Modeling of Policy Frames for Morality Detection on Twitter

Kristen Johnson and Dan Goldwasser
Department of Computer Science
Purdue University, West Lafayette, IN 47907
{john1187, dgoldwas}@purdue.edu

Abstract

Previous works have shown correlations in text between political ideologies and moral foundations expressed in the text. Additional work has shown that policy frames, which are used by politicians to bias the public towards their stance on an issue, are also correlated with political ideology. Based on these associations, we are interested in developing models which combine features of the language used on social media microblogs, specifically Twitter, as well as how politicians frame issues on Twitter, in order to predict the moral foundations used by politicians to express their stances on issues. This abstract presents the details of our annotation process and the resulting dataset, annotated for use in future morality-based prediction tasks. We also present our initial steps towards accurately modeling political discourse on Twitter in terms of language, ideology, and message framing, as well as how these components relate to the moral foundations expressed in tweets.

1 Introduction

Social media microblogging platforms, specifically Twitter, have become highly influential and relevant to current political events. Such platforms allow politicians to communicate with the public as events are unfolding. Further, politicians are able to express their stances on issues and by selectively using certain political slogans, also reveal their underlying political ideologies and moral views on an issue. Due to the association of morality with political ideology, it would be beneficial to understand, detect, and apply features of morality to political discourse analysis models.

We are therefore interested in exploring how political ideology, language, and framing interact to represent morality on Twitter and can be combined together for the analysis of real world political behavior. Previous works have studied framing in longer texts, such as congressional speeches and presidential debates (Fulgoni et al., 2016; Tsur et al., 2015; Card et al., 2015; Baumer et al., 2015; Tan et al., 2018), as well as on Twitter (Johnson et al., 2017). Ideology measurement (Iyyer et al., 2014; Bamman and Smith, 2015; Sim et al., 2013; Djemili et al., 2014) and polls based on Twitter political sentiment (Bermingham and Smeaton, 2011; O'Connor et al., 2010; Tumasjan et al., 2010) have also explored framing. The association between Twitter and framing in molding public opinion of issues (Burch et al., 2015; Harlow and Johnson, 2011; Meraz and Papacharissi, 2013; Jang and Hart, 2015) has also been studied. The connection between morality and political ideology has also been explored in the fields of psychology and sociology (Graham et al., 2009, 2012). Applications of the Moral Foundations Dictionary, a set of unigrams expected to indicate each moral foundation, have also been studied specific to Twitter (Garten et al.; Lin et al., 2017).

Different from these works, we: (1) study the tweets of politicians in which the content is carefully crafted; (2) explore how ideology, framing, and morality interact on a fine-grained level; (3) propose that political slogans are more indicative of the moral foundation of a tweet. To date, most works follow a key word based approach (using the Moral Foundations Dictionary provided by Haidt and Joseph), which we observed falls short when dealing with political tweets as politicians craft their messages to carefully reveal morality via associations to specific aspects of issues. Studying political messaging through slogans can help capture the context and moral aspects being discussed in politicians' tweets.

We present our first steps towards examining the interplay of political slogans, for example *"repeal and replace"* when referring to the Affordable Care Act, and policy framing techniques (Boydstun et al., 2014; Johnson et al., 2017) as features for predicting the underlying moral values which are expressed in politicians' tweets. The moral values we are interested in predicting are the five moral foundations described in the Moral Foundations Theory (Haidt and Joseph, 2004; Graham et al., 2009). Classifying these foundations in tweets, as opposed to longer texts, presents unique challenges due to the short length of tweets and the resulting lack of context. Furthermore, due to the highly dynamic nature of political discourse on Twitter, we propose weakly supervised feature extraction models to isolate relevant information, including language and political ideology features. In future works, this information will be incorporated into a probabilistic relational model and used to analyze political tweets at scale.

Moral Foundation and Brief Description
1. Care/Harm: Care for others, generosity, compassion, ability to feel pain of others, sensitivity to suffering of others, prohibiting actions that harm others.
2. Fairness/Cheating: Fairness, justice, reciprocity, reciprocal altruism, rights, autonomy, equality, proportionality, prohibiting cheating.
3. Loyalty/Betrayal: Group affiliation and solidarity, virtues of patriotism, self-sacrifice for the group, prohibiting betrayal of one's group.
4. Authority/Subversion: Fulfilling social roles, submitting to authority, respect for social hierarchy, leadership, fellowship, respect for traditions, prohibiting rebellion against authority.
5. Purity/Degradation: Associations with the sacred and holy, disgust, contamination, underlies religious notions of striving to live in an elevated way, prohibiting violating the sacred.
6. Non-moral: Does not fall under any other foundations.

Table 1: Brief Descriptions of the Moral Foundations.

2 Dataset Annotation

For this work, annotators manually annotated the Congressional Tweets Dataset (Johnson et al., 2017) using the moral foundations shown in Table 1. We chose this dataset because we are interested in studying the language, framing, and morality of U.S. *politicians*, as opposed to the general public. To the best of our knowledge, this is the first dataset of U.S. politicians on Twitter to be annotated for the morals described in the Moral Foundations Theory. We initially attempted to use Amazon Mechanical Turk, but found that most workers would choose the Care/Harm or Fairness/Cheating label a majority of the time. Therefore, we chose two annotators to manually annotate a subset of tweets, agree on general guidelines, and then label the remaining tweets of the dataset. To achieve a neutral, unbiased annotation, we chose annotators with different self-reported political ideologies, i.e., one liberal and one conservative annotator.

Labeling tweets, and thus classifying their frames or morals, presents several challenges. First, tweets are short and thus lack the context that is often necessary for choosing a moral viewpoint. Tweets are often ambiguous, e.g., a tweet may express care for people who are being harmed by a policy. Another challenge was overcoming the political bias of the annotator. For example, if a tweet discusses opposing Planned Parenthood, the liberal annotator typically viewed this as Harm (i.e., taking services away from women and thus hurting them), while the conservative annotator tended to view this as Purity (i.e., all life is sacred). To overcome this bias, annotators were given the political party of the politician who wrote the tweets and instructed to choose the moral foundation *from the politician's perspective*. Finally, as noted in Johnson et al., tweets present a compound problem: tweets often present two thoughts, some of which can even be contradictory. This results in one tweet having multiple moral foundations. Annotators chose a primary moral foundation whenever possible, but were allowed a secondary foundation if the tweet presented two differing thoughts.

The resulting labeled dataset has an inter-annotator agreement of 79.2% using Cohen's Kappa coefficient. Table 2 presents a summary of the statistics of the original dataset and the distributions of the moral foundations present in the labeled portion of the dataset.

Several recurring themes continued to appear throughout the dataset including "thoughts and prayers" for victims of gun shooting events or rhetoric against the opposing political party. The annotators agreed on the following general guidelines for these repeating topics: (1) The Purity label is used for tweets that relate to prayers or the fight against ISIL/ISIS. (2) Loyalty is for tweets that discuss "stand(ing) with" others, American values, American troops or allies, or reference a demographic that the politician belongs to, e.g., if the politician tweeting is a woman and she discusses women-related issues. (3) At the time the dataset was collected, the President was Barack Obama and the Republican party controlled Congress. Therefore, any tweets specifi-

CATEGORY	OVERALL	PARTY	
		REP	DEM
All	92457	48504	43953
Labeled	2050	894	1156
Care	524	156	368
Harm	355	151	204
Fairness	268	55	213
Cheating	82	37	45
Loyalty	303	63	240
Betrayal	53	25	28
Authority	192	62	130
Subversion	419	251	168
Purity	174	86	88
Degradation	66	34	32
Non-moral	334	198	136

Table 2: Distributions of Dataset. Overall is across the entire dataset. Party is the Republican (Rep) or Democrat (Dem) specific distributions. All represents all tweets in the dataset after filtering. Labeled represents the portion of tweets labeled for policy frame. The remaining categories are the number of tweets for each moral foundation from the Labeled portion of the dataset.

cally attacking Obama or Republicans (the majority party) were labeled as Subversion. (4) Tweets discussing health or welfare were labeled as Care. (5) Tweets which discussed limiting or restricting laws or rights were labeled as Cheating. (6) Sarcastic attacks, typically against the opposing political party, were labeled as Degradation.

3 Initial Results

We have designed initial weakly-supervised feature extraction models to identify representations of unigrams, political slogans, and frames from politicians' tweets. The unigrams used as the initial source of supervision for the global models are either taken directly from the Moral Foundations Dictionary (MFD) or from unigrams that the annotators stated they found most useful for choosing a moral foundation (AN). These features are then combined into global Probabilistic Soft Logic (PSL) models which are used to predict the moral foundation expressed in a tweet.

Table 3 presents the results of our supervised experiments using these initial models. The first column lists the features of the PSL model. The second column presents the results of the model when using the MFD unigrams as the basis of the initial PSL model. The final column shows the results when the AN unigrams are used as the initial source of supervision.

One interesting finding is that the AN unigrams produce better average performances when only unigrams are used for features. Models that incorporate more abstract textual representations, in the form of political slogans and frames, tend to perform better when using the MFD-based unigrams. This suggests that the AN unigrams, which are tuned specifically to the political Twitter domain may be more useful than dictionary-based unigrams, when only unigrams are available. Conversely, because the MFD unigrams are designed to capture the *idea of morality*, these models have weaker results when using only unigrams as features, but higher results when combined with more abstract features.

PSL MODEL	MFD	AN
MAJORITY VOTE	12.5	10.86
UNIGRAMS ONLY	7.17	8.68
UNIGRAMS + SLOGANS	67.93	66.50
UNIS + SLOGANS + FRAMES	72.49	69.38

Table 3: Overview of Macro-Average F_1 Scores of PSL Models. Majority Vote represents the traditional baseline of using a majority count of presence of unigrams.

4 Future Work

In our continuation of this work, we will conduct more detailed feature analysis experiments to better understand the effects of different unigrams and slogans on moral foundation prediction in tweets. We will also explore the usefulness of jointly modeling policy frame and moral foundation prediction. Our ultimate goal is to apply these models for various political discourse analysis tasks, such as political ideology or stance prediction.

Acknowledgments

We thank Professor Lyle Ungar for his insight and discussion which inspired this work. We also thank the anonymous reviewers for their thoughtful comments and suggestions. Finally, we thank the conservative annotator for helping us to compile this new dataset.

References

David Bamman and Noah A Smith. 2015. Open extraction of fine-grained political statements. In *Proc. of EMNLP*.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *In Proc. of NAACL*.

Adam Bermingham and Alan F Smeaton. 2011. On using twitter to monitor political sentiment and predict election results.

Amber Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

Lauren M. Burch, Evan L. Frederick, and Ann Pegoraro. 2015. Kissing in the carnage: An examination of framing on twitter during the vancouver riots. *Journal of Broadcasting & Electronic Media*, 59(3):399–415.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proc. of ACL*.

Sarah Djemili, Julien Longhi, Claudia Marinica, Dimitris Kotzinos, and Georges-Elia Sarfati. 2014. What does twitter have to say about ideology? In *NLP 4 CMC*.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preotiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proc. of LREC*.

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. Morality between the lines: Detecting moral sentiment in text.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Jesse Graham, Brian A Nosek, and Jonathan Haidt. 2012. The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PloS one*, 7(12):e50092.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Summer Harlow and Thomas Johnson. 2011. The arab spring— overthrowing the protest paradigm? how the new york times, global voices and twitter covered the egyptian revolution. *International Journal of Communication*, 5(0).

Iyyer, Enns, Boyd-Graber, and Resnik. 2014. Political ideology detection using recursive neural networks. In *Proc. of ACL*.

S. Mo Jang and P. Sol Hart. 2015. Polarized frames on ”climate change” and ”global warming” across countries and states: Evidence from twitter big data. *Global Environmental Change*, 32:11–17.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on twitter. In *Proc. of ACL*.

Ying Lin, Joe Hoover, Morteza Dehghani, Marlon Mooijman, and Heng Ji. 2017. Acquiring background knowledge to improve moral value prediction. *arXiv preprint arXiv:1709.05467*.

Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics*, 18(2):138–166.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of ICWSM*.

Sim, Acree, Gross, and Smith. 2013. Measuring ideological proportions in political speeches. In *Proc. of EMNLP*.

Chenhao Tan, Hao Peng, and Noah A. Smith. 2018. ”you are no jack kennedy”: On media selection of highlights from presidential debates. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 945–954, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proc. of ACL*.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*.