mapping automatically predicted scores on continuous dimensions to discrete emotions introduces a new opportunity for errors, but interpreting (VAD- or) appraisal-scores without such a mapping is a complex task on its own and not always conducive to further explorations, as is the case in this thesis. Domain and annotation scheme will thus narrow the choice of emotion resources that is described in detail in section 3.3.

2.2.3 Hedging

In linguistic research, *epistemic logic* is a field of semantics concerned with constructions that explicitly or implicitly convey a degree of knowledge or certainty about a proposition (Lyons, 1977). While stating either (3) or (4), the speaker commits to the truth of the same proposition (3), though this commitment is made both uncertain and explicit in (4):

(3) Paris is the capital of France.

(4) I believe that Paris is the capital of France.

The first statement is thus called epistemically non-modal, as any commitment

to the proposition's truth is implicit. Conversely, hedges or hedge terms like *believe* in (4) are defined as words and expressions that address epistemic modality, or as (Lakoff, 1973; p. 271) put it, "words whose job it is to make things fuzzier or less fuzzy."

In their study of hedge usage in discussions between physicians, Prince et al. (1982) distinguished two types of hedges, *approximators* and *shields*, later called *propositional* and *relational* hedges. Approximators introduce uncertainty into the propositional content (5), while shields introduce uncertainty into the speaker-commitment (6), i.e., the speaker's relation to the propositional content.

(5) Paris is kind of the capital of France.

(6) a. According to Clara, Paris is the capital of France.

b. As far as I can tell, Paris is the capital of France.

Shields can be further distinguished into those that hedge the speaker-commitment by attributing the proposition to another person (6a) or by implying that the speaker arrived at the proposition not by deductive logic but through plausible reasoning (6b). Hedge terms thus include adverbs and markers introduced to the proposition like *a few, about, little, approximately, most* or *slightly*, as well as epistemic verbs like *believe, suggest, appear, seem* and phrases like *in my opinion, more or less* or *to some extent*.

In their case study of physician-physician discourse, Prince et al. (1982) speculate that the use of hedges may be a strategy to save face professionally, or, as doctors interact with lay people and are seen as omniscient in life-or-death scenarios, a demonstration of scientific conduct to give a correct representation of their knowledge. Similarly, in constructing a logical framework for politeness through indirect speech, Ardissono et al. (1999) include hedges in their examples of indirect speech as a politeness strategy saving face as the own positive self-image. Vasilieva (2004) shows how in computer-related instructions, hedging is used to appear cooperative, more so in men than women (though Xiao and Tao, 2007; show an inverse gender distribution for a larger, more diverse dataset). Hedging and the associated uncertainty are thus used as either a conscious strategy or unconsciously to appear less

self-assured and thus according to the situation more approachable, less infallible and more polite.

How uncertainty manifests in lexical hedges is however dependent on the domain and mode of language. Informal speech is usually produced spontaneously without script or rehearsal, thus uncertainty and hesitation are generally more commonplace than in written text. Thus, the working definition of hedges in spoken language often includes fillers (*uhh*, *hmm*) and smallwords or discourse markers (*like*, *you know*), and is subsumed into a larger group of uncertainty markers, that include para-verbal cues such as pauses, disfluency, articulation rate, repetition or self-repair (Rosanti and Jaelani, 2015; Prince et al., 1982; Chatterjee et al., 2014). These cues are not easily divided from lexical cues, as smallwords like *well*, *so* do not meaningfully alter the proposition (Hasselgren et al., 2002) as hedges do, even though they are verbal. Non- or para-verbal cues on the other hand are also investigated away from hedging, e.g, as markers psychological distress (DeVault et al., 2013). When investigating uncertainty explicitly, research often addresses multiple or all of these uncertainty markers, for example as part of more natural, fluent speech in language learners (Hasselgren et al., 2002; Wang, 2021), or as linguistic gender differences (Vasilieva, 2004; Xiao and Tao, 2007; Rosanti and Jaelani, 2015). When addressing hedges in particular, the scope of phrases is most often the main focus rather than detecting the hedge itself (e.g., Kärkkäinen, 2010). Uncertainty in spoken language is thus not limited to lexical hedges, and neither all verbal nor all para-verbal cues are equally applicable or inapplicable to similarly informal written language. As encountered by Prokofieva and Hirschberg (2014) when developing guidelines and methods to annotate hedging automatically in both speech and text data.

In written language, following a similar logic as Prince et al. (1982) identifying hedges with accurately verbalizing scientific inquiry and knowledge processes, the phenomenon is traditionally studied in scientific and biomedical writing. This was motivated in 1998 by Hyland from two sides. Firstly, despite the assumption that scientific writing consists of "impersonal statements of facts that add up to the truth" (p. 6), hedging is a vital component to accurately verbalize scientific inquiry in academics. Earlier on, Bryant and Norman (1979) had physicians rank

the certainty of diagnosis sentences that included shields like (6b) with inconsistent results between subjects and between different hedges. Secondly, as professional scientific writing follows hypotheses, which are tested and reasoned about through experimental evidence that usually does not allow complete certainty, this domain should be fruitful for hedging research. While these reasons steered computational hedging research in only this direction until recently, they may just as well be used to motivate hedging research on arguments – which are assumed to be statements of conviction, but aside from completely objective, statistical evidence, all reasoning includes a component of making claims (resp. hypotheses) based on plausible deductions (resp. experiemental evidence).

In the field of computational linguistics, Light et al. (2004) provided the first manual annotation study for abstracts in the National Library of Medicine database MEDLINE, labeling 3,429 sentences from 1,110 abstracts as highly speculative, low speculative or definitive and training an SVM-based text classifier on the resulting dataset. In 2007, Medlock and Briscoe annotated sentences from 5,579 full-text papers on fruit fly genome with a weakly supervised approach and more refined annotation guidelines for the initial annotation set. These guidelines show the difference between previously discussed informal and verbal hedges and those in scientific literature, as they include, e.g., statements of knowledge paucity (7a) or speculative hypotheses (7b) and recounting hedges from previous works (7c):

(7) a. How endocytosis of Dl leads to the activation of N remains to be elucidated.

b. To test whether the reported sea urchin sequences represent a true RAG1-like match, we repeated the BLASTP search against all GenBank proteins.

c. Dl and Ser have been proposed to act redundantly in the sensory bristle lineage.

This area of hedging work was extended by Vincze et al. (2008) releasing *BioScope*, an expert annotated corpus of over 20,000 sentences from medical texts and biological papers and abstracts for hedge and negation clues and their scope, which Agarwal and Yu (2010) used for their hedge detection experiments with conditional random fields. These texts are domain-specific not just in the types of possible

hedges but also the general contents. A similarly formal but less specialized domain is that of Wikipedia articles, where editors are advised to avoid *weasel words* like *some people say*, which are tagged for improvement by the editors and thus allowed Ganter and Strube (2009) to aggregate an already annotated corpus and employ shallow linguistic features such as numerical underspecification in *many, some* or passive constructions like *it is believed* to detect hedges automatically with an F_1 of .70. Data from both BioScope and Wikipedia was used in the CoNNL-2010 Shared Task (Farkas et al., 2010) for hedge (and scope) detection, where systems mainly consisted of BoW approaches or identifying and then disambiguating hedge cues via token classification or sequential labelling. Approaches based on any of these data are however inapplicable to any form of more informal text. All biomedical resources contain very specialized language and hedging constructions, and while the language in the Wikipedia Weasel corpus is less domain-specific, the article still follow guidelines of formality and annotations using an editing tag for expressions to avoid and correct leads to a very unbalanced dataset with only 437 in 168,923 sentences.

There are however newer computational linguistics contributions that address hedging in informal contexts like web forums. The first of these is Mamani Sanchez and Vogel's 2013 *exploratory study of hedging in web forums*. In their dataset of 172,253 posts from a forum belonging to the customer support service of a software company, they study the influence of the use and type of hedges on other user's perception of the author. To this end, they devise a hedging scheme that distinguishes epistemic phrases from non-phrasal hedges and annotated the forum posts automatically by greedily matching terms from lexicons aggregated from previous, non-computational work on hedging in speech (e.g., the already mentioned Kärkkäinen, 2010) or text. After aggregating these lexicons of 76 and 109 terms respectively, they explicitly include abbreviations as a feature of user-generated web content, with *IMHO* or *AFAIK* standing in for the epistemic phrases *in my humble opinion* and *as far as I know*. Thus sorting posts by differently combining measures for no hedging and the two types of hedges, they investigate the link between different hedging behavior and perception of other users by correlating their annotations

about	If token `t` has part-of-speech `IN`, `t` is non-hedge. Otherwise, hedge. **Hedge:** There are *about* 10 million packages in transit right now. **Non-hedge:** We need to talk *about* Mark.

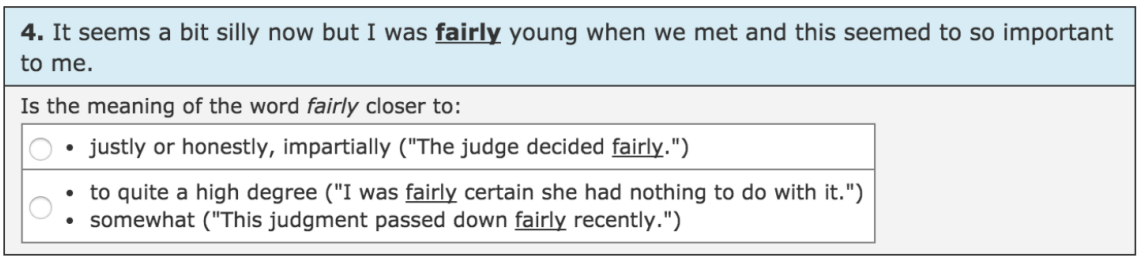


Figure 4: Example of a hedge disambiguation rule from Ulinski et al. (2018) and crowdsourcing definition from Ulinski and Hirschberg (2019).

with the amount of *kudos* scores per post, a voting system in the forum that allows users to award useful contributions. They show that posts with mainly epistemic phrases have the highest kudos scores while non-hedged posts have the lowest. This is in line with verbal hedging research which attributes strategies for cooperation and politeness with hedging. In 2015 they extended this study into a detailed annotation scheme for epistemic hedges in informal text, including a hedge term's scope, the epistemic source (similar to the distinction between (6a) and (6b)), and the type of hedge (Sanchez and Vogel, 2015). These first forays into the idiosyncracies of informal text were followed by Ulinski et al. (2018) and Ulinski and Hirschberg (2019), who both use a similar approach of a hedge term lexicon, though they group hedges into propositional (5) and relational (6) terms and further disambiguate the true hedged meaning. Ulinski et al. (2018) are the first to introduce syntactic and morphological rules for specific terms disambiguating hedge use from non-hedge use, while Ulinski and Hirschberg (2019) reformulate these rules as simple definitions with corresponding example use-cases, which are then supplied to crowd workers who disambiguate automatically detected hedges (figure 4).

This simple approach of detecting terms and then disambiguating them via rules is also adapted by Islam et al. (2020), who use a lexicon not only for hedges but also discourse markers and words that boost certainty, which they count as hedges

if negated (e.g., *not sure*, *without certainty*). These few works on hedges in informal language all share the use of lexicons and simple, rule-based algorithms, if at all. This is in part due to the lack of big training corpora, which are all focused on the previously described biomedical or Wikipedia domain. Another reason against the use of complex models to detect hedging however is a lack of need, as hedges are explicitly terms, i.e., tokens or token n-grams that may be extracted without abstracting from surface representations to sentence meaning. While this is not verifiable due to the lack of annotated informal corpora, the recent rise of informal hedge detection research not demanding for more complex methods is promising for applying a similarly simple approach to arguments. As stated above, argument research including hedges either does not state the detection process like Luu et al. (2019) or includes only a subcategory not identified as hedging like modal verbs in Wei et al. (2016).

3 Methods

Investigating the link between argument strength and the subjective features of storytelling, emotions and hedging requires argument data that is annotated not only for argument strength but also for each of these features. As there is currently no such dataset available, a suitable corpus has to be aggregated through automatic means. This section introduces the argument data used as a foundation for the subsequent analysis (3.1), as well as the resources and methods used to automatically annotate the data with each of the features (3.2, 3.3, 3.4).

3.1 Argument Strength

There are few argument corpora annotated for argument strength on the instance level, and those that also include one of the other needed features (e.g., the *ChangeMyView* Subreddit dataset aggregated by Tan et al., 2016) are very small. Therefore, I am using corpora solely annotated for argument strength as the basis of the new aggregated datasets.

To approximate the diverging conceptualizations of argument strength explicated above (2.1), I am choosing two datasets that differ in collection method, domain, argument length and annotation procedure.

IBM ArgQ 5.3k This argument quality dataset by Toledo et al. (2019) was created as part of *IBMRank* which also includes the same data grouped into pairs annotated with relative quality labels. It consists of 5.3k arguments generated by debate club members of varying skill level and a general audience who were asked to submit as many short, impersonal arguments for or against a given topic as they wished. The topics span 11 controversial discussions such as privacy laws, gambling or vegetarianism with explicit stances, e.g., *We should adopt cryptocurrency* and *We should abandon cryptocurrency* (for an overview of all topics, see table A.1). The participants submitted their arguments into the *Speech by Crowd* UI which showed the topics, a guideline submit arguments without any information about an identifiable person and an example argument by a professional debater. Additionally, the UI accepted only arguments with 8 – 36 words. This collection procedure resulted in a dataset of very short arguments employing impersonal, rational rhetoric without verbose use of personal experience or statistical evidence.

To obtain the argument strength annotation, Toledo et al. (2019) asked crowdworkers to judge *Disregarding your own opinion on the topic, would you recommend a friend preparing a speech supporting/contesting the topic to use this argument as is in the speech? (yes/no)*. This constitutes rather vague guidelines, as the annotators must employ their own concept and hierarchy of quality dimensions, e.g., topic relevance, linguistic clarity or sound rhetoric, and the single binary judgement makes reconstruction of these dimensions impossible. While not explicitly stated, by invoking an argumentative speech and having debate club members take part in the argument generation, the argument strength conceptualization of this dataset falls in line with the traditional argument mining field and assumes argument strength as an objective measure of arguments employing a *Logos* strategy for persuasion. Toledo et al. (2019) moreover cite only argument strength taxonomies using either these rational features or a relative approach in which an argument's individual per-

Argument	Score
A government mandate on flu vaccines will standardize vaccine policy across the nation where there is currently variability depending on the state that prevents people from being able to predict what they should be doing.	0.43
Aggregate benefits of vaccination always outweigh negative individual results.	0.8

Table 1: Two arguments from IBM ArgQ with their aggregated argument quality scores (Toledo et al., 2019).

suasiveness emerges from its relative persuasiveness compared to other arguments (i.e., Simpson and Gurevych, 2018; Gleize et al., 2019). The further annotation process strengthens this impression, as the binary annotation decisions are aggregated for 15–17 annotators per instance to approximate an objective, universal argument quality score denoting the ratio of positive judgements for each instance. In the final dataset, each instance is annotated with a score between 0 and 1 which allows for comparisons of quality between two arguments for the same topic and stance as depicted in table 1. In the following analysis, this dataset is referred to as IBM ArgQ and represents argument strength as conceptualized by the traditional argument mining field.

Cornell ChangeMyView To depict a diverging view of argument strength, a corpus aggregated by Tan et al. (2016) from comments on an internet forum is used as the second dataset. The data is collected from the Reddit forum *ChangeMyView*[2], where users feeling that they hold a "wrong" opinion begin conversations by stating their viewpoint with detailed background on their thought process to be challenged and persuaded to change their view by other users through constructive debate. Thus, for each new original post (OP), multiple users argue for – or rather against – the same position until the first person is persuaded and awards a *delta point* Δ to those answers that persuaded them. Tan et al. (2016) collected all discussion

[2] `https://www.reddit.com/r/changemyview/`

trees from between the forum's creation in January 2013 and August 2015 for their dataset (henceforth Cornell CMV). Following the unique setup of the forum, quality control measures for individual arguments are already in place from the posting guidelines and the delta point system provides a natural argument strength annotation denoting complete persuasion away from the OP view. The forum is actively moderated both for civility and for maintaining a constructive discussion in which comments must advance the conversation as "Comments that are only links, jokes, or 'written upvotes' will be removed", and the OP author is asked to be mindful of their use of delta points and "must include an explanation of the change along with the delta so we know it's genuine" (r/ChangeMyView, 2023). This ensures that the delta points can be used as a gold label in the dataset as they are vetted by other users and moderators of the forum and the posts arguing against the OP contain concrete stances with meaningful argumentation.

Apart from these considerations of data quality however, the domain properties make for much longer texts sometimes containing multiple premises and stances forming a rhetoric argumentative sequence, or direct quotes from the OP which are addressed point by point. Table 2 shows two texts arguing against the same OP stating that driving is the most dangerous activity one can do with similar rhetoric of recounting other habits more deadly than driving but differing in the level of detail (the persuasive text includes concrete statistics) and the level of formality (the non-persuasive text is considerably more informal in its laconic description of suicide and use of curse words). In aggregating the dataset used here, Tan et al. (2016) structure the posts as one OP parent with a pair of corresponding comments, one with and one without a delta point awarded by the OP author. Extracting the thus annotated comments, Cornell CMV includes 11,567 argumentative texts with a balanced distribution of the binary persuasiveness label.

Given all above differences between IBM ArgQ and Cornell CMV, it is apparent that the two datasets conceptualize arguments themselves as well as argument strength in very different ways. While the number of differences disallows a comparison of pure argument strength conceptualization without any confounding factors, including both corpora in the investigation covers idiosyncrasies across the spectrum

CMV: Driving a car is insanely risky and probably the most dangerous thing you do in your everyday life.

I find it difficult to understand how so many people enjoy driving a car or can even relax while doing it. I am almost continually tense while on the road thinking about what's at stake (and I've been driving for almost 20 years).

While I have never been in an accident, I often find myself thinking how dangerous even small motions of a driver can be. For example, a sudden small jerking movement of an arm on the steering wheel leading the car into oncoming traffic can lead to almost certain instant death. I cannot think of any other action in my daily life where so many small actions (of me or other people) can be lethal.

Even leaving accidents and catastrophic scenarios out of consideration, driving a car seems extremely risky to me: For many, maybe most people their car is the most expensive single item that they own. Even small mistakes like a lack of concentration or a tiny miscalculation while parking into a small space, can lead to high damage and expensive repairs.

Δ 1	Δ 0
Mortality for drivers in the US is roughly 50 per millions. Death while working in construction in 2006 was 108 per millions. Driving is not the most dangerous thing these workers do in their everyday life. (edit. The more i'm looking into it the more I find that stats regarding this subject varies a lot.)	By the death rate, eating unhealthy is the most dangerous thing that you can do. Cellular reproduction is up there are well. Then there's realizing your worthless and life is futile, then taking your own life. Looking at the CDC, suicide isn't on there. But breathing shit other than oxygen and nitrogen is up there. So is, the fatty food thing again.

Table 2: Example of an original post from Cornell CMV with the user-given title in bold and two counter-arguments. The left answer has persuaded the OP and the right has not.

of the argument mining field on what argument strength means. To illustrate the diverging concepts, in the following analysis (4.1) argument strength is called *quality* when investigating only IBM ArgQ and *persuasiveness* for Cornell CMV.

However, as the datasets have no annotations for the here investigated features, it is necessary to annotate these retroactively. As a manual annotation with gold labels is a complex, resource-intensive process, the labels are annotated automatically. The following sections elaborate on the annotation process of each feature and the resulting statistics on the two argument datasets.

3.2 Storytelling

The storytelling annotation process follows closely from Falk and Lapesa (2022). While previous research on personal testimony or narratives often takes the form of a small case study where the data is limited to online discussions or public hearing transcripts on a singular topic (cf. e.g., Maia et al., 2020; Black, 2013), Falk and Lapesa (2022) explicitly aim to make storytelling investigations accessible through datasets and, leading from those, robust automatic methods to compile storytelling annotations for new data. I adapt their methodology of training and comparing storytelling classifiers using two different settings: firstly I am training a storytelling classifier on a single domain (*one-domain*), and secondly I am adopting their mixed-domain approach using three different datasets for more robust classification (*mixed-domain*).

3.2.1 Training data

In order to account for both classification settings, I consider all three datasets provided by Falk and Lapesa (2022) for training, as they vary by domain and annotation scheme.

RegulationRoom The first dataset consists of 725 comments from a discussion on consumer debt collection practices (CDCP) hosted on the eRulemaking platform

Regulation Room. The corpus (henceforth RegRoom) is collected and by Park and Cardie (2018) and span-annotated for elementary units (EUs) and support relations, the former including *facts*, *testimony*, *value*, *policy* and *reference*, pairs of which are linked in *reason* or *evidence* support relations. The relevant feature for this thesis is *testimony*, which Park and Cardie (2018; p. 1625) define as an "objective proposition about the author's personal state or experience". The sub-instance level annotation leads to instances with more than one *testimony* EU and 1,117 *testimony* EUs overall. However, following Falk and Lapesa (2022), the testimony annotation is collapsed into a binary variable indicating the presence of at least one *testimony* span in the instance, thus leading to 302 positive instances (for an overview on all used storytelling corpora, see table 3).

ChangeMyView The second dataset is aggregated from a smaller subset of 344 posts from the *ChangeMyView* Subreddit[3] and annotated with a similar scheme of elementary units and support relations (Egawa et al., 2019). Unlike RegRoom, the annotation scheme includes the EU type *rhetorical statement* instead of *reference* and the two support relations *support* and *attack*. *Testimony* is defined exactly as in RegRoom and similarly, the 354 *testimony* annotations were collapsed into instance-level annotations.

EuroPolis Finally, the EuroPolis corpus (Gerber et al., 2018) consists of transcribed and translated contributions to a transnational discussion about the EU and immigration. The corpus consists of 856 speech contributions that are translated to English from German, French and Polish and annotated on the contribution level with dimensions of deliberative quality. These dimensions include the rationality of the justification, orientation toward the common good, inquisitiveness, respect for the *immigrants* group under discussion and for other speakers' arguments and finally storytelling. This latter dimension is the relevant annotation for this thesis and is measured by Gerber et al. (2018; p. 1101) as whether "participants use personal

[3]Note that, while this corpus (henceforth CMV) is collected from the same platform as Cornell CMV, both corpora are distinct.

Corpus	Topic	|i|	Size	# Story	% Story
RegRoom	CDCP	129.7	725	302	41.6
CMV	diverse	290.2	344	130	37.8
EuroPolis	immigration	157.5	856	303	35.4

Table 3: Differences between the three corpora used for training a storytelling classifier. **|i|** denotes the average instance length in words and the last two columns denote the number # and ratio % of instances containing the relevant annotation.

narratives or experiences".

As apparent from table 3, all three corpora are comparatively small for training purposes and storytelling is the minority class. As both RegRoom and EuroPolis include only one topic, they seem ill-fitted as training data for the two multi-topic argument strength corpora. The remaining CMV corpus has the advantage of matching the domain of Cornell CMV, although the small size still remains as a hurdle and the domain-matching introduces a new problem. Namely, annotating *storytelling* for both IBM ArgQ and Cornell CMV with a model trained on data matching the domain of the latter but not the former introduces a new bias into the annotation, as it will likely be better for in-domain use than on the IBM ArgQ corpus. These domain considerations as well as the better instance length match between IBM ArgQ and the two other storytelling corpora serve as motivation to adopt Falk and Lapesa's method of *mixed-domain* training.

3.2.2 Annotation implementation

In their exploration of different constellations of training data and training method, Falk and Lapesa (2022) show that for in-domain (training and testing on the same dataset) and cross-domain applications (training on two datasets and testing on the remaining one), a BERT model without domain-specific finetuning of the language model performs best. The best results overall are achieved when training on all datasets and with a domain-adapted BERT model. Extrapolating from these results to the best model for IBM ArgQ and Cornell CMV, all models would mean cross-

Corpus	Storytelling
RegRoom	I was never informed by Bank of America that they sold my credit card and closed the card. When I realized it, I paid it off immediately. During that quarter, after long illnesses, my Father and Mother both passed (within 31 days of each other) and frankly, credit card payments were not in the forefront of my thinking. ALSO, just because a bank or credit card company has been exempted from Usury laws does not mean they do not commit the violation! THAT needs to be stopped!
CMV	I used to work at an aquatic center that had women's only hours once a week during which only female lifeguards would cover the pool. As it was explained to me, the primary purpose of these hours was to give Muslim and Orthodox women a place to swim without violating their religion. It was common for non-religious women to swim during these times because they felt more comfortable not having to swim in front of men. I don't know what the rationale is at your gym. I would argue that yes, the women's only hours there may be sexist, but they also allow women to partake in an activity that would otherwise be prohibited to them during normal hours
EuroPolis	In Slovenia, we have a lot of immigrants from the non-EU countries, especially in the health care sector, because we need specialists in Slovenia. Slovenians do not want to work in this sector so of course people from other countries are coming to work there.

Table 4: Example of an instance containing the relevant annotation (testimony or storytelling) for each of the three corpora used for training a storytelling classifier (Falk and Lapesa, 2022; p. 5533).

domain application for IBM ArgQ, while there is a possibility for in- or mixed-domain application for Cornell CMV. As the performance difference on the CMV data between the best adapted mixed-domain models (F_1 =.82) and the non-adapted in-domain BERT model (F_1 =.81) is negligible, one model is trained only on CMV data to harness this in-domain advantage. This model is applied to both argument strength corpora and the resulting annotation is called *one-domain* in the following analysis. Since there is no such possibility for the IBM ArgQ data, which might, as previously stated, bias the annotation performance and subsequent analyses, a second storytelling model is trained. The cross-domain experiments do not profit from – and are sometimes hindered by – the domain adaptation; thus, another

standard transformer model is trained on all three storytelling datasets, henceforth called *mixed-domain*. Both models use the newer RoBERTa transformer variant (Liu et al., 2019) as the base embedding and are finetuned for five epochs on ten different splits for the training data to ensure a robust classification result. Only the model steps that improve on both training and validation data are saved and the resulting ten predictions for IBM ArgQ and Cornell CMV are then consolidated into a binary storytelling annotation with 0 for positive results in less than 6 splits and 1 for positive results in at least 6 splits. This finally results in four feature variants for storytelling.

3.2.3 Annotation results

Table 5 already shows a notable difference between the storytelling prediction in the two datasets. As previously speculated because of the focus on debate and objectivity during the creation of IBM ArgQ, the dataset includes much fewer instances of storytelling than Cornell CMV does, with the difference between 45 (46) and 2,288 (1,936) instances predicted after *mixed-domain* (*one-domain*) training marking 20 times fewer instances. While the in-domain evaluation on a held-out set of training data (cf. table A.2) is in line with the original results by Falk and Lapesa (2022), the regression analysis and thus the exploration of the research question relies on accurate independent variables for meaningful results.

With the differences in annotation behavior on IBM ArgQ and Cornell CMV, the in-domain evaluation might not equate to a good unsupervised annotation. In order to evaluate the performance on the actual dataset, a manual evaluation is conducted on a subset of 150 instances for each corpus, including positive and negative samples of both *mixed-domain* and *one-domain* prediction. As depicted in table 6, the *mixed-domain* and *one-domain* training setups result in vastly different annotation quality. In line with Falk and Lapesa (2022), an improvement in quality can be observed from diversifying the training domain, which results in an increase in F_1 of $\Delta F_1 = .12$ for IBM ArgQ and unexpectedly, as it has the same domain as the *one-domain* classifier, an increase of $\Delta F_1 = .31$ for Cornell CMV. The result-

ing annotation quality scores for both corpora are a good basis for the following regression analysis, though the sparseness in IBM ArgQ might still pose a challenge. Therefore, the average prediction probability of the positive class for all 10 models is used as an alternate feature for the regression analysis. This probability can be interpreted as the model's certainty that an instance includes storytelling, thus any effect that storytelling has is still reduced for instances of low probability, but without completely missing these instances due to their 0 score. Observing these continuous probabilities also reveals a difference in model certainty between the training variants: while the ratio of positive instances is 20 times higher for Cornell CMV, the average probability is only 10 times higher for the *mixed-domain* model (table 5). This average probability score also reveals more about the model behavior, as the *one-domain* training results in only 2.5 times higher probabilities in Cornell CMV than IBM ArgQ. This suggests that the *one-domain* model is less certain about the discrete labels, as the annotation is almost identical for both training variants (45 vs. 46 positive IBM ArgQ instances), thus showing higher overall probabilities, which means the probabilities are not converging as much toward the binary of 0 and 1.

	# Story		**% Story**		**⌀P(Story)**	
Corpus	mix	one	mix	one	mix	one
IBM ArgQ	45	46	0.8	0.9	.02	.11
Cornell CMV	2288	1936	19.8	16.7	.22	.28

Table 5: The automated storytelling predictions on IBM ArgQ and Cornell CMV with the number (#) and percentage (%) of storytelling instances (discrete annotation), and corpus-wide averaged storytelling probability (⌀P, continuous annotation). Each statistic is reported for the annotation following *mixed-domain* and *one-domain* training.

3.3 Emotion

When automatically annotating emotionality in text, there are a variety of annotation schemes and training data to choose from (cf. section 2.2.2). This choice is

variant	Precision	Recall	F_1
one-domain	.61	.78	.68
mixed-domain	.77	.94	.85

(a) IBM ArgQ

variant	Precision	Recall	F_1
one-domain	.33	.27	.30
mixed-domain	.61	.62	.61

(b) Cornell CMV

Table 6: Precision, recall and F_1 scores obtained through a manual evaluation of *storytelling* annotations from the *one-domain* and *mixed-domain* training setup on 150 samples from each IBM ArgQ and Cornell CMV.

particularly relevant to the application of annotating argumentative text to investigate emotionality and argument strength.

Between the main three annotation types of appraisal dimensions, discrete emotion categories and continuous VAD dimensions, discrete categories seem the best choice for a first analysis of emotions in arguments. Unlike the previous feature (section 3.2), there are no emotion corpora from the argument domain, and datasets are very different from IBM ArgQ or Cornell CMV. Many corpora are aggregated from Twitter (Mohammad, 2012; Mohammad and Bravo-Marquez, 2017), or use news headlines or literary text (Alm et al., 2005; Strapparava and Mihalcea, 2007), with the text form of these corpora exacerbating the domain-mismatch that already exists between emotion and argument text.

3.3.1 Training data

Therefore, I am using the Crowd-enVENT dataset (Troiano et al., 2023), that consists of event descriptions given for a specific emotion prompt, as this method leads to emotional information being conveyed implicitly through the situational contents (see table 7 for examples). Between this aggregation procedure, the inclusion of a neutral *no-emotion* class and measures to avoid the overuse of stereotypical emotional events, using Crowd-enVENT as training data might mitigate some domain-adaptation problems with the argument corpora.

Emotion	Generated Text
sadness	I felt . . . when I graduated high school because I remember that I'm growing up and that means leaving people behind.
pride	I baked a delicious strawberry cobbler.
fear	I felt . . . when there was a power outage in my home. That day, my wife and I were cuddling in the sitting room when a thunderstorm started. Then . . . filled me when thunder hit our roof and all the lights went off.
surprise	I got a dog for my birthday

Table 7: Examples of generated event descriptions for three emotions from the CROWD-ENVENT corpus (Troiano et al., 2023) with the emotion words masked.

CROWD-ENVENT consists of 6,600 instances compiled from 550 event descriptions of 10 emotions, the *no-emotion* class and 275 instances each for *guilt* and *shame*, which is motivated by the affinity and unclear distinction between the two labels (Troiano et al., 2023). The emotion classes include the six basic Ekman emotions of *anger*, *disgust*, *fear*, *guilt*, *joy*, and *sadness*, as well as the more complex emotions *boredom*, *pride*, *relief*, *surprise* and *trust* and the already mentioned *shame*. The inclusion of complex emotions is motivated by their relation to appraisal dimensions, 21 of which the author of each event description rated on a Likert-scale. As stated above, this thesis does not make use of appraisals, but the inclusion of complex emotions is in itself advantageous when analyzing arguments, as the *pride* and *trust* classes might capture positive argument strategies not built on *joy* as a basic emotion.

3.3.2 Annotation implementation

In the annotation process for the argument datasets, the annotations in CROWD-ENVENT are adapted to binary labels for each individual emotion, as the single-label restriction in the original data does not hold for the target domain. During this process, the *guilt* and *shame* instances are treated as one for an equal label

ratio in all models, and the resulting 550 positive instances are extended by 1650 instances sampled from all other categories to keep a balanced dataset of 25% positive labels instead of 8% as in the complete CROWD-ENVENT. This is not done for the *no-emotion* label as it is implicitly included in the multi-label setup in instances where all emotion annotations are negative. To investigate different domain-adaptation methods, this process is performed with two different versions of the training data, one including the original generated event descriptions (*original*) and one with salient emotion words masked by "..." (*masked*) to obtain a more robust classifier that learns non-trivial information.

Finally, with two versions for eleven emotion models, the annotation is done analogous to that of the storytelling feature, training ten ROBERTA models on different splits of the data for five epochs and predicting each time with the best model according to the CROWD-ENVENT validation data for each feature variant (11 emotions x 2 training data variants). The ten results are aggregated into a discrete and probability annotation for each emotion (table 10). IBM ARGQ is annotated with only these results, but, as the argumentative texts in CORNELL CMV are much longer than those of IBM ARGQ or CROWD-ENVENT, this annotation misses emotions present in the end of an instance due to the ROBERTA cut-off at 512 (sub-word) tokens. Instead, the corpus provides an opportunity to model emotional progression over the course of the text. Thus, each CORNELL CMV instance is split in half with a three word overlap to then apply the same classifiers to both halves individually and obtain a difference (table 11) and aggregate (table 12).

3.3.3 Annotation result

When observing the results of the emotion annotation, what becomes immediately apparent is the underrepresentation of most emotions in both argument corpora, appearing only in up 5% of the instances and sometimes not at all (*surprise* in IBM ARGQ) or below 1%. Some of this sparseness is expected due to the domain-mismatch between training and argument data making correct detections more difficult. Additional sparseness in the annotation of whole CORNELL CMV instances (see table A.4)

	Precision		Recall		F_1	
Emotion	*mask*	*orig*	*mask*	*orig*	*mask*	*orig*
anger	.52	.45	.77	.41	.62	.43
boredom	.10	.00	1.00	.00	.18	.00
disgust	.48	.42	.82	.65	.61	.51
fear	.70	.50	.50	.21	.58	.30
guilt/shame	.22	.25	.67	.67	.33	.36
joy	.90	.83	.64	.36	.75	.50
pride	.57	.20	1.00	.75	.73	.32
relief	.25	.00	.67	.00	.36	.00
sadness	.62	.29	.89	.22	.73	.25
surprise	.00	.00	.00	.00	.00	.00
trust	.33	.12	1.00	.50	.50	.20
no emotion	.75	.49	.49	.34	.59	.40
average	.45	.30	.70	.34	.50	.27

Table 8: Manual evaluation of instances from IBM ArgQ sampled to include positive annotations for each emotion. The manual annotation was compared to those of the model using *masked* and *original* training data. *no emotion* includes scores for instances with no annotation to confirm the validity of annotation sparseness and the last row contains the average scores over all emotions.

might come from cutting off information during prediction, as stated above. Notable exceptions to this behavior are *anger* and *disgust*, which occur in around a third and half of the IBM ArgQ corpus and half of the Cornell CMV instances respectively. Therefore, a manual evaluation is again conducted on 150 samples each from IBM ArgQ and Cornell CMV, trying to include enough positive annotations from each training variant where possible (i.e., not for *surprise*, which is not annotated in IBM ArgQ). With the Crowd-enVENT data collection of emotional event descriptions, the labels are expert-annotated writer emotions. This annotation scheme is not possible in the manual evaluation, thus a more general notion is adopted, where instances are annotated as they convey an emotion, either from the writer

	Precision				Recall				F_1			
Emotion	*m*	*o*	*m-A*	*o-A*	*m*	*o*	*m-A*	*o-A*	*m*	*o*	*m-A*	*o-A*
anger	.46	.38	.40	.41	.56	.51	.72	.51	.51	.44	.51	.49
boredom	.00	.00	.06	.05	.00	.00	1.00	1.00	.00	.00	.11	.10
disgust	.40	.46	.34	.34	.50	.55	.90	.90	.44	.50	.49	.49
fear	.55	.62	.54	.40	.43	.36	.93	.29	.48	.45	.68	.33
guilt/shame	.73	.40	.50	.53	.62	.15	.85	.69	.67	.22	.63	.60
joy	.80	.90	.83	.87	.22	.50	.83	.72	.35	.64	.83	.79
pride	.50	.69	.47	.37	.58	.75	.67	.92	.54	.72	.55	.52
relief	.33	.25	.32	.46	.33	.17	1.00	1.00	.33	.20	.48	.63
sadness	1.00	.75	1.00	.54	.27	.27	.73	.64	.43	.40	.84	.58
surprise	.25	.00	.25	.40	.12	.00	.25	.50	.17	.00	.25	.44
trust	.67	.40	.58	.50	.75	.25	.88	.62	.71	.31	.70	.56
no emotion	.45	.42	.64	.33	.42	.44	.13	.08	.44	.43	.22	.12
average	.51	.44	.49	.43	.40	.33	.74	.66	.42	.36	.53	.47

Table 9: Manual evaluation of instances from Cornell CMV sampled to include positive annotations for each emotion. The manual annotation was compared to those of the model using *masked* (*m*) and *original* (*o*) training data and the two corresponding aggregated annotations (*m-A, o-A*). The last two rows denote a neutral category to confirm the validity of annotation sparseness and an average of the scores over all emotions.

explaining their feelings or trying to appeal to reader's emotions.

The results in tables 8 and 9 firstly show that there is no consistent annotation performance between the emotions. Instead, F_1-scores vary between 0 and .84. With *surprise* and *boredom* having the lowest scores, any regression results obtained later on from these emotions should not be seen as conclusive evidence for the emotion's influence even with statistically significant effects, as the low evaluation scores suggest that the classifier found some other, related pattern in the data that is annotated instead. Removing the scores of these emotions improves the average F_1 of the best overall annotation method from .50 to .53 for IBM ArgQ and from .53 to .63 for Cornell CMV. Looking at these averages reveals a clear advantage of masking salient emotion words over training on the original texts: in IBM ArgQ,

the performance doubles (cf. table 8), while in Cornell CMV, it improves by .06 (table 9). Another improvement of .11 comes from aggregating Cornell CMV annotations from two halves instead of the whole instance. As the models using *masked* training data furthermore systematically predict more positive instances, it seems like restricting the model's access to explicit emotion words improves generalization. This effect is more pronounced in IBM ArgQ, while the results on Cornell CMV are much more mixed. While the overall best method is masking emotion words and aggregating the annotations, on an individual level some emotions profit from using the original text or annotating the whole instance as one (e.g., *disgust, pride*, table 9). There seems to be no clear reason for these differences, and as the sample size is small and contains very few positive instances of each emotion, I am using the annotations from the overall best methods for all regression experiments to keep the analysis consistent. As such, the following observations are based on the *masked* variant for IBM ArgQ and the *masked, aggregated* variant for Cornell CMV. Further, apart from the outliers of *boredom* and *surprise*, the classification results are on par with those reported by Troiano et al. (2023), as well as matching or even exceeding the results of cross-domain emotion classification reported by Bostan and Klinger (2018).

When comparing the annotations between IBM ArgQ and Cornell CMV, though both corpora are extremely sparse, Cornell CMV has considerably more positive instances for each emotion (compare tables 10 and 11). As the F_1 scores between the two corpora are comparable, this might be characteristic of domain differences. While IBM ArgQ contains short, standalone arguments, those in Cornell CMV are longer and formed as part of a mutual discussion, which leaves more room for emotions. Furthermore, the negative emotions (especially, but even apart from, *anger* and *disgust*) are more frequent than positive emotions, which might imply a bias that arguments appeal to different emotion in different capacities. Lastly, the differences of emotion annotations in the first and second half of Cornell CMV instances shown in table 12 include the much higher ratio of *guilt/shame*, *sadness* and almost all positive emotions in the second half of the text, e.g., with 84.1% more joy in the end of an instance than at the beginning. *Surprise* has an inverse emotion

	#		%		∅P	
Emotion	*mask*	*orig*	*mask*	*orig*	*mask*	*orig*
anger	1,814	1,140	34.2	21.5	.39	.31
boredom	116	69	2.2	1.3	.06	.04
disgust	2,920	2,733	55.1	51.6	.54	.53
fear	347	202	6.6	3.8	.14	.08
guilt/shame	107	42	2.0	0.8	.12	.09
joy	47	23	0.9	0.4	.07	.06
pride	80	319	1.5	6.0	.10	.12
relief	64	27	1.2	0.5	.06	.05
sadness	175	65	3.3	1.2	.14	.06
surprise	0	0	0.0	0.0	.03	.02
trust	112	131	2.1	2.5	.07	.08

Table 10: Results of the automated emotion annotation on IBM ArgQ. The columns denote the number (#) and ratio (%) of positive instances and average probability output (∅P) for each emotion. Each statistic is reported for the annotation following training with *masked* and *original* emotion texts.

progression with 50% less emotion in the second halves' annotations. These big ratios are however misleading in that the emotions are generally sparse, thus, the big 84.1% difference in *joy* only includes 53 instances. This caveat equally holds for the tentative interpretations of the annotation study, and to alleviate some effects of the sparseness, analogous to the storytelling feature, I will also include the classification probabilities in the regression analysis alongside the discrete annotations.

3.4 Hedging

Following traditional hedge detection methods, I am annotating the argument corpora with a lexicon- and rule-based algorithm. Hedging is tied to specific words and phrases, which obviates complex machine learning. The right lexicon to apply to

	#		%		∅P	
Emotion	*mask*	*orig*	*mask*	*orig*	*mask*	*orig*
anger	6,467	5,864	55.9	50.7	.43	.40
boredom	538	523	4.7	4.5	.07	.07
disgust	5,111	5,002	44.2	43.3	.37	.37
fear	822	260	7.1	2.3	.11	.04
guilt/shame	631	473	5.5	4.1	.14	.09
joy	208	162	1.8	1.4	.05	.03
pride	615	1,207	5.3	10.4	.12	.14
relief	256	157	2.2	1.4	.06	.06
sadness	429	389	3.7	3.4	.12	.08
surprise	53	26	0.5	0.2	.04	.04
trust	159	142	1.4	1.2	.04	.03

Table 11: Results of the automated emotion annotation on Cornell CMV aggregated from the annotation for the first and second half of each instance. The columns denote the number (#) and ratio (%) of positive instances and average probability output (∅P) for each emotion. Each statistic is reported for the annotation following training with *masked* and *original* emotion texts.

the argument data must handle informal speech well, as neither corpus is similar to the scientific domain typical for early hedge research and Cornell CMV includes internet terms and abbreviations. In the informal domain, hedging includes uncertainty markers such as *might* or *probably* as well as more subjective phrases such as *I believe* and online acronyms like *AFAIK/as far as I know* or *IMHO/in my honest opinion* (Mamani Sanchez and Vogel, 2013).

3.4.1 Annotation implementation

Thus, I am combining the resources from Ulinski and Hirschberg (2019), Islam et al. (2020) and Mamani Sanchez and Vogel (2013) into a unified lexicon of hedge terms. The lexicon includes individual hedge tokens and multi-word hedges (e.g., *in my*

	Δ		%	
Emotion	*mask*	*orig*	*mask*	*orig*
anger	−17	13	−0.4	0.3
boredom	26	3	9.1	1.0
disgust	103	−4	2.9	−0.1
fear	11	12	7.7	2.4
guilt/shame	94	58	45.4	18.2
joy	53	60	84.1	70.6
pride	205	92	33.0	28.0
relief	24	36	33.8	30.5
sadness	55	51	28.6	23.5
surprise	−9	−12	−50.0	−35.3
trust	24	37	36.4	54.4

Table 12: Absolute (Δ) and relative (%) difference of emotion counts between the halves of CORNELL CMV instances calculated as (*end* − *beginning*), i.e., positive values denote more and negative values denote fewer instances in the second half. The sub columns denote the differences between the annotations obtained after training on the unaltered training data *orig* or data with masked emotion words (*mask*).

opinion) and is complemented by a lexicon of words that express high certainty and are in their negated form (e.g., *I am not certain*) analogous to Islam et al. (2020). The annotation procedure starts by preprocessing each argument instance with stanza (Qi et al., 2020), a Stanford CoreNLP adaptation for Python. The text is tokenized and split into sentences which are then annotated with universal POS tags and morphological features, and lastly parsed for dependency relations. The preprocessed instances are then sentence-wise annotated for hedges by first matching hedge terms from the lexicon and applying syntactical rules based on the POS and dependency information to disambiguate certain terms that also carry a non-hedge meaning. Table 13 shows some of these rules adopted from Islam et al. (2020) and Ulinski and Hirschberg (2019) and expanded for other ambiguous terms and the negation of certainty terms. Such detected hedges are tallied for each sentence and

Term	Rule
about, *around*	If the token is used as an adjective (part-of-speech IN), it is a non-hedge. Otherwise, it is a hedge. **Hedge:** There are *around* 10 million packages in transit right now. **Non-hedge:** We need to talk *about* Mark.
pretty	If the token is used as adverbially, it is a hedge. **Hedge:** I am *pretty* certain about this statistic. **Non-hedge:** She has a really *pretty* cat.
impression	If the token has a 1. person possessive pronoun as dependent or its head has a 1. person nominal subject as a second dependent, it is a hedge. **Hedge:** I get the *impression* that we have to wait longer for official information. **Non-hedge:** The protagonist's performance left a lasting *impression* on everyone.

Table 13: Examples for the syntactical hedge disambiguation rules, the first of which is lifted from Islam et al. (2020).

then saved in multiple features for each instance.

3.4.2 Annotation results

The features include the number and ratio of hedge words in the first and last sentence, the number of hedge words in the whole instance and the average ratio of hedge words in all sentences of an instance (table 14). The annotation results mainly show the difference in instance length between IBM ArgQ and Cornell CMV, with the highest number of hedges in a single sentence being 5 for IBM ArgQ and 9 for Cornell CMV, and the overall number of hedges per instance being nine times higher in Cornell CMV although the ratio of hedges per sentence does not diverge as much. As this feature is annotated not though prediction but a deterministic term-matching, a manual evaluation of the annotation quality is not necessary.

	# First		% First		# Final		% Final		# All		% All	
Corpus	∅	max	∅	max	∅	max	∅	max	∅	max	∅	max
Cornell CMV	0.72	8	0.04	0.5	0.92	9	0.04	0.43	9.15	93	0.04	0.33
IBM ArgQ	0.65	1	0.10	0.2	0.58	5	0.03	0.33	1.35	7	0.06	0.22

Table 14: The average hedge features annotated for Cornell CMV and IBM ArgQ including the absolute number (#) and ratio (%) of hedge words for the first and final sentence, as well as the number of hedge words in the whole instance (# All) and the average hedge ratio of all sentences (% All). For each feature, the first column (∅) contains the corpus average while the second (max) contains the maximum score of any instance.

4 Analysis

The automatic annotation of argument data from Cornell CMV and IBM ArgQ allows to investigate the link between subjective features and argument strength. Toward this goal, I am employing linear and logistic regression in an incremental process allowing for systemic comparison. The basis for this method is explained in the following section, to then first investigate each feature separately and finally arrive at an analysis including the full set of features.

4.1 Regression as analysis

Regression is at its core a machine learning method that learns the best weights to combine different independent variables and predict a dependent variable. As such, the explanatory quality of the model and the values of the different independent variables' (IV) weights can be used to analyze their impact on the dependent variable (DV).

Linear Regression In simple linear regression, i.e., regression with just one independent variable, a model fit with n observations has the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

for each observation i, thus the model fits the line most close to all n observations of the form (x, y). The β coefficients are the weights optimized to predict all y, with β_0 being a constant, also called intercept, and ϵ_i denoting the residual error of each sample i. The assumptions of linear regression define that the individual error estimate, that is the variance of each observation and not the whole model, is 0 (Rencher and Schaalje, 2008). When introducing multiple (k) independent variables, the model has the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i \tag{2}$$

for every observation i, which for the whole sample of n observations can be simplified to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \equiv \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \vdots \\ 1 & x_{n1} & \ldots & x:nk \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{3}$$

To fit the model to a sample of observations, the β coefficients are optimized such that the plane or hyperplane of estimated $\hat{\mathbf{y}}$ that is defined by $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is as close to all observed $\mathbf{y}$ by minimizing the sum of squares of the error $y_i - \hat{y}_i$:

$$\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2. \tag{4}$$

To obtain individual β coefficients, the partial derivative for each β_j $(0 \leq j \leq k)$ is set equal to 0. This partial derivative constitutes each β_j as a *partial* regression coefficient that indicates the change in y_i when x_{ji} increases by 1 unit when the other elements of x_i are constant (Rencher and Schaalje, 2008). For multiple linear regression, using the same notation as in 3, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is equivalent to calculating each derivative simultaneously.

Minimizing the sum of least squares results in the best regression $\hat{y}$ given all k independent variables x, though it gives no indication on the explanatory power of this best model. The coefficient of determination r^2 denotes the ratio of variance

in y that is explained by the regression model (Rencher and Schaalje, 2008) and is calculated from the regression sum of squares (SSR) and the total sum of squares (SST),

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5}$$

where the total sum of squares includes the regression sum of squares and the error sum of squares also used in the least squares method. As such, in a model with $r^2 = 1$, the independent variables explain the variance in the dependent variable perfectly, while in a model with $r^2 = 0$, the independent variables explain none of the dependent variable.

Logistic regression If the dependent variable y is not continuous but rather dichotomous, as is the persuasiveness label in Cornell CMV, a linear regression model is unfit, since the variable has only two manifestations of 0 and 1 and thus violates the assumption of linear regression, that the variance or error of $\hat{y}_i$ is constant and does not depend on x_i (Rencher and Schaalje, 2008). Instead of a linear estimate $E(y_i) = \hat{y}_i$, the regression model rather estimates the probability of y_i being 1, i.e., $E(y_i) = p(y_i = 1)$. Logistic regression keeps this probability estimate bound by 0 and 1 and can be described as

$$p_i = \frac{1}{1 - e^{-\mathbf{X}\boldsymbol{\beta}}}. \tag{6}$$

As the model is no longer linear with a discrete DV, the β coefficients are estimated with maximum likelihood which calculates the joint density of y's. Given a normal distribution of $\mathbf{y}$ and the normal joint density function $L(\boldsymbol{\beta}, \sigma^2)$ however, it results in

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

analogous to the calculation for multiple linear regression (Rencher and Schaalje, 2008). The β coefficients in logistic regression do not express the same as in linear regression. Where in the latter a coefficient of $\beta_i = -0.5$ denotes a reduction of the dependent variable by 0.5 for each unit increase of the corresponding x_i, logistic

regression models probabilities and as such the coefficients denote changes in the odds of an outcome. Namely, the same $\beta_i = -0.5$ is the logarithm of the *odd ratio* of the corresponding feature and can be exponentiated such that $e^{\beta_i} = 0.61$ denotes the odds of the positive class of y_i, i.e., $p(y_i = 1)$ are 0.61 when x_i is 1. The odds ratios can thus have values between 0 and infinity, with 1 denoting completely equal odds between the two outcomes of y_i. As a second consequence of changing to a discrete DV, the coefficient of determination r^2 is no longer usable to determine model fit. A prediction error on discrete variables does not result in a small variance which can be measured against the perfect model, thus, multiple approximations of r^2 are available, from which I use the pseudo-r^2 by McFadden (1973), which compares the log-likelihood of the fitted model to a constant, intercept-only model. As this measure does not reach values of 1 like r^2, a pseudo-r^2 of $0.2-0.4$ already constitutes an excellent fit (McFadden, 1973), thus I'm using the measure mostly to compare between different logistic regression models instead of determining exact model fit.

Analysis Setup Employing linear regression on the IBM ArgQ data and logistic regression on Cornell CMV, I am first investigating each feature separately through simple regression to compare the different feature variants. The resulting best features are then added to consecutively more complex models through stepwise regression. In this process, for each step up in complexity, the best feature to add at that time is chosen through comparing improvements in r^2 (pseudo-r^2 for the logistic regression case) of all remaining unused features. When introducing interaction, I am substituting the explained variance score r^2 for the Akaike Information Criterion, which weighs the explanatory improvement against the increase in model complexity, to arrive at the best model with the least features. Lastly, while stepwise adding features, I am keeping track of the significance of the explanatory improvement by comparing the new model to the smaller model with an ANoVA.

Anno	adj. r^2	Coef	p
discrete	0.0042	-0.148	0.0 ***
probability	0.0047	-0.182	0.0 ***

(a) Linear regression on IBM ArgQ

Anno	pseudo-r^2	Odds	p
discrete	0.00019	1.084	0.084
probability	0.00037	1.148	0.015 *

(b) Logistic regression on Cornell CMV

Table 15: Results of the simple regression on IBM ArgQ argument quality and Cornell CMV persuasiveness as DVs respectively with *discrete* and *probability* annotations of storytelling as IV. Reported are the adjusted r^2, the coefficient of the feature variant and its significance given a t-test assuming the same outcome with the coefficient set to 0.

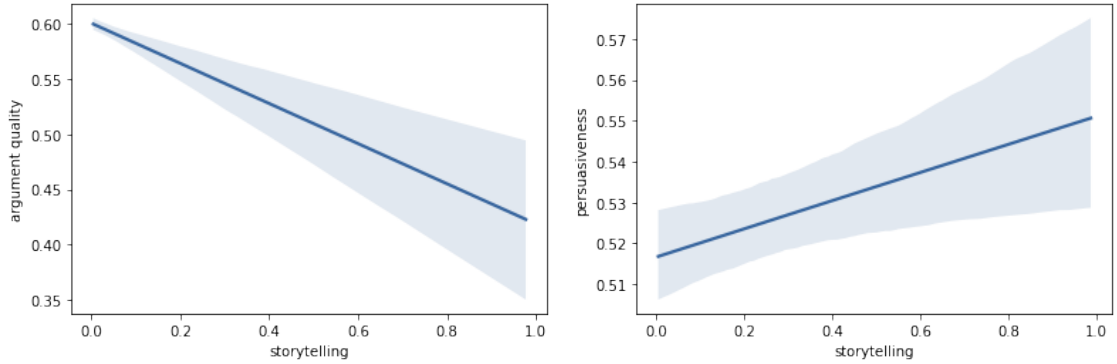
4.2 Investigation of individual features

4.2.1 Storytelling

The storytelling feature has four variants combined from the two training data variants of *mixed-domain* and *one-domain* and, due to the sparseness of the *discrete* annotation, a second annotation variant using *probabilities*. As the *mixed-domain* annotation performed best on both IBM ArgQ and Cornell CMV, the following experiments use only this annotation.

An immediate observation that follows through with all other features is the low r^2 and pseudo-r^2 value of each regression (IBM ArgQ $r^2 = 0.47\%$ and Cornell CMV pseudo-$r^2 = 0.037\%$, cf. table 15). This is somewhat expected from the previously stated complexity of a measure like argument strength, as no feature alone will have a deciding impact on quality or persuasiveness. The scores can nonetheless be used to obtain relative explanatory power comparisons between features, and the p-values of the coefficients shed light on the significance of effects regardless of size. Comparing the variants of storytelling then, the results confirm the choice to include of *probability* scores alongside the normal *discrete* storytelling label, as the probability features systematically explain more variance than the *discrete* labels (cf. table 15).

In all other aspects, the results on IBM ArgQ and Cornell CMV demonstrate diverging behavior. While both annotation types have a statistically highly signifi-



(a) Linear regression on IBM ArgQ. **(b)** Logistic regression on Cornell CMV.

Figure 5: Results of regressing *mixed-domain storytelling probabilities* on IBM ArgQ argument quality and Cornell CMV persuasiveness respectively. The graph includes the confidence interval of the model.

cant effect on IBM ArgQ argument quality, on the Cornell CMV data, only the effect of *probability* storytelling labels on persuasiveness is statistically significant (p=.015).

Furthermore, storytelling seems to interact differently with argument strength in the two argument corpora. As apparent from figure 5a, in IBM ArgQ storytelling has a highly significant, negative effect on the argument quality, with the *probability* feature reducing argument quality by .18. Even *discrete* storytelling reduces quality by .15, which means at least a 15% decrease in argument quality for positive instances of storytelling (if starting from a perfect argument with a score of 1). The effect on Cornell CMV persuasiveness (figure 5b) is however positive – with the *probability* feature increasing the odds of persuasiveness by 14.8%.

4.2.2 Emotion

As shown in the manual evaluation of the emotion annotations, while the feature variant with the best F_1 varies from emotion to emotion, the overall best variant uses *masked* training data and, in the case of Cornell CMV, is *aggregated* from annotations of the two halves of each instance. While it might be argued that the

best regression results are obtained when selecting the feature variant combination (*masked/original*; *aggregated/whole*) individually based on the F_1 of each emotion, sampling bias resulting from the sparseness of the data and the small sample size of the manual evaluation complicates decisions made from individual emotions' results. Thus, the following regression experiments will use the overall best variant (IBM ArgQ: *masked*, Cornell CMV: *masked, aggregated*) to ensure results are comparable.

Similar to storytelling, the annotation variant of emotion *probabilities* is more informative than *discrete* emotion labels for almost all emotions. Notable exceptions are *pride* and *relief* on IBM ArgQ and *anger, boredom, sadness* and *trust* on Cornell CMV. Comparing p-values for these emotions shows that apart from IBM ArgQ's *relief*, neither the more informative *discrete* nor the *probability* annotation have significant effects. While the effect of *discrete relief* is highly significant (p=.0009), the *probability* effect is also significant (p=.02; table 16). Therefore, the in depth investigation will use the *probability* annotation.

Regressing each emotion on argument strength (table 16) reveals a higher number of significant effects on argument quality in IBM ArgQ than on persuasiveness in Cornell CMV. The most informative emotion for IBM ArgQ is *guilt/shame*, explaining 0.97% of the variance in argument quality with a highly significant negative effect. In general, the effect directions of emotions on IBM ArgQ follow emotion polarity lines, with *anger, boredom, disgust* and *guilt/shame* having significant to highly significant negative effects and *joy, relief* and *trust* having significant to highly significant positive effects (the positive effect of *pride* is not significant). The exceptions to this are *sadness, fear* and *surprise*, with the latter having an insignificant negative effect while the former two emotions have significant and highly significant positive effects on argument quality. Thus, before combining features, the most informative emotions for IBM ArgQ argument quality are *guilt/shame, trust* and *joy*, while the least informative emotions are *surprise, pride* and *boredom*.

On Cornell CMV, the emotion effects are smaller, though the same mapping of emotion polarity and effect direction exists here – notably also including the negative effect of *surprise* and positive effects of *fear* and *sadness*. While the effect of *sadness*

	IBM ArgQ			Cornell CMV		
	r^2	P-value	Coef	pseudo-r^2	P-value	Odds
anger	0.0011	0.009 **	−0.026	0.0000	0.377	0.928
boredom	0.0006	0.042 *	−0.050	0.0000	0.487	0.897
disgust	0.0022	0.00 ***	−0.031	**0.0010**	0.0 ***	0.751
fear	0.0026	0.0 ***	0.056	0.0003	0.035 *	1.307
guilt/shame	**0.0097**	0.0 ***	−0.139	0.0005	0.006 **	0.640
joy	0.0065	0.0 ***	0.173	0.0001	0.149	1.397
pride	0.0003	0.091	0.037	0.0003	0.042 *	1.365
relief	0.0008	0.023 *	0.063	0.0005	0.007 *	1.749
sadness	0.0007	0.031 *	0.044	0.0000	0.470	1.138
surprise	0.0003	0.111	−0.182	0.0003	0.042 *	0.489
trust	0.0067	0.0 ***	0.140	0.0000	0.654	0.886

Table 16: Results of regressing IBM ArgQ argument quality and Cornell CMV persuasiveness on the individual emotion *probability* using the *masked* and *masked, aggregated* annotation respectively. The reported values are the (pseudo-) r^2 score of each regression model, the emotion's coefficient or odds (equivalent to exp(coefficient) for logistic regression) and its p-value and significance ($p<.001$ – ***; 0.01 – **; 0.05 – *) given a two-sided t-test.

is not statistically significant, the other two effects are. However, with an odds ratio of 0.654, *trust* is joining *surprise* as a positive emotion impacting persuasiveness negatively. The most informative feature here is *disgust*, which results in a highly significant decrease of persuasiveness odds. Apart from *disgust*, the only other negative emotion with a significant (**) negative effect is *guilt/shame*, with significant (*) positive effects holding for *pride* and *relief*. The last emotion with a statistically significant effect on persuasiveness is *fear*, which influences persuasiveness positively. Thus, the most informative emotions for Cornell CMV persuasiveness are *disgust, guilt/shame* and *relief*, while the remaining emotion effects of *anger, boredom, joy, sadness* and *trust* do not approach significance on their own.

Notable similarities between IBM ArgQ and Cornell CMV are the highly sig-

nificant negative effects of *disgust* and *guilt/shame* and the significant positive effect of *fear* on both measures of argument strength, suggesting a universal pattern to keep in mind during the following combinatory experiments.

Emotional progression As mentioned above (section 3.3), the process of aggregating the annotations of an instance's first and second half for each emotion in Cornell CMV allows to also model the broad progression of each emotion, i.e., if the emotion is constant throughout the instance or if it goes up or down during the course of the argument. This feature again uses the best, i.e., *masked*, training variant and is calculated for both *discrete* and *probability* annotations as $\Delta(emo) = emo_{end} - emo_{start}$. However, as table 17 shows, the progression of all emotions is insignificant with either annotation type and with the progression of *sadness* and *surprise probability* nearest to significance (p=.09 and p=.11 respectively). Both features negatively effect persuasiveness, i.e., when *sadness* or *surprise* occur at the end but not the beginning of an instance, the odds of persuasiveness are lowered, while they are improved if either emotion occurs only in the beginning but not the end of an instance. As none of the other features' effects are close to statistically significance, I am not including the emotion progression on the combined regression models. Reasons for the overall insignificant results apart from a genuine lack of effect are discussed further in section 5.

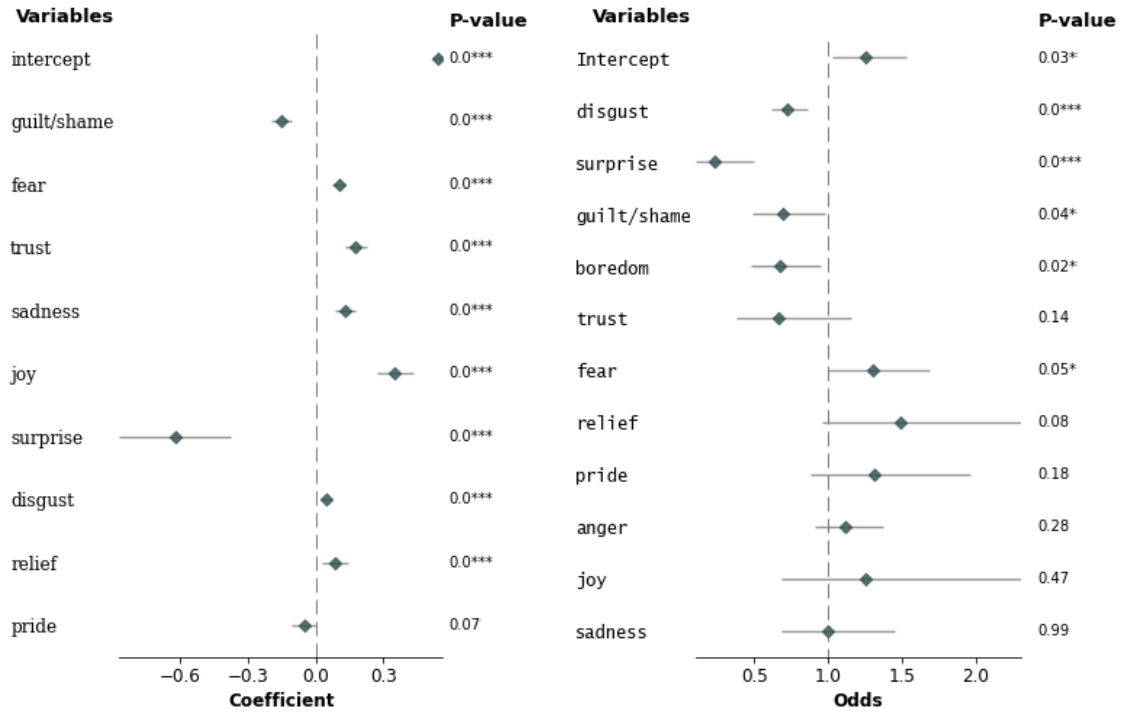
Combined emotion effects Concluding the investigation of emotions alone, to fully grasp the effect of emotions on argument strength, a stepwise regression of emotions without interactions is fitted and shown in figure 6 and table 18.

In line with the higher number of individually significant emotion effects in IBM ArgQ, all but two emotions, *anger, boredom*, are added by the stepwise regression before the explanatory improvement is too small, and all effects except for *pride* are significant in the presence of the other emotion IVs. The behavior of *fear* and *sadness* improving and *surprise* lowering argument quality can still be observed, though *disgust* and *pride* now also influence argument quality contrary to their polarity. However, the large confidence interval of the negative effect of *surprise* indicates that

	discrete			*probabilities*		
	pseudo-r^2	Odds	p	pseudo-r^2	Odds	p
anger	0.0000	1.004	0.90	0.0000	1.056	0.46
boredom	0.0000	1.022	0.81	0.0001	1.179	0.18
disgust	0.0000	0.994	0.88	0.0000	0.997	0.97
fear	0.0000	1.053	0.51	0.0000	0.970	0.79
guilt/shame	0.0000	0.928	0.37	0.0000	0.964	0.77
joy	0.0000	0.906	0.50	0.0001	0.834	0.32
pride	0.0000	0.977	0.80	0.0001	1.148	0.32
relief	0.0000	1.067	0.62	0.0000	0.982	0.90
sadness	0.0000	0.997	0.98	0.0002	0.783	0.09
surprise	0.0001	0.707	0.23	0.0002	0.643	0.11
trust	0.0000	1.098	0.58	0.0000	0.949	0.79

Table 17: Results of regressing the progression of each emotion separately as IV on Cornell CMV persuasiveness as DV. The reported values are the pseudo-r^2 of each regression model, the odds (exp(coef)) of each emotion and its coefficient's p-value given a two-sided t-test.

the exact coefficient is not representative. This may be the result of the low frequency of the emotion (no positive instances in IBM ArgQ with *discrete* annotation) but also the result of collinearity. To confirm that none of the independent variables are dependent on each other and that the confidence interval is not the result of such unaddressed multicollinearity, the Variance Inflation Factor (VIF) is calculated for the whole feature set of both corpora (i.e., including *storytelling* and *hedge*). IBM ArgQ (table B.7a) shows the highest values for *anger* (4.70), *disgust* (5.98) and *first hedge* (4.49), which, while expected from the three features present in around half of IBM ArgQ (cf. tables 10 and 14), approaches the threshold of multicollinearity for *disgust*. As *anger* is not added during the stepwise regression, the multicollinearity is not a problem in this regression and the VIF values are lowered by removing *anger* from the feature set available to the full interaction model. Thus, the first interpretation of the behavior of *surprise* can be maintained. Apart from *surprise*,



(a) IBM ArgQ argument quality **(b)** Cornell CMV persuasiveness

Figure 6: Emotion IV effects in stepwise linear and logistic regression on IBM ArgQ and Cornell CMV respectively, showing the effect size and confidence interval of each feature. IBM ArgQ effects are displayed as coefficients, and Cornell CMV effects are exponentiated coefficients, i.e., the odds-ratios.

the largest effects on argument quality are the negative effect of *guilt/shame* and the positive effect of *joy*, which coincides with the most informative features during simple regression ($r^2(guilt/shame) = .97\%$, $r^2(joy) = .65\%$).

Unlike IBM ArgQ, all emotions are added in the Cornell CMV combined model, i.e., the model improves with each added emotion, but the improvement from the last seven emotions and their individual effect sizes are not significant. Instead, only *disgust, surprise, guilt/shame, boredom* and *fear* have significant effects in the company of other emotion IVs. With the previously significant effects of *pride* and *relief* (p=.042 and p=.007 respectively) rendered insignificant by the inclusion of more informative emotions, Cornell CMV persuasiveness is significantly influenced

Emotion IV	adj. r^2	sign.
guilt/shame	0.97%	
+ fear	1.38%	***
+ trust	1.88%	***
+ sadness	2.32%	***
+ joy	2.95%	***
+ surprise	3.40%	***
+ disgust	3.62%	***
+ relief	3.75%	**
+ pride	3.80%	.
− anger	−	
− boredom	−	

(a) IBM ArgQ argument quality

Emotion IV	pseudo-r^2	sign.
disgust	0.0010	
+ surprise	0.0015	**
+ guilt/shame	0.0019	*
+ boredom	0.0024	**
+ trust	0.0026	.
+ fear	0.0027	
+ relief	0.0029	
+ pride	0.0030	
+ anger	0.0031	
+ joy	0.0031	
+ sadness	0.0031	

(b) Cornell CMV persuasiveness

Table 18: Adjusted r^2 (IBM ArgQ) and pseudo-r^2 (Cornell CMV) for each regression on the argument quality (IBM ArgQ) and persuasiveness (Cornell CMV) with stepwise added emotion IVs (the last two emotions in IBM ArgQ were not added). The significance of adding each emotion is tested via ANOVA for IBM ArgQ and nested F-test for Cornell CMV.

only by negative and negatively correlated (*surprise*) emotions. Of this set, only *guilt/shame* and *disgust* had significant effects alone. Further, the effect of *trust*, which is insignificant in both simple and combined regression, changes direction, i.e., while *trust* on its own improves the odds of persuasiveness, in the presence of the more informative emotions it is harmful for persuasiveness. The VIF scores for Cornell CMV (table B.7b) do not have a similarly high outlier as IBM ArgQ. The overall larger confidence intervals of the latter half of stepwise added features is thus a result mostly of the insignificance of the effect, i.e., the low explanatory power of the variables.

In conclusion, there seem to be three main emotions to take note of: *guilt/shame* has a negative effect on argument strength that is consistent and statistically in both corpora and in the presence of other emotions. *Fear* as a similarly consistent positive

effect that remains significant through both simple and combined regression, similar to the non-significant behavior of *sadness*. Lastly, *disgust* on its own has a highly significant negative effect in both corpora, which remains when combined with other emotions in Cornell CMV but in IBM ArgQ is reversed and diminished in effect size in combination with other emotions.

4.2.3 Hedging

The hedging feature has six variants that denote the absolute and relative number of hedge terms in the first, final and all sentences.

As tables 19 and 20 show, hedging is similar to storytelling in how its effect differs between the two argument corpora. When individually regressing the hedging variants on argument quality in IBM ArgQ, all effects but the absolute number of hedges in the final sentence are significant. Furthermore, of all hedging variants, only this insignificant effect on argument quality is positive, when all other hedging variants negatively influence argument quality in IBM ArgQ. The two most informative variants are the absolute number of hedge terms in the first sentence and the average hedge ratio of all sentences. As the first sentence in IBM ArgQ always includes only one or no hedge terms (cf. table 14), the feature is binary, thus the occurrence of a hedge in the beginning of an argument reduces its quality score by $\beta = .029$.

In Cornell CMV however, only two hedging variants, the overall absolute number of hedges and the average hedge ratio of all sentences, have a significant effect with none of the other variants approaching significant p-levels. The overall absolute number of hedges per instance has highly significant odds of 1.03 for persuasiveness. As the annotation results showed (table 14), a Cornell CMV instance includes 9 hedge terms on average and up to 93 terms at maximum, thus even with an odd ratio close to 1, the practical effect size of this feature is large. Surprisingly, the average hedge ratio of the whole instance is the only other significant variant and as the second most informative feature has a negative effect, as these features are not only collinear but also describe the same information in different ways, it would be

Score	Sent	r^2	Coef	P-value	
absolute	*first*	**0.0044**	-0.029	**0.0**	***
	final	-0.0002	0.001	0.894	
	all	0.0027	-0.011	0.0	***
ratio	*first*	0.0036	-0.160	0.0	***
	final	0.0007	-0.159	0.026	*
	all	0.0036	-0.296	0.0	***

Table 19: Results of individually regressing each hedging variant as IV on IBM ArgQ argument quality as DV. The variants are divided by **Score** type and for which **Sentence** the score is calculated. Reported are the adjusted r^2, the coefficient of the feature variant and its significance given a t-test assuming the same outcome with the coefficient set to 0.

expected to find similar effects.

Score	Sent	pseudo-r^2	Odds	P-value	
absolute	*first*	0.00005	1.018	0.358	
	final	0.0	0.999	0.947	
	all	**0.01056**	1.030	**0.0**	***
ratio	*first*	0.00002	1.235	0.565	
	final	0.00012	0.579	0.174	
	all	0.00035	0.124	0.018	*

Table 20: Results of individually regressing each hedging variant as IV on Cornell CMV persuasiveness as DV. The variants are divided by **Score** type and for which **Sentence** the score is calculated. Reported are the pseudo-r^2 and **Odds** (exponentiated coefficient) of each feature variant and its significance given a t-test assuming the same regression outcome with the coefficient set to 0.

4.3 Combining all subjective features

Analogous to the combined emotion models, all three subjective features are first used in stepwise regression without interaction. This approach allows to distinguish

IVs	adj. r^2	sign.
guilt/shame	0.97%	
+ storytelling	1.49%	***
+ sadness	1.88%	***
+ trust	2.48%	***
+ joy	2.98%	***
+ fear	3.56%	***
+ surprise	3.97%	***
+ disgust	4.17%	***
+ first hedge	4.34%	**
+ relief	4.57%	***
− anger	–	
− boredom	–	
− pride	–	

(a) Explained variance in IBM ArgQ argument quality

IVs	pseudo-r^2	sign.
disgust	0.0010	
+ surprise	0.0015	**
+ guilt/shame	0.0019	*
+ boredom	0.0024	**
+ storytelling	0.0027	*
+ fear	0.0029	.
+ trust	0.0030	
+ relief	0.0032	
+ pride	0.0033	
+ anger	0.0034	
+ joy	0.0034	
+ final hedge	0.0034	
+ sadness	0.0034	

(b) Explained variance in Cornell CMV persuasiveness

Table 21: Adjusted r^2 (pseudo-r^2) for each regression on the argument quality (persuasiveness) on IBM ArgQ (Cornell CMV) with stepwise added subjective feature IVs (*anger*, *boredom* and *pride* were rejected in IBM ArgQ stepwise regression). The significance of adding each feature is tested via ANOVA for IBM ArgQ and via F-test for Cornell CMV.

between how the individual effects of all features are influenced by the inclusion of others, to then introduce two-way interaction and observe the effect of these interactions separately.

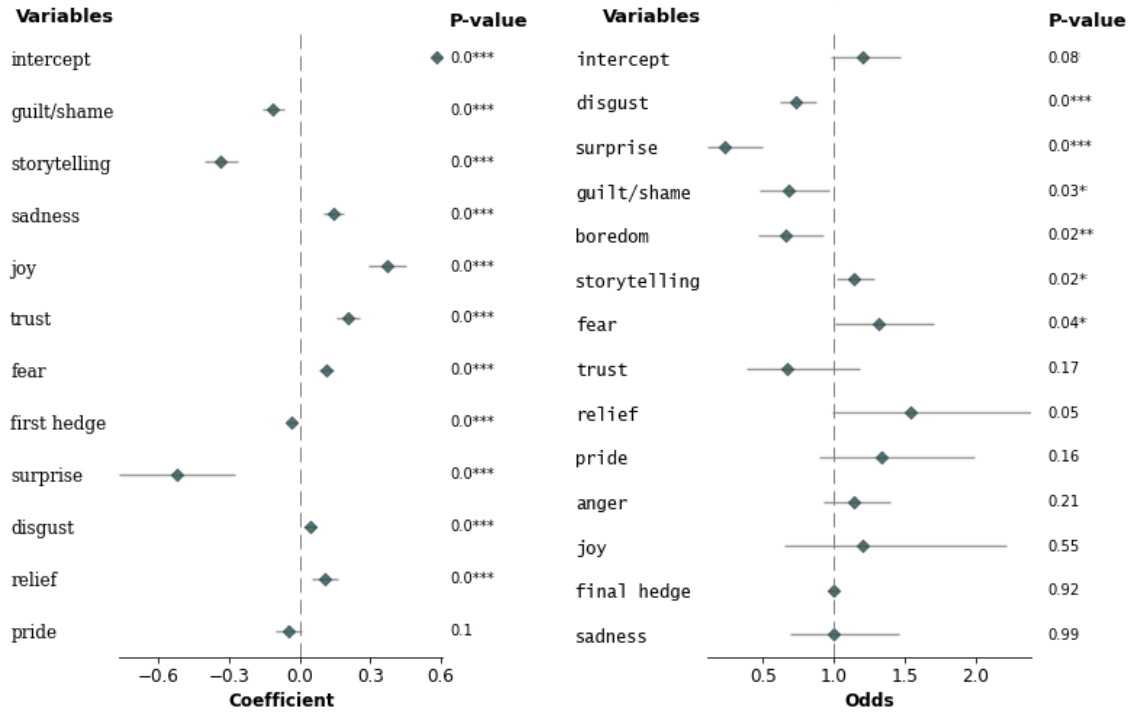
The results of the stepwise regression with all features mirrors the those of the emotion features insofar as for IBM ArgQ, three features are not part of the full model, but all added features significantly improve the model fit, while in the full Cornell CMV model, all available features are added to the model but only the first five are significant (cf. table 21).

For IBM ArgQ, stepwise regression with all results in a model with 9 IVs that

explains 4.57% of argument quality variance, which is the highest coefficient of determination thus far. All 9 features are statistically significant in both additional variance explained by the model and the effect size of the coefficient. The most informative, i.e., first selected feature is *guilt/shame*, followed by *storytelling* (cf. table 21a). Besides *surprise*, the largest negative effect comes from these two most informative features with $\beta = -0.21$ for *storytelling* and $\beta = -0.11$ for *guilt/shame*. While the negative effect of *surprise* on IBM ArgQ again has a uniquely large confidence interval as in all previous experiments, the high significance levels and small confidence intervals for all other coefficients attest to meaningful effects regardless of the small coefficient of determination. Further, these effects are in line with the features' individual regression results, as are the largest positive effects on argument quality, *joy* ($\beta = 0.32$) and *trust* ($\beta = 0.20$), and the positive effects of *fear* and *sadness*.

The features significantly improving regression on Cornell CMV persuasiveness are *disgust, surprise, guilt/shame, boredom* and *storytelling*, with *disgust* being the most informative feature and *storytelling* the last feature to statistically significantly improve the model (cf. table 21b). As more than half of the features are not significant in either added explanatory power or effect size, the large confidence intervals are expected. Of the feature with significant effects, the largest positive influence comes from *fear* (odds= 1.31), while the largest negative influence comes from *surprise* (odds= 0.23), i.e., a certain ($p(emo) = 1$) occurrence of *fear* increases the odds of persuasiveness by 31% and the occurrence of *surprise* decreases them by 77% given fixed values for all other features. Thus, *guilt/shame, surprise, storytelling* and *fear* are the features with significant standalone effects, regardless of the presence of other IVs.

With this, the last question is that of the interaction of the subjective features, which is tested by fitting a final full model for each dataset. For the full model including two-way interactions, the stepwise regression includes all individual features and their two-way interactions (bar *anger* in IBM ArgQ to avoid effects of multicollinearity) as available IVs to add in each step. As this constitutes an overly large amount of variables (12/13 features + 66/78 interactions), the stepwise algorithm



(a) IBM ArgQ argument quality **(b)** Cornell CMV persuasiveness

Figure 7: IV effects in stepwise regression of all subjective features, showing the effect size and confidence intervals of each feature and the significance of the coefficient. IBM ArgQ effects are displayed as coefficients, and Cornell CMV effects are exponentiated coefficients, i.e., the odds-ratios.

using only r^2 increase as its metric may inflate the amount of features meaningful explanatory improvement to the model. Thus, the metric is changed to the *Akaike Information Criterion*, which balances the increase in explanatory power with the increase in complexity, resulting in the model explaining the most DV variance with the least amount of IVs. The stepwise regression process stops once the more complex model's improvement in AIC over the less complex model falls below 0.5, which happens after 17 IVs for IBM ArgQ (cf. table 22a) and after just 7 IVs for Cornell CMV (cf. table 22b).

The full model for IBM ArgQ reaches 5.8% explained argument quality variance and all variables bar one interaction have significant effects on argument quality

IVs	adj. r^2	sign.
guilt/shame : surprise	1.512%	x
+ fear : sadness	2.166%	***
+ trust	2.660%	***
+ boredom : trust	3.122%	***
+ storytelling	3.575%	***
+ joy	3.999%	***
+ disgust : sadness	4.314%	***
+ surprise : trust	4.558%	***
+ fear	4.785%	***
+ first hedge : pride	4.986%	***
+ pride : relief	5.272%	***
+ first hedge : boredom	5.353%	*
+ boredom : sadness	5.460%	**
+ fear : relief	5.535%	*
+ guilt/shame : trust	5.600%	*
+ first hedge : trust	5.698%	*
+ pride : storytelling	5.759%	*

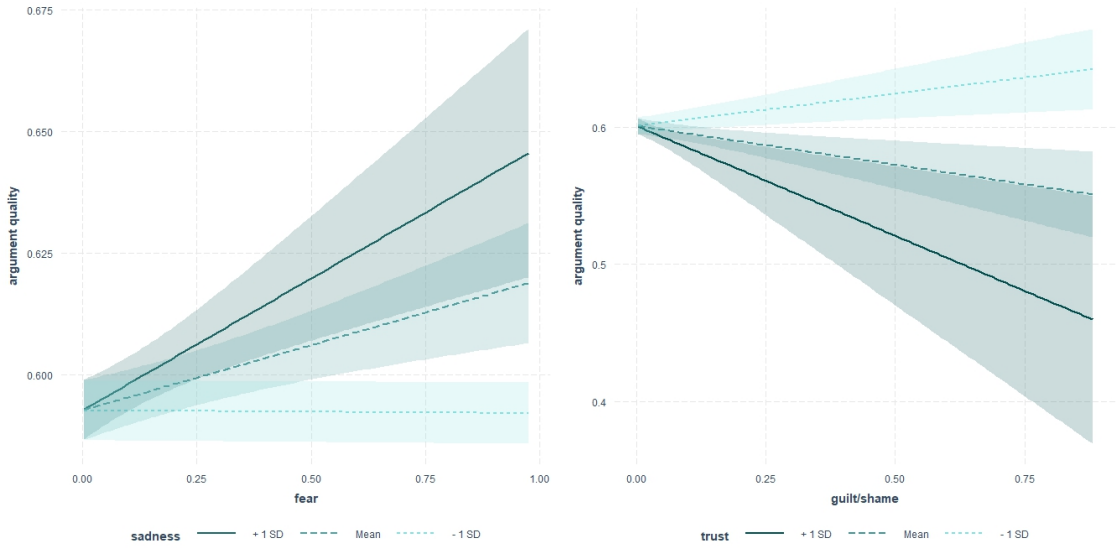
(a) StepAIC regression on IBM ArgQ.

IVs	pseudo-r^2	sign.
disgust : guilt/shame	0.0012	
+ fear : pride	0.0019	***
+ surprise	0.0026	**
+ anger : relief	0.0031	**
+ pride : storytelling	0.0035	**
+ boredom : storytelling	0.0039	**
+ pride : trust	0.0042	*

(b) StepAIC regression on Cornell CMV.

Table 22: Adjusted/pseudo-r^2 for each regression on IBM ArgQ argument quality and Cornell CMV persuasiveness with stepwise added subjective feature IVs sampled from all individual features and two-way interactions. The significance of adding each feature/interaction is tested via ANOVA for IBM ArgQ and via F-test for Cornell CMV.

(table 22a). The only individual features selected in the stepwise process are *trust, joy, fear* and *storytelling*, whose effects are consistent with the size and direction in individual and combined regression as well as highly statistically significant ($p < .001$). Not included individually are the *hedge* feature and most of the emotions (*boredom, disgust, guilt/shame, pride, relief, sadness* and *surprise*). Though neither the previously most informative (by r^2) feature *guilt/shame* nor the most influential feature (by β) *surprise* are included individually, their interaction is the full model's most informative feature and has a highly negative effect ($\beta = -5.97$) on argument quality. The individual regression of the interaction (figure 10a) shows that while the

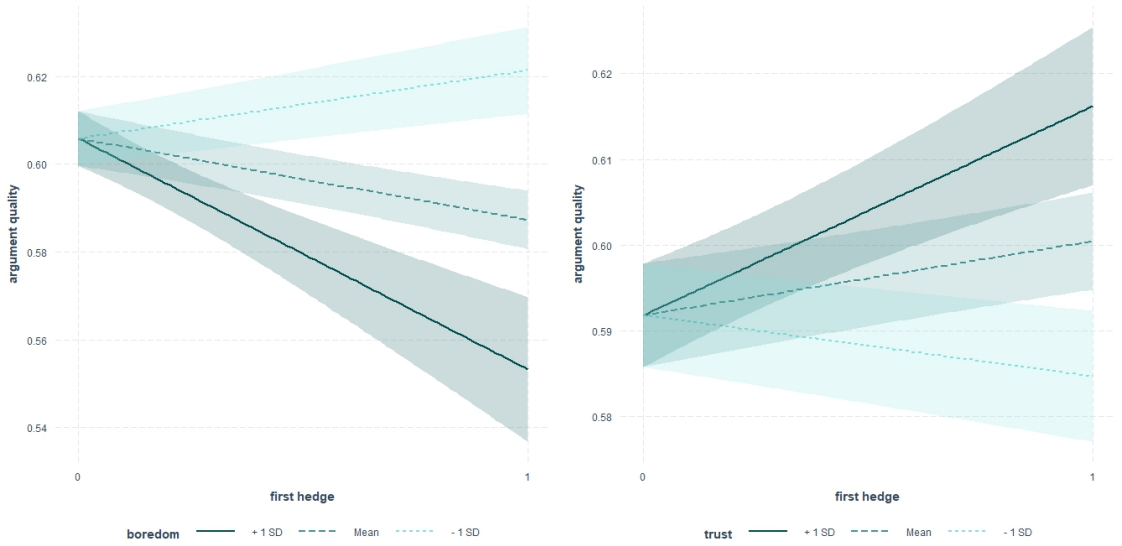


(a) *fear* : *sadness* **(b)** *guilt* : *trust*

Figure 8: Individual regression estimates of three emotion interactions included in the full IBM ArgQ model. The interactions are plotted as the second IV mean (mid tone, dashed line), +1 standard deviance (dark, solid line) and −1 standard deviance (light, dotted line) with confidence intervals.

probability of surprise is low (mean, −1 standard deviance), the negative effect of high *guilt/shame* probabilities is mitigated, while argument quality decreases if both emotions have high probabilities. This effect is however misleading, as the discrete *surprise* prediction resulted in no positive instances, thus the highest probabilities for *surprise* lie below .5, thus changing the interpretation of the interaction to have the highest negative impact when both emotions are at their highest probabilities that may not be high at all.

The next most informative interaction is that of *fear* and *sadness*, which previously both had positive effects on argument quality in IBM ArgQ and interact similar to *guilt/shame* : *surprise*. While the interaction effect is not significant in the presence of other variables in the full model, the interaction effect shown in figure 8a is positive, with *fear* not influencing argument quality much if *sadness* probabilities are low, while the co-occurrence of high *fear* and *sadness* probabilities significantly



(a) *first hedge* : *boredom* **(b)** *first hedge* : *trust*

Figure 9: Individual regression estimates of the three *first hedge* interactions in the full model of IBM ArgQ. The interacted emotion is plotted for the mean ± standard deviance.

increases argument quality. Most other interactions between emotions observe the same principle, i.e., for two variables with the same individual effect direction, the effect of the first IV is either mitigated or disappears completely unless the second variable also has a high probability. An exception is *guilt/shame* : *trust*, even though the two emotions have individually highly significant opposite effects, in instances where *guilt/shame* occurs without *trust*, it has a positive effect on argument quality and as soon as *trust* has a higher probability, the effect is reversed to a negative one. This is notable because on its own, *guilt/shame* has a negative effect, suggesting the assumption that the effect should be negative when *trust* is not involved, while the individually positive effect of *trust* would suggest the negative effect of *guilt/shame* to be mitigated. The *hedge* feature of absolute number of hedge words in the first sentence is interacted with *pride, boredom* and *trust*, with the effect of the latter being positive and the former two negative. When interacted with *trust*, while the emotion probability is low, *hedging* behaves as individually, though as the probability if *trust* rises, the occurrence of a hedge in the beginning of an instance positively

influences argument quality (figure 9b). In the interaction with *boredom*, this effect is reversed: the individually negative effect of *hedging* is reversed in instances with low *boredom* and only occurs when the emotion's probabilities are high (figure 9a). Lastly, *storytelling* is included both individually and interacted with *pride*. This latter interaction is notable as *storytelling* influences IBM ArgQ argument quality negatively on its own as well as interacted with *pride*, though this effect is reversed in the full model, i.e., a high probability of both *pride* and *storytelling* improves significantly argument quality ($\beta = 0.39$, *).

In Cornell CMV, only one variable, *surprise* is included alone in the stepwise full model (cf. table 22b). Furthermore, while all IVs have significant effects on the odds of persuasiveness, the model only includes 6 interactions (cf. **??**), leaving out *hedging* and the emotions *joy* and *sadness* altogether. The most informative feature is the interaction *disgust* : *guilt/shame*, which are the most informative features individually. The interaction effects are mitigated unless both *disgust* and *guilt/shame* probabilities are high. The second most informative feature and that with the largest effect ($\beta = 5.31$, odds ratio $= 202.34$) is the interaction *fear* : *pride* which has a significant positive effect on persuasiveness, though the very high odds ratio and large confidence interval of $[2.28, 8.34]$ disallow inferring much from this result. Next, the individual effect of *surprise* mirrors that in previous Cornell CMV regression experiments, i.e., it leads to a statistically significant decrease in persuasiveness odds. The two remaining emotion interactions are *anger* : *relief* and *pride* : *trust*. The former has a significant positive effect, with anger having no affect alone, i.e., in instances with low *relief* probability, but significantly increasing the odds of persuasiveness in instances where both emotions have high probabilities. The latter interaction then follows the same principle as most interactions in IBM ArgQ, as both *pride* and *trust* have individually positive effects, the interaction mitigates the effect if only one emotion is probable while significantly increasing the odds of persuasiveness in instances with high *pride* and *trust*. Lastly, while *storytelling* is not included individually in the full model, two interactions, with *pride* and *boredom*, are present. These follow the general trend of the interacted emotion, whose effect is mitigated for low *storytelling* and high for instances with high probabilities of the

emotion and *storytelling*. As *boredom* individually decreases persuasiveness odds, the combined effect is negative and conversely, the combined effect for *storytelling* and *pride*, which individually increases persuasiveness odds, is positive.

5 Discussion

5.1 Storytelling

When observing the individual regression results of each feature – storytelling, emotions and hedging – the most notable influence on argument strength in the two corpora seems to be *storytelling*. Not only does it have a highly significant effect on argument quality and persuasiveness, but it is included in all models combining features, and has highly significant and significant effects in both combined IBM ArgQ models and the non-interacted Cornell CMV model respectively. This is in line with findings by Falk and Lapesa (2022) about the influence of storytelling on quality dimensions like *emotional appeal*, *sufficiency* or *appropriateness*. While the standalone results on Cornell CMV corroborate these findings, the effect on IBM ArgQ constitutes a clear divergence. Where arguments in Cornell CMV profit from writers' including personal narratives, in IBM ArgQ they diminish argument quality, which might reflect on the quality definition of the corpus. The arguments were collected from debate audiences and club members, and the quality scores are aggregated from judgements of the arguments' appropriateness for debate speeches. In debate, logical rhetoric and objectivity are held up as attributes of good debaters, and personal anecdotes might consequently be dismissed as unsound evidence.

Beyond these findings on the standalone effect of storytelling however, I did not obtain conclusive evidence in the interaction experiments. While storytelling should help with emotional appeal (cf. Maia and Hauber, 2020) and thus bolster argument strength, the interactions between storytelling and different emotions seem to be mostly influenced by the respective emotion's individual effect and often are not significant. Apart from errors introduced by the automatic annotations, this might have three different reasons. Firstly, storytelling might not be as connected

to emotions as previously thought, which is improbable given previous work on the emotional impact of personal narratives. Secondly, storytelling might be more connected to overall emotionality than any single emotion in particular, which is not captured in the design of this analysis. Lastly, this result might be a case of storytelling enforcing the emotion effect, e.g., the low strength of an argument based on disgust is not increased by the inclusion of a narrative but rather reinforced as the narrative bolsters the emotion and thus its negative effect.

When addressing the question of the validity of this automated approach and its different variants, the storytelling annotation has been proven successful. The classifiers generalized well to the new domain of IBM ArgQ, obtaining even better results than Cornell CMV in the manual evaluation. Furthermore, as IBM ArgQ is further removed from both Cornell CMV and all training corpora in argument style, length and domain, the clear advantage of *mixed-domain* training over *one-domain* training already demonstrated by Falk and Lapesa (2022) was shown to extend to a new, different argument domain. *Mixed-domain* training obtained higher results in both corpora, with no advantage of same training domain in Cornell CMV, but instead a doubled F_1 (.30 vs. .61) for the *mixed-domain* training setup and even higher results on IBM ArgQ.

5.2 Emotion

The other two features are less clearly interpretable in their impact on argument strength. In the absence of other features, a general trend of the correlation between polarity and argument strength emerges, though this trend includes effects that are not statistically significant, while some individual, significant effects counter this notion. *Fear* and *sadness* are positively correlated with the argument strength variable in both corpora and in all experiments where they are included. This might at first seems counterintuitive, but the correct instances from the manual evaluation (cf. tables 23 and 24) point towards a strategic use of the two emotions.

These instances from the corpora are exemplary of the emotional rhetoric of *pathos* strategies, appealing to extroversive emotions by evoking empathy, as in

ID	Text
CMV1	*CMV: Instead of firearms, police should use/carry tranquilizer guns.* A huge problem is that there are a large number of guns in the hands of criminals in the US, and they do not hesitate to use them if necessary. If a police officer is being shot at with a real gun, it is imperative that they are able to retaliate to defend themselves. It would make more sense to add a weapon capable of incapacitating a suspect in a non- lethal manner, but to remove their primary tool of self defence is going too far.
CMV2	*CMV:As a society, we should not use public resources to cure or treat children with chronic illnesses.* There are two major flaws with your point of view. The first is the utter lack of empathy. Having a child with sever medical problems is incredibly taxing. It could happen to anyone through no fault of their own, and as such we have decided as a society to spread that risk around to limit its severity. The second flaw is ignoring the fallout from not assisting in any way. You're taking one of the most powerful motivating forces in a humans life, the well being of their child, and placing them in a situation where they have no *legal* recourse to saving it. That would potentially result in lots of illegal activity as the only means of providing support. Desperate people do desperate things.
CMV3	*I think people who post about their dead relatives on Reddit are barbarians. CMV.* Maybe people don't want to bother their friends with discussing the death of a loved one. Maybe they feel uncomfortable burdening friends with such topics and photographs. Maybe the rest of their friends/family are a wreck and they feel like they have to be strong, so an anonymous internet community is the best place for them to come to terms with the death. Maybe they don't even HAVE anyone to share thoughts and feelings with.
CMV4	*I think the boys in that steubenville rape case got way too much crap throughout the whole trial....CMV.* They were rapists. They deserved to be hung, honestly, or spend life in prison. They should never play professional sports, and should be stuck in a MickyD's the rest of their life. They comitted a horrible atrocity. They ruined the girl's life. She will never go a day without being tramatized for it. The Media protected them, they didn't do anything to make the boys feel an ounce of pain for what they did.
CMV5	Analog clocks give you the ability to visualize time. They are actually *more* intuitive than digital clocks. Will 8:48 - 9:12 give me enough time to pick up the kids? With analog clocks, no mental math is required. It's all visual. Analog clocks are also classier and more visually appealing. Digital clocks require artificially lit displays. Analog clocks can complement other pieces in the room. Digital clocks almost always detract from the interior design.

(a) Example instances from Cornell CMV.

ID	emotion	pers.	story.
CMV1	*disgust*, *fear*, *anger*	1	0
CMV2	*disgust*, *fear*, *anger*, *guilt/shame*	1	0
CMV3	*sadness*	1	0
CMV4	*disgust*, anger, *guilt/shame*	0	0
CMV5	*boredom*	1	0

(b) Example annotations from Cornell CMV

Table 23: Example instances from Cornell CMV with automatically annotated emotion, persuasiveness (pers.) and storytelling (story.) annotations. The first sentence of each instance is the title of the OP.

ID	Text
IBM1	*We should not ban fossil fuels.* A ban on fossil fuels would lead to clandestine organizations taking over this sector.
IBM2	*We should promote autonomous cars.* Autonomous cars will free up time to do other things such as reading, working, doing emails. This time freed up will improve productivity and our ability to learn more.
IBM3	*We should abandon cryptocurrency.* Cryptocurrency mining and transferring is energy inefficient and takes up a huge amount of data, this leads to higher cost of energy overall.
IBM4	*Social media brings more harm than good.* I mean its fine to have lots of opportunity in life ,but if we depend our life on social media , I am afraid one day we are going to lose our loved ones mentally.
IBM5	*Gambling should be banned.* Gambling can be addictive and those who become addicted face severe financial and personal consequences such as bankruptcy, jail (from financial crimes as stealing or embezzlement to support the addiction), divorce and suicide.

(a) Example instances from IBM ArgQ.

ID	emotion	pers.	story.
IBM1	*anger*, *disgust*	.43	0
IBM2	*boredom*	.62	0
IBM3	*anger*, *boredom*	.76	0
IBM4	*fear*, *sadness*	.80	0
IBM5	*fear*, *sadness*	1.00	0

(b) Example annotations from IBM ArgQ.

Table 24: Example instances from IBM ArgQ with automatically annotated emotion, persuasiveness (pers.) and storytelling (story.) annotations. The first sentence of each instance is the stance of the argument.

(CMV3) for people seeking emotional support online, or to introversive emotions by invoking a hypothetical threat like criminals with guns in (CMV1) or one's own child falling ill as in (CMV2). For IBM ArgQ, *fear* : *sadness* is furthermore a highly informative (second stepwise feature) interaction, working in the same way as exemplified by (IBM4, IBM5) both evoking empathy for gambling addicts and alienation through social media, and invoking the threat of such a fate. Through observing this systematic, statistically significant effect during the regression experiments and seeing it mirrored in the argumentation strategies in example data, considerable evidence for *fear* and *sadness* as uniquely powerful emotions in argumentation has been brought forth. Emotional appeal was shown advantageous in persuasive efforts (Benlamine et al., 2017) as well as in minorities' efforts to gain a voice in discussions (Maia and Hauber, 2020). As such, these emotions are interpreted as those conveyed by the text to the reader. Conversely, reading *guilt/shame* and *disgust* not as writer- but as reader-emotions, they become adverse attacks on the opponent (cf. CMV2, "utter lack of empathy") and their significant negative effect in all feature and corpus combinations shows how not to use *pathos* strategies.

Boredom and *surprise* both demonstrate significant and large negative effects on argument strength. This matches with intuitive expectations: a boring argument is seldom a good one and while new information might lead to shifts in opinion, it is not the unexpectedness of the information that contributes to the argument's persuasiveness but rather the information itself. Considering the results of the manual evaluation however, the results might be skewed by erroneous automatic annotations. Although the assumption that classification performance profits from masking salient words from the source domain held true (with the *masked* emotion annotation outperforming the *original* features in the manual evaluation), the quality of annotation differed greatly between the individual emotions. *Boredom* in both corpora appears to be overpredicted and conflated with a neutral tone, as evidenced by the large gap between precision (.1 in IBM ArgQ and .05 in Cornell CMV) and recall (1 in both corpora, tables 8 and 9). While instances might include boring tasks like doing emails (IBM2), and while cryptocurrency (IBM3) or the advantages of analog clocks (CMV5) might constitute boring topics to some, the topic is in-

herent to the discussion and as such not intended to induce boredom by the writer. With an F_1 of 0 and a significant effect on IBM ArgQ, *boredom* thus exemplifies the problem of cross-domain emotion prediction. The effect observed during the analysis for *boredom* might thus not be as reliable. Similarly unreliable scores in the manual evaluation were seen for *surprise*, which is likely due to differences in the use of *surprise* in the different domains of training and argument data. As such, *surprise* is assumed to be especially infrequent in arguments, thus leading to less positive instances to fit the regression on. The significant negative effect of surprise could thus be explained by both the quality and sparsity of the annotation as well as its minor role in argumentation intuited above.

Other emotions, although just as sparse, might still be relevant for argumentation. As an example, *joy* has a positive effect on argument strength in both corpora, but as it is only included in 1-2% of all instances, conclusions drawn from the observed regression effect have to consider the chance of the regression directly learning argument strength scores of certain instances instead of meaningfully depicting the effect of *joy*. As almost all emotions are extremely sparse on the datasets, this constraint on extrapolations only grows for interactions, where two features may only be observed together in a handful of instances. The observation of interactions being more informative than individual emotions might thus show that overall emotionality is a bigger indicator of argument strength but might also show the limitations of regression on such sparse data. To mediate the sparseness, the *discrete* annotations were replaced with *probability* scores, helping those instances where the classifier is uncertain in the negative annotation, but still skewing towards the two extremes of 0 and 1.

A reverse effect can be seen in the high frequency of *disgust* and *anger* in both argument corpora. Looking at the more informal, personal arguments in Cornell CMV like (CMV2, CMV5), the prediction makes sense. The large number of *anger* and *disgust* annotations and the resulting many instances annotated with both however demonstrate something intrinsic to the argument domain. When including other examples like (CMV1, CMV6) or (IBM1), the instances seem less like expressions of anger and disgust but rather a different, mid-level emotion like indignation. The

latter is associated with less excitement than genuine anger, and when considering the targeted nature of arguments, the prevalence of *anger* and *disgust* as the components of the more complex emotion of *indignation* about the topic under discussion is understandable. This however poses a problem for the automatic approach. While Maia and Hauber (2020) investigated fear, anger, indignation and compassion in arguments, this set of emotions is untypical for emotion analysis, which signifies a lack of suitable resources for automatic emotion classification in the argument domain. This problem is exacerbated by the domain-mismatch of annotation-scheme. While the CROWD-ENVENT annotation is necessarily single-label writer-emotions because of the data collection method, to transfer this scheme to the argument domain without consideration for the new domain leads to misclassifications. With natural, long texts as in CORNELL CMV, a single-label assumption cannot be upheld, and as debaters use emotions to evoke empathy and compassion or outrage and fear it is necessary to model emotions that are attributed to the writer as well as emotions they convey and try to instill in the reader.

In general, when adopting automatic emotion classification to a new domain like arguments, the resource should be chosen carefully to match style, length, and use of emotions as closely as possible. An advantageous approach might employ a training dataset already using a multi-label annotation scheme, by which a classifier might learn diverging patterns in how emotions manifest alone and in combination. In the end however, with a task as complex as emotion classification, an annotation that is aligned with the domain characteristics is vital for generalizable analyses and as such, one has to consider a manual annotation process.

5.3 Hedging

For the last feature of hedging, the regression results are mixed. The most informative feature on IBM ARGQ is negatively correlated with argument quality and the most informative feature on CORNELL CMV is positively correlated, which suggests an interpretation analogous to that of storytelling is possible. This is however implausible when considering the idiosyncrasies of the data. The most informative

feature on IBM ArgQ is the absolute number of hedge terms in the first sentence. As subjects however often repeat or rephrase the argument stance, this first sentence only depicts the topic of the argument, most of which are worded to include a hedge term: *We **should** ban fossil fuels.* Only four stances of two topics do not include hedges, thus, the regression likely encountered topic-specific argument quality variance. Conversely, the most informative feature in Cornell CMV is the absolute count of overall hedge terms, which has a positive effect on persuasiveness. The second most informative feature however is the corresponding average, which is negatively correlated with persuasiveness. As these features are not only collinear but calculated from the same base counts, this behavior discredits any interpretation of hedging effects on Cornell CMV data. More likely than an inverse effect of hedges in the two argument domains is the interpretation that in Cornell CMV, the feature also picks up on a different pattern. As the negatively correlated average hedge ratio is balanced in relation to the length and number of sentences, the absolute count most probably depicts the overall length of an instance, and in regression shows that longer instances are more persuasive. With this background, the negative effects of overall average hedge counts in both corpora serve as tentative evidence that verbalizing uncertainty diminishes argument strength. The methodology can however be improved upon, as the automatic annotation for hedges showed the limitations of pure deterministic counts as a feature. It might however be fruitful to explore hedging in arguments if modeled differently, e.g., by weighing it by the scope of each hedge or including a theory-based classification into relational and propositional hedges, and filler words (cf. section 2.2.3), similar to Mamani Sanchez and Vogel (2013).

When considering the results of all regression experiments, argument strength seems to consist of a lot of different features and dimensions instead of being highly correlated to a single feature. This follows from the overall very low coefficient of determination, and the tendency on both IBM ArgQ and Cornell CMV, to prefer feature interactions over individual features, with interactions being selected for 13 out of 17, and 6 out of 7 independent variables in the full stepwise IBM ArgQ and Cornell CMV models. Further, IBM ArgQ with its aggregated quality score shows

clearer patterns, while in Cornell CMV, with persuasiveness indicating the changed opinion of the OP author, a single person that differs between instances, the number of significant effects is lower. As such, argument corpora with aggregated quality scores like IBM ArgQ have their use in systematic investigations into argument quality features, even if the universality assumptions of judgements only hold up to a certain point. This also however shows the importance of addressing subjective bias explicitly when investigating corpora compiled like Cornell CMV, for example by including the original poster ID in all experiments.

6 Conclusion

This thesis investigated the influence of three subjective features on argument strength in two different domains. To this end, two existing argument strength corpora, IBM ArgQ and Cornell CMV, were annotated automatically with the three features of *storytelling*, *emotions* and *hedging*. The aggregated dataset was then used to fit regression models of increasing complexity, first observing each feature independently, then combining and lastly interacting the features to gain insight into their individual and combined effects on argument strength. This was done in parallel for both corpora to compare different notions of argument strength in different domains.

In the regression analysis, I observed an impact of *storytelling*, emotion polarity and the individual emotions of *fear/sadness* and *guilt/shame*. The influence of *storytelling* is systematically reversed between the two argument domains, with the argument quality of the shorter, standalone, debate-focused arguments in IBM ArgQ impacted negatively by narratives, while the persuasiveness score of the longer, dialogic arguments from Cornell CMV profited from personal anecdotes. I hypothesize this effect to demonstrate a difference in the argument strength notions between corpora, with IBM ArgQ valuing objective reasoning and Cornell CMV valuing mutual exchange. Further, both domains exhibited emotion effects along polarity lines, i.e., guilt or disgust decreasing argument strength while joy or relief increased it. The notable exceptions to this rule constitute *fear* and *sadness*, which are systematically positively associated with argument strength in both corpora and all experiments.

This is hypothesized to capture their unique appeal in the *pathos* strategy, as they combine appeals to extroversive emotions of empathy and introversive emotions of safety.

Beyond the question of how argument strength and these features are correlated, a further research question addressed here concerns the methodology of conducting such an analysis without existing data or a dedicated annotation study. To this end, I predicted each feature using different training and annotation setups, which I then evaluated manually. By comparing storytelling classification using mixed-domain training data and one-domain training data, I found results to be mostly on par with previous approaches (cf. Falk and Lapesa, 2022) and showed that these approaches can extend to the – in the context of storytelling research – novel domain of IBM ArgQ without any drops in annotation quality. I further confirmed that similar assumptions about cross-domain classification hold true for adapting emotion prediction to the argument domain: the masking of salient emotion words improves overall cross-domain performance in both precision and recall. This is however variable between the emotion classes with some low annotation performance overall and in the better, *masked* training setup. Similarly, I encountered the limitations of simple rule- and lexicon-based hedge detection, as this feature occurs in the context of unrelated surface features like text length and argument topic.

A limiting factor in the regression analysis was the data sparseness, as almost all emotions are severely underrepresented and hedging too is a minority class. While I mitigated some problems of this sparseness by converting discrete annotation into probabilities, future work should include the experimentation with further domain-adaptation methods like zero-shot learning. I did however also encounter emotion concepts that diverge from traditional emotion analysis standards, such as the complex emotion of indignation being captured by co-occurrences of anger and disgust. This might limit the possibilities of this automatic approach, though the promising and unexpected results I did observe from storytelling and some emotions prove it fruitful to address such limitations systematically, e.g., by devising an annotation scheme for emotions that is tailored to the specifics of argumentation, modeling the differences between writer emotion, reader emotion and targeted emotion use in

strategies of emotional appeal and being open to untraditional emotion categories like indignation or compassion (cf. Maia and Hauber, 2020).

Lastly, while I observed many notable effects, this analysis consists of two corpora and is thus limited in its generalizability. To test extrapolations on notions of argument strength between, e.g., more informal dialogic debate forums and more formal, rationale-focused argument mining corpora, it should be easy to extend the especially interesting storytelling analysis to observe if the differences I found are consistent in other corpora sharing IBM ArgQ's characteristics, e.g., essays or transcripts of actual professional debates.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Shashank Agarwal and Hong Yu. 2010. Detecting hedge cues and their scope in biomedical text with conditional random fields. *Journal of Biomedical Informatics*, 43(6):953–961.

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, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark. Association for Computational Linguistics.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Liliana Ardissono, Guido Boella, and Leonardo Lesmo. 1999. Politeness and speech acts. In *Proc. Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, pages 41–55. Citeseer.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse.* (Kennedy, G.A., translator), Oxford University Press.

Mohamed Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. Emotions in argumentation: an empirical evaluation. In *International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 156–163.

Mohamed Sahbi Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive argumentation and emotions: An empirical evaluation with users. In *Human-Computer Interaction. User Interface Design, Development and Multimodality*, pages 659–671, Cham. Springer International Publishing.

Laura W Black. 2008. Deliberation, storytelling, and dialogic moments. *Communication Theory*, 18(1):93–116.

Laura W. Black. 2013. Framing Democracy and Conflict Through Storytelling in Deliberative Groups. *Journal of Public Deliberation*, 9(1):art. 4.

Laura W. Black and Ron Lubensky. 2013. 6 deliberative design and storytelling in the australian citizens' parliament. In Lyn Carson, John Gastil, Janette Hartz-Karp, and Ron Lubensky, editors, *The Australian Citizens' Parliament and the Future of Deliberative Democracy*, pages 81–94. Penn State University Press, University Park, USA.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

G D Bryant and G R Norman. 1979. The communication of uncertainty. In *Proceedings of the Eighteenth Annual Conference on Research in Medical Education*.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016*, pages 1114–1122. IOS Press.

Sven Buechel and Udo Hahn. 2017a. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017b. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Moitreya Chatterjee, Sunghyun Park, Han Suk Shim, Kenji Sagae, and Louis-Philippe Morency. 2014. Verbal behaviors and persuasiveness in online multimedia content. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 50–58, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Gerald L. Clore and Andrew Ortony. 2013. Psychological construction in the occ model of emotion. *Emotion Review*, 5(4):335–343.

David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert Skip Rizzo, and Louis-Philippe Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*, pages 193–202, Metz, France. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. 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 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

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

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Computational Models of Argument*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 299–310, Netherlands. IOS Press.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

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

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

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computa-*

tional Natural Language Learning – Shared Task, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Michael Fromm, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy Faerman, and Thomas Seidl. 2022. Towards a holistic view on argument quality prediction.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 173–176, USA. Association for Computational Linguistics.

Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118.

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.

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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, (AAAI 2020)*, pages 7805–7813. AAAI Press.

Kathrin Grosse, Maria P Gonzalez, Carlos I Chesnevar, and Ana G Maguitman. 2015. Integrating argumentation and sentiment analysis for mining opinions from twitter. *AI Communications*, 28(3):387–401.

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

Angela Hasselgren, S Granger, J Hung, and S Petch-Tyson. 2002. Learner corpora and language testing. *Computer learner corpora, second language acquisition and foreign language teaching*, pages 143–173.

Ken Hyland. 1998. *Hedging in scientific research articles*. John Benjamins.

Jumayel Islam, Lu Xiao, and Robert E. Mercer. 2020. A lexicon-based approach for detecting hedges in informal text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3109–3113, Marseille, France. European Language Resources Association.

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. SWAT-MP:the SemEval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313, Prague, Czech Republic. Association for Computational Linguistics.

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

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 453–462, New York, NY, USA. Association for Computing Machinery.

Florian Krebs, Bruno Lubascher, Tobias Moers, Pieter Schaap, and Gerasimos Spanakis. 2018. Social emotion mining techniques for facebook posts reaction prediction. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (1): ICAART*, pages 211–220. INSTICC, SciTePress.

Elise Kärkkäinen. 2010. Position and scope of epistemic phrases in planned and unplanned american english. In *New Approaches to Hedging*, pages 203–236, Leiden, Netherlands. Brill.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508.

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.

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

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, USA. Association for Computational Linguistics.

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.

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.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550.

John Lyons. 1977. *Modality*, volume 2, page 787–849. Cambridge University Press.

Rousiley C. M. Maia, Danila Cal, Janine Bargas, and Neylson J. B. Crepalde. 2020. Which types of reason-giving and storytelling are good for deliberation? assessing the discussion dynamics in legislative and citizen forums. *European Political Science Review*, 12(2):113–132.

Rousiley C. M. Maia and Regiane L. O. Garcêz. 2014. Recognition, feelings of injustice and claim justification: a case study of deaf people's storytelling on the internet. *European Political Science Review*, 6(3):359–382.

Rousiley C. M. Maia and Gabriella Hauber. 2020. The emotional dimensions of reason-giving in deliberative forums. *Policy Sciences*, 53:33–59.

Liliana Mamani Sanchez and Carl Vogel. 2013. IMHO: An exploratory study of hedging in web forums. In *Proceedings of the SIGDIAL 2013 Conference*, pages 309–313, Metz, France. Association for Computational Linguistics.

Daniel McFadden. 1973. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–.142. Academic Press New York, New York, NY.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 992–999, Prague, Czech Republic. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

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.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumen-*

tation Mining, pages 29–38, Baltimore, Maryland. 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).

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. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Robert Plutchik. 2001. The Nature of Emotions. *American Scientist*, 89(4):344.

Francesca Polletta and John Lee. 2006. Is telling stories good for democracy? rhetoric in public deliberation after 9/11. *American Sociological Review*, 71(5):699–721.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan. The COLING 2016 Organizing Committee.

E.F. Prince, J. Frader, and C. Bosk. 1982. On hedging in physician discourse. *Linguistics and the Professions, Alex Publishing Corporation*, pages 83–97.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *Proceedings of the 5th International Workshop on Emotions, Social Signals, Sentiment & Linked Open Data (ES3LOD 2014)*, pages 10–13.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Zahra Rahimi, Diane J. Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. 2014. Automatic scoring of an analytical response-to-text assessment. In *Intelligent Tutoring Systems*, pages 601–610, Cham. Springer International Publishing.

r/ChangeMyView. 2023. *Forum guidelines and rules*. `https://www.reddit.com/r/changemyview/wiki/rules/` [Last accessed: 15.10.2024].

Chris Reed. 2006. Preliminary results from an argument corpus. In *Linguistics in the Twenty First Century*, pages 185–196. Cambridge Scholars Press.

A.C. Rencher and G.B. Schaalje. 2008. *Linear Models in Statistics*. Wiley.

Elfira Dwi Rosanti and Alan Jaelani. 2015. The use of lexical hedges in spoken language by female and male students. *English Journal*, 9(1):29–39.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

David M. Ryfe. 2006. Narrative and deliberation in small group forums. *Journal of Applied Communication Research*, 34(1):72–93.

Liliana Mamani Sanchez and Carl Vogel. 2015. A hedging annotation scheme focused on epistemic phrases for informal language. In *Proceedings of the Workshop on Models for Modality Annotation*, London, UK. Association for Computational Linguistics.

Andrea Scarantino and Ronald de Sousa. 2021. Emotion. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Klaus R. Scherer. 2001. Appraisal considered as a process of multi-level sequential checking. 92:92–120.

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Mark D Shermis and Jill C Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Craig A. Smith and Phoebe C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–38.

Wei Song, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Anecdote recognition and recommendation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2592–2602, Osaka, Japan. The COLING 2016 Organizing Committee.

Manfred Stede. 2020. Automatic argumentation mining and the role of stance and sentiment. *Journal of Argumentation in Context*, 9(1):19–41.

Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for 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. *CoRR*, abs/1602.01103.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

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. 2003. *The Uses of Argument*, 2 edition. Cambridge University Press.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1):1–72.

Morgan Ulinski, Seth Benjamin, and Julia Hirschberg. 2018. Using hedge detection to improve committed belief tagging. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 1–5, New Orleans, Louisiana. Association for Computational Linguistics.

Morgan Ulinski and Julia Hirschberg. 2019. Crowdsourced hedge term disambiguation. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 1–5, Florence, Italy. Association for Computational Linguistics.

I. Vasilieva. 2004. Gender-specific use of boosting and hedging adverbs in english computer-related texts – a corpus-based study. In *International Conference on Language, Politeness and Gender*, pages 2–5.

Serena Villata, Elena Cabrio, Imène Jraidi, Sahbi Benlamine, Maher Chaouachi, Claude Frasson, and Fabien Gandon. 2017. Emotions and personality traits in argumentation: An empirical evaluation. *Argument & Computation*, 8(1):61–87.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1–9.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of*

the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Yibei Wang. 2021. A study on the use of hesitation markers in varied-level efl learners' l2 speaking process. *Open Journal of Modern Linguistics*, 11(5).

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.

Richard Xiao and Hongyin Tao. 2007. A corpus-based sociolinguistic study of amplifiers in british english. *Sociolinguistic studies*, 1(2):241–273.

A Information on the aggregation process

A.1 Argument corpora

Topic	Pro	Con
Flu vaccines	Flu vaccination should be mandatory	Flu vaccination should not be mandatory
Gambling	Gambling should be banned	Gambling should not be banned
Online shopping	Online shopping brings more good than harm	Online shopping brings more harm than good
Social media	Social media brings more good than harm	Social media brings more harm than good
Cryptocurrency	We should adopt cryptocurrency	We should abandon cryptocurrency
Vegetarianism	We should adopt vegetarianism	We should abandon vegetarianism
Violent video games	We should allow the sale of vvg to minors	We should ban the sale of vvg to minors
Fossil fuels	We should ban fossil fuels	We should not ban fossil fuels
Doping	We should legalize doping in sport	We should ban doping in sport
Autonomous cars	We should promote autonomous cars	We should limit autonomous cars
Information privacy laws	We should support information privacy laws	We should discourage information privacy laws

Table A.1: The eleven topics given during the argument collection for *IBMRank* with their two opposing stances as reported in Toledo et al. (2019)

A.2 Storytelling annotation

training variant	Precision	Recall	F_1
one-domain	$.84 \pm .04$	$.83 \pm .04$	$.83 \pm .04$
mixed-domain	$.86 \pm .02$	$.86 \pm .02$	$.86 \pm .02$

Table A.2: In-domain evaluation of the storytelling classifiers using *one-domain* or *mixed-domain* training data. The scores are averaged over the results from the held-out data in each of the ten splits and reported with standard deviance between different splits.

A.3 Emotion annotation

	Precision		**Recall**		**F_1**	
Emotion	*mask*	*orig*	*mask*	*orig*	*mask*	*orig*
anger	$.79 \pm .02$	$.84 \pm .02$	$.80 \pm .04$	$.84 \pm .03$	$.79 \pm .02$	$.84 \pm .02$
boredom	$.90 \pm .01$	$.92 \pm .02$	$.90 \pm .02$	$.92 \pm .02$	$.89 \pm .01$	$.92 \pm .02$
disgust	$.84 \pm .03$	$.88 \pm .03$	$.83 \pm .03$	$.87 \pm .02$	$.83 \pm .03$	$.88 \pm .02$
fear	$.83 \pm .02$	$.87 \pm .02$	$.81 \pm .02$	$.87 \pm .03$	$.82 \pm .02$	$.87 \pm .02$
guilt/shame	$.87 \pm .02$	$.91 \pm .02$	$.86 \pm .01$	$.89 \pm .02$	$.86 \pm .01$	$.90 \pm .02$
joy	$.80 \pm .02$	$.84 \pm .02$	$.82 \pm .02$	$.86 \pm .02$	$.80 \pm .01$	$.85 \pm .02$
pride	$.83 \pm .02$	$.88 \pm .02$	$.83 \pm .02$	$.89 \pm .02$	$.83 \pm .02$	$.88 \pm .02$
relief	$.80 \pm .14$	$.88 \pm .02$	$.82 \pm .11$	$.88 \pm .02$	$.81 \pm .13$	$.88 \pm .02$
sadness	$.82 \pm .03$	$.88 \pm .03$	$.82 \pm .02$	$.87 \pm .02$	$.82 \pm .02$	$.87 \pm .02$
surprise	$.78 \pm .03$	$.85 \pm .02$	$.78 \pm .02$	$.85 \pm .02$	$.78 \pm .02$	$.85 \pm .01$
trust	$.87 \pm .03$	$.91 \pm .01$	$.86 \pm .01$	$.89 \pm .02$	$.87 \pm .02$	$.90 \pm .01$

Table A.3: In-domain evaluation of all emotion classifiers using *masked* or *original* training data. The scores are averaged over the results from the held-out data in each of the ten splits and reported with standard deviance between different splits.

	#		%		∅P	
Emotion	*mask*	*orig*	*mask*	*orig*	*mask*	*orig*
anger	5,457	5,464	47.2	47.2	.44	.46
boredom	150	145	1.3	1.3	.06	.06
disgust	2,830	2,390	24.5	20.7	.34	.29
fear	249	206	2.2	1.8	.07	.05
guilt/shame	281	161	2.4	1.4	.12	.07
joy	29	64	0.2	0.5	.03	.03
pride	418	486	3.6	4.2	.14	.11
relief	44	39	0.4	0.3	.05	.06
sadness	144	101	1.2	0.8	.13	.08
surprise	12	11	0.1	0.1	.05	.03
trust	37	23	0.3	0.2	.03	.02

Table A.4: Annotation results of the automated emotion annotation on Cornell CMV whole instances. The columns denote the number (#) and ratio (%) of positive instances and average probability output (∅P) for each emotion. Results are reported for the two annotation variants of training on text with masked emotion words (*mask*) and the unaltered generated text (*orig*).

B Regression results

B.1 Simple regression results of storytelling

B.2 Full models

B.2.1 Independent variables

Annotation	Training	r^2	Coef	P-value
discrete	*mixed-domain*	0.0042	-0.148	0.0 ***
	one-domain	0.0045	-0.153	0.0 ***
probability	*mixed-domain*	0.0047	-0.182	0.0 ***
	one-domain	0.0074	-0.229	0.0 ***

Table B.5: Results of the simple linear regression of storytelling on IBM ArgQ argument quality. The variants are divided by **Annotation** type and **Training** setup. Reported are the adjusted r^2, the coefficient of the feature variant and its significance given a t-test assuming the same outcome with the coefficient set to 0.

Annotation	Training	pseudo-r^2	Odds	P-value
discrete	*mixed-domain*	0.00019	1.084	0.084
	one-domain	0.00001	0.986	0.78
probability	*mixed-domain*	0.00037	1.148	0.015 *
	one-domain	0.00033	1.215	0.02 *

Table B.6: Results of the simple logistic regression with storytelling variants as IV and Cornell CMV persuasiveness label as dependent variable. The variants are divided by **Annotation** type and used **Training** data. Reported are the pseudo-r^2 and **Odds** (exponentiated coefficient) of each feature variant and its significance given a t-test assuming the same regression outcome with the coefficient set to 0.

The individual independent variables available for the full model of each corpus consist of:

- IBM ArgQ
  - Storytelling – *mixed-domain probability*
  - Hedging – *first sentence absolute counts*
  - Boredom – *masked probability*
  - Disgust – *masked probability*
  - Fear – *masked probability*
  - Guilt/shame – *masked probability*
  - Joy – *masked probability*
  - Pride – *masked probability*
  - Relief – *masked probability*
  - Sadness – *masked probability*
  - Surprise – *masked probability*
  - Trust – *masked probability*
- Cornell CMV
  - Storytelling – *mixed-domain probability*
  - Hedging – *overall absolute counts*
  - Anger – *masked probability*
  - Boredom – *masked probability*
  - Disgust – *masked probability*
  - Fear – *masked probability*
  - Guilt/shame – *masked probability*
  - Joy – *masked probability*
  - Pride – *masked probability*
  - Relief – *masked probability*
  - Sadness – *masked probability*
  - Surprise – *masked probability*
  - Trust – *masked probability*

With *anger* removed from IBM ArgQ for mulitcolinearity, there are 12 individual IVs in the IBM ArgQ model and 13 for Cornell CMV. With all possible two-way interactions, this results in 78 features available in the IBM ArgQ step-wise regression and 89 for Cornell CMV.

B.3 Validity of effects

IV	*VIF*
storytelling	2.8937
first hedge	4.4963
anger	4.9310
boredom	1.3933
disgust	5.9959
fear	1.8397
guilt/shame	2.7251
joy	2.9567
pride	2.7455
relief	1.5240
sadness	2.4923
surprise	2.7330
trust	1.4991

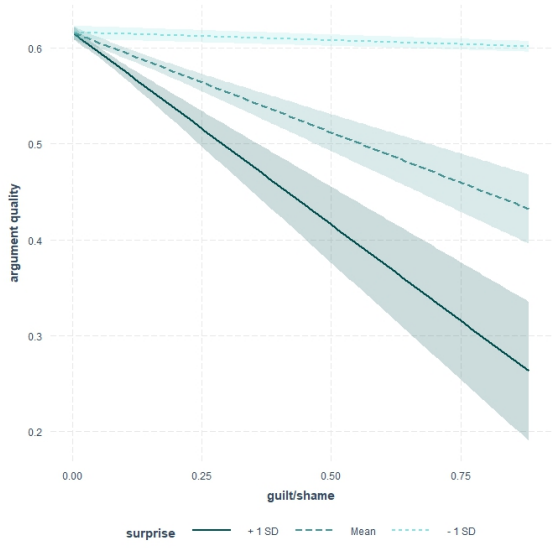
(a) VIF of all features on IBM ArgQ

feature	*VIF*
disgust	3.6382
surprise	1.7695
guilt/shame	2.5316
boredom	1.3788
storytelling	1.4541
fear	1.6016
trust	1.3421
relief	1.4560
pride	2.7428
anger	4.8713
joy	2.3998
final hedge	1.6541
sadness	2.1536

(b) VIF of all features on Cornell CMV

Table B.7: Variance Inflation Factor of IV features used in the regression models on both argument corpora, with feature variants *masked probability* for emotion, *one-domain probability* for storytelling, and the absolute number of hedges in the first sentence for IBM ArgQ and the last sentence for Cornell CMV.

B.4 Interactions on IBM ArgQ



(a) *guilt/shame* : *surprise*

Figure 10: Individual regression plots of all interactions selected in the stepwise regression on IBM ArgQ argument quality.