Literature Review: Framing and Understanding Argumentation and Effective Persuasion

Adrian Apaza
Stanford Graduate School of Business
apaza@stanford.edu
github repo:$https://github.com/adrianapaza/Change_and_Framing$

Abstract

This paper propose to explore a social science phenomena, framing, and relates it to the literature in natural language processing regarding argumentation and persuasion. I discuss work done by sociologists and psychologists on the framing, what it is and what makes for an effective frame. I also discuss computational work done on argumentation primarily with respect to identifying persuasive arguments: methods developed, datasets used, and their results. I identify that further researching emotions is needed based upon past research and analyze a unique dataset

1 Motivational Problem: Framing and Persuading

Understanding how organizations frame themselves, their products, and their actions and engage in persuasion is a key part of the literatures in Sociology, Political Science, Organizational Behavior, and Marketing. When organizations and individuals such as marketing firms, political parties, and social movements frame what they are doing is attempting to make "perceived reality... more salient in a communicating in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation" (Entman 1993). An organization may wish to frame a product as solving a particular problem in its advertising, a social movement may convince you that its policy prescriptions are beneficial to you in some way, or that a politician has diagnosed the correct source of some ailment to the populace.

2 Summaries of Computational Work in Persuasion

The work conducted in natural language understanding and processing by computational linguistics and computer scientists that most closely resembles the phenomenon of framing in the social world is work on argumentation. Research work on argumentation and argumentation mining has touched many areas including identifying claims and premises, identifying valence or position of a piece of discourse, identifying argument similarity, and attempting to predict the efficacy or persuasiveness of an argument. For this literature review I shall primarily focus most on argument persuasiveness as I believe that is most relevant to social science research on framing.

The papers I discuss below are all theoretically motivated to study as an outcome whether some discourse is persuasive or not. Many of them take some level of persuasion as a dependent variable that may be operationalized as indicative of whether the discourse convinced someone to change their mind or act in a measurable way (such as by giving money to a venture). Although others works also identify intermediary mechanisms (such as the rhetorical tactics used, including emotional appeals) that are demonstrably predictive of persuasive discourse.

2.1 Mitra and Gilbert (2014)

One of the earlier works trying to study persuasion online was by Mitra and Gilbert (2014). Their context was Kickstarter and the specific measure they sought to predict was project success. In their work they identified key phrases that were predictive of crowdfunding success and later released these phrases as a public dataset. Their work to identify predictive phrases was inspired by other work in sentiment analysis, but instead of identifying words and phrases indicative of affective states or positive or negative sentiment, they sought phrases that were indicative of being effective at convincing people to fund a Kickstarter campaign.

Their dataset consisted of over 45,000 project descriptions web-scraped from kickstarter.com, about half of which had reached their funding goal and most of which had finished their campaign. Besides project description text, project promised rewards were also scraped. Additionally they gathered data on the project category (e.g. Art, Games, etc.) and whether a video was present, and other non-textual features used when building models to explain project funding success. To build a dataset of phrases, they first identified every unigram bigram and trigram, accumulating over 9,000,000 unique phrases. Those which occurred less than 50 times were dropped.

Their core model utilized penalized logistic regression to identify predictive phrases. The phrases (here counts, akin to a bag of words) and control variables were the variables in Lasso regression models. To prevent overfitting, they also employed ten fold cross validation. They found that in the model with only controls compared to the one with phrases and controls that the cross validation model error rate drops from 17% to 2.4%.

With the phrases that were identified as predictive, the authors then examined and classified them. They utilized the Linguistic Inquiry and Cord Count program to classify the text. They found that predictive phrases fell into linguistic categories of Cognitive Process, Social Process, and Affect (words indicative of positive sentiment are predictive of being more likely to be funded). Additionally they classified (by hand) phrases identified as predictive according to persuasive styles, these styles included reciprocity, scarcity, and authority.

2.2 Habernal and Gurevych (2016a; 2016b)

A pair of studies by Habernal and Gurevych (2016a; 2016b) built up a database of arguments, which included pairwise labels of arguments indicating which one is more persuasive/convincing and annotated descriptions of arguments. The dataset they created consists of online debates from createdebate.com and procon.org. These two websites allow people to propose debate topics, such as whether books or tv is better or whether pornography is wrong, and then people may respond with their own arguments on either side of the debate topic. From these websites they were able to compare arguments on both sides of a debate topic. They then crowdsourced Amazon Mechanical Turkers to evaluate pairs of arguments and determine which one of the two is more persuasive or convincing. They further annotated the data by having the annotators label why an argument is convincing or not. Labeled explanations include "Argument X is not an argument/is only opinion/ is rant", "Argument X is clear, crisp, to the point/ well written", and "Argument X has better credibility / reliability / confidence"; in total there are 19 distinct labels applied.

Habernal and Gurevych (2016a) first sought to build a model to predict which argument, in a pair of arguments, is more convincing. They built two distinct models to compare. Their first model was a support vector machine (SVM). The SVM consisted of a variety of features. The features included unigrams and bigrams, ratio of adjective to adverb endings, a measure of contextuality (i.e. formal vs informal contexts), punctuation usage (exclamations and quotation marks), verb tenses, name entity types, readability measures, sentiment scores, superlative usage, and word counts among some other measures. As we can see this SVM model included both features directly extractable from the text such as the word length or usage of exclamation points and some features that require some further preprocessing such as readability measures. The accuracy of the SVM on the dataset was .78.

The second model was a bidirectional LSTM (BLSTM). For this model they utilized pre-trained word embeddings (*GloVe* specifically). The inputs to this model concatenate both argument embedding representations together. Notably model does not utilize include attention. The BLSTM model has an accuracy 0.76. While this model did perform slightly worse, the authors note that the SVM needed significant data pre-processing. Because of this, the authors draw the conclusion that even traditional models, like SVM, which require intense feature extraction can have their performance matched by simple neural network models.

Habernal and Gurevych (2016b) later sought to predict the annotated explanations/labels of why an argument is or is not convincing/persuasive. Again they utilized BLSTM models, one with an attention mechanism and one without, and they used another SVM model. Again pre-trained *GloVe* embeddings are utilized. The SVM model

uses a similar set of features as their earlier work (see above). However unlike in their earlier work, the BLSTM models were significantly better at predicting appropriate labels to the arguments, though the model with attention performed only marginally better (statistically insignificant) as compared to the BLSTM without attention.

Habernal and Gurevych (2016b) improves the work in measuring persuasiveness by including a labelling schema for argumentation. Although their models were far from perfect (Macro F-1 for the best model was 0.353) they did inspire later work in the field for labeling argumentation.

2.3 Yang et al. (2019)

Yang et al. (2019) built a computational model to identify persuasive strategies at the sentence level in text. They utilized data from Kiva.org; Kiva is a crowdfunding micro-lending platform which typically contains requests for crowdfunding ventures from individuals in the developing world. Their dataset of focus though is with regards to communication between lenders who are a part of teams or communities within the platform (such as those based on school affiliation or hobbies). Here a message by one person to their team is deemed persuasive text when it encourages them to donate to a particular borrower. The outcome measure of the persuasiveness of the text is operationalized as the amount of team members who then pledge loans to that borrower. They used Amazon mechanical Turkers to label the dataset.

They attempted to identify arguments based upon persuasive tactics including *Emotion, Scarcity, Commitment, Social Identity, Concreteness* and *Impact and Value* using a neural network architecture. In their model they first attempted to identify, using some labeled training data, the persuasive tactic used for a given sentence. Here they use a GRU to encode each word (using an embedding matrix) in a sentence, utilizing context, to a hidden state. An attention mechanism is then applied to the words, and finally we get a representation of the sentence that is a weighted sum of the hidden state of the word given the word's attention. Finally a softmax is applied to the sentence representation and the output is a vector of probabilities for various persuasive strategies. The sentence-level softmax vectors are then used in another neural network, very similar to the afore one used to create sentence representations, but now at the sentence level (rather than word), to create a document vector which is then fed into a neural net to give a final prediction.

Their best model at identifying rhetorical strategies at the sentence level had an accuracy of approximately 57% and F1 score of .52. Notably including other contextual information (that is details about the lending team or loan) did not improve accuracy. They also achieved a root mean squared error of .87 when predicting the number of team members who contribute to a project.

They also measured the relative importance of the various rhetorical strategies they identified. They utilized linear regression where the independent variables are the ratios of each strategy used in a document and the dependent variable again is teammate contributions. They find that identity and emotion based rhetorical strategies predict more contributions from teammates, but scarcity and commitment rhetoric are negatively influential in convincing teammates to contribute. Furthermore, they compared their regression model including predicted rhetorical strategies to one with just controls for team and loan characteristics. They found that when including their predictions of rhetorical strategies in the text that they increase the R^2, of a model predicting loan funding from teammates, from 0.152 (linear model with characteristics but not including predicted rhetoric ratios) to 0.17 (a significant increase).

2.4 Hidey and McKeown (2018)

Hidey and McKeown (2018) added to the literature on identifying persuasive text by modeling sentence-level argument sequencing. Hidey and McKeown utilized a dataset of discourse on Reddit (an online forum website) from a subreddit (forum) called *Change My View*[1]. This subreddit is particularly interesting; in it a user prompts discussion by posting on a specific issue with their own stance on the issue and encourages other forum participants to post arguments that change the original user/poster's view. Here, the persuasiveness of a user's post as a reply to the original poster is marked as to whether or not it managed to "change a view". Thus rather than measure some later outcome such as funding success or sales we may measure whether the argument is directly persuasive.

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

Hidey and McKeown create sentence level representations of discussion replies in a variety of ways. They create a sentence vector by concatenating three different vector representations: words, frames, and inter-sentence relation (note that frame as used here is distinct from framing as I discussed earlier) . The word vector is created by a weighted sum of the words (word embeddings) in the sentence where attention is used to weight. A frame vector is created by utilizing the FrameNet parser (Das et al. 2010). FrameNet is similar to WordNet in that it attempts to identify the meaning of the word when it may have multiple meanings. An attention weight is calculated for each word in a sentence and each potential frame for that word, then we get a weighted average of the frames for a sentence level vector representation. Lastly inter-sentence relation utilizes the Penn Discourse Tree Bank to identify the relation of one sentence to another (e.g. causal linkages or comparisons), and this is based upon prior work by Biran and McKeown (2015). Then this concatenated sentence vector is fed into a highway network.

They calculate a representation of the reply post based upon a weighted sum of the sentence representations (attention multiplied by a hidden state). Hidey and McKeown used various representations to calculate the hidden state of the sentence and then concatenated them. A bidirectional LSTM hidden state of a given sentence is created, and a hidden state based upon the original post (the person who's view is solicited to be changed). That is, a representation for the entire original post is created as done earlier for the replies, and a memory representation is also created. In the model, the memory representation is the previous iteration's representation of a reply post's representation (at initialization it is just the average of the hidden states of the sentences in the reply post). A final sentence representation is created by concatenating the representation of the reply, the original post representation, and the memory representation. This concatenated vector representation is then passed to a multi-layer perceptron to predict the dependent variable of influence (whether a view was changed by that particular post or not).

What is interesting from this paper is the ablation comparison of the models with and without LSTM as described above. They find that the LSTM greatly improves performance (an F1 score increase from 50.3 to 53 and accuracy increase from 68.8 to 75). This suggests that sentence ordering or argument ordering greatly matters. That is, sequential presentation of sentences is influential in persuading someone may be considered a takeaway of this paper. Furthermore the authors crowdsourced workers to identify, in a pairwise task, which argument is persuasive or not and found that their model outperformed the workers (though the authors caution that experts may perform better)

2.5 Fluency and Attitude; Lee and Labroo (2004)

I find the work above, especially Hidey and McKeown's especially encouraging for social scientists as it maps onto work done marketing relating to the construct of *fluency*. Fluency is an important construct in framing, marketing, and persuasion as highly fluent discourse and stimuli induces positive affective states as compared to low fluency owing to the ease in interpretability of fluent objects (Reber, Winkielman, and Schwarz 1998; Seamon et al. 1995). . Advertisers seek high fluency in their messages since by doing so they may induce positive affect in people towards their message and brand, leading to higher sales. Fluency may be thought of as a continuous variable which is the ability of someone, cognitively, to recognize, identify, and make sense of a target object (Jacoby and Dallas 1981). Fluency is often times sequentially induced, that is whether or not positive or negative fluency is induced objects depends upon the sequential presentation of stimuli (imagery or language).

Lee and Labroo (2004), in replicating Whittlesea (1993) used a variety of experiments to study fluency, including both language and imagery in their experiments. In their first set of experiments, which I will focus on, they recruited participants to read various sentences and then recorded participant affective reactions towards the sentence.

They hypothesized, and found, that participants reacted more favorably to a word when they had just read a sentence were the word was mentioned, expected by the context, or semantically similar to words presented in a previous sentence. In the one condition, for example, a sentence may read "All the neighbors gathered together to talk about the book" the participants react favorably when later asked to evaluate the word 'book', and they also reacted favorably to the unmentioned but seman-

tically similar word 'read'. However when participants are asked to rate how pleasant they found the word 'napkin' upon reading the same sentence, they reacted less pleasantly.

Another experimental condition analyzed disposition when a word is predicted given the prior context, as in the following sentence and paired word 'book': "The librarian reached for the top shelf and pulled down a book." When participants are then asked how pleasant they find the words 'book' and 'read' they react favorably, even more so compared to the 'neighborhood' context sentence. Here we may easily predict and process the word 'book' (and semantically similar word 'read), as it is expected given the prior words mentioned (such as 'librarian') which was not true for the neighborhood sentence.

What Lee and Labroo have shown here (and in their other experiments) is that the predictive and conceptually (or semantically) similar sequential presentation of stimuli can induce positive affect in consumers. These stimuli can be either language or imagery. Thus in understanding framing and persuasion we may expect that discourse which is presented in sequences which are semantically similar or whereby discourse is expected based upon the prior language context will be evaluated more favorably by readers. This favorability or positive affective state should then lead to the reader to be more likely to engage in behavior the writer of the discourse desires.

3 Synthesizing Research

Overtime the literature in persuasion modeling seems to have increasingly used neural networks rather than other machine learning methods (such as svm and regression). Notably earlier work did quite well in identifying phrases that were associated with effective persuasion (Mitra and Gilbert 2014) or predicting what discourse is persuasive without utilizing neural networks or word embeddings (Habernal and Gurevych 2016a); this illustrated that creating features from the text, such as bigrams or measures of readability, is fruitful in identifying whether or not an argument is persuasive.

Later work using LSTM and neural networks were effective in labelling the persuasive strategies employed by text (Habernal and Gurevych 2016b; Yang et al. 2019). These works engaged in supervised learning to label and explain the argumentation tactics/styles used. Yang et al. 2019 further utilized labeled predictions of argumentation style/tactic to predict whether the discourse was persuasive. Again, as in earlier work, Yang et al. treated the outcome prediction for the rhetorical strategies as a feature that was then fed into a regression.

Hidey and McKeown's (2018) work differs quite a bit from the others in that it did not seek to create features that were then further used in a regression model. Instead these authors relied primarily on word embeddings and neural networks. This paper also used discourse from Reddit, and discussions on this website (in terms of words in a post) were notably longer than the ones used by Habernal and Gurevych in their work; this may have allowed them to identify, using an LSTM model, that argument order (at the sentence level) matters. Such a finding may not have been found when previous work focused upon creating features of a whole document. However a potential drawback of this work is that it does not help too much in identifying why a piece of text is persuasive in the way that feature extraction and argument labeling in previous work did. Nonetheless, that the context of some rhetoric (defined at the sentence level) matters is supported by Marketing/Psychology research which demonstrates that attitude is influenced by context (Lee and Labroo 2004). Lee and Labroo demonstrated that the predicative context a word appears in, based upon based words, can affect individuals' favorability towards that word and their affective state. I further discuss the potential role of affect and sentiment in understanding effective persuasion in the future work section.

3.1 Datasets

The research work conducted on persuasion is quite varied with respect to the datasets used. Creating corpuses of persuasion may not be easy and the data varies as persuading occurs in many domains in life. Persuading is still recognized in the context of marketing with regards to product descriptions (Pryzant, Chung, and Jurafsky 2017), online crowdfunding (Yang et al. 2019; Mitra and Gilbert 2014), and annotated corpora of arguments scraped from the argument-specific websites (Hidey and McKeown 2018; Habernal and Gurevych 2016a, 2016b). However for future research I think it may be best to focus on the Red-

dit corpus, as it is clearly annotated as to whether someone was or was not persuaded (as opposed to say less direct measures such as product sales or fundraising), it contains thousands of unique arguments across a variety of topics, and the responses themselves tend to have many more words in them. Furthermore the Reddit data has been made publicly available and is already in a clean analyzable format.[2]

4 Future Directions

I believe that there are a variety of ways to enrich our ability to understand whether discourse is persuasive or not. Previous research has identified that affect and emotions play a role in persuading (Yang et al. 2019; Mitra and Gilbert 2014). Further work that identifies emotional arguments or arguments which generate affective reactions in people, such as those creating fluency (Lee and Labroo 2004), may be insightful. Thus I think it may be fruitful to include some representation of the emotional content of discourse or sentiment.

Utilizing models for detecting sentiment and emotions may be fruitful. Past work has sought to identify individuals' attitudes (or sentiment), often valanced as positive or negative, towards objects (Socher et al. 2013; Pang and Lee 2008). However other research has sought to specifically identify emotions, Chatterjee et al. (2019) identified and reviewed various methods for detecting emotions such as anger, happiness, and sadness from text. Using pre-trained models, such as DeepMoji (Felbo et al. 2017), and applying them to identify the emotional content of text could be useful in creating features.

While a model such as DeepMoji is created on a different set of discourse, we may use it at the sentence level of discourse in arguments to identify emotional or affective content at that level. One could then treat the outputted softmax vector of potential emotions/emojis/affective states of a sentence as an input to a neural network or regression. As in Hidey and McKeown (2018), one could concatenate the emotional representation of a sentence with a vector representation of the word embeddings.

Additionally one could create a measure of fluency at the sentence level. One could create a measure as to whether the current sentence is predictive, or not, of the following sentence. This measure could then be treated as a feature and concatenated into a vector representation of the sentence. Indeed this sort of task is one of those which BERT has demonstrated its capabilities (Devlin et al. 2018).

By creating more features to measure emotions and to measure fluency, I believe models created to understand whether a piece of discourse is persuasive or not may be meaningfully improved. Including such features has grounded basis in social science, and previous work computationally analyzing persuasion has noted that these seem to matter but have not devoted special attention to them. Identifying that such features matter and then being able to predict the efficacy of persuasive text based upon those features may then be of great benefit to organizations. Organizations then may then be able to better frame persuasive messages so that they develop more effective marketing campaigns, social movements, or raise more capital when fundraising by taking into account this measurable features of discourse.

5 References

Biran, O., & McKeown, K. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 96?104. Prague, Czech Republic: Association for Computational Linguistics.

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019, June). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 39-48).

Das, D.; Schneider, N.; Chen, D.; and Smith, N. A. 2010. Prob- abilistic frame-semantic parsing. In Human Language Technolo- gies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 948?956. Los Angeles, California: Association for Computational Linguistics.

Devlin, J., Chang, M. W., Lee, K.,& Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2] https://chenhaot.com/data/cmv/cmv.tar.bz2 (Tan et al. 2016)

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of communication, 43(4), 51-58.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524.

Lee, A. Y., & Labroo, A. A. (2004). The effect of conceptual and perceptual fluency on brand evaluation. Journal of Marketing Research, 41(2), 151-165.

Habernal(a), I., & Gurevych, I. (2016, August). 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) (pp. 1589-1599).

Habernal(b), I., & Gurevych, I. (2016, November). 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 (pp. 1214-1223).

Hidey, C. T., & McKeown, K. (2018, April). Persuasive influence detection: The role of argument sequencing. In Thirty-Second AAAI Conference on Artificial Intelligence.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. Journal of Experimental Psychology: General, 110(3), 306.

Mitra, T., & Gilbert, E. (2014, February). The language that gets people to give: Phrases that predict success on kickstarter. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 49-61).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?. Personality and social psychology review, 8(4), 364-382.

Seamon, J. G., Williams, P. C., Crowley, M. J., Kim, I. J., Langer, S. A., Orne, P. J., & Wishengrad, D. L. (1995). The mere exposure effect is based on implicit memory: Effects of stimulus type, encoding conditions, and number of exposures on recognition and affect judgments. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(3), 711.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016, April). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In Proceedings of the 25th international conference on world wide web (pp. 613-624).

Whittlesea, B. W. (1993). Illusions of familiarity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(6), 1235.

Yang, D., Chen, J., Yang, Z., Jurafsky, D., & Hovy, E. (2019, June). Let?s Make Your Request More Persuasive: Modeling Persuasive Strategies via Semi-Supervised Neural Nets on Crowdfunding Platforms. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 3620-3630).