Abstract

The study of rhetoric has been a favorite occupation of humanity since the time of ancient Greece. We study the application of appeals to emotion (*pathos*) as a rhetorical strategy by building on the work of Tan et al (2016) to conclude that evoking emotion is a powerful rhetorical strategy. Tan et al (2016) conclude that (1) negative sentiment is more persuasive than positive sentiment and (2) the topic of discourse has less effect on persuasion than rhetorical strategies. The original paper uses Reddit's *ChangeMyView* community for its data source, and we obtain additional data over a 3 year period, from 3.5 years ago to 0.5 years ago as of March 11, 2022. Our focus is on understanding strategies for how one might *evoke emotions* during an argument. We limit our study to language use given that our data is from a text-based medium. We test persuasive efficacy using emotional appeals to fear, confidence, belligerence, qualification, flattery, and excitement.

Theoretical Underpinnings

Our interest in persuasion led us to cognitive theories on persuasiveness discussed by Cameron (2009) and DeMers (2016). Such cognitive theories on persuasion introduced theories on parallel response and subjective utility models. Parallel response theory is centered on *fear appeal* and how people respond to comments which evoke emotions of fear. We test this theory by questioning how fear-based emotion words influence persuasive efficacy. Subjective utility models (i.e., subjective expected utility) center on decision making with regard to the perceived *risk* of making that decision. We investigate the following questions: How do potential consequences influence what choices a decision-maker makes? How does the receptivity of the decision-maker affect the use of *pathos* as a strategy? Is appealing to emotion conducive to a persuasive argument? If so, which appeals (emotions) are effective?

If a persuadee is fearful of public opinion/response based on an opinion they express online, will the persuadee be less likely to express that they have been persuaded? Further, if the persuader makes an argument to the persuadee with belligerence, would the persuadee match the aggression and calcify their opinions, or would their stances degrade and crumble, as well as their convictions?

Balance theory suggests that the effectiveness of persuasion, and how receptive the listener is to alternative viewpoints, relies on three relationships: the persuadee's attitude toward the persuader, the persuader's attitude toward the attitudinal object, and the persuadee's attitudes toward the attitudinal object. The relationship is balanced if all three relationships are positive, or if two are negative. If the relationship is unbalanced, then balance theory suggests this "often motivat[es] one to alter one of the three relationships (Duerr et al). Duerr relates this to 'trustworthiness', and suggests that "a persuader wants the persuadee to have a positive attitude." This suggests that flattery could be used as an effective strategy for persuasion, which is supported by DeMers.

Confidence may also play a role in persuasion. In a study by London, Meldman, and Lanckton, undergraduate students were asked to serve on a jury case. Students received a summary of facts about the case, judge's instructions to the jury, and a legal analysis of the case. Students were paired, and members of each pair were giving competing legal analyses. Students were asked to create a verdict before and after discussing the case in their pair, without knowing that they were given competing legal arguments. Much like our work, the data provides self-determined ratings on how much the persuadee's opinion changed, but on a continuous scale. Also like our paper, confidence and doubt were measured by taking note of confidence-indicating and doubt-indicating expressions. The paper finds that "the single significant behavioral difference between persuaders and persuadees was in the expression of confidence" (p. 182), where persuaders are determined to be those who succeeded in persuading the other party of their viewpoint, and persuadees are determined by those who changed their minds.

As such, we expected emotional elements of back-and-forth exchanges to play a significant role in persuasive efficacy in the *ChangeMyView* community on Reddit.

Data and Methods

Data Filtering

Our initial dataset was a stream of comments in which each set is a sample of 500 comments made within an hour's window. The comments are organized by the thread into posts; each post being a request by the original poster to have their view on a specific topic challenged by responders. In keeping with Tan et al, we organize these posts into trees. The root is the original post.. The rules of the subreddit specify that a commenter must respond to a previous comment in the thread with the goal being to change the original poster's (OP) view on the topic of discourse. If the OP finds a challenger's argument persuasive, they award deltas signifying that their view has been changed. Each OP can award as many deltas as they like, but each comment may get only a single delta. To this data, we apply the following filters.

Post Count	**Record Count**	**Filter Condition Applied**	**Issue Addressed From Previous Tier**
548 K	2.7 M	Dates: comments from 3.5 yrs. ago to 0.5 yrs.	

		ago Limits: max of 500 comments in each hour-long window	
43 K	1.2 M	Organize comments into posts; discard any post with fewer than 10 comments	Posts with few comments are often inactive and pollute the dataset with negative examples.
6.6K	19K	Pair each deltaed comment with a unique un-deltaed comment that is most similar to it within the same post.	~99.5% of comments are without deltas, so the only viable strategy for the classification task is to predict no delta.

Dictionary

We define a *dictionary* as a set of words appealing to a specific emotion. We construct dictionaries related to each of the 6 emotions, fear, belligerence, qualification, flattery, confidence, and excitement. Our process for this involves selecting a number of 'root' words related to each emotion (for instance, some of fear's root words include "nightmare" and "death", and some of confidence's include "always" and "sure"). Then, we find related words by using relatedwords.org, a website that uses WordNet and ConceptNet to find words related to the query. We repeat the process for every emotional category. Some statistics are shown below. It is important to keep track of how common each dictionary is in the corpus. An approach to finding appropriately balanced dictionaries might involve tuning these dictionaries so that each one occupies the same proportion of the corpus. However, we recognize that some emotional appeals may be more frequent than others. We discuss alternative better ways to extract emotional meaning from sequences in the conclusions.

Emotion Dictionary and Corpus Statistics

	Number of Words in	**Number of Tokens in**	**Share of Corpus**

	Vocab	Corpus	
Fear	168	11637	.4%
Belligerence	1110	12325	.4%
Qualification	47	20760	.7%
Flattery	188	26809	.9%
Confidence	223	115372	3.8%
Excitement	21	3118	.1%
Total Corpus	45939	2962316	100%

Process Sample

Fix some emotion for which we have some vocabulary V represented as a dictionary D mapping from words in V to a random index. We process each sample:

1) remove each word w not in D, and
2) drop the sample if it's empty after step 1.

Process Dataset

Then, for the full data set, we

1. split the corpus randomly into training (70%), validation (20%), and test sets (10%).
2. Train BERT on the training set for 5 epochs, generating and saving 5 models. Choose the model with the best validation accuracy for testing.

Baseline: Randomly Sampled Dictionary

We need to use dictionaries of randomly sampled words from the corpus each of comparable size, frequency, and distribution to the emotion-vocabularies as our control. A large factor in model performance is string length; thus correcting for length by sampling strips the effects of length and isolates the effects of appeal to emotion.Given the small size of our vocabulary, order is potentially important thus we choose BERT rather than a bag-of-words model. So we evaluate this control vocabulary the same way we evaluate the ones that are related to emotional appeals running the "*Process Sample*" procedure described above.

Our procedure for generating randomly chosen dictionaries is comparable to top-p sampling, whereby a threshold is set, and we sample words into the dictionary until the dictionary constitutes a certain proportion of the corpus. Words below a certain number of occurrences in the corpus are not eligible for sampling into the dictionary, and the sampling is done uniformly across the vocabulary.

Random Sample Vocabulary Generation

The process has 2 hyperparameters, M denoting the minimum frequency of a token in the full corpus and P denoting the proportion of the corpus P that a given word constitutes. Here is our procedure for generating random dictionaries:

1) Generate a comprehensive vocabulary V of the corpus.
2) Produce $V' = \{v \mid v \in V \wedge \mathit{frequency}(v) \geq M\}$ where $|V'| < 500$.
3) Randomly sample from V' to produce a vocabulary R until $\mathit{sumFrequencies}(R)/|V'| \cong P$.
4) As a validation measure, measure the number of tokens from the training set that are in the randomly sampled dictionary. If this number is within 10,000 of 62904, keep the random dictionary.

Our goal for step 4 is to achieve a number of around **62904** for the number of random dictionary words in the training set, which matches the number or words from the confidence dictionary that are in the training set for confidence. Our assumption here is that larger dictionaries will be more predictive, since more words will be passed to BERT. Therefore, if our random dictionary occupies an equal or larger proportion of the corpus than all of our emotional dictionaries, it should be at least equally as predictive as the emotion dictionaries under the null hypothesis. Generating one random dictionary for each emotional dictionary that matches the frequency of the emotional dictionary was another option, however, we preferred having a uniform baseline for comparisons between dictionaries.

We set the hyperparameter M around 0.019, which is experimentally set to about half of the proportion of our corpus occupied by words in the confidence dictionary. We vary P between 3 and 6, as this allows us to sample from more frequently and less frequently used words. The important measurement of random dictionary quality is done in step 4, which measures the approximate proportion of the corpus being represented by the random dictionary.

The reason we reject randomly generated dictionaries of length 500 or longer is because we want the distribution of random dictionaries to be comparable to the distribution of words from our emotional dictionaries. We do not want dictionaries that only consist of very infrequent words, as the overall frequency of words in the dictionary may be an intervening variable.

We sample uniformly because the emotional dictionaries are not biased towards extremely frequent words. We create 10 randomly generated dictionaries using this procedure, for the sake of getting an accurate baseline.

Here are some statistics from our randomly generated vocabularies. The target for the number of tokens in the training set that are part of the random dictionary is 62904.

Dictionary	**Number of Words**	**Number of Tokens in Respective Training Sets**
Random 1	291	67154
Random 2	373	63229
Random 3	376	63139
Random 4	208	56231
Random 5	411	68852
Random 6	451	67165
Random 7	439	60648
Random 8	478	66466
Random 9	347	65496

Random 10	495	62370

Analysis

Emotion / Vocabulary	**Test Accuracy**	**Z-Score w.r.t Average Random Sample**
Fear	0.72	5.53
Confidence	0.56	-4.13
Belligerence	0.70	4.78
Qualification	0.67	2.50
Flattery	0.67	2.88
Excitement	0.64	1.00
Baseline Average Random Sample	0.627 (with std. error = 0.0159 over 10 runs)	0
Baseline All Tokens in Negatively Sampled Set	0.608	N/A

The results should not be interpreted as saying that certain types of appeals are more effective than others based on their predictivity scores. Rather, the predictivity scores are essentially measuring the how predictive the variations are in deploying a given emotion.

Of the 6 emotions we tested against, fear most often indicates persuasion in *ChangeMyView*. Surprisingly, confidence does not appear to influence persuasion; its test accuracy is 56%, lower than the average test accuracy random-vocabulary baseline of 63%. This is surprising, considering Pulford et al's assertion that "the single significant behavioral difference between persuaders and persuadees [in dyadic interactions] was in the expression of confidence" (Pulford et al, citing London, Meldman, and Lanckton, 1970). One potential reason for this is our assumption that our dictionaries for each fear appeal category should be of the same size. In reality, it is more likely that certain categories are represented by fewer words than others in the English language. In further extensions of the project, we might allow for variable sizes for the dictionaries we constructed that represent each emotion.

Our dictionary of confidence words is worse than a randomly sampled dictionary of approximately the same size and distribution. There are many possible reasons why confidence is not as predictive as we initially thought.

1) Confidence is expressed through non-verbal means. In the study of students in a jury by London, Meldman, and Lanckton, students met face-to-face. Expressions of confidence like body-language, tone, cadence, and gestures might be more important signals than language, and may be more predictive as well.
2) Confidence is more effective when the interaction is private. Another characteristic of the study by London, Meldman, and Lanckton is that the two students who were asked to deliberate were meeting in private. This could potentially amplify expressions of confidence, and perceived intimidation tactics. On a public forum like the changemyview subreddit, the persuadee may feel more protected by the presence of viewers. They are also less likely to be fooled by extreme confidence and 'gaslighting', if other commenters are able to subdue the effect of extremely confident persuaders.
3) Confident language is used both by persuasive and unpersuasive commenters. Perhaps comments that come off as extremely confident also display dogmatism and closed-mindedness, which indicate that they cannot be trusted.

We compare this to our predictivity (accuracy) value for qualification, which is higher. Perhaps maintaining ambiguity and open dialogue, and conceding certain things to the OP is necessary to make oneself trustworthy and credible. Perhaps qualifying one's statement is an indication of listening, and demonstrating understanding increases perceived intelligence. Balance Theory could also play a role here, as it would suggest strengthening the relationship between persuader and persuadee, which could happen through concessions. Maybe qualification actually has the opposite effect - remember that our BERT model does not indicate the direction of the relationship, but just the predictivity of the dictionary. The direction of any given relationship is also likely complex and multifaceted, as we are not using BERT to detect the presence of any given emotion, but rather to make predictions that reason about how it is deployed.

This could explain the high predictivity value for belligerence. This score does not indicate that belligerence is good for persuasion, but rather that, when we analyze comments with belligerence in them, we can use the variance in these comments' deployment of belligerence to make predictions. Excitement does not perform significantly higher than a random baseline, but this could be due to a small vocabulary and vocabulary frequency. The fact that flattery and qualification perform equally well could support the idea that Balance Theory has some predictive power.

There are many reasons why we could see these results, and possible sources of error. First, after we have filtered the corpus by our dictionary, we remove empty strings. This has the effect of removing examples in the corpus where no words from that example were in the dictionary. This alters our sample space to include only comments with words from the dictionary. Effectively, the question we are answering is, given that a certain emotional appeal is attempted, how well does the variation in the ways in which this appeal is deployed predict effectiveness of persuasion.

This design choice was necessary, first because we did not think it useful to attempt to use BERT on empty inputs, but also because it serves as a form of normalization for the differently sized dictionaries.

Confounding Results

Furthermore, a larger dictionary size does not necessarily result in more predictive power. We note a few interesting trends in the results above. First we offer two baselines; one of the randomly sampled set described above. But the latter is the set of all tokens in our negatively sampled set (see description for 19K comments in Data Filtering section above). We see significantly lower accuracy when including all tokens. Moreover, we also see that when all tokens are masked, we see accuracies circa ~70% from BERT based on comment length alone!

Investigating these surprising trends is a hopeful next step, but in the interim we offer the following preliminary explanations. We observed that training with the "All Tokens" baseline takes time. In each redacted emotion-vocab, we train only for 5 epochs with 3 layers. Our maximum token length is a mere 10. In contrast, BERT suffers with such long tokens as we must use for the 'All Tokens' version. More epochs, higher embedding dimension for the hidden layers, and setting the maximum token length 10 100 aren't enough to speed BERT up significantly, nor increase its accuracy. The low accuracy might have more to do with the lower number of layers and epochs relative to size of the lines in the redacted emotion or random sample data. A better baseline might be a longformer model which can handle such lengths.

A second surprise was confidence performing so poorly relative to the random sampled baseline. According to the result, the model would have us conclude that confidence and bluster do not make for persuasive arguments. But much psychological research contradicts this (CITE HERE). Our suspicion here is that bluster is effective in oral and in-person persuasion. Text based persuasion doesn't register these features, and as such, we hypothesize they are less significant than other emotions to an audience. We would have liked to investigate these ideas further if we had the time.

Conclusions, Challenges and Possible Extensions of the Project

Our project measures the predictivity of different kinds of emotional appeals in persuasion in the changemyview context. There are many possible ways to extend our research.

One additional signal that we did not consider is the receptivity score, which we computed but did not use in our final report. For each comment, we include information on the delta score and upvote score. Further, we calculated individual receptivity scores as follows. First, we constructed a dictionary from LIWC by combining sets of words in the "Assent" and "Insight" categories (set union) to indicate receptivity. Let this combined set be called X. Then, for each comment in our data, we tokenized the data and maintained counts for the number of words in that tokenized set that also exist in X. We normalize this count within the bounds [0, 1] to compute the "receptivity score" for a given comment. We then round this score such that each comment has an associated label that exists in {0, 1}. We then divide the comments into 2 batches - those with labels 0 and those with labels 1 - to indicate which comments are described as "non-receptive" and which comments are described as "receptive", respectively. We could imagine running experiments where we predict the receptivity label instead of the deltas, and see how similar they are.

As noted earlier, one possible extension of our project would be to allow for variable sizes for the dictionaries categorizing each emotion we tested against. This would take into account the fact that the number of words that represent different emotions in the English language is not consistent across emotions. Another extension is we may allow for weighted vocabularies, giving greater importance for certain words over others in their effectiveness in evoking a specific emotional response. Finally, we may pre-train the dataset on a multi-category sentiment classification task, to answer the following questions: How do interactions between multiple emotions influence persuasiveness in back-and-forth exchanges? Which combinations of emotional appeals are most effective in persuading others?

Something we learned while we were doing this project was how challenging getting good data can be. We were banned and kicked from several APIs before we found a strategy that worked for collecting data.

This involved running code with delays to avoid triggering alarms. The data took about 30 hours to collect this way.

Works Cited

Cameron, Kenzie. (2009). A practitioner's guide to persuasion: An overview of 15 selected persuasion theories, models and frameworks. Patient education and counseling. 74. 309-17. 10.1016/j.pec.2008.12.003.

DeMers, Jayson. "6 Ways to Persuade Anyone of Anything." Business Insider, Business Insider, 16 July 2016, https://www.businessinsider.com/6-ways-to-persuade-anyone-of-anything-2016-7.

Duerr, Sebastian & Gloor, Peter. (2021). Persuasive Natural Language Generation -- A Literature Review.

London, H., Meldman, P.J., & Lanckton, A.V.C. (1970). The jury method: How the persuader persuades. Public Opinion Quarterly, 34, 171-183.

Pulford, Briony, and Andrew M Colman. "Confidence and Communication in an E-Fit Identification Task Fit Identification Task." Le.ac.uk, University of Leicester, 1 Jan. 2002,

Tan, Chenhao & Niculae, Vlad & Danescu-Niculescu-Mizil, Cristian & Lee, Lillian. (2016). Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. WWW '16: Proceedings of the 25th International Conference on World Wide Web. 10.1145/2872427.2883081.