Investigating Subjective Factors of Argument Strength: Storytelling, Emotions, and Hedging

Carlotta Quensel
Leibniz University Hannover
c.quensel@ai.uni-hannover.de

Neele Falk
University of Stuttgart
neele.falk@ims.uni-stuttgart.de

Gabriella Lapesa
Leibniz Institute for the Social Sciences - GESIS & Heinrich Heine University Düsseldorf
gabriella.lapesa@gesis.de

Abstract

In assessing argument strength, the notions of what makes a good argument are manifold. With the broader trend towards treating subjectivity as an asset and not a problem in NLP, new dimensions of argument quality are studied. Although studies on individual subjective features like personal stories exist, there is a lack of large-scale analyses of the relation between these features and argument strength. To address this gap, we conduct regression analysis to quantify the impact of subjective factors – emotions, storytelling, and hedging – on two standard datasets annotated for objective argument quality and subjective persuasion. As such, our contribution is twofold: at the level of contributed resources, as there are no datasets annotated with all studied dimensions, this work compares and evaluates automated annotation methods for each subjective feature. At the level of novel insights, our regression analysis uncovers different patterns of impact of subjective features on the two facets of argument strength encoded in the datasets. Our results show that storytelling and hedging have contrasting effects on objective and subjective argument quality, while the influence of emotions depends on their rhetoric utilization rather than the domain.

1 Introduction

Argument Mining describes 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). This notion of argumentation as primarily reason-giving, paired with the prominent domains of academic writing, student essays, or professional debate, necessitating objectivity for judging and automatic essay scoring, led to a narrow conceptualization of argument quality. Quality assessment, as emerged from argument mining, observes **objective aspects** such as clarity and argument organization (Persing et al., 2010), use of evidence (Rahimi et al., 2014), or a combination of those (Ong et al., 2014). In the past years, however, a clear need for a shift towards a more subjective notion of argument quality has emerged, driven by the entry of laypeople into the debate space through online forums and citizen participation programs, as well as insights contending the link between objective quality and persuasive strength (Benlamine et al., 2017). This paper contributes to a better empirical understanding of the impact of subjectivity on argument quality.

Sports offer a lot more than you'd think. . . 1) It gives children a sense of being a part of something (crucial for kids without stable families) 2) Sports are a GREAT source of exercise [. . .] There's **many more** reasons but this is all I **can think** of for now. As for my own experiences, baseball and football has helped me come out of my shell and meet some of the best people I've ever met in my life. I **don't know** where I'd be without these sports. (Δ1, *joy, story*, $\varnothing$*hedges*=0.051)

Table 1: Annotated CORNELL CMV instance with positive labels listed at the end and boldened hedge terms.

More specifically, we focus on three subjective features, namely emotions, storytelling (personal and/or anecdotal narratives), and hedging (terms marking uncertainty, e.g., *probably, I think, likely*). While these aspects have already been investigated individually, i.e., in works investigating the use of personal narratives in argumentation (Falk and Lapesa, 2022), emotional progression (Benlamine et al., 2017), or human values (Kiesel et al., 2022), the crucial element of novelty of this work is the fact that we consider the (joint) impact of such subjective features on argument strength as opposed to previous work that considers them in isolation. Table 1 shows an argument appealing to *joyful* emotions and personal experiences, while recognizing knowledge gaps. The argument originates from the online forum *r/ChangeMyView*, where the user was successful (Δ1) in the forum's goal of persuading the discussion's initiator, showing the importance

Proceedings of the 12th Argument Mining Workshop, pages 126–139

of investigating these features and their impact on argument quality more rigorously.

Toward this end, we carry out a parallel analysis on two datasets containing argument quality annotations which approximate the diverging conceptualizations of argument strength related to the function of argumentations: for the reason-giving function we selected IBM ArgQ (Toledo et al., 2019), whose annotations encode **objective argument quality**; for the persuasion function, we selected the Cornell CMV dataset (Tan et al., 2016) aggregated from the previously mentioned *r/ChangeMyView* forum, whose metadata (i.e., the presence of a delta indicating that the originator of a discussion changed their opinion following a specific answer) encode **individualized persuasion**. Differing not only in collection method, domain, argument length, and annotation procedure, these two datasets also lend themselves as the perfect pairing for a contrastive analysis of the impact of subjective features.

Our work proceeds in two steps. As a first step, we automatically enrich the two datasets with one annotation layer per subjective feature. To this end, we compare and evaluate alternative annotation methods (cf. Sec. 4) and reflect on their properties and suitability for our domains of interest. In our second step (Sec. 5), we address the main research goal of the paper: the impact of subjective features on argument strength. We employ regression analysis and address two research questions: **RQ1:** Do subjective features impact argument strength? **RQ2:** Do the patterns of their impact differ in the comparison between objective argument quality and individualized persuasion?

The contributions of our work are accordingly twofold. At the level of novel insights on the phenomenon of argument quality, our work is the first one that targets the *joint* impact of storytelling, emotion, and hedging on argument quality. At the level of contributed resources, we release and share with the community the datasets with the new annotation: this will enable further research on the interplay of these phenomena.[1]

2 Related Works

2.1 Argument Strength

The question of what makes a good argument has been studied since Aristotle (2007), who devised three main strategies of *ethos* or appeal to authority (of experience or persona), *pathos* or appeal to emotions, and *logos* or appeal to logic. The latter strategy maps onto the notion of argumentation as *reason-giving*, which has historically been favored in research. In both computational argumentation and the social sciences, a primary view of argumentation as a rational, somewhat mechanistic process of finding the objectively best claim through a combination of premises and evidence narrowed the notion of argument quality into one of successful *logos* rhetoric. In the predominant domains of student essays and professional debate, this is necessary, but limits the features and dimensions investigated in relation to argument quality to the objective and logical. As such, there are several investigations into clarity, use of evidence, or organization (Persing et al., 2010; Persing and Ng, 2013; Rahimi et al., 2014), with multiple argument quality corpora using corresponding definitions: ease of understanding (Swanson et al., 2015) or the general suitability as part of a larger thesis (Toledo et al., 2019; Gretz et al., 2020). These datasets are usually annotated by merging crowdsourcing labels, which further affirms the notion of argument quality as an, if not explicitly objective, then explicitly universal measure.

The inclusion of user-generated arguments in informal online settings shifted this focus at a similar time as the *affective turn* in the political sciences (Hoggett and Thompson, 2012), reorienting argument strength notions toward the persuasion function of argumentation as well as *ethos* and *pathos* strategies. This shift produced multiple studies of features related to *ethos*, mainly codifying meta-information such as prior beliefs, personal characteristics, and human values (Lukin et al., 2017; Al-Khatib et al., 2017; Kiesel et al., 2022), or, only recently, personal narratives as a form of non-traditional expertise (Falk and Lapesa, 2022, 2023). *Ethos*-related works mainly looked at emotional appeal (Benlamine et al., 2017) or fallacious emotions (Ziegenbein et al., 2023). While multiply new datasets were published in parallel to these studies, targeting *convincingness* and *persuasion* (Habernal and Gurevych, 2016; Simpson and Gurevych, 2018; Gleize et al., 2019), or aiming to codify all existing dimensions of argument quality into a cohesive taxonomy and annotation hierarchy (Wachsmuth et al., 2017; Ng et al., 2020), many of these datasets similarly encode argument quality as a universal average of multiple crowdworkers,

[1]Data and code are available at: https://github.com/CarlottaQuensel/subjective-argument-strength

thus blurring the distinction between objective and subjective dimensions.

Thus, a gap becomes apparent in the understanding of features relating to *ethos* and *pathos*, such as the establishment of personal authority through *storytelling* or *hedging* and the direct investigation of individual *emotions*. Though these three features hold promise for argument assessment, they are largely understudied in Computational Argumentation.

2.2 Subjective Argument Features

Storytelling Research on personal testimonies or *storytelling* originates from the field of deliberative research, where it has long been recognized as a tool to convey empathy and lived experience (Black, 2008, 2013; Esau, 2018). By establishing personal expertise, personal narratives aid in the construction of *ethos*, though Maia et al. (2020) show how narratives enrich debates in public hearings, incorporating *logos* and *pathos* in complex ways. Thus, storytelling serves as an alternative evidency type for non-experts and allows for disagreements without direct conflicts of facts. These observations, however, stem from small case studies and in Computational Argumentation, storytelling only recently gained attention. El Baff et al. (2020) included the number of anecdote sentences in news editorials, but do not address the feature separately. Falk and Lapesa (2022, 2023) consolidate multiple small social science datasets to allow for computational investigations of the phenomenon and argue that integrating personal narratives into argument mining helps include voices often excluded by logos-centric models. Their exploratory findings suggest that storytelling may positively correlate with several quality dimensions in an annotated corpus, but the effects on overall argument quality remain underexplored in a large scale or systematic analysis.

Emotion There are multiple investigations into the impact of emotions on arguments, though investigations of multiple discrete emotions are scant, small, and very recent. Most Computational Argumentation approaches collapse *emotion* and *emotional appeal* into one feature modeled as stance, polarity (e.g., Grosse et al., 2015; Stede, 2020; El Baff et al., 2020), intensity, or the general presence of any emotion (Fromm et al., 2022). Further, *emotional appeal* is historically seen as a fallacy in rational discourse, leading to multiple works investigating emotions as a negative feature (e.g., toxic emotions, Ziegenbein et al., 2023). The argument quality taxonomy and dataset by Wachsmuth et al. (2017) also includes emotional appeal in its 15 labels. 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. Benlamine et al. (2015, 2017) showed the link between emotions and argumentation behavior and found that, from Aristotle's rhetoric strategies, emotional appeal (*pathos*) is most persuasive. Only recently, the first (to our knowledge) small dataset of 1031 German arguments annotated for convincingness and 10 discrete emotions was released by Greschner and Klinger (2024). Despite this encouraging first step, there are, however, neither other (English) datasets nor large-scale analysis of emotions and argument strength available as of yet.

Hedging is one of multiple strategies to verbalize the epistemic modality of a proposition (Lyons, 1977), i.e., convey its degree of certainty (*likely*) or speaker-commitment (*according to* . . .). In academic writing, it reflects the precision and caution of the scientific inquiry process, anticipating objections and gaining community acceptance (Hyland, 1998; Martín, 2003). In the fields of medicine and law, hedging serves as a professional face-saver, to build rapport with colleagues, patients, or a jury, and to avoid misinterpretation, thus enhancing speaker credibility (Bryant and Norman, 1979; Prince et al., 1982; Zaitseva, 2023). Informally, hedging is investigated as a strategy of politeness and positive self-image (Ardissono et al., 1999), and as a cooperative strategy to indicate openness to corrections and change (Vasilieva, 2004; Jordan et al., 2012).

Thus, with the rhetoric strategy of *ethos* encompassing recognized expertise, hedging is directly tied to this strategy. Wielded purposely, it appeals to the honest conduct and credibility of a speaker, similar to storytelling, although apparent uncertainty may just as well hamper recognized expertise. Despite this relevance, hedging is rarely studied in Computational Argumentation: Existing works link hedging to debaters' improvement (Luu et al., 2019), predict persuasiveness with paraverbal hesitation cues (Chatterjee et al., 2014) or modal verbs (Wei et al., 2016), but few address the size and direction of any observed effects. Habernal and

Gurevych (2017) show an uneven distribution of hedges skewed toward constructive, nonpolarized discussions. Only Tan et al. (2016) directly observe a positive effect on persuasiveness. The mixed findings highlight a gap: Given its surface-level detectability and interpretive flexibility, hedging is a promising but overlooked feature for capturing subjective argument quality. Hedging might enhance argument strength by boosting credibility, or weaken it by implying doubt – yet no systematic study explores this trade-off.

3 Data

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 must be aggregated automatically. Multiple corpora are suitable as a base dataset that includes a gold annotation for the target variable (DV) of argument strength. To approximate the diverging conceptualizations of argument strength explicated above, we chose two datasets that differ in collection method, domain, argument length, and annotation procedure, categorized below as objective argument quality and individualized persuasion.

Objective argument quality IBM-ARGQ 5.3k (Toledo et al., 2019) consists of 5.3k short, standalone arguments generated at formal debate events by debate club members of varying skill levels and the general audience. Participants were asked to produce short arguments (max. 36 words) after seeing a professional example argument and choosing one of 11 controversial topics, such as privacy laws, gambling, or vegetarianism with two opposing stances, e.g., *We should adopt vegetarianism* and *We should abandon vegetarianism*. Participants were advised to keep arguments impersonal to avoid privacy concerns in the final dataset.
The argument strength annotation is an average of binary crowd judgments: for each argument, 15-17 annotators judged its adequacy as part of a debate speech,[2] which was averaged for the final score to model the ratio of positive judgments. This procedure attests to a rather unspecific conceptualization of generalized 'overall' argument strength, as the annotators must employ their own concept and hierarchy of relevant features, e.g., topic relevance, linguistic clarity, or sound rhetoric, and the single binary judgment paired with the averaging makes reconstruction of these features impossible. As such, while the utilized notion of argument strength is not explicitly stated 'objective', the domain, style, and annotation process of IBM-ARGQ 5.3k invoke an argument strength conceptualization in line with the traditional *logos* focus of the argument mining field, by removing subjective context and aggregating judgements to approximate a generalized, universal, and thus more objective, argument quality score. Thus, in the following analysis, this dataset is referred to as IBM ARGQ and represents argument strength as conceptualized by the traditional argument mining field.

Individualized persuasion CORNELL CMV was aggregated by Tan et al. (2016) from 11567 comments posted to the Reddit forum *ChangeMyView*[3] between January 2013 and August 2015, where users state their viewpoint with detailed background on their thought process to engage in constructive discussion that aims at changing their view. Thus, in one comment thread, multiple users argue against the same position until the original poster (OP) awards a *delta point* (Δ) to one or more answers that persuade them. The unique setup of the forum provides an inherent annotation and ensures data quality, with the delta point system that denotes the OP's persuasion and posting guidelines that are actively moderated by volunteers both for civility and for maintaining a constructive discussion in which comments must advance the conversation and decisions for delta points must be explained. The resulting label stands in contrast to the score of IBM ARGQ, as it encodes the subjective change in opinion of one person from a specific argument, in the context of a mutual discussion and multiple alternative arguments. The domain properties further 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. In the dataset used here (henceforth CORNELL CMV), the posts are structured as contrasting pairs of comments addressing the same OP, one with and one without a delta point, making for a balanced distribution of the binary persuasiveness label.

Given all the above differences between IBM

[2] *'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?'*

[3] https://www.reddit.com/r/changemyview/

ArgQ and Cornell CMV, it is apparent that the two datasets conceptualize arguments as well as argument strength in very different ways. Although the number of differences disallows a comparison of pure argument strength conceptualization without any confounding factors, the inclusion of both corpora in the investigation covers idiosyncrasies across the spectrum of the argument mining field on what argument strength means. Tab. 7 shows examples from both datasets. To illustrate the diverging concepts, in the following analysis, argument strength is called *quality* when investigating IBM ArgQ and *persuasiveness* for Cornell CMV.

4 Automatic Annotation of Subjective Features

As the two datasets do not have annotations for the investigated features, it is necessary to enrich the datasets with the corresponding annotation layers as a first step. Thus, an automated annotation model is devised for each of the three features. In what follows, we describe the computational methods we used to achieve this goal separately for each feature. For storytelling and emotions, an ensemble consisting of ten transformer-based classifiers is trained on annotated data. As hedging is a surface feature dependent on individual terms, it is annotated using a simple rule-based algorithm. The following sections 4.1, 4.2, and 4.3 elaborate on the annotation process of each feature and the resulting statistics on the two argument datasets.

4.1 Storytelling

Training Data As most storytelling research is comprised of small case studies from the political sciences, we combine multiple datasets from different sources following the approach of Falk and Lapesa (2022). Falk and Lapesa (2022) use a collection of different datasets and domains covering diverse topics, such as expert-moderated discussions on immigration (Gerber et al., 2018) and consumer debt collection (Park and Cardie, 2018) and a subset of the online debate forum *r/ChangeMyView*. They consolidate different original annotations indicating wether an argument contains a personal experience or story (1) or not.

Training Setup We fine-tune RoBERTa transformers (Liu et al., 2019) using a 10-fold cross-validation ensemble, where the full dataset is split into ten parts and ten separate models are trained, each on a different combination of training and validation folds. This ensemble approach is used to produce more robust and stable predictions, as it mitigates variance due to random initialization and training data fluctuations (cf. e.g., Lakshminarayanan et al., 2017; Mohammed and Kora, 2023). For annotation, we apply the majority vote across the ten ensemble models to assign labels to our two target datasets. This setup follows Falk and Lapesa (2022), both to replicate the results of the original paper and to harness the identification of mixed-domain training as the most robust configuration for cross-domain generalization, making it most suitable for our IBM ArgQ data. As their reported same-domain performance for the *ChangeMyView* subset is on par with the mixed-domain classifier, we additionally train a classifier on only this subset to potentially harness this effect for Cornell CMV.

Results As apparent from the test performance on a heldout dataset (cf. Appendix Tab. 5), the mixed-domain ensemble prevails over the same-domain classifier, both in terms of performance (F_1 =.82 vs. F_1 =.78) and lower variance, which is in line with findings by Falk and Lapesa (2022). Otherwise, the performance is on par with the results of the best-performing models of the original experiments (Falk and Lapesa, 2022) (F_1 between .76 and .92), allowing us to continue with the analysis using the *mixed-domain* annotations. The resulting predictions are, however, very sparse for both corpora (cf. Tab. 2), especially so IBM ArgQ (0.8% positive), which can be attributed in part to the unbalanced distribution in the training data (storytelling is the minority class), but more importantly to the brevity and impersonality of IBM ArgQ instances. To mitigate the sparseness, we follow Lakshminarayanan et al. (2017) and interpret the average classification probability as a certainty measure of the binary annotation, thus introducing a richer source of information in the next step.

4.2 Emotion

Training Data As expanded in section 2, while there are multiple works on *emotionality* (intensity, polarity, etc.) in arguments, there are no works and related datasets modeling discrete emotions in English arguments. As such, our approach has to bridge a gap from the emotion domain to the argument domain. Though recent works showed the capabilities of LLMs in emotion classification (cf. last year's WASSA shared task; Maladry et al.,

Feature	IBM ArgQ			Cornell CMV		
	#	%	∅P	#	%	∅P
anger	1,814	34.2	.39	6,467	55.9	.43
boredom	116	2.2	.06	538	4.7	.07
disgust	2,920	55.1	.54	5,111	44.2	.37
fear	347	6.6	.14	822	7.1	.11
guilt/shame	107	2.0	.12	631	5.5	.14
joy	47	0.9	.07	208	1.8	.05
pride	80	1.5	.10	615	5.3	.12
relief	64	1.2	.06	256	2.2	.06
sadness	175	3.3	.14	429	3.7	.12
surprise	0	0.0	.03	53	0.5	.04
trust	112	2.1	.07	159	1.4	.04
storytelling	45	0.8	.02	2288	19.8	.22

Table 2: Feature distribution according to the best ensembles for emotion (*masked/aggregated*) and storytelling (*mixed*) on IBM ArgQ and Cornell CMV, including the number (#) and ratio (%) of positive instances, and the corpus-wide average classification probability (∅P).

2024), the zero-shot approach necessitated by our lack of in-domain examples is still outperformed by traditional fine-tuning, given a sufficient amount of high-quality training data (Kazakov et al., 2024). With no emotion-annotated datasets in the argument domain, we selected our training data to best match the register and style of our target data. This precludes both very informal and formal datasets aggregated from Twitter or from novels and news headlines, as well as data collected through emotion-specific emojis, words, hashtags, or forums to avoid surface-level emotion representations with low cross-domain adaptability. Thus, we chose Crowd-enVENT (Troiano et al., 2019) as our training data, a crowdsourced dataset of event descriptions for eleven different emotions,[4] which allows for an implicit emotion representation.

Training Setup In line with the setup for the *storytelling* feature, we employ an ensemble consisting of RoBERTa classifiers (Liu et al., 2019) fine-tuned on a 10-fold data split and aggregate the predictions into a majority vote. The dataset is originally single-label, with 550 event descriptions generated separately for one emotion. For our target data, we cannot assume a single-label distribution. Thus, we trained a separate classifier for each emotion and downsampled 1650 instances from all other emotion instances for a balanced training set with diverse negative instances.[5] Similar to the *storytelling* annotation, we compare two strategies for cross-domain robustness: the event descriptions are available in their original form as well as with salient emotion terms masked. We trained models on both versions to compare the impact of harnessing lexical surface features (*original*) with that of learning more implicit emotion representations (*masked*) and thus gaining more robust performance. As the arguments in Cornell CMV are longer than both the texts in the training data and the model's cutoff token length, we additionally split these instances in half and then aggregate the annotations for both halves.

Results As the test performance from the training process shows, using *masked* training data improves classification performance significantly (avg. F_1 increase: 0.074) and exceeds the benchmark performance reported by Troiano et al. (2023). The resulting label distribution of the best ensemble is reported in Tab. 2. Apart from *anger* and *disgust*, which occur in almost half of all instances, the data – especially IBM ArgQ– emotions are very sparse, with a ratio of positive instances below 10% for all other emotions and *surprise* missing entirely from IBM ArgQ. Thus, we can observe a higher use of emotions in the more subjective Cornell CMV data, together with a general skew towards 'indignation-adjacent' emotions like *anger* and *disgust*. While argument-specific emotion use is further analyzed later on (see Sec. 5), at this point, we observe that very low performance might be related to disuse in argumentation: arguments might intuitively stem from anger or appeal to pride, though arguing from a point of boredom or surprise (our two worst results) might be unusual.

Thus, we continue with annotations from the *masked* and *masked-aggregated* classifiers for our analysis, discarding *surprise* due to its absence in IBM ArgQ and replacing the binary annotation by averaged classification probabilities in further experiments. We thereby combat data sparseness and leverage prediction confidence (to have indications of 'weaker' or 'stronger' signs of emotion), making sure that the statistical model can account for robustness.

4.3 Hedging

As a surface-level feature, hedges can be extracted through a simple lexicon matching approach. We adapt and combine multiple lexicons from approaches outside the argument domain (Islam et al., 2020; Sanchez and Vogel, 2015; Ulinski and

[4] Generated as, e.g., *I felt fear when:* ... and analogously.
[5] The full dataset would result in 8% positive instances.

	IBM ArgQ				Cornell CMV			
IV	r^2	p		Coef	pseudo-r^2	p		Odds
storytelling	0.0047	0.0	***	−0.182	0.0004	0.015	*	1.148
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	0.0	***	−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
trust	0.0067	0.0	***	0.140	0.0000	0.654		0.886
# hedges	0.0027	0.0	***	−0.011	**0.0106**	0.0	***	1.030

Table 3: Individual regression results including the explained variance (adjusted r^2), respectively, pseudo-r^2 for logistic regression, the p-value and significance of the effect (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$) and the coefficient, respectively, the logistic odds.

Hirschberg, 2019), which are targeted toward similar semi-formal domains (i.e., internet forums) and thus include domain-specific abbreviations and colloquialisms like *AFAIK* (*As far as I know*). Our pipeline first tokenizes and parses the arguments, then matches tokens to a hedging lexicon and further disambiguates terms with simple syntax rules, an example of which can be found in table 6. We were thus able to obtain the number of hedges per argument and create different feature variants, i.e., the overall number of hedges in the first and last sentence, versus in the whole argument instance, as well as the hedge-token ratio for each absolute variant. By including multiple, relative variants of the feature, we are able to abstract from the difference in instance length between the two corpora and accurately portray differences in the usage of hedges. Overall, our automated annotation approach proves successful, with increased robustness stemming from our generalization strategies: we find that mixed domain training, masking superficial lexical cues, and employing a deep ensemble is helpful. Although the performance on the argument data is expectably lower than in the training domain, it is nonetheless sufficient for our subsequent analysis and must be seen in relation to the very sparse label distribution in the argument domain.

5 Regression Analysis

Following the successful automated annotation procedure, we implement a regression analysis to investigate the impact of all 16 features (1 *storytelling*, 9 emotions excluding *boredom, surprise*, 6 *hedging*) as independent variables on the dependent variables of *quality* score in IBM ArgQ and *persuasion* label in Cornell CMV. We use the Python *statsmodels* library (Seabold and Perktold, 2010) to implement OLS linear regression with t-testing for significance on the *quality* score of IBM ArgQ and logistic regression with z-testing for significance on the binary *persuasiveness* label of Cornell CMV. To measure how much variance can be explained by individual features and how much additional variance can be explained by combining features, we compare regression models that employ a single feature as IV to richer models with multiple IVs and two-way interactions.

In comparing individual regression results of all features (see Tab. 3), two major divergences between the two corpora emerge. Firstly, both *storytelling* ($\beta = -.182$) and the absolute *hedging* count ($\beta = -.011$, for hedging in all variants, see Appendix Tab. 8) are highly significant negative predictors of argument quality in IBM ArgQ, but significantly improve persuasiveness in Cornell CMV ($\beta_{story} = 0.138$, $\beta_{hedge}) = 0.030$, cf. Fig. 1), with *hedging* constituting the most informative feature for this dataset. Secondly, an overall trend of greater and more frequent significant effects can be observed for IBM ArgQ argument quality than for Cornell CMV persuasion. This trend comes along with a greater predictive power of the IBM ArgQ

There is a difference between a fear of being killed by a terrorist (very small likelihood) and the fear of being *terrorized*. I was in Boston when the marathon bombings happened. Terrorism affected everyone on the streets, even though only 3 people were killed. The scope of an act of terrorism is much greater than the strict number of casualties. It has a psychological and traumatizing effect on people even in its periphery. That being said I am much more afraid of police than an act of terrorism. This is because after the bombings, when Tsarnaev was hiding in a boat about a quarter mile from my apartment at the time, militarized police with bomb dogs searched my house without announcing themselves, came to my door with assault rifles, and kept me locked in my house for a whole day while bomb vans and squad cars raced up and down my street. It was one of the most terrifying days of my life. **I felt** more electric fear answering the door to what looked like a 9-man SWAT team in full tactical gear and AK-47s than I did in the several previous days of news coverage following the bombing. I don't **necessarily** agree that having other things to be afraid of, like the abuse of power by the police, makes being afraid of things like acts of terrorism (which are designed to frighten) unreasonable. Fear is real and you don't always have a choice in the matter when it comes to whether or not it will infiltrate your life. (Δ0, *fear, storytelling*, $\varnothing$*hedges*=0.007)
Don't mean to be harsh, but that thinking is very dumb. There's a fine line between eating other animals, and cannibalism. Cannibalism is morally wrong because you are practical eating yourself. (Δ0, *guilt/shame, disgust*, $\varnothing$*hedges*=0.0)
Social media brings more good than harm. Social media helps reconnect with past friends. I was able to reconnect with a childhood best friend not seen in years shortly before he died. For that I am grateful. (*score*=0.6, *joy, sadness, storytelling*, $\varnothing$*hedge*=0.0)
Social media brings more harm than good. facts are not checked on social media platforms, allowing public shaming of different figures, hurting them and their career immensely even without them doing anything wrong (*score*=0.47, *disgust, anger*, $\varnothing$*hedge*=0.0)
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. (*score*=1.0, *fear, sadness*, $\varnothing$*hedge*=0.11)
Flu vaccination should not be mandatory. While **I believe** that flu vaccines are beneficial to people, I do not **believe** they **should** be mandatory because I **should** have a right to decide if I want to take a risk with my health.(*score*=0.8, $\varnothing$*hedge*=0.12)

Table 4: Fully annotated examples from CORNELL CMV and IBM ARGQ, with all positive labels listed below the post text and hedge terms rendered bold.

models,[6] and is continued in the best multiple regression model, which includes more IVs for IBM ARGQ than for CORNELL CMV.

In contrast to these domain differences, the impact of emotions on argument strength is largely domain-independent, with direction and magnitude of effects comparable between IBM ARGQ and CORNELL CMV for all emotions but *trust*. As such, the emotions with the highest impact on argument strength are *guilt/shame* and *disgust*, which both significantly decrease argument strength. For these emotions, as for most others, emotion polarity matches effect direction, including the significant emotions of *relief* (both corpora), *pride* (CORNELL CMV), and *joy* (IBM ARGQ). Two emotions contradict this trend: opposite to their polarity, *fear* ($***$ IBM ARGQ; $*$ CORNELL CMV) and *sadness* ($*$ IBM ARGQ) improve argument strength in both corpora.

To further investigate the interplay between different argument features, we implemented two multiple regression analyses with and without interaction. We used stepwise multiple regression, where individual IVs or two-way feature interactions are added incrementally according to their AIC value (predictive improvement relative to model size), while ensuring the significance of added IVs compared to the smaller model through ANOVA (IBM ARGQ) and F-test (CORNELL CMV).

The full models reveal the consistency of most effects on argument strength, as the most informative features of *guilt/shame* retain their salience, and notable observations like the diverging effect of *storytelling* on persuasion vs. quality are present in the full model as well. Interactions show a general trend of same-directed features combining to an effect of greater magnitude, as seen with the individually positive features of *fear* and *sadness* interacting on IBM ARGQ argument quality to form a highly positive combined effect while their individual effects are neutralized (Fig. 2). The full models with interaction further show the persistant importance of *storytelling*, which (in contrast to the individual IBM ARGQ regression) has a positive effect in both datasets. The final explained variance is 3.96% adjusted r^2 for IBM ARGQ and 1.36% pseudo-r^2 for CORNELL CMV. Although generally low, these values are reasonable and expected for a regression on the complex notion of argument strength, considering the exclusion of contextual information (e.g., topic, demographics of the annotators/OPs) and overall low values (and thus error margins) for both independent and target variables (between 0 and 1).

[6]While the adjusted r^2 of the IBM ARGQ models can be interpreted as the percentage of explained variance, this cannot be compared directly to the pseudo-r^2 of the logistic CORNELL CMV models. The general difference in magnitude nonetheless holds.

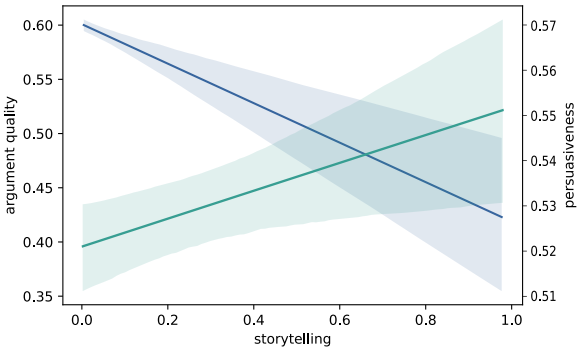

Figure 1: *Storytelling* effect on IBM ArgQ argument quality (teal, left y-axis) and on Cornell CMV persuasion (blue, right y-axis), with confidence intervals.

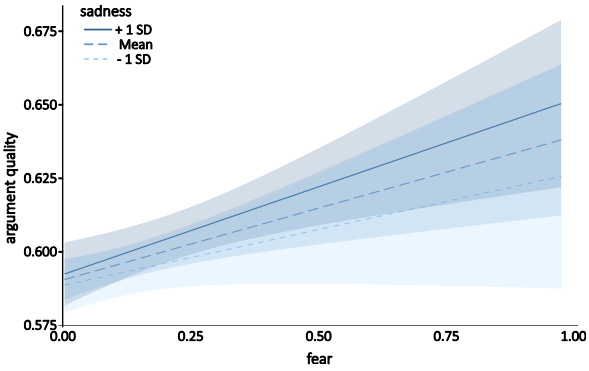

Figure 2: Interaction between *fear* (x-axis) and *sadness* (standard deviations shown through hue and dashing) on IBM ArgQ with confidence intervals.

Discussion The diverging effects of *hedging* and *storytelling* show the importance of domain-aware rhetoric: harnessing such subjective features significantly improves the odds of subjective persuasion, but in the objective domain of IBM ArgQ, they hinder argumentative success (cf. Fig. 1). As all subjective features are infrequent in IBM ArgQ, where arguments were mandated as short and impersonal, their successful use in Cornell CMV seems intuitive, indicating their importance for non-experts.

When viewing the results of our two steps side by side, it is apparent that emotions are utilized differently in argumentation than in their original context. While *disgust* and *anger* are overrepresented compared to all other features, a qualitative analysis (see Tab. 4) shows their idiosyncratic appearance in arguments. Both emotions seem closer to indignation or 'righteous' anger, and occur, with the similarly impactful *guilt/shame*, almost always explicitly targeted towards either another participant ('*that thinking is very dumb*') or the topic under discussion ('*allowing public shaming*'). The very beneficial emotions of *fear* and *sadness*, on the other hand, seem reframed as an appeal to universal concerns instead of individual experiences, even when combined with personal experiences: '*personal consequences*', '*whether or not it will infiltrate your life*'. Therefore, we hypothesize that discrete emotions are utilized in two diverging strategies of *emotional attacks* and *emotional appeals*. While the latter are highly efficient in persuasion, the former hinder argument strength but are much more frequent in the data.

6 Conclusion

This paper has investigated the impact of a number of subjective features on two diverging facets of argument strength. To that end, we first determined the feasibility of large-scale automated annotation of our subjective features, to then systematically reveal correlations through a regression analysis. We could reveal a significant effect of almost all observed features on argument strength, thus affirming **RQ1**. We moreover demonstrated the importance of argument context for subjective features, as personal anecdotes and uncertainty indicate a lack of rhetoric proficiency in objective settings, but strengthen arguments in the subjective domain, thereby affirming **RQ2**. Further qualitative assessment shows frequent *emotional attacks* with righteous indignation impeding argument strength, while less frequent *emotional appeals* to empathy and universal fears seem to strengthen arguments. This finding reveals an avenue for continuing argument-specific emotion research, a research gap that is further emphasized by the results of our automated modeling. We could successfully model *storytelling* and most emotions automatically due to our robustness strategies of employing a deep ensemble based on training data from mixed-domains and with masked surface lexical cues. Thus, in situations where large-scale gold data is neither available nor easily attainable, these strategies constitute an acceptable alternative. However, the unbalanced and idiosyncratic distribution of emotions also reveals the limits of cross-domain approaches, as some emotions are used extremely seldomly, or appear changed from their original definition. We thus highlight once more the need for emotion data and definitions directed at argumentation, a research gap that has recently been addressed for German text by Greschner and Klinger (2024) and should receive further attention on a larger scale.

Acknowledgments

We would like to thank the anonymous reviewers whose feedback helped us improve this paper. This research has been partially funded by the Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation)

Limitations

Apart from the obvious constraint of English-only modeling, automatically annotating the independent variables bears the risk of modeling the influence of features that differ from the named features. For the features of storytelling and hedging, our success in recreating results from existing works leads us to believe that the annotations are acceptable even on unseen data. For our emotion features, we rely on our strategies of masking salient surface features and aggregating predictions for long instances to lead to an acceptable performance based on the good results on the heldout training data. Thus, we believe our regression to realistically model the influence of the remaining investigated features. This influence is very small, as denoted by the low r^2 and pseudo-r^2 scores of the regression models. However, while this shows that the features investigated here cannot fully explain argument strength, the high significance of most features nonetheless shows their importance for argument strength. As previous research shows, argument strength is a complex and subjective feature. We thus expect that a model regressing argument strength to a higher degree must include context, such as prior beliefs and demographic features of the annotators/OP and the author, topic information, or discussion history. The significance of our results constitutes one step in a growing field of research aiming to explore argument strength as a multi-faceted complex feature.

Ethical Considerations

As always in the analysis of argument strength, our results may potentially be exploited in the persuasion strategies of bad actors. However, we observed significant but very small effects that may be less impactful than demographic and contextual features, which we omitted. Further, features like emotions or uncertainty are likely used intuitively and, as shown elsewhere (cf. e.g., Vasilieva, 2004), used differently depending on demographic factors. While reporting negative influence might discredit argument strategies used by already disadvantaged groups, we believe that our features bear no inherent demographic inclination and understanding such effects is the first step to encourage thoughtful argumentation.

References

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.

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.

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

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.

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

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.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. 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. *arXiv preprint*.

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.

Lynn Greschner and Roman Klinger. 2024. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and llms. *Preprint*, arXiv:2412.15993.

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.

Paul Hoggett and Simon Thompson. 2012. *Politics and the Emotions: The Affective Turn in Contemporary Political Studies*. Bloomsbury Publishing.

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.

Michelle E. Jordan, Diane L. Schallert, Yangjoo Park, SoonAh Lee, Yueh hui Vanessa Chiang, An-Chih Janne Cheng, Kwangok Song, Hsiang-Ning Rebecca Chu, Taehee Kim, and Haekyung Lee. 2012. Expressing uncertainty in computer-mediated discourse: Language as a marker of intellectual work. *Discourse Processes*, 49(8):660–692.

Roman Kazakov, Kseniia Petukhova, and Ekaterina Kochmar. 2024. PetKaz at SemEval-2024 task 3: Advancing emotion classification with an LLM for emotion-cause pair extraction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1127–1134, Mexico City, Mexico. 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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

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

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. *arXiv preprint*.

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

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the WASSA 2024 EXALT shared task on explainability for cross-lingual emotion in tweets. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 454–463, Bangkok, Thailand. Association for Computational Linguistics.

Pedro Martín. 2003. The pragmatic rhetorical strategy of hedging in academic writing. *Vigo International Journal of Applied Linguistics (VIAL)*, 0.

Ammar Mohammed and Rania Kora. 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2):757–774.

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. 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.

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

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.

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.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

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

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

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.

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.

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.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. 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.

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.

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.

Margaryta Zaitseva. 2023. Some observations on altering hedging phenomenon in courtroom discourse. *LINGUISTICS*, 1(47):152–162.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Supplementary Material

A.1 Data and Annotation

Table 7 shows exemplary instances of the base corpora we used, while tables 5 and 6 explicate the further annotation process.

A.2 Regression Results

Reported below are the regression results for all hedging variants (Tab. 8), the results of full stepwise regression model with interaction (Tab. 9), and two exemplary regression plots (Fig. 1, 2).

	Training variant		**Benchmark**
Feature	*masked*	*orig*	
anger	0.76(±0.03)	0.69(±0.04)	0.53
boredom	0.88(±0.03)	0.84(±0.02)	0.84
disgust	0.82(±0.03)	0.75(±0.04)	0.66
fear	0.81(±0.03)	0.72(±0.03)	0.65
guilt/shame	0.85(±0.03)	0.80(±0.02)	0.48/0.51
joy	0.77(±0.03)	0.71(±0.02)	0.45
pride	0.83(±0.03)	0.75(±0.03)	0.54
relief	0.82(±0.03)	0.70(±0.25)	0.63
sadness	0.81(±0.04)	0.73(±0.03)	0.59
surprise	0.78(±0.02)	0.67(±0.04)	0.53
trust	0.85(±0.02)	0.80(±0.02)	0.74
	mixed	*one*	
storytelling	0.82(±0.03)	0.78(±0.05)	0.76-0.94

Table 5: F_1 performance of the ensemble classifiers on the heldout test set of their respective training data with standard deviance reported in brackets. The last column lists the originally reported benchmark: Troiano et al.'s (2023 text-based classifier (multilabel versus our single label classifiers) and the best overall approach by Falk and Lapesa (2022, performance is reported separately for three subsets, thus ranging between values).

Term	**Rule**
about, *around*	If the token is an adjective, it is a non-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 6: Examplary hedge disambiguation rules, the first of which is lifted from Islam et al. (2020).

IBM ARGQ	
We should ban fossil fuels. fossil fuels are bad for the country because of your country dont have them they have to be in an inferior position to ather countrys. (score=0.18)	*We should ban fossil fuels.* Fossil fuels destabilize the ecosystem which will harm future generations. (score=1.0)
CORNELL CMV	
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). [...]	
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. (Δ0)	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.) (Δ1)

Table 7: Examples from IBM ARGQ and CORNELL CMV of a bad (left) and good (right) argument about the same topic, with the shortened original post from CORNELL CMV given above the two answering arguments.

score	**sent**	**r²**	**Coef**	***p***	
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	***

(a) IBM ARGQ

score	**sent**	**pseudo-r²**	**Odds**	***p***	
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	*

(b) CORNELL CMV

Table 8: Individual regression results of each hedging variant as IV on IBM ARGQ argument quality and CORNELL CMV persuasiveness. The variants are listed by **score** (absolute or ratio values) and the **sent**ence for which the score is calculated. Reported are the adjusted r^2 percentage, respectively, pseudo-r^2 for logistic regression, the coefficient/odds of the feature variant and the effect's p-value/significance.

IVs	**adjusted r^2**	**sign.**
guilt/shame	0.971	x
+ *all hedge×storytelling*	1.723	***
+ *fear×guilt/shame*	2.273	***
+ *joy*	2.602	***
+ *disgust×sadness*	3.082	***
+ *boredom×pride*	3.484	***
+ *pride×relief*	3.579	*
+ *pride×sadness*	3.715	**
+ *disgust×fear*	3.774	*
+ *sadness*	3.845	*
+ *storytelling*	3.904	*
+ *fear×relief*	3.962	*

(a) IBM ARGQ

IVs	**pseudo-r^2**	**sign.**
# hedge	0.0106	x
+ *disgust×guilt/shame*	0.0113	***
+ *fear×pride*	0.0119	**
+ *anger×relief*	0.0123	**
+ *# hedge×anger*	0.0128	**
+ *disgust×pride*	0.0132	**
+ *# hedge×guilt/shame*	0.0136	*

(b) CORNELL CMV

Table 9: Features and explained variance of the interactive multiple regression on IBM ARGQ and CORNELL CMV. The model is built stepwise by adding features/interactions with the highest AIC (Akaike Information Criterion relating predictive power to model size) and stops if no improvement is observed. The significance (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$) of adding each new feature is tested via ANOVA for IBM ARGQ and via F-test for CORNELL CMV.