Trust Me, I Can Convince You: The Contextualized Argument Appraisal Framework and the ContArgA Corpus

Lynn Greschner, Sabine Weber, Roman Klinger
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{firstname.lastname}@uni-bamberg.de

Abstract

Emotions that somebody develops based on an argument do not only depend on the argument itself – they are also influenced by a subjective evaluation of the argument's potential impact on the self. For instance, an argument to ban plastic bottles might cause fear of losing a job for a bottle industry worker, which lowers the convincingness – presumably independent of its content. While binary emotionality of arguments has been studied, such cognitive appraisal models have only been proposed in other subtasks of emotion analysis, but not in the context of arguments and their convincingness. To fill this research gap, we propose the Contextualized Argument Appraisal Framework to model the interplay between the sender, receiver, and argument. We adapt established appraisal models from psychology to argument mining, including argument pleasantness, familiarity, response urgency, and expected effort, as well as convincingness variables. To evaluate the framework and pave the way for computational modeling, we develop a novel role-playing-based annotation setup, mimicking real-world exposure to arguments. Participants disclose their emotion, explain the main cause, the argument appraisal, and the perceived convincingness. To consider the subjective nature of such annotations, we also collect demographic data and personality traits of both the participants and ask them to disclose the same variables for their perception of the argument sender. The analysis of the resulting ContArgA corpus of 4000 annotations reveals that convincingness is positively correlated with positive emotions (e.g., trust) and negatively correlated with negative emotions (e.g., anger). The appraisal variables particularly point to the importance of the annotator's familiarity with the argument.

Keywords: emotions, appraisals, arguments, convincingness, persuasion, implicit language

1. Introduction

People are frequently exposed to argumentation, where the main goal is persuasion (Habernal and Gurevych, 2016a). Argument quality plays a key role in this process, depending on logical structure (logos), speaker credibility (ethos), emotional appeal (pathos), and contextual relevance (kairos) (Aristotle, 1991; Schiappa and Nordin, 2013). While there is a substantial amount of research on the overall argument quality and convincingness (Habernal and Gurevych, 2016b; Gleize et al., 2019; Wachsmuth et al., 2017; Lauscher et al., 2020), the emotional appeal of arguments remains underexplored in the natural language processing (NLP) community.

One potential reason is that capturing emotions in text is a challenging task due to their contextualized and implicit nature (Casel et al., 2021; Klinger et al., 2018; Koga et al., 2024; Lee and Lau, 2020). This is particularly true in arguments, which rarely directly express an emotion that is intended to be caused. Figure 1 shows an example in our framework. While the argument appears non-emotional, various factors influence the emotional response: prior beliefs of the receiver, their personality, the relationship with the sender and their traits, or the content of the argument itself. Appraisal theories serve as a suitable basis to model the associated cognitive component (Ellsworth and Scherer, 2003). While they received some attention in event-centered emotion analysis (Klinger, 2023), there is neither a psychological framework of argument appraisal, nor is there work in NLP on fine-grained modeling of emotions in arguments.

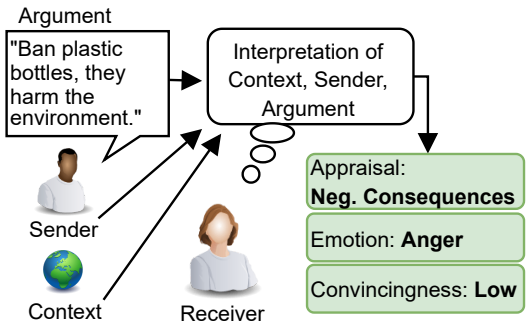


Figure 1: Illustration of the Contextualized Argument Appraisal Framework. The perception of the argument, context and sender are cognitively evaluated and assessed as having a negative consequence for the receiver (e.g., losing a job if plastic bottles were banned).

We close this gap by proposing the Contextualized Argument Appraisal Framework which focuses on the subjective perception of arguments. It encompasses *emotion categories*, a dedicated set of *appraisal variables* such as argument familiarity, expected effort, or the consequences on the

own goals, and the perceived *convincingness*, all in context of variables of the (perceived) argument sender, the receiver, and the argument text itself. The framework therefore is an emotion argument analysis approach under the paradigm of perspectivism (Frenda et al., 2025).

In this paper, we use the proposed framework and data from the corresponding study to answer the following research questions:

- **RQ1:** How do discrete emotions and appraisals correlate with argument convincingness?
- **RQ2:** Which contextual factors shape the emotional response to arguments?

Next to answering these questions, the main contributions are the novel framework and the associated corpus ContArgA, consisting of 4000 individual annotations for 800 arguments[1].

2. Related Work

We now summarize relevant research on argument mining, emotion analysis, and appraisal theories to contextualize our work.

2.1. Argument Mining and Convincingness

Previous work has looked in depth at the textual qualities that make an argument convincing (Habernal and Gurevych, 2016a,b). However, the argument itself is not the only driver of the argument's convincingness; both sender and receiver play a part. Al Khatib et al. (2020) and Lukin et al. (2017) examine the personality of the receiver using the Big Five personality traits, finding that certain personality types are more receptive to certain types of arguments. Lukin et al. (2017) also show that the prior belief is important, with receivers who hold only weak beliefs about the argument topic being more easily convinced. Durmus and Cardie (2018) show that the prior belief of the receiver plays a more important role for the prediction of convincingness than linguistic features of the argument. In line with that, Rescala et al. (2024) demonstrate that LLMs perform on par with humans when predicting which arguments would be convincing for individuals with specific demographics and beliefs.

We follow this work by collecting not only assessments of argument properties, but also properties of the annotators (age, gender, education level and Big Five personality traits). We specifically collect the stance of the receiver toward the argument topic beforehand and check for stance change after an argument was displayed. Additionally, we collect information about the sender that annotators imagine after reading the argument (age, education level, personality traits and a free text description).

2.2. Emotion Analysis and Appraisals

There are two main types of emotion models for emotion analysis: categorical and dimensional. Ekman's basic emotion model, as an example for categorical models, consists of six discrete emotions: anger, surprise, disgust, joy, fear, and sadness (Ekman, 1992). Dimensional models represent emotions along continuous axes in a multidimensional space. The most prominent dimensional model in NLP is the Circumplex Model of Affect (Posner et al., 2005), evaluating emotions in terms of valence and arousal. Appraisal theories also constitute a dimensional model, in which emotions are considered in terms of the cognitive evaluation (appraisal) of an event (Scherer et al., 2001). While multiple frameworks of appraisal theories exist (Roseman, 1984; Roseman and Smith, 2001; Scherer, 2009), there is no consensus on the concrete set, leading to specialized framework for comparably niche areas, such as conspiracy theories (Pummerer et al., 2024). Common variables in appraisal frameworks do, however, encompass aspects of agency, pleasantness, consequences on the self, responsibility, or expected effort and novelty.

While most work on emotion analysis is on categorical models, there is considerable work that employs appraisal theories (Klinger, 2023), including work on coping strategies (Troiano et al., 2024), social media analysis (Stranisci et al., 2022), or emotion event self reports (Hofmann et al., 2020). The largest set of appraisal variables for emotion analysis has been proposed by Troiano et al. (2023), which we use as inspiration in our work. There is no appraisal framework for arguments in psychological research (but the recognition that it would be important, Dillard and Peck (2000)). Therefore, our contribution is both on the conceptual side and the empirical natural language processing side.

2.3. Emotions in Arguments

A substantial number of psychological studies point out the role of cognitive argument evaluations for convincingness (Bohner et al., 1992; Petty et al., 1993; Pfau et al., 2006; Worth and Mackie, 1987; Benlamine et al., 2015). Particularly, *pathos* plays an important role in stance changes (Benlamine et al., 2017). In natural language processing, research on emotions in arguments is limited, where emotional appeal is one factor of many (Habernal and Gurevych, 2016a) or is considered fallacious (Evgrafova et al., 2024).

[1]The data is available under a Creative Commons license here: https://www.uni-bamberg.de/en/nlproc/projects/emcona/.

Dimension	Description
Suddenness	the argument appears sudden or abrupt to the receiver
Suppression	the receiver tries to shut the argument out of their mind
Familiarity	the argument is familiar to the receiver
Pleasantness	the argument is pleasant for the receiver
Unpleasantness	the argument is unpleasant for the receiver
Consequencial Importance	the argument has important consequences for the receiver
Positive Consequentiality	the argument has positive consequences for the receiver
Negative Consequentiality	the argument has negative consequences for the receiver
Consequence Manageability	the receiver can easily live with the unavoidable consequences of the argument
Internal Check	the consequences of the argument clash with the receiver's standards and ideals
External Check	the consequences of the argument violate laws or socially accepted norms
Response urgency	the receiver urges to immediately respond to the argument
Cognitive Effort	processing the argument requires a great deal of energy of the receiver
Argument Internal Check	statements in the argument clash with the receiver's standards and ideals
Argument External Check	statements in the argument violate laws or socially accepted norms

Table 1: Appraisal dimensions used to measure argument evaluation in the context of the self.

The convincingness and emotion are related, though studies with large language models show the challenging nature of automatically detecting these properties (Chen and Eger, 2025). There is very little work that studies not only binary emotionality but differentiates emotion categories. Greschner and Klinger (2025), as one of the few cases, demonstrate that the discrete emotion category enhances LLM predictions for emotionality in arguments. Habernal and Gurevych (2016a,b) constructed and analyzed a corpus for convincingness strategies, including an emotion layer. In line with that, Wachsmuth et al. (2017) find a weak positive correlation between the emotional appeal of arguments and convincingness when analyzing 15 dimensions of arguments. Lukin et al. (2017) focus on audience factors and argument convincingness and find that information about personality traits of the audience improves the automatic prediction of belief change for emotional arguments.

Our work considers sender and receiver properties in the context of argument convincingness, focusing on the implicitly evoked emotions in receivers. To the best of our knowledge, this work is the first to introduce appraisal theories to arguments in NLP.

3. Contextualized Argument Appraisal Framework

We illustrate the Contextualized Argument Appraisal Framework (CAAF) in Figure 1. CAAF considers arguments within a situational context with the sender and world knowledge, and the self, the receiver. The receiver cognitively evaluates, in a synchronized but not necessarily sequential manner, the appraisal, emotion category, and convincingness of the argument.

3.1. Emotions

The cognitive appraisal of the argument leads to a subjective experience that corresponds to a discrete emotion category. This emotion category, along with the perceived emotion intensity, are subjective variables of the receiver and need to be assessed by them (i.e., cannot be externally reconstructed by annotators).

Arguments are situated within context and presented by a sender. We include these factors in CAAF to study the major source of the emotion stimulus: *how* is the emotion evoked and *why* is the emotion evoked.

3.2. Appraisals

To reconstruct the subjective cognitive evaluation process of a presented argument in context, we introduce appraisal variables for arguments, the main and novel component in our framework. We base the appraisals used in our study on the work by Troiano et al. (2023), adapting a subset (*suddenness*, *familiarity*, *pleasantness*, *unpleasantness*, *internal check*, *external check*) to match the appraisal of arguments, as described in Table 1. Furthermore, we design nine additional appraisal dimensions: *suppression*, *consequencial importance*, *positive consequentiality*, *negative consequentiality*, *consequence manageability*, *response urgency*, *cognitive effort*, *argument internal check*, and *argument external check*.

Appraisal dimensions allow us to capture the underlying cognitive evaluations that give rise to emotional responses. They enable a systematic study of subjective argument properties.

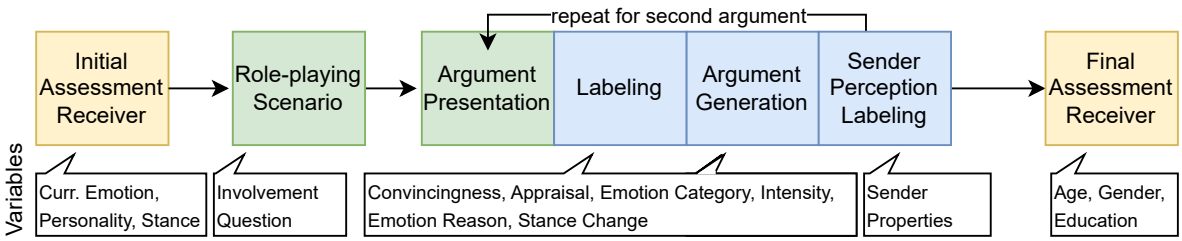


Figure 2: Flowchart of the annotation workflow. Yellow fields indicate information collected about the receiver (the annotator), green fields indicate argument and scenario display and blue fields indicate argument specific annotation.

3.3. Convincingness

Central to the CAAF is the convincingness assessment of arguments. Convincingness refers to the quality of an argument that leads its audience to believe the content of the argument is factually true or morally right. In our framework, we acquire this variable by a self-assessment of the argument receiver, but, in addition, obtain the psychologically more valid property of an observed stance change. We do, however, use the convincingness variable in our studies due to pragmatic design choices in the data collection described later.

3.4. Context

The variables above can only be assessed in the situational context, which therefore requires information about the sender and the receiver. Text is rarely produced in a contextual vacuum. Rather it is influenced by demographic properties, personality traits and preconceived notions of the people involved. Capturing these features allows us to create models that take individual nuances into account, both in the ground-truth properties of the receiver (e.g., their education level) and the perceived properties of the sender (e.g., does the argument sound like it was said by an educated person).

4. Corpus Creation

We now describe our annotation setup based on CAAF, to study the correlations of emotion, appraisal, and convincingness, and pave the way for computational modeling. Screenshots of the annotation study are in Appendix F.

4.1. Overview

Figure 2 illustrates our annotation setup. The study begins with assessing the participant's emotional state, personality, and stance toward the topic. Subsequently, participants are immersed in a role-playing scenario simulating a town-hall meeting to approximate a real-world argumentative context. The exact scenario descriptions and the involvement questions are Appendix E.

To ensure that participants do engage with the role-playing, we ask them to describe where they would sit and what they would see in their surroundings. Next, we tell participants that a speaker presents an argument, which they annotate. They then generate an argument themselves to increase involvement (this data is not evaluated in this paper). Participants further answer demographic and personality trait questions about the speaker. This process is repeated for a second argument within the same role-playing scenario. Lastly, the participants disclose own demographics. Seven attention and sanity checks are included in the study; see Appendix F for details.

4.2. Variables

We now turn to a detailed description of the annotated variables.

4.2.1. Argument

The following section describes the collected variables on the argument level in detail. All variables are assessed on a 5-point Likert scale if not otherwise stated.

Emotions. Our emotion label set comprises Ekman's basic emotions *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, expanded by *shame*, *guilt*, *pride*, (Greschner and Klinger, 2025), *trust* and *relief* (Troiano et al., 2023). We refer to the emotion with the highest rating as the dominant emotion. In the case of ties, participants are requested to chose one. We additionally assess the perceived interest and boredom of the argument, see Appendix F for the formulations.

Emotion Reason. Participants are instructed to provide a short textual explanation of the evoked emotions using a free text field.

Emotion Influence. Participants answer three questions about the influence of the sender, argument content, and receiver. The questions are

ID	Argument	Topic	Annotations	
			Emo.	Conv.
1	"people simply can't compete with others in the same professional setting once they reach a certain age, mandatory retirement is necessary for fair competition in a limited job market"	Retirement	*guilt*	4
			disgust	1
			relief	3
			sadness	2
			surprise	1
2	"women in polygamous marriages have been shown to be at much greater risk of abuse than those in monogamous marriages"	Polygamy	*disgust*	3
			sadness	2
			anger	3
			sadness	3
			surprise	2
3	"there are direct links to drinking bottled water and certain types of disease, even cancer"	Plastic bottles	*fear*	4
			trust	1
			anger	1
			guilt	2
			surprise	2

Table 2: Examples from the ContArgA corpus (argument text source: Habernal and Gurevych, 2016b; Gretz et al., 2020). Emotion and convincingness assessments are from five individual annotators.

phrased in the format "[The speaker identity, the argument itself, who I am] was important for the emotion I developed".

Appraisals. We phrase each appraisal dimension as an affirmation, e.g., for response urgency we formulate: "I feel the urge to immediately respond to the argument".

Manipulation. For each argument, participants provide their perceived degree of manipulative intent. We formulate the question: "Do you feel manipulated by this argument?"[2].

Convincingness. Similar to the manipulation variable, participants annotate the perceived convincingness for the question "How convincing is this argument for you?"

4.2.2. Receiver

Given that individual differences among receivers can affect the perceived persuasiveness of arguments, as demonstrated by Lukin et al. (2017), our study accounts for these variations by employing the following measurements.

Stance in Relation to Argument Topic. Participants are asked to provide their stance towards a the topic of the displayed town hall discussion. This question is asked once prior to argument exposure and once after each argument, allowing us to measure potential change in the stance after specific arguments. Additionally, participants provide their familiarity with the topic (not to be confounded with the appraisal of familiarity with the argument).

Demographics. Previous work on argument quality has shown that the lack of information about the annotators of corpora poses an impediment to perspectivist approaches to the topic (Romberg et al., 2025). To avoid this shortcoming, we collect the age, gender and education level of the participants.

Personality Traits. Lukin et al. (2017) demonstrate that more conscientious people are rather convinced by emotional arguments. Since we hypothesize that evoked emotions are one of the primary drivers of the argument's convincingness, we use the Ten-Item Personality Inventory questionnaire (TIPI, Gosling et al., 2003) to assess the extraversion, agreeableness, conscientiousness, emotional stability, and openness to experiences of the participants.

4.2.3. Sender

In our study setup, the age (old/young) and gender (male/female) of the argument-presenting sender are systematically manipulated. While we are aware that neither age nor gender are binary variables, this simplified setup allows us to investigate whether these properties influence persuasiveness. Each argument is attributed to one single sender, whose age and gender are assigned according to a balanced randomization scheme across topics to ensure approximately equal representation of all four demographic combinations. We present participants with a description of the sender and the argument in the format "A [speaker age] [speaker

[2] In preliminary studies, we found perceived manipulation to decrease convincingness. In ContArgA, there is a weak, significant negative correlation between manipulation and convincingness, which we leave for future work for a more in-depth analysis.

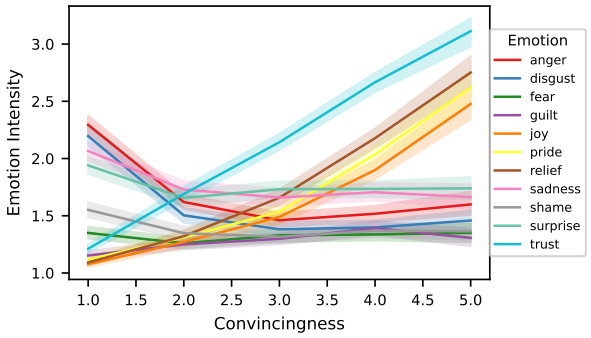


Figure 3: Emotion Intensities and Convincingness.

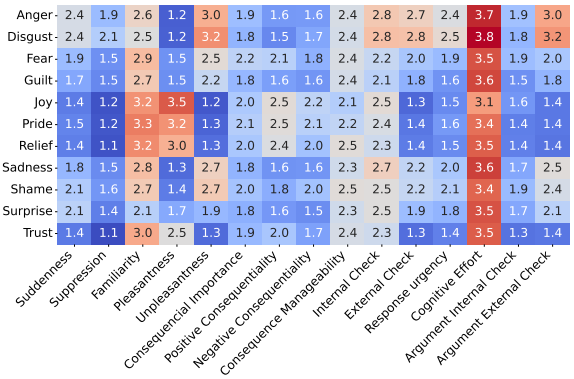


Figure 4: Analysis of discrete emotion categories and appraisal dimensions in arguments. Each row is an appraisal, each column is an emotion, and each cell is the average appraisal value for that emotion.

gender] approaches the microphone and makes a statement: [argument]".

Following the argument exposure, participants answer the same demographic and TIPI questionnaire questions described in Section 4.2.2 about their perception of the sender. This approach allows us to examine the interplay between speaker characteristics and audience perception within argumentative discourse. Additionally, the participants can give details about the imagined sender in a free text field. We leave a more in-depth analysis of these variables for future work.

4.3. Crowd-sourcing Details

We use the platform Prolific[3] with a self-implemented study framework based on Streamlit[4]. Participants are required to live in the United Kingdom or Ireland, have English as their first, native, and primary language, and have an approval rate of 95–100%. Each participant answers the survey for two arguments. We pay each participant 3.48 £ for one survey, which on average takes 22.5 minutes. The study includes 7 attention checks. Participants can participate up to 40 times, and therefore annotate up to 80 arguments. In total, the cost of the study amounts to 9,404 £ . The contributing participants in our studies were on average 40.4 years old (18 minimum, 81 maximum). From this set, 1,715 identified as female, 1,541 as male, 374 as non-binary, and 30 as genderfluid/non-conforming. Further, 14 specified their own gender, 12 preferred not to answer the question, and 8 as questioning[5].

4.4. Datasets

The *UKPConvArgv1* (Habernal and Gurevych, 2016b) and *IBM-Rank-30k* (Gretz et al., 2020) corpora, hereafter UKP and IBM, serve as sources for the arguments. Both datasets provide isolated arguments and their respective discussion topic. Isolated arguments allow us to investigate the initial argument appraisal without adding noise due to lengthy debates containing multiple arguments. We sample 14 and 26 topics from UKP and IBM, respectively. For each topic, an equal number of pro and con arguments is sampled.

Arguments are manually annotated for comprehensiveness, persuasive attempt, and stance alignment. Details on the manual filtering are in Appendix A. Importantly, we exclude arguments that point toward the age or gender of the sender, since we manipulate these variables in our study (see Section 4.2.3). We randomly select 800 arguments from the 953 valid ones for our study.

5. Data Analysis

Our study results in a corpus of 4,000 individual argument annotations of 800 arguments from 39 topics with 5 individual annotations (170 annotations excluded due to failed attention checks. Details about the topics can be found in Appendix D.

Table 2 displays example arguments with emotion and convincingness labels. The average argument length is 32.1 tokens, 1.7 sentences. The average value of convincingness is 2.7 (min=1, max=5); the values vary across topics, with *retirement* being most (3.2) and *atheism* being least convincing (2.0). The frequencies of the dominant emotions are, in descending order: *trust* (709), *surprise* (564), *sadness* (561), *anger* (521), *relief* (382), *disgust* (292), *pride* (198), *joy* (193), *shame* (100), *fear* (90), *guilt* (84). The average emotion intensity value across emotions is 1.61, with *trust* showing the highest (2.02), and *guilt* the lowest (1.27) value.

[3] https://www.prolific.com/

[4] https://docs.streamlit.io/

[5] Note that on the Prolific platform, participants cannot provide their gender in such detail. We distributed our study as 40% female, 40% male, 20% non-binary and obtained the gender variable in the study.

Emotions	r	Appraisals	r
Trust	.58***	Pleas	.58***
Relief	.51***	Pos Cons	.40***
Pride	.47***	Fam	.34***
Joy	.45***	Neg Cons	.22***
Guilt	.11***	Cons Imp	.15***
Fear	.01	Cons Man	−.02
Shame	−.05*	Cog Eff	−.08***
Surprise	−.09***	Int Ch	−.11***
Sadness	−.13***	Arg Int Ch	−.12***
Anger	−.22***	Resp Ur	−.25***
Disgust	−.23***	Sup	−.30***
Other	r	Sud	−.34***
Interest	.69***	Unpleas	−.36***
Manipulation	−.10***	Ex Ch	−.37***
Boredom	−.49***	Arg Ex Ch	−.50***

Table 3: Pearson correlations (r) of emotions and appraisals with the convincingness variable. Significance levels (with Bonferroni correction): * $p < .05$, ** $p < .01$, *** $p < .001$.

We now turn to the example arguments from ContArgA, displayed in Table 2, in more detail. The first argument evokes 5 different emotions and convincingness labels from 1–4. In contrast, the second argument evokes negative emotions (*disgust*, *sadness*, *anger*) with similar convincingness values, with the exception of *surprise*. The second argument discusses the greater risk of abuse of women in polygamous marriages, intuitively evoking negative emotions. *Surprise* may arise if the receiver is unfamiliar with the topic.

For the third argument, we find negative emotions (*fear*, *anger*, *guilt*). *Fear* may result from the risk of developing cancer, while a person regularly using plastic bottles could feel guilty. The argument evoking trust is less intuitively clear. These examples illustrate the wide variety of emotions evoked in different participants by the same arguments.

Inter-Annotator Agreement. The contextualized argument appraisal framework is designed to capture individual argument appraisals on divisive topics. We intentionally construct it to elicit differences between receivers of arguments rather than to derive one 'gold-label', leading to 5 individual argument appraisals for each argument. As expected, the inter-annotator agreement on single variables is therefore low, with a Krippendorff's Alpha of 0.06 for convincingness and Fleiss' Kappa of 0.17 for the most dominant emotion category of an argument. These numbers do not suggest a low quality of the data. In contrast, they justify the need for modeling and analysis under the paradigm of perspectivism.

6. Results

In the following, we answer our posed research questions individually.

6.1. RQ1: How do discrete emotions and appraisals correlate with argument convincingness?

We analyze ContArgA to investigate the interplay between evoked emotions and convincingness.

Emotions. We aim at understanding the specific emotions that are evoked in the receivers of arguments to examine their interplay with perceived convincingness. To this end, we analyze ContArgA arguments for relationships between discrete emotion categories and convincingness labels.

Table 3 displays all correlation values of the variables of interest and convincingness with Bonferroni correction. All correlations between emotion and convincingness are significant, except for *fear*. The strongest positive effects are with *trust* (.57), *relief* (.51), *pride* (.46), and *joy* (.44). In contrast, *anger* (−.22) and *disgust* (−.23) show the strongest negative correlations with convincingness.

Emotion Intensities. Figure 3 allows a closer look at these correlations, based on intensity values. For *trust*, the emotion intensity and convincingness have an almost linear relationship, similarly for *relief*, *pride*, *joy*. For *fear*, *shame*, and *guilt*, the emotion intensity stays relatively stable across convincingness values. In contrast, for *anger* and *disgust*, the highest intensity values are found for the lowest convincingness values, with a drop at value 2 (little convincing), continuing in a moderate intensity (around 1.4) across higher convincingness values. Similar, yet weaker trends are found for *sadness* and *surprise*, with on average higher emotion intensity values (around 1.7).

Appraisals. We now focus on the analysis of emotions and appraisals in arguments, illustrated in Figure 4. For each emotion, the average appraisal value is displayed in the heatmap. To the best of our knowledge, we are the first to apply appraisal theory to arguments. However, the appraisal dimensions of *suddenness*, *familiarity*, *pleasantness*, *unpleasantness*, *internal check*, *external check* values are in line with other applications of appraisal theories to emotion analysis in natural language processing (Troiano et al., 2023, Figure 8), confirming the validity of the approach.

Turning to the analysis of appraisal dimensions in arguments, we find that *unpleasantness*, *internal check* and *external check* appraisals consistently show the highest ratings for *anger* (3.0, 2.8, 2.7) and *disgust* (3.2, 2.8, 2.9). Intuitively, arguments evoking negative emotions are being appraised as unpleasant, clashing with the receiver's ideals,

Emotion	Sender	Argument	Receiver
Anger	1.86	3.78	3.23
Disgust	1.82	3.72	3.25
Fear	1.78	3.68	3.11
Guilt	1.65	3.27	2.81
Joy	2.13	3.73	3.21
Pride	2.24	3.79	3.24
Relief	1.98	3.70	3.14
Sadness	1.84	3.41	3.09
Shame	2.00	3.34	2.86
Surprise	1.86	3.39	2.69
Trust	1.82	3.47	2.97
Avg.	1.89	3.56	3.05

Table 4: Average values of how important the sender, the argument, and the receiver are for the emotional response of the receiver. The importance of each dimension is rated on a 1–5 scale.

and violating social norms. In contrast, *joy*, *pride*, and *relief* show high values for *pleasantness* (3.5, 3.2, 3.0), *familiarity* (3.1, 3.3, 3.2), and *positive consequentiality* (2.5, 2.5, 2.5). This indicates that arguments being appraised as beneficial, familiar, and rewarding evoke positive emotions.

Beyond categorical emotions, appraisals are also related with convincingness. Table 3 displays all correlations. We find significant correlations for all appraisal dimensions except *consequence manageability*. *Pleasantness* shows a strong positive (.57), *positive consequentiality* a moderate (.39), and *familiarity* a moderate (.33) positive correlation with convincingness. In contrast, we find strong negative correlations with *argument external check* ($-$.50), *unpleasantness* ($-$.39), as well as *external check* ($-$.36).

Both discrete emotion and appraisal analyses demonstrate significant correlations between positive emotions and convincingness, and negative correlations between negative emotions and convincingness. The results with respect to discrete emotions align with Greschner and Klinger (2025) (showing negative correlations between negative emotions and convincingness using 300 German arguments), but contradict that *fear* increases convincingness (Dillard and Anderson, 2004).

6.2. RQ2: Which contextual factors shape the emotional response to arguments?

After analyzing *which* emotions are developed and how they are correlate with the argument's convincingness, we now turn to understanding *how* these emotions are developed. We analyze the participant's (argument receiver's) textual explanations about the evoked emotions and the strongest influence on the emotion (sender, argument, receiver). We also analyze how strong (1–5) the influence (sender, argument, receiver) was for the development of the evoked emotion.

Table 4 shows the results. The argument (3.6) is the primary driver of the emotional response, followed by the receiver themselves (3.1) and the sender (1.9). Receivers rarely attribute *guilt* and *shame* to being evoked by the sender. *Pride* is often attributed to the argument itself, which is also the most frequent emotion and has the highest correlation to convincingness. Similar patterns are found for *anger*, which has, however, the strongest negative convincingness correlation. Interestingly, the sender is the least important component when it comes to the evoked emotion of *guilt*. In summary, the argument itself is the primary driver of the emotional response.

7. Conclusion

In this paper, we introduce the Contextualized Argument Appraisal Framework (CAAF). It is a novel approach to explain how arguments are cognitively appraised. In a synchronized process, individuals develop an emotion and perceive a particular convincingness after argument exposure. The analysis of the resulting corpus ContArgA, the data resulting from a role-playing annotation study based on CAAF, shows that appraisals are appropriate for explaining the relationship between emotions and convincingness. Furthermore, our results validate the framework's contextual approach: while the arguments themselves generally exert the strongest influence on emotional responses, emotion-specific patterns emerge regarding the relative contributions of sender, argument, and receiver factors. We release the data, enabling research on arguments with focus on such contextual properties.

Our work demonstrates that affective responses are not merely superficial reactions but are integrally tied to how arguments are processed. Future work shall study CAAF from a computational perspective. We hypothesize that the integration of appraisal and categorical emotion variables improves the performance of automatic convincingness estimates and make the decision process transparent. Further, models that automatically assess the subjective appraisal and emotion are valuable in itself for corpus and argumentation studies.

Particularly, CAAF helps to disentangle the influences that shape a person's reaction to arguments, because it captures the role of properties of the sender, receiver, and the argument itself. In downstream systems, these factors can contribute to argument mining and content moderation systems, for example to support studies on desinformation and argumentation strategies. The framework lays the groundwork for context-sensitive and emotion-

ally aware computational systems that reflect how people engage with persuasive text.

Limitations

One of the limitations of this work is the focus on the English language and a relatively limited geographical spread of the annotators involved in the creation of our corpus. We make this choice to ensure higher data quality, limiting a possible mismatch between dialects or regional language usage. On the other hand, a wider linguistic and geographic spread would allow us to make statements about the nature of emotions in arguments that are more comprehensive. While we see these shortcomings in the annotation itself, the Contextualized Argument Appraisal Framework stands independent of its implementation in this specific annotation scenario and can be used across various languages.

Previous work has also examined multi-turn argumentative conversations. While this allows an observation of how emotions and appraisals develop over the course of a longer exchange, this is out of the scope of our current setting of the framework. Adapting it for multi-turn interactions is another avenue for future work.

Regarding our study, each participant annotates two arguments from one topic. The argument selection from the topic is randomized, which can lead to biases due to the prior stance of the participant and the order of pro and con arguments. We acknowledge that future work can take a deep dive into the role of argument presentation according to stances, since we report the exact argument in displayed order within our available ContArgA corpus. Further, we allow participation up to 39 times, which ensures that one participant can only annotate two arguments per topic, however, there might be annotators that do in fact participate in all 39 studies (i.e, contributing 78 individual argument annotations out of the 4,000 arguments in ContArgA), therefore, introducing annotator bias.

To address the emerging challenge of automated bot participation in crowdsourcing studies, we manually verfy all free-text response to ensure authenticity. Participants whose completion times fell significantly below the study average were systematically excluded from analysis. Recognizing the inherent quality control challenges in crowdsourced research, we incorporated seven attention and saliency checks throughout the study. Only data from participants who successfully passed all validation measures were retained for analysis. These quality assurance protocols, while necessary to mitigate common crowdsourcing limitations, enabled us to leverage the key advantage of crowdsourcing: access to a demographically diverse participant pool that would be difficult to achieve through other recruitment methods.

Ethical Considerations

The study is part of a larger project that has been approved, including a description of the presented study, by the ethics board of the University of Stuttgart. Following recommendations with respect to ethical challenges in emotion analysis (Mohammad, 2022), we point out the following facts about our work: The purpose of the CAAF is to study the interplay of emotions in argument texts while taking contextual factors into account, focusing on the emotions evoked in receivers. We choose categorical emotion labels from emotion theories in accordance with previous work, and expand appraisal labels from dimensional appraisal theories to be applicable to arguments.

We employed crowd-sourcing to collect human annotations for the cognitive appraisal process of arguments. For a given argument, participants provide emotional responses, convincingness, stances toward and familiarity with the topic. In addition, participants provide demographic information, personality traits, and their emotional state before starting the study. We informed the participants that the intent of the study is research with the goal of a scientific publication. We obtained consent from all participants to use their data to analyze emotions and convincingness in arguments as well as publishing their annotation in anonymized from. While there is one free text field in our study, we manually check all of these responses for any hints that would allow tracing back the identity of the annotator, for instance, if a participant unconsciously disclosed personal information. We explicitly make annotators aware that they will not be compensated if they fail more than 2 attention checks. Details about the payment are described in Section 4.3. We are aware that being exposed to arguments can be upsetting. We manually select a comprehensive set of trigger words and warn the study participants about the potentially upsetting nature of arguments. The participants can drop out of the study at any point without consequences.

For this study, we chose to use crowd-sourcing since the study is closer to a study in psychology than NLP. Our data collection can be seen more as an online experiment. The platform we use, Prolific, is designed for that. However, we do provide fair compensation to our study participants and report all relevant details about the online experiment. The quality of our data is ensured by multiple attention and saliency checks, which are all reported. With proper attention and saliency checks as well as a careful study design, crowdsourcing is the best option to collect lots of data from a broad variety of people with different demographics.

8. Acknowledgements

This project has been conducted as part of the EMCONA (The Interplay of Emotions and Convincingness in Arguments) project, which is funded by the German Research Foundation (DFG, project KL2869/12–1, project number 516512112). We thank Meike Bauer for her contribution to the data preparation and the fruitful discussions.

9. Bibliographical References

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.

Aristotle. 1991. *On Rhetoric: A Theory of Civic Discourse*, 2nd edition. Oxford University Press, New York.

Mohamed 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*.

Mohamed Benlamine, Ramla Ghali, Serena Villata, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive argumentation and emotions: An empirical evaluation with users. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Gerd Bohner, Kimberly Crow, Hans-Peter Erb, and Norbert Schwarz. 1992. Affect and persuasion: Mood effects on the processing of message content and context cues and on subsequent behavior. *European Journal of Social Psychology*, 22:511–530.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Yanran Chen and Steffen Eger. 2025. Do emotions really affect argument convincingness? a dynamic approach with LLM-based manipulation checks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24357–24381, Vienna, Austria. Association for Computational Linguistics.

James Price Dillard and Eugenia Peck. 2000. Affect and Persuasion: Emotional Responses to Public Service Announcements. *Communication Research*, 27(4):461–495. Publisher: SAGE Publications.

James Prince Dillard and Jason Anderson. 2004. The role of fear in persuasion. *Psychology & Marketing*, 21(11):909–926.

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.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Phoebe Clemencia Ellsworth and Klaus Rainer Scherer. 2003. *Appraisal Processes in Emotion*. Oxford University Press, Oxford, UK.

Natalia Evgrafova, Veronique Hoste, and Els Lefever. 2024. Analysing pathos in user-generated argumentative text. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 39–44, Torino, Italia. ELRA and ICCL.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746. Publisher: Springer Science and Business Media LLC.

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.

Samuel Gosling, Peter Rentfrow, and William Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Lynn Greschner and Roman Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 628–646, Albuquerque, USA. 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. 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.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Roman Klinger. 2023. Where are we in event-centric emotion analysis? bridging emotion role labeling and appraisal-based approaches. In *Proceedings of the Big Picture Workshop*, pages 1–17, Singapore. Association for Computational Linguistics.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

Yurie Koga, Shunsuke Kando, and Yusuke Miyao. 2024. Forecasting implicit emotions elicited in conversations. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 145–152, Tokyo, Japan. Association for Computational Linguistics.

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

Sophia Yat Mei Lee and Helena Yan Ping Lau. 2020. An event-comment social media corpus for implicit emotion analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1633–1642, Marseille, France. European Language Resources Association.

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.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Richard Petty, David Schumann, Steven Richman, and Alan Strathman. 1993. Positive mood and persuasion: Different roles for affect under high and low-elaboration conditions. *Journal of Personality and Social Psychology*, 64:5–20.

Michael Pfau, Erin Alison Szabo, Jason Anderson, Joshua Morrill, Jessica Zubric, and Hua-Hsin Wan. 2006. The role and impact of affect in the process of resistance to persuasion. *Human Communication Research*, 27:216 – 252.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734.

Lotte Pummerer, Theofilos Gkinopoulos, Karen M. Douglas, Daniel Jolley, and Kai Sassenberg. 2024. The Appraisal Model of Conspiracy Theories (AMCT): Highlighting Core Concepts and Potential Extensions. *Psychological Inquiry*, 35(3-4):233–242. Publisher: Informa UK Limited.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837, Miami, Florida, USA. Association for Computational Linguistics.

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. 2025. Towards a perspectivist turn in argument quality assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7458–7485, Albuquerque, New Mexico. Association for Computational Linguistics.

Ira Roseman. 1984. Cognitive determinants of emotion: A structural theory. *Rev. Pers. Soc. Psychol.*, 5.

Ira Roseman and Craig Smith. 2001. *Appraisal Theory: Overview, Assumptions, Varieties, Controversies*. Oxford University Press, Oxford, UK.

Klaus Rainer Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7):1307–1351.

Klaus Rainer Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Edward Schiappa and John Nordin. 2013. *Argumentation: Keeping Faith with Reason*, 1st edition. Pearson UK, Harlow.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. APPReddit: a corpus of Reddit posts annotated for appraisal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.

Enrica Troiano, Sofie Labat, Marco Stranisci, Rossana Damiano, Viviana Patti, and Roman Klinger. 2024. Dealing with controversy: An emotion and coping strategy corpus based on role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1634–1658, Miami, Florida, USA. 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).

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.

Leila Worth and Diane Mackie. 1987. Cognitive mediation of positive affect in persuasion. *Social Cognition*, 5:76–94.

A. Manual Data Filtering of UKP and IBM.

The *UKPConvArgv1* (Habernal and Gurevych, 2016b) and *IBM-Rank-30k* (Gretz et al., 2020) corpora, further referred to as UKP and IBM, serve as the source for the arguments. We aimed for 1000 arguments in total. We approximate needing 40 various topics to capture topical differences; hence, we select all 16 topics from UKP and randomly select 24 topics from IBM. For each topic, we randomly sample 14 arguments per stance (pro/con). For this set, we perform standard data cleaning and pre-processing steps.

We manually check if the argument matches our definition of arguments: *The argument must provide reasons supporting a specific claim (standpoint or point of view). The argument must be complete and comprehensive in itself. The reasons must at least be attempted to be logical and reasonable. The argument must follow the goal of persuasion (convince others of the supported standpoint) and must be logical in itself.* Further, we exclude arguments using the following criteria. (1) The argument is not in line with our definition of an argument. (2) The argument shows hints toward author characteristics (e.g., “As a woman, I would argue that...”). (3) The argument is a duplicate. (4) The argument does not have a clear stance toward the topic.

For the resulting arguments, we check the stance of each argument for correctness, re-annotate the stance if necessary, and normalize all stances to be either “yes” if the argument supports the topic or “no” if the argument opposes the topic. We normalize the arguments’ topics according to this structure: “We should [verb] [object].” For instance, we normalize the topic “gay-marriage-right-or-wrong” to “We should allow gay marriage”.

Variable	r	p
Anger	−0.222***	0.000
Disgust	−0.233***	0.000
Fear	0.006	1.000
Guilt	0.113***	0.000
Joy	0.447***	0.000
Pride	0.465***	0.000
Relief	0.510***	0.000
Sadness	−0.125***	0.000
Shame	−0.052*	0.028
Surprise	−0.085***	0.000
Trust	0.578***	0.000
Suddenness	−0.345***	0.000
Suppression	−0.296***	0.000
Familiarity	0.336***	0.000
Pleasantness	0.576***	0.000
Unpleasantness	−0.363***	0.000
Consequential Importance	0.151***	0.000
Positive Consequentiality	0.400***	0.000
Negative Consequentiality	0.217***	0.000
Consequence Manageability	−0.018	1.000
Internal Check	−0.105***	0.000
External Check	−0.366***	0.000
Response Urgency	−0.249***	0.000
Cognitive Effort	−0.078***	0.000
Argument Internal Check	−0.124***	0.000
Argument External Check	−0.498***	0.000
Manipulation	−0.104***	0.000
Interest	0.691***	0.000
Boredom	−0.492***	0.000

Table 5: Pearson correlations (r) and corresponding p values with Bonferroni correction for all variables discussed in the main paper. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$. All numbers are rounded to three decimal points.

This manual effort results in 953 arguments, from which we randomly select 800 for the annotation.

B. Correlations

To answer our research questions, we compute correlations between all emotion variables, interest, boredom, manipulation, all appraisal variables, and convincingness. The results are displayed in Table 5. The correlations are contextualized, interpreted, and analyzed in the main part of the paper.

C. Emotion Influence

In Section 6.2 we analyze the importance of the sender, argument, and receiver in detail to understand the drivers of emotional responses. We analyze how important (1-5) the influence (sender, argument, receiver) was for the development of the evoked emotion, individually for each topic and display the results in Table 6. While the main results are discussed and contextualized in Section 6.2,

Topic	Sender	Argument	Receiver
Weapons	1.67	4.08	3.30
Farming	1.48	3.76	3.37
Zoos	1.51	3.71	3.05
Libertarianism	1.54	3.69	3.18
Gay Marriage	1.85	3.97	3.71
Guantanamo	1.65	3.76	3.16
Cosmetic Surgery	1.78	3.73	2.89
Exec. Compensation	1.64	3.51	3.02
Gender	2.04	3.87	3.30
Naturopathy	1.73	3.56	2.93
Plastic Bottles	1.82	3.63	2.67
TV Books	1.66	3.47	2.62
Int. Property	1.69	3.48	2.66
Entrapment	1.79	3.56	3.34
Personal Pursuit	1.80	3.55	2.93
ZTP Schools	1.76	3.46	3.04
PE Schools	1.74	3.43	3.12
Atheism	1.78	3.47	3.00
Journalism	1.76	3.43	2.81
Holocaust	1.82	3.48	2.96
Economic Sanctions	1.82	3.46	3.06
Prostitution	2.09	3.72	3.24
Spouse	2.03	3.66	3.45
Creationism	1.56	3.17	3.10
Education	1.98	3.57	3.07
Spanking	2.04	3.61	3.33
Polygamy	1.90	3.44	2.98
India	1.92	3.44	2.98
School Uniform	2.00	3.51	2.78
Abortion	2.35	3.85	3.00
Women Combat	2.31	3.77	3.39
Student Loans	1.93	3.36	2.82
Judicial Activism	2.08	3.50	2.83
Fastfood	1.95	3.36	3.09
Father	2.23	3.44	3.20
Retirement	2.38	3.56	2.84
Subsidize Dads	2.18	3.35	2.98
Porn	1.92	2.83	3.08
Firefox	2.22	3.14	2.56

Table 6: Average emotion influence scores across topics.

we display the results for all topics here.

D. Topics

ContArgA contains 39 distinct topics. We display the short names for the topics, along with a description description of each topic, the number of pro, con, and total arguments in Table 7.

E. Role-Playing Scenario

To mimic the exposure to an argument in the wild and to increase emotional engagement, we create the role-playing scenario of a town hall meeting that the annotator is part of. The exact role-playing scenario is described in Figure 5. The involvement

Name	Description	Corpus	Pro	Con	Total
Abortion	We should legalize abortion	IBM	9	9	18
Atheism	We should adopt atheism	UKP	7	7	14
Cosmetic Surgery	We should ban cosmetic surgery for minors	IBM	7	7	14
Creationism	We should believe in creationism	UKP	13	13	26
Economic Sanctions	We should end the use of economic sanctions	IBM	11	11	22
Education	We should subsidize vocational education	IBM	13	13	26
Entrapment	Entrapment should be legalized	IBM	13	13	26
Exec. Compensation	We should limit executive compensation	IBM	13	13	26
Farming	We should ban factory farming	IBM	11	11	22
Fastfood	We should ban fast food	IBM	14	14	28
Father	It is better to have a lousy father than to be fatherless	UKP	13	13	26
Firefox	We should choose firefox over internet explorer	UKP	11	11	22
Gay Marriage	We should allow gay marriage	UKP	4	4	8
Gender	We should adopt gender-neutral language	IBM	12	12	24
Guantanamo	We should close Guantanamo Bay detention camp	IBM	14	14	28
Holocaust	Holocaust denial should be a criminal offence	IBM	13	13	26
India	India has the potential to lead the world	UKP	6	6	12
Int. Property	We should abolish intellectual property rights	IBM	12	12	24
Journalism	We should subsidize journalism	IBM	12	12	24
Judicial Activism	We should limit judicial activism	IBM	13	13	26
Libertarianism	We should adopt libertarianism	IBM	12	12	14
Naturopathy	We should ban naturopathy	IBM	14	14	28
PE Schools	Physical education should be mandatory in schools	UKP	9	9	18
Personal Pursuit	We should prioritize the personal pursuit over advancing the common good	UKP	13	13	26
Plastic Bottles	We should ban plastic water bottles	UKP	6	6	12
Polygamy	We should legalize polygamy	IBM	13	13	26
Porn	Porn is wrong	UKP	8	8	16
Prostitution	We should legalize prostitution	IBM	14	14	28
Retirement	We should end mandatory retirement	IBM	13	13	26
School Uniform	The school uniform is a good idea	UKP	9	9	18
Spanking	We should allow spanking of children	UKP	10	10	20
Spouse	One should turn in their spouse if they committed murder	UKP	12	12	24
Student Loans	We should subsidize student loans	IBM	13	13	26
Subsidize Dads	We should subsidize stay-at-home dads	IBM	13	13	26
TV Books	TV is better than books	UKP	7	7	14
Weapons	We should fight for the abolition of nuclear weapons	IBM	11	11	22
Women Combat	We should prohibit women in combat	IBM	12	12	24
ZTP Schools	We should adopt a zero-tolerance policy in schools	IBM	13	13	26
Zoos	We should abolish zoos	IBM	13	13	26

Table 7: List of topics in ContArgA, along with original data source (IBM-Rank-30k (Gretz et al., 2020) or UKPConvArgv1 (Habernal and Gurevych, 2016b)), pro arguments, con arguments, and total number of arguments.

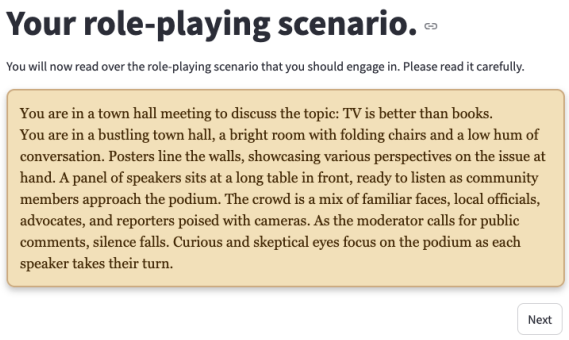


Figure 5: Description of the role-playing scenario in our annotation study.

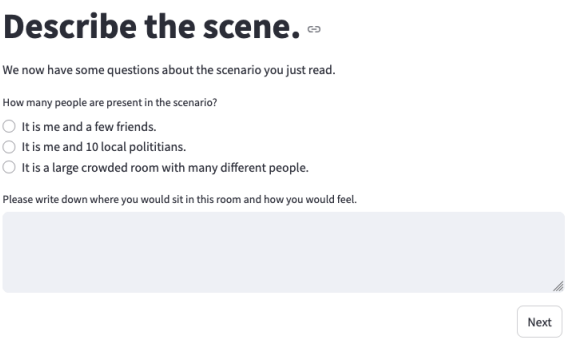
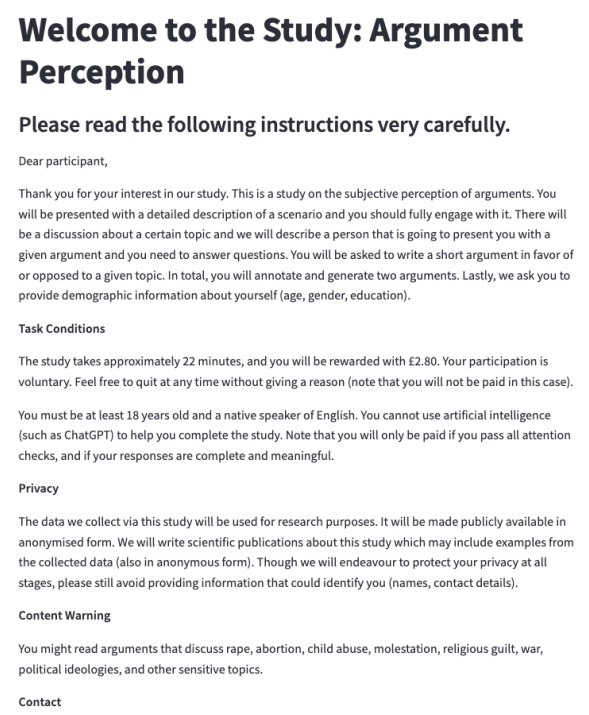


Figure 6: Involement question for the role-playing scenario in our annotation study.

questions that help the annotator engage with the argument can be seen in Figure 6.



Welcome to the Study: Argument Perception

Please read the following instructions very carefully.

Dear participant,

Thank you for your interest in our study. This is a study on the subjective perception of arguments. You will be presented with a detailed description of a scenario and you should fully engage with it. There will be a discussion about a certain topic and we will describe a person that is going to present you with a given argument and you need to answer questions. You will be asked to write a short argument in favor of or opposed to a given topic. In total, you will annotate and generate two arguments. Lastly, we ask you to provide demographic information about yourself (age, gender, education).

Task Conditions

The study takes approximately 22 minutes, and you will be rewarded with £2.80. Your participation is voluntary. Feel free to quit at any time without giving a reason (note that you will not be paid in this case).

You must be at least 18 years old and a native speaker of English. You cannot use artificial intelligence (such as ChatGPT) to help you complete the study. Note that you will only be paid if you pass all attention checks, and if your responses are complete and meaningful.

Privacy

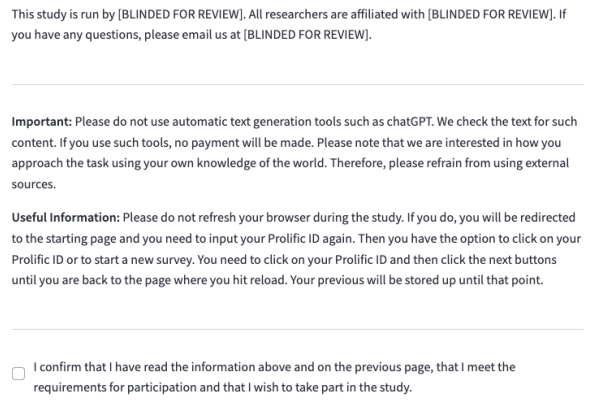
The data we collect via this study will be used for research purposes. It will be made publicly available in anonymised form. We will write scientific publications about this study which may include examples from the collected data (also in anonymous form). Though we will endeavour to protect your privacy at all stages, please still avoid providing information that could identify you (names, contact details).

Content Warning

You might read arguments that discuss rape, abortion, child abuse, molestation, religious guilt, war, political ideologies, and other sensitive topics.

Contact

Figure 7: Screenshot of our study, Page No. 1



This study is run by [BLINDED FOR REVIEW]. All researchers are affiliated with [BLINDED FOR REVIEW]. If you have any questions, please email us at [BLINDED FOR REVIEW].

Important: Please do not use automatic text generation tools such as chatGPT. We check the text for such content. If you use such tools, no payment will be made. Please note that we are interested in how you approach the task using your own knowledge of the world. Therefore, please refrain from using external sources.

Useful Information: Please do not refresh your browser during the study. If you do, you will be redirected to the starting page and you need to input your Prolific ID again. Then you have the option to click on your Prolific ID or to start a new survey. You need to click on your Prolific ID and then click the next buttons until you are back to the page where you hit reload. Your previous will be stored up until that point.

☐ I confirm that I have read the information above and on the previous page, that I meet the requirements for participation and that I wish to take part in the study.

Figure 8: Screenshot of our study, Page No. 2

F. Study Design

The following section shows the screenshots of out study as described in Section 4.3. After Figure 26 the study loops back to Figure 14 to display the second argument.

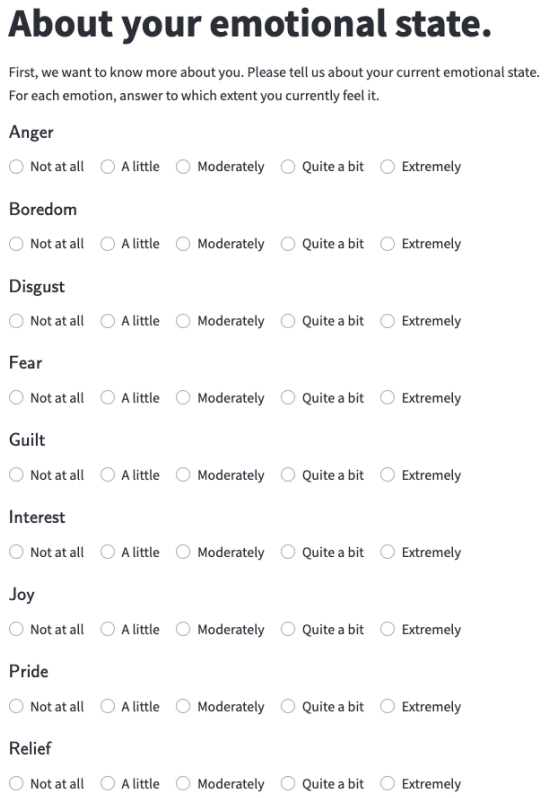


Figure 9: Screenshot of our study, Page No. 3

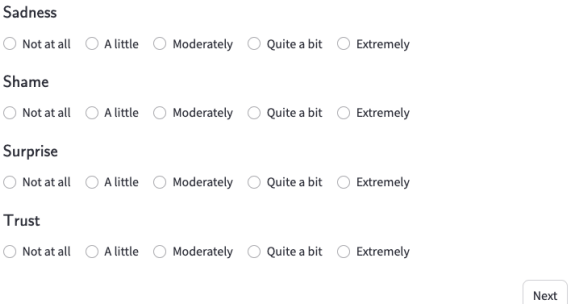


Figure 10: Screenshot of our study, Page No. 4

Figure 11: Screenshot of our study, Page No. 5

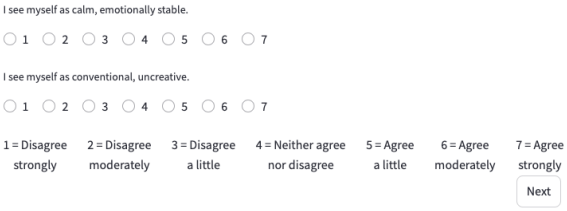


Figure 12: Screenshot of our study, Page No. 6

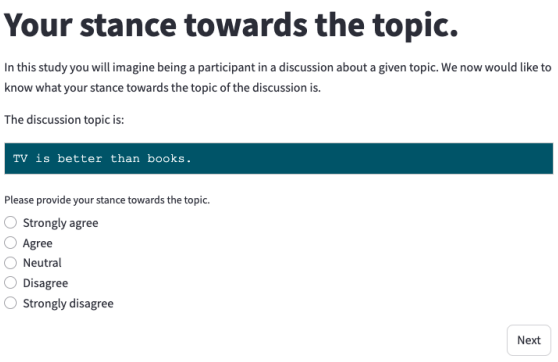


Figure 13: Screenshot of our study, Page No. 7

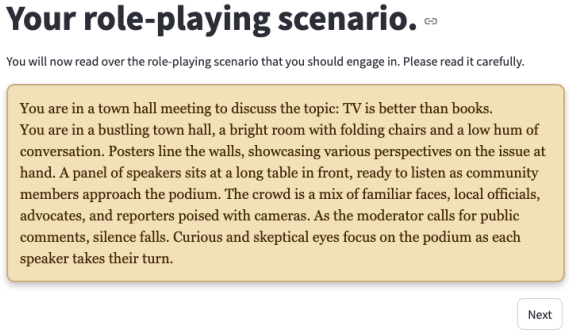


Figure 14: Screenshot of our study, Page No. 8

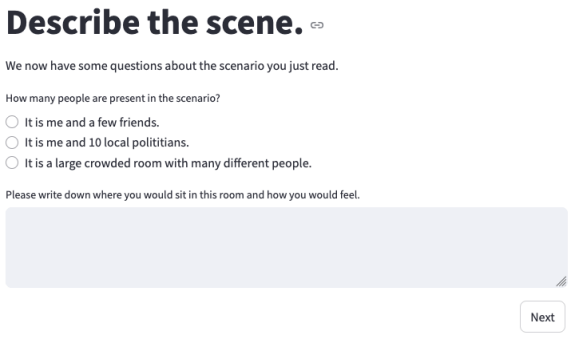


Figure 15: Screenshot of our study, Page No. 9

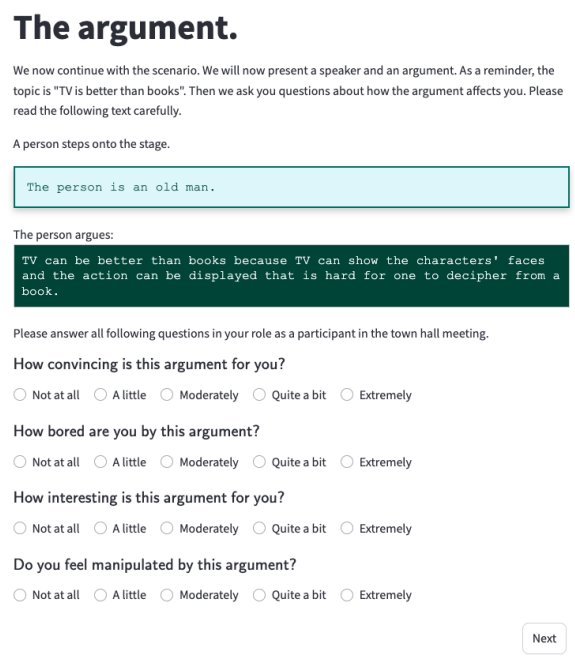


Figure 16: Screenshot of our study, Page No. 10

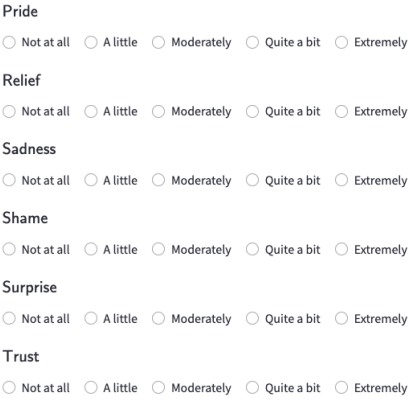


Figure 17: Screenshot of our study, Page No. 11

Figure 18: Screenshot of our study, Page No. 12

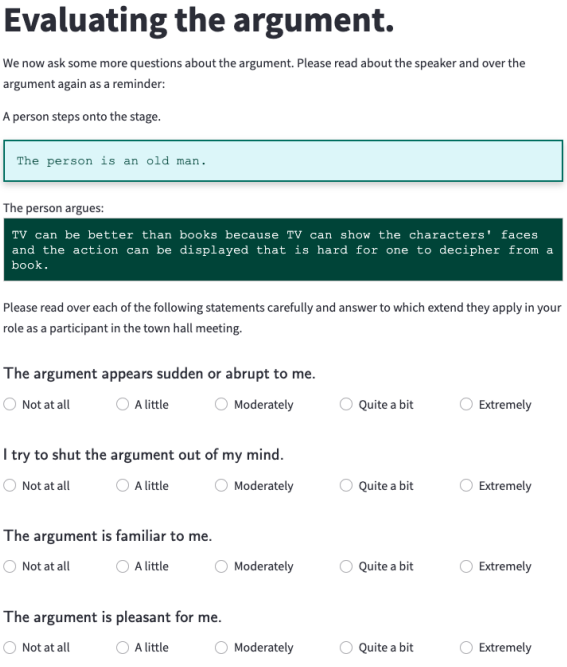


Figure 19: Screenshot of our study, Page No. 13

Figure 20: Screenshot of our study, Page No. 14

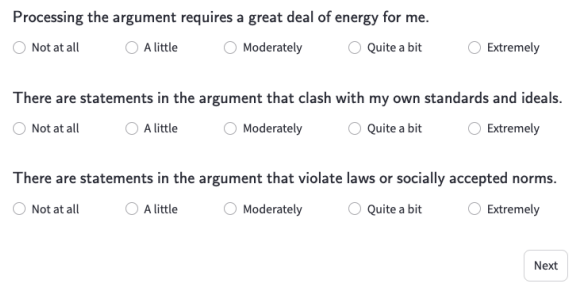


Figure 21: Screenshot of our study, Page No. 15

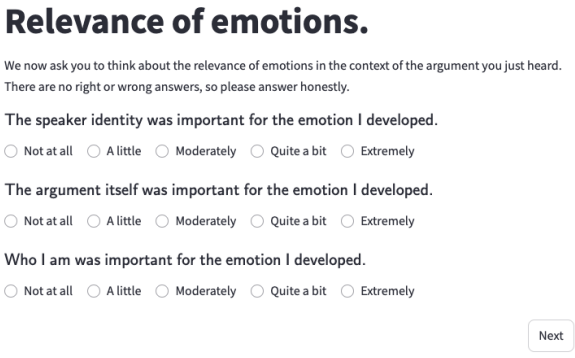


Figure 22: Screenshot of our study, Page No. 16

Figure 23: Screenshot of our study, Page No. 17

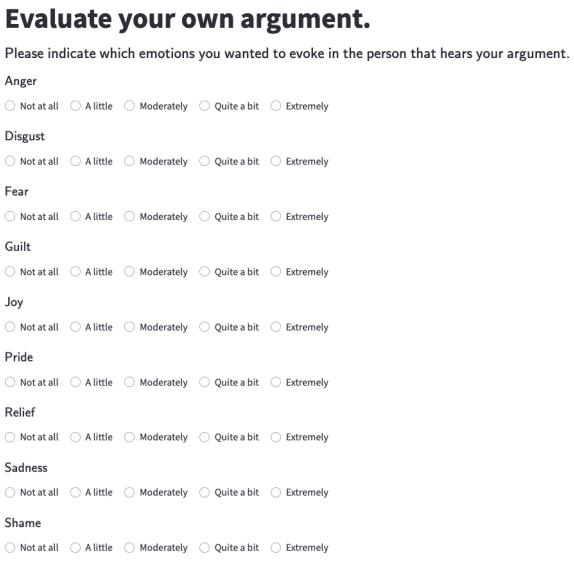


Figure 24: Screenshot of our study, Page No. 18

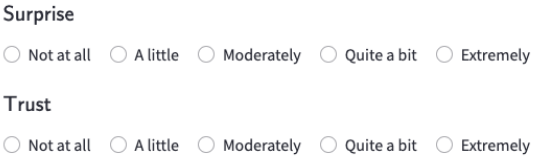


Figure 25: Screenshot of our study, Page No. 19

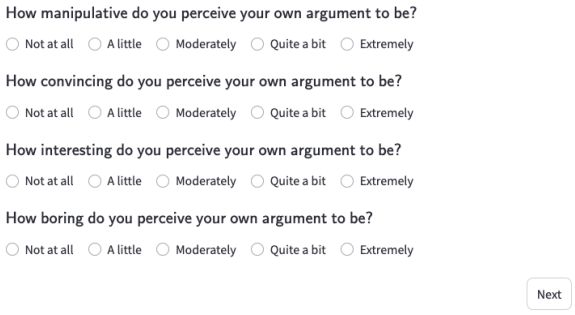


Figure 26: Screenshot of our study, Page No. 20

Figure 27: Screenshot of our study, Page No. 21

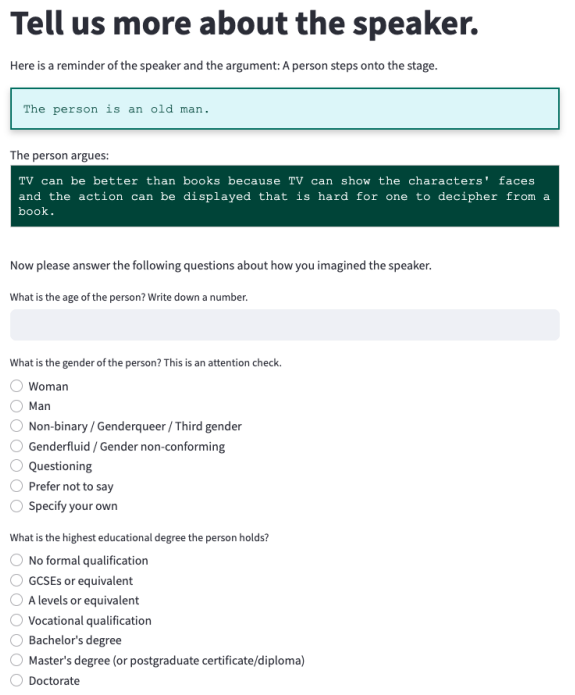


Figure 28: Screenshot of our study, Page No. 22

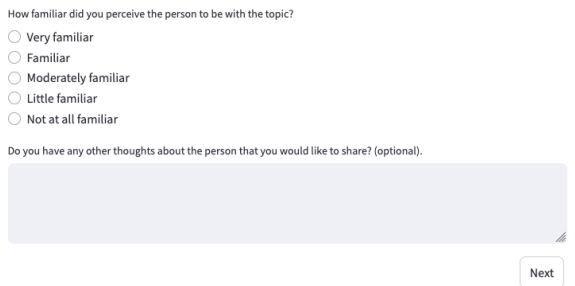


Figure 29: Screenshot of our study, Page No. 23

Figure 30: Screenshot of our study, Page No. 24

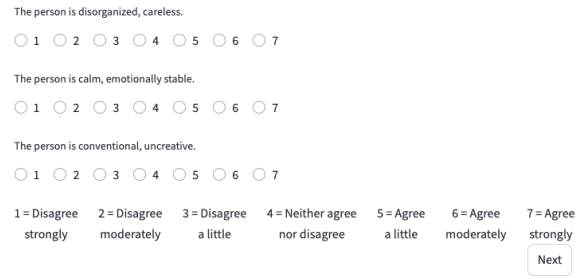


Figure 31: Screenshot of our study, Page No. 25

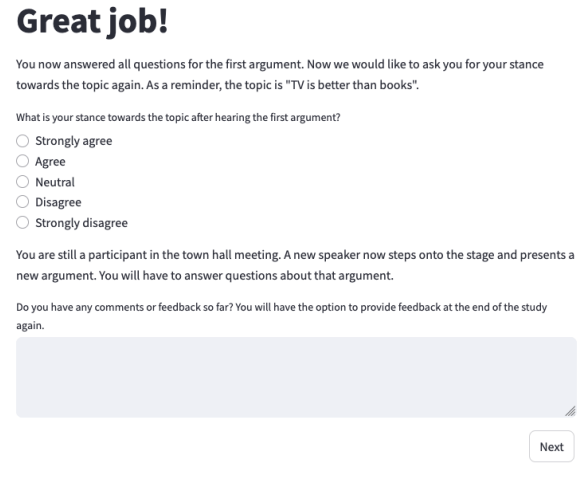


Figure 32: Screenshot of our study, Page No. 26

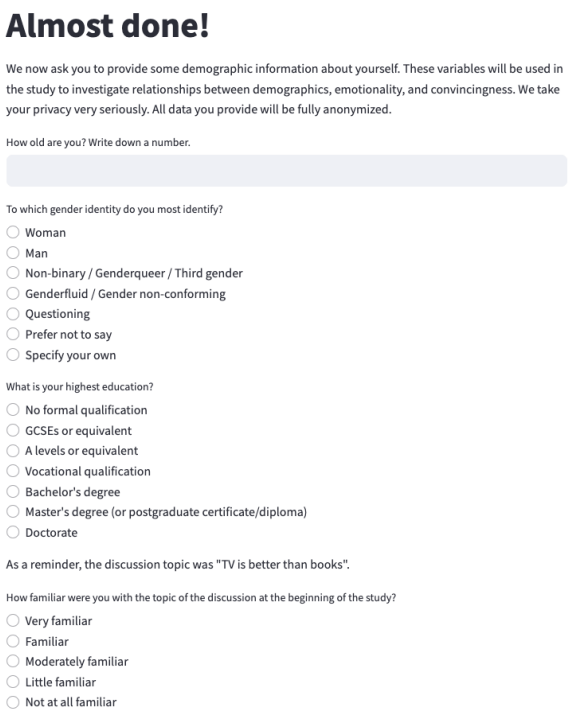


Figure 33: Screenshot of our study, Page No. 27

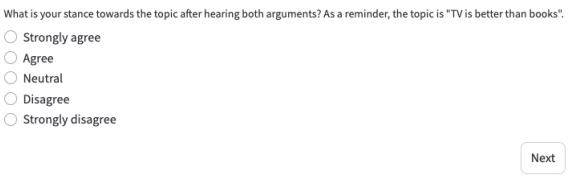


Figure 34: Screenshot of our study, Page No. 28

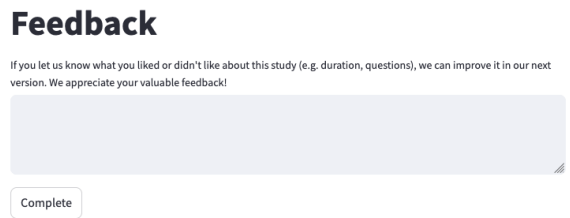


Figure 35: Screenshot of our study, Page No. 29