# Exploring the Role of Emotion and Sentiment in Online Persuasion

**Kayson Hansen**
Stanford University
kayson@stanford.edu

**Yixing Wang**
Stanford University
yixingw@stanford.edu

**Asef Islam**
Stanford University
ai25@stanford.edu

## Abstract

Persuasion plays a pivotal role in human communication, and understanding the factors that contribute to the success of persuasive arguments is of great importance. In this project, we investigate the influence of emotion and sentiment on argument persuasiveness in online discussions. Leveraging the ChangeMyView Corpus from Reddit, we explore the interplay between emotional cues, sentiment, and the strength of persuasive appeals. We first classified emotion and sentiment features using fine-tuned transformer models. Then, we tried to use our learned features to predict the persuasiveness of each challenger's reply. We later trained 3 other classifiers for each model type, one using emotion alone as features, one using sentiment alone, and one using both emotion and sentiment. Surprisingly, our experiments reveal that none of the models utilizing emotion or sentiment features outperform a baseline model based solely on the length of the reply. We conclude that reply length is a stronger predictor of persuasiveness than emotional cues, suggesting that longer replies are more likely to change the original poster's view.

## 1 Introduction

The art of persuasion has forever been at the heart of human communication. Whether in political debates, advertising campaigns, or everyday conversations, individuals strive to convince others of their viewpoints, ideas, or products. While traditional methods of persuasion often rely on logical reasoning and factual evidence, recent research has unveiled the remarkable influence of emotion and sentiment in shaping the success of persuasive arguments. Understanding the interplay between emotional expression, sentiment, and argument persuasiveness has immense practical implications, ranging from marketing strategies to political discourse.

Natural language processing (NLP) has emerged as a powerful tool to dissect and analyze various aspects of human language, enabling researchers to gain deeper insights into the complexities of persuasive communication. By leveraging computational techniques, NLP offers a unique opportunity to delve into vast quantities of text data, uncover hidden patterns, and extract meaningful information regarding emotional cues and sentiment.

This paper aims to explore the role of emotion and sentiment in the persuasiveness of arguments using advanced NLP methodologies. We will delve into the existing literature on argumentation and persuasive discourse, highlighting the key findings that have established emotion and sentiment as significant factors in the effectiveness of persuasive communication. Building upon this foundation, we will present our research framework, which integrates sentiment analysis and emotion detection to systematically investigate the impact of emotion and sentiment on argument persuasiveness.

By conducting extensive experiments on the r/changemyview dataset, we aim to unveil the underlying mechanisms through which emotional cues and sentiment contribute to the overall persuasiveness of arguments. The hypotheses for this project center on the relationship between the emotion of a comment and its persuasiveness in changing an original poster's view. The core hypothesis is that there exists a correlation between the emotion expressed in the comments and the likelihood of successfully altering the OP's stance. We determine the persuasiveness of a comment based on its delta character count. We hypothesize that comments with emotions that are generally regarded as positive will have more persuasiveness – and thus, more delta counts – than those with neutral and negative emotions. We group the emotions into different sentiment categories and also hypothesize that comments with positive sentiment are more persuasive. Lastly, we postulate that the strength of the sentiment used in the comments will play a role in influencing persuasiveness. By analyzing the sentiment distribution, sentiment strength, and delta feature correlation, we aim to validate these hypotheses and gain insights into the relationship between emotion and the persuasive impact of comments in the ChangeMyView Corpus from Reddit.

## 2 Prior Literature

In our literature review, first we will discuss the previous findings from the r/changemyview dataset, then we will investigate findings on emotion and sentiment from other papers in the field more broadly.

To begin with, the original authors of the r/changemyview dataset did a number of analyses on the data (Tan et al., 2016). They looked at what led to arguments being more persuasive, how interplay between original posts and replies worked, and what features led to more open-mindedness in original posts.

The first linguistic factors the authors examined in determining what led to more persuasive arguments were features that captured the interplay between the original post and the replies. Through features like number of common words, Jaccard fraction, OP fraction, and reply fraction (the latter two look at words in common but normalize by length) the authors found that replies that were more dissimilar in wording to the original post were more likely to be awarded a delta.

Next, the authors examined features that were specific to only the replies, not the original posts. They looked at number of words, number of paragraphs, and number of sentences, and found that across all metrics, longer replies were more likely to be successful. They also looked at word category-based features (positive words, negative words, arguer-relevant personal pronouns, links, hedging, examples, question marks, and quotations), producing a number of findings, including that persuasive arguments use more personal pronouns, cite more links as evidence, include more examples, and use more hedging. Finally, the authors also considered word score-based features, which they did using four scalar word-level attributes: arousal (emotional intensity), concreteness (the degree to which a word denotes something perceptible), dominance (degree of control expressed by a word), and valence (how pleasant a word's denotation is). They found no significant effects for concreteness and dominance, but found negative effects for both arousal and valence (so calmer, less pleasant language is more persuasive).

The authors trained logistic regression models to see what features were the most predictive of persuasiveness. They chose bag-of-words features, part-of-speech tags, interplay (features that captured the interplay between the original post and the replies), and style (features that were specific to only the replies). Using the number of words in a reply as a baseline metric ($\sim 66\%$ accuracy), they found that interplay and style both outperformed the baseline (while bag-of-words and part-of-speech either performed equally as good or worse), and specifically, interplay led to a $\sim 5\%$ improvement in accuracy.

Finally, the authors examined what features led to OPs being more or less open-minded. Interestingly, humans do not perform well on such a task, achieving only 50% accuracy on such a test in a pilot study. The authors employed the same set of features as above, and this time, bag-of-words, part-of-speech, and style features all significantly outperformed the baseline, number of words (by 2-3%). Specifically, using the bag-of-words features, they found that resistant views tend to be expressed using more decisive words such as anyone, certain, ever, nothing, and wrong, while help and please turned out to be more malleable words. Through the part-of-speech features, the authors found that comparative adjectives and adverbs are signs of malleability, while superlative adjectives suggest stubbornness.

Moving onto other papers in the field that deal with how emotion and sentiment affect persuasiveness, the next paper we will discuss is from a team at the USC Institute for Creative Technologies (Chatterjee et al., 2014). The authors hypothesize that both lexical features and markers of hesitation are predictors of persuasive power in online multimedia content. They also wanted to test whether the sentiment polarity of a movie review can help in the prediction of the persuasiveness of the speaker. Examples of verbal behaviors include a speaker's stuttering or having pauses in their speech and filler words such as "um" and "uh". They use a corpus of 1000 movie review videos from ExpoTV.com to test their hypotheses. Each of the reviews in the dataset consists of a video of the speaker talking about a particular movie as well as an integer rating of the movie ranging from 1 to 5 stars. Thus, the rating is indicative of persuasion in that a speaker who rates a movie 5 stars would be trying to persuade the audience to watch the movie whereas a speaker who rates a movie 1 star would be trying to persuade the audience against watching the movie. The polarity is also indicated by the star rating as a 5-star review indicates a positive rating whereas a 1- or 2-star review indicates a negative rating. The authors chose to score persuasiveness using a subjective rating system by evaluation using Amazon Mechanical Turk in which for each video in the corpus 3 crowdsourced workers were asked to evaluate the persuasiveness of the speaker on a scale of 1 to 7. Then, a binary threshold was applied where scores greater than or equal to 5.5 were considered to be persuasive and those less than or equal to 2.5 were considered to be unpersuasive. The authors performed transcriptions of the videos and then used a bag-of-words representation which resulted in around 4500 unigrams and 24000 bigrams. They looked at 4 paraverbal descriptors of hesitations: pause-fillers such as "um" and "uh", disfluency markers such as stuttering, articulation rate i.e. rate of speaking, and the mean span of silence. Using these features, they performed classification experiments in which they found the features as well as sentiment to be predictive of persuasiveness.

The next paper we will discuss is by Hong et al., (Hong et al., 2020), which researches the factors that influence how persuasive online reviews are. They utilize data collected from the JD website, one of the most popular e-commerce platforms in China; specifically, they scraped approximately 10,000 reviews from six major brands of phones. They hypothesized that pathos, logos, ethos, and feature statements (when reviewers describe the product or service features) all contribute to the persuasiveness of online reviews. In order to determine which words fit under each category, Hong et al. randomly selected 100 reviews, and manually classified words as falling under each category. They also used Python code to include synonyms for the words they found, eventually creating a vocabulary with over 125,000 words in it. The authors then conducted a Tobit regression analysis to determine what features affected the persuasiveness of online reviews. Using the vocabulary they created, they were able to use the keyword frequency of pathos, logos, ethos, and feature statements as features, as well as brand, picture, video, membership, and rating. From the regression analysis, Hong et al. discovered that pathos, logos, and feature statements all have a significant effect on facilitating the persuasiveness of online reviews; however, ethos didn't have a significant effect. They also found that picture, video, and membership also promoted review persuasiveness, while brand and rating have no significant promotion effect.

Samad et al 2022, from a group of researchers at IIT Patna, (Samad et al., 2022) takes an approach to model empathy within conversations using a fine-tuned MLE-based language model in an RL-based framework where a conversational agent is rewarded for demonstrating acts of certain emotions which are denoted as being empathetic. they used an existing dataset called PERSUASIONFORGOOD which they extended and annotated for their work. They manually annotated this dataset with 23 different emotions. Then, they fine-tuned pretrained transformer models to be able to recognize and classify persuasion within this dataset. The RL-based dialogue generation agent is driven by two generic and two task-specific rewards based on fluency, non-repetition, empathy, and persuasiveness. This is then compared to a state-of-the-art model called ARDM and is shown to generate a response that is more consistent, fluent, empathetic, and persuasive. They developed a policy and then used a proximal policy optimization process to generate sequences which they scored based on evaluation metrics based on the aforementioned rewards. They found that the empathetic factor led to engaging users more in the dialogue. Specifically, they referred to scores in the two task-specific categories related to persuasiveness and empathy, in which their model achieved scores of 3.91 and 3.51 specifically, which outperformed the model which they compared to. In addition, they measured performance in a task in which the model attempted to convince 20 users to donate and found that the model was able to successfully convince 68 percent of users to do so. The authors conclude that current state-of-the-art dialogue agents are lacking in their ability to generate empathetic responses and that generative models can use an RL-based framework to reward more empathetic responses to better connect with end users.

## 3 Data

Our project will use the ChangeMyView dataset from the (Tan et al., 2016) paper. This dataset includes conversations from the "r/changemyview" subreddit on the Reddit website. The mechanics of the site are as follows. Users that “accept that they may be wrong or want help changing their view” submit original posts, and readers are invited to argue for the other side. The original posters (OPs) explicitly recognize arguments that succeed in changing their view by replying with the delta ($\Delta$) character and including “an explanation as to why and how” their view changed. A Reddit bot called the DeltaBot confirms deltas and maintains a leaderboard of per-user $\Delta$ counts. This dataset is inordinately useful for researching the persuasiveness of arguments, because the "deltas" serve as explicit persuasion labels that are provided by the actual participants and at the fine-grained level of individual arguments, as opposed to mere indications that the OP’s view was changed.

In this project, we use the version of the dataset found here. This version of the ChangeMyView dataset converts the data into the "ConvoKit" format, allowing for direct download using pip (a Python package manager), and easy access to important metadata attributes.

In the original paper introducing the ChangeMyView dataset, the authors did a number of analyses on the data. First, they found that the earlier a reply was, the more likely it was to successfully change the OP’s mind and be awarded a delta ($\sim 4.5\%$ for the first reply vs $\sim 1.5\%$ for the tenth reply). Next, they found that as the number of people replying to an OP increased, so did the likelihood of the OP changing their mind and thus awarding a delta to someone. This increase was quite significant, going from $\sim 13\%$ for 18 challengers to $\sim 40\%$ for 40 challengers.

The authors went to extensive lengths to control for alternative explanations when researching linguistic factors and how they affect persuasiveness. They looked at posts where the OP awarded a delta, and then considered all the replies in the thread by the succcessful challenger. They then computed the Jaccard similarity between the set of words in all the replies by the successful challenger and all the other commenters to find the most similar thread to the successful one. This way, the authors could compare two lexically similar sets of replies to a post, one with a delta and one without a delta, in order to identify what factors contribute to persuasiveness.

In this paper, instead of looking at pairs of comments that were lexically similar where one had a delta and one didn’t like the original authors, we consider original posts and comments as a whole. Using the deltas awarded to replies as outputs indicating whether or not they were persuasive, we train machine learning models to try to predict reply persuasiveness using emotion and sentiment.

## 4 Model

For our emotion classification of the original posts and replies, we used a model from Hugging Face called Emotion English DistilRoBERTa-base by Jochen Hartman (Hartmann, 2022), which is a finetuned checkpoint of DistilRoBERTa-base. This model can be used to classify utterances in English text into scores of 7 different emotions: anger, disgust, fear, joy, neutral, sadness, and surprise. This model is trained on 6 datasets of English text from diverse sources including Twitter, Reddit, student reports, and TV dialogue. In total there are 2,811 observations per emotion and nearly 20k observations in total that went into training this model, 80 percent of which was used for training and 20 percent for evaluation, with an evaluation accuracy of 66 percent. In classifying the sentiment of the replies, we used another model from Hugging Face called Twitter-roBERTa-base-sentiment which is tuned on 58 million tweets for sentiment analysis with the TweetEval benchmark and gives a sentiment label of 0 for negative, 1 for neutral or 2 for positive. Finally, we used a number of different machine learning models from scikit-learn in order to predict persuasiveness from these emotion and sentiment features, namely logistic regression, multi-layer perceptron (MLP), support vector machine (SVM), decision trees, and Adaboost.

## 5 Methods

To start off, there were some pre-processing steps that were necessary in order to get the data ready for analysis. The data is represented as a sequence of utterance IDs along with associated metadata. We separated utterances into 5 different lists: the original posts, challenger replies to the original post that were successful in changing the original poster’s view, challenger replies that were unsuccessful, original poster’s replies to challengers in which the challenger was successful, and original poster’s replies to challengers in which the challenger was unsuccessful. Given the ID for an utterance, we are able to determine whether it is the original post based on whether the ID matches the

| Model | Number of Words | Emotion | Sentiment | Emotion + Sentiment |
|---|---|---|---|---|
| Logistic Regression | 0.5486 | 0.3942 | 0.3536 | 0.4066 |
| MLP | 0.5634 | 0.3925 | 0.3536 | 0.4386 |
| SVM | 0.5498 | 0.4003 | 0.3536 | 0.3895 |
| Decision Trees | 0.5232 | 0.5081 | 0.3536 | 0.5117 |
| Adaboost | 0.5514 | 0.4551 | 0.3536 | 0.4575 |

Table 1: Macro-averaged F1 scores for all experiments.

ID of the "root" field within the metadata. Therefore, any utterances whose ID does not match the root ID can be identified as a reply and not an original post. Success in changing the original poster's view can be determined based on the "success" field within the metadata. There is also a speaker ID field that can be used to determine whether the author of a post is the original poster or a challenger. After organizing the utterances according to whether they were an original post or a reply, whether they belonged to a thread in which the original poster's view was successfully changed or not and whether the speaker was the original poster or a challenger, we then proceeded to run the Hugging Face transformer models to obtain the emotion and sentiment scores for all of the utterances. Because there were a total of 22,765 utterances this was a computationally intensive process and took several hours to run. In order to prevent the loss of this data from the Colab runtime once the process was completed we used the pickle module to store the sentiment and emotion features in external data files which we could reload easily for analysis.

We began with some statistical analyses of calculating the average emotion and sentiment for original posts, as well as for successful and unsuccessful replies and original poster replies to successful and unsuccessful replies to see whether any noticeable differences would arise. Then, we proceeded to the machine learning portion of our project, in which we used the emotion and sentiment features to try and predict the persuasiveness of each challenger's reply, i.e. whether or not they were successful in changing the original poster's view. For each type of classification model that we tested, we first trained a classifier as a baseline model that would predict persuasiveness using only the length of the reply in words as a feature. This was because this was something tested by the original authors of the dataset and we wanted to see whether using emotion and sentiment would surpass this baseline. We then trained 3 other classifiers for each model type, one using emotion alone as features, one using sentiment alone, and one using both emotion and sentiment. We obtained precision, recall, and F1 scores for each of these experiments to be compared in the results section.

## 6 Results

Table 1 summarizes the macro-averaged F1 scores across all of the different models that we trained using either number of words in the reply as a feature: emotion, sentiment, or both emotion and sentiment.

Figures 1 and 2 on the right summarize the average scores in each of the seven emotion categories, splitting into successful and unsuccessful groups. Figure 1 corresponds to data in challenger replies and Figure 2 to replies to challenger replies. Across both successful and unsuccessful groups and the same emotion category, the emotion scores sum up to 1.

## 7 Analysis

None of our models using emotion or sentiment features outperformed the baseline model of using the number of words as an estimator. This is likely because the length of a reply is actually a fairly good predictor of whether it is likely to be persuasive or not, whereas expressing more emotion is not necessarily predictive of more persuasiveness. Performance using sentiment as a feature alone was poor and the same across all models. There were modest gains in performance from adding sentiment to emotion as features for logistic regression, MLP, decision trees, and Adaboost but not for SVMS, and in the case of decision trees performance almost reached the level of the baseline model. In investigating further why the F1 score happened to be the same for all of the models in the case of using sentiment alone as a predictor, we discovered that the recall in those cases for the negative class is always 0, and thus the model is only predicting persuasive replies, and it is unclear

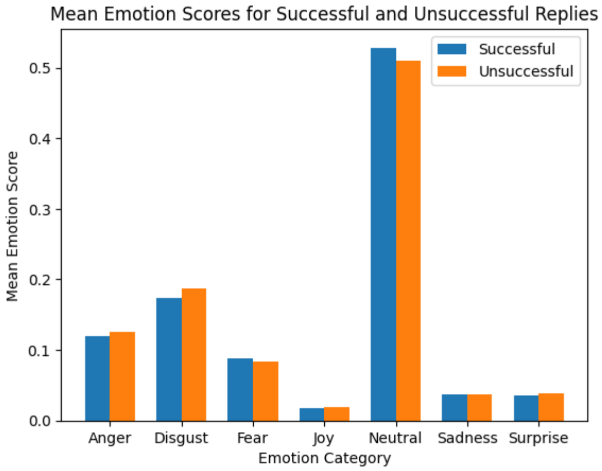

Figure 1: Average scores for replies in seven emotion categories: anger, disgust, fear, joy, neutral, sadness, surprise

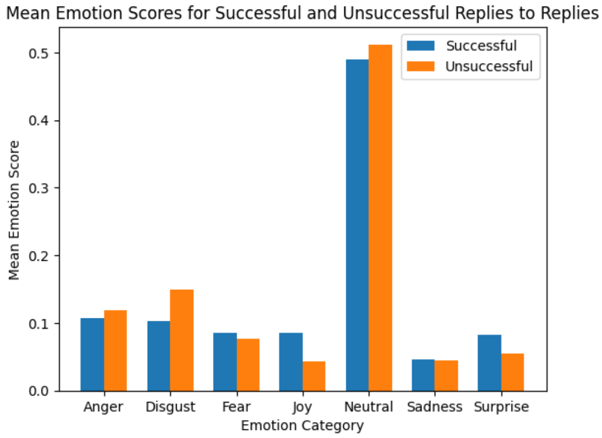

Figure 2: Average scores for replies to replies in seven emotion categories: anger, disgust, fear, joy, neutral, sadness, surprise

to us why this result was obtained.

From our statistical analyses of the datasets, we realized that there isn't a significant difference between emotion scores received for the successful versus unsuccessful replies. This explains why our models utilizing emotion features did not outperform the baseline. In line with our initial hypothesis, we expected that comments featuring emotions generally perceived as positive will exhibit greater persuasiveness, resulting in higher delta counts compared to comments with neutral or negative emotions. Despite this, our analysis revealed that the emotion scores received by successful and unsuccessful replies were not significantly different. Therefore, relying solely on emotion as a predictor proved ineffective in distinguishing persuasive responses.

Ultimately, our findings highlight the significance of reply length as a predictor of persuasiveness, as shown in the original paper. The number of words consistently outperformed models utilizing emotion or sentiment features. This suggests that the length of a reply serves as a reliable indicator of its persuasiveness, while the expression of emotions alone does not necessarily correlate with persuasive efficacy.

## 8 Conclusion

Contrary to our initial hypothesis, none of our models utilizing emotion or sentiment features outperformed the baseline model that relied solely on the length of the reply in words. We validated that the length of a reply emerged as a strong predictor of persuasiveness, whereas expressing more emotion did not necessarily lead to increased persuasiveness. The performance of sentiment features alone was consistently poor across all models, with the recall for the negative class consistently being 0, indicating that the models were only predicting persuasive replies. The reason behind this result remains unclear and warrants further investigation. Statistical analyses also revealed that there was no significant difference in emotion scores between successful and unsuccessful replies. This suggests that the expression of emotions alone does not play a decisive role in persuading the original poster to change its view.

Although our initial hypotheses regarding the relationship between emotion, sentiment, and persuasiveness were not fully supported, our study contributes valuable insights to the understanding of persuasive communication strategies. By highlighting the importance of reply length and the limitations of relying solely on emotion or sentiment features, our findings can inform future research in designing more effective persuasive techniques.

Moving forward, it would be beneficial to explore additional factors that may contribute to persuasive arguments, such as the structure of the argument, the use of evidence, or the rhetorical strategies employed. By incorporating these factors alongside emotion and sentiment, we can gain a more comprehensive understanding of the dynamics behind persuasive communication.

In conclusion, our study underscores the complex nature of persuasive arguments and the need for a multifaceted approach when examining the factors that influence persuasiveness. Through continued research and investigation, we can uncover new insights and develop more nuanced models for understanding and enhancing persuasive communication.

## Acknowledgements

We would like to express our special thanks for the instructions of Professor Christopher Potts and Siyan Li who inspired and advised on this project.

## Known Project Limitations

The limitation of our project includes a drawback in the transformer models we used to classify the sentiment and emotion of the utterances in the r/changemyview dataset. These DistilRoBERTa and RoBERTa-based models have a maximum length of 512 tokens for input texts, but a significant number of the posts and comments in the dataset are longer than 512 tokens. Thus, we used the "truncate" option on the transformer models to truncate the texts, then classify their sentiment or emotion, but this is clearly a flawed approach because we were likely missing out on lots of very important text that could've affected the sentiment and emotion of the posts. This likely led to our machine learning models using emotion and sentiment as inputs performing less well than they could've if we didn't have to truncate the texts. Also, the dataset itself may have inherent biases, as it consists of user-generated content from a specific online community. This can limit the generalizability of the findings to other contexts or platforms. The emotion classification model used in the study categorizes utterances into a limited set of emotions. This oversimplification may fail to capture the nuances and complexities of emotional expression, potentially leading to incomplete or inaccurate emotion analysis. What's more, we determined the persuasiveness of a reply by looking at the delta count. Since the delta count is an indicator variable, it is not indicative of the nuanced strength of the persuasiveness. Lastly, the choice of machine learning models used for prediction may impact the results. Different models have varying strengths and weaknesses, and the selected models may not be optimal for capturing the complexities of persuasive arguments.

## Authorship Statement

Kayson wrote the introduction, prior literature, and data sections. He also wrote the code for preprocessing the data. Additionally, he wrote the code to import the transformer models to classify the emotion and sentiment of utterances, ultimately storing the results in pickle files. He programmed to train logistic regression models that used the number of words as a feature to get a baseline score for comparison. Finally, he trained multi-layer perceptron models to classify the persuasiveness of replies based on sentiment, emotion, and both sentiment and emotion.

Yixing wrote the abstract, results, and analysis sections, and served as the proofreader for the paper. Additionally, she performed statistical analysis on the datasets by comparing the emotion scores between the two groups of replies, from which she made visualizations.

Asef wrote the model and methods sections and ran code to obtain the classification results of SVMs, Decision Trees and Adaboost. He also gathered and tabulated the results for all of the classification models and analyzed them.

## References

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.

Jochen Hartmann. 2022. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`.

Wei Hong, Zemin Yu, Linhai Wu, and Xujin Pu. 2020. Influencing factors of the persuasiveness of online reviews considering persuasion methods. *Electronic Commerce Research and Applications*, 39:100912.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States. 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. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.