Classification of Moral Foundations in Microblog Political Discourse

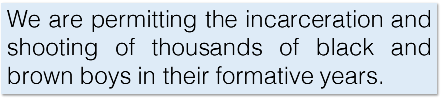
Kristen Johnson and Dan Goldwasser
Department of Computer Science
Purdue University, West Lafayette, IN 47907
{john1187, dgoldwas}@purdue.edu

Abstract

Previous works in computer science, as well as political and social science, have shown correlation in text between political ideologies and the moral foundations expressed within that 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, this work takes a first step towards modeling both the language and how politicians frame issues on Twitter, in order to predict the moral foundations that are used by politicians to express their stances on issues. The contributions of this work includes a dataset annotated for the moral foundations, annotation guidelines, and probabilistic graphical models which show the usefulness of jointly modeling abstract political slogans, as opposed to the unigrams of previous works, with policy frames for the prediction of the morality underlying political 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 and shape public discourse on various issues. Furthermore, politicians are able to express their stances on issues and by selectively using certain political slogans, reveal their underlying political ideologies and moral views on an issue. Previous works in political and social science have shown a correlation between political ideology, stances on political issues, and the moral convictions used to justify these stances (Graham et al., 2009). For example, Figure 1 presents a tweet, by a prominent member of the U.S. Congress, which expresses concern about the fate of young individuals (i.e., *incarceration, shooting*), specifically for vulnerable members of minority groups. The Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Haidt and Graham, 2007) provides a theoretical framework for explaining these nuanced distinctions. The theory suggests that there are five basic moral values which underlie human moral perspectives, emerging from evolutionary, social, and cultural origins. These are referred to as the moral foundations (MF) and include *Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion*, and *Purity/Degradation* (Table 1 provides a more detailed explanation). The above example reflects the moral foundations that shape the author's perspective on the issue: *Harm* and *Cheating*.



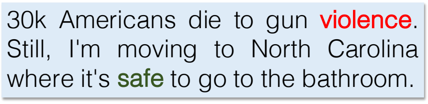
We are permitting the incarceration and shooting of thousands of black and brown boys in their formative years.

Figure 1: Example Tweet Highlighting Classification Difficulty.

Traditionally, analyzing text based on the MFT has relied on the use of a lexical resource, the Moral Foundations Dictionary (MFD) (Haidt and Graham, 2007; Graham et al., 2009). The MFD, similar to LIWC (Pennebaker et al., 2001; Tausczik and Pennebaker, 2010), associates a list of related words with each one of the moral foundations. Therefore, analyzing text equates to counting the number of occurrences of words in the text which also match the words in the MFD. Given the highly abstract and generalized nature of the moral foundations, this approach often falls short of dealing with the highly ambiguous text

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 720–730
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

politicians use to express their perspectives on specific issues. The following tweet, by another prominent member of the U.S. Congress, reflects the author's use of both the *Harm* and *Cheating* moral foundations.



30k Americans die to gun violence. Still, I'm moving to North Carolina where it's safe to go to the bathroom.

Figure 2: Example Tweet Highlighting Classification Difficulty.

While the first foundation (*Harm*) can be directly identified using a word match to the MFD (as shown in red), the second foundation requires first identifying the sarcastic expression referring to LGBTQ rights and then using extensive world knowledge to determine the appropriate moral foundation. [1] Relying on a match of *safe* to the MFD would indicate the *Care* MF is being used instead of the *Cheating* foundation.

In this paper, we aim to solve this challenge by suggesting a data-driven approach to moral foundation identification in tweets. Previous work (Garten et al., 2016) has looked at classification-based approaches over tweets specifically related to Hurricane Sandy, augmenting the textual content with background knowledge using entity linking (Lin et al., 2017). Different from this and similar works, we look at the tweets of U.S. politicians over a long period of time, discussing a large number of events, and touching on several different political issues. Our approach is guided by the intuition that the abstract moral foundations will manifest differently in text, depending on the specific characteristics of the events discussed in the tweet. As a result, it is necessary to correctly model the relevant contextualizing information.

Specifically, we are interested in exploring how political ideology, language, and framing interact to represent morality on Twitter. We examine 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. Additionally, we identify high-level themes characterizing the main point of the tweet, which allows the model to identify the author's perspective on specific issues and generalize over the specific wording used (for example, if the tweet mentions `Religion` or `Political Maneuvering`).

This information is incorporated into global probabilistic models using Probabilistic Soft Logic (PSL), a graphical probabilistic modeling framework (Bach et al., 2013). PSL specifies high level rules over a relational representation of these features, which are compiled into a graphical model called a hinge-loss Markov random field that is used to make the final prediction. Our experiments show the importance of modeling contextualizing information, leading to significant improvements over dictionary driven approaches and purely lexical methods.

In summary, this paper makes the following contributions: (1) This work is among the first to explore jointly modeling language and political framing techniques for the classification of moral foundations used in the tweets of U.S. politicians on Twitter. (2) We provide a description of our annotation guidelines and an annotated dataset of 2,050 tweets.[2] (3) We suggest computational models which easily adapt to new policy issues, for the classification of the moral foundations present in tweets.

2 Related Works

In this paper, we explore how political ideology, language, framing, and morality interact on Twitter. Previous works have studied framing in longer texts, such as congressional speeches and news (Fulgoni et al., 2016; Tsur et al., 2015; Card et al., 2015; Baumer et al., 2015), as well as issue-independent framing on Twitter (Johnson and Goldwasser, 2016; Johnson et al., 2017). Ideology measurement (Iyyer et al., 2014; Bamman and Smith, 2015; Sim et al., 2013; Djemili et al., 2014), political sentiment analysis (Pla and Hurtado, 2014; Bakliwal et al., 2013), and polls based on Twitter political sentiment (Bermingham and Smeaton, 2011; O'Connor et al., 2010; Tumasjan et al., 2010) are also related to the study of framing. The association between Twitter and framing in molding public opinion of events and issues (Burch et al., 2015; Harlow and Johnson, 2011; Meraz and Papacharissi, 2013; Jang and

[1] The tweet refers to legislation proposed in 2016 concerning transgender bathroom access restrictions.

[2] The data will be available at `http://purduenlp.cs.purdue.edu/projects/twittermorals`.

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/traditions, leadership, prohibiting rebellion against authority.
5. Purity/Degradation: Associations with the sacred and holy, disgust, contamination, religious notions which guide how to live, prohibiting violating the sacred.
6. Non-moral: Does not fall under any other foundations.

Table 1: Brief Descriptions of Moral Foundations.

Hart, 2015) has also been studied.

The connection between morality and political ideology has been explored in the fields of psychology and sociology (Graham et al., 2009, 2012). Moral foundations were also used to inform downstream tasks, by using the MFD to identify the moral foundations in partisan news sources (Fulgoni et al., 2016), or to construct features for other downstream tasks (Volkova et al., 2017). Several recent works have looked into using data-driven methods that go beyond the MFD to study tweets related to Hurricane Sandy (Garten et al., 2016; Lin et al., 2017).

3 Data Annotation

The Moral Foundations Theory (Haidt and Graham, 2007) was proposed by sociologists and psychologists as a way to understand how morality develops, as well as its similarities and differences across cultures. The theory consists of the five moral foundations shown in Table 1. The goal of this work is to classify the tweets of the Congressional Tweets Dataset (Johnson et al., 2017) with the moral foundation implied in the tweet.

We first attempted to use Amazon Mechanical Turk for annotation, but found that most Mechanical Turkers would choose the *Care/Harm* or *Fairness/Cheating* label a majority of the time. Additionally, annotators preferred choosing first the foundation branch (i.e., *Care/Harm*) and then its sentiment (positive or negative) as opposed to the choice of each foundation separately, i.e., given the choice between *Harm* or *Care/Harm and Negative*, annotators preferred the latter. Based on these observations, two annotators, one liberal and one conservative (self-reported), manually annotated a subset of tweets. This subset had an inter-annotator agreement of 67.2% using Cohen's Kappa coefficient. The annotators then discussed and agreed on general guidelines which were used to label the remaining tweets of the dataset. The resulting dataset has an inter-annotator agreement of 79.2% using Cohen's Kappa statistic. The overall distribution, distributions by political party, and distributions per issue of the labeled dataset are presented in Table 2. Table 3 lists the frames that most frequently co-occured with each MF. As expected, frames concerning Morality and Sympathy are highly correlated with the *Purity* foundation, while *Subversion* is highly correlated with the Legal and Political frames.

Labeling tweets presents several challenges. First, tweets are short and thus lack the context 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 major challenge was overcoming the political bias of the annotator. For example, if a tweet discusses opposing Planned Parenthood because it provides abortion services, the liberal annotator typically viewed this as *Harm* (i.e., hurting women by taking away services from them), while the conservative annotator tended to view this as *Purity* (i.e., all life is sacred and should be protected). 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*. To further simplify the annotation process, all tweets belonging to one political party were labeled together, i.e., all Republican tweets were labeled and then all Democrat tweets were labeled. Finally, tweets present a compound problem, often expressing two thoughts which can further 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.

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 to use the following moral foundation labels for these repeating topics as follows: (1) The *Purity* label is used for tweets that relate to

Morals	Overall	Party		Issue					
		Rep	Dem	Abo	ACA	Gun	Imm	LGBTQ	Ter
Care	524	156	368	37	123	215	33	34	113
Harm	355	151	204	26	64	141	19	34	101
Fairness	268	55	213	41	81	19	11	86	39
Cheating	82	37	45	14	27	11	10	9	13
Loyalty	303	63	240	28	29	128	36	38	58
Betrayal	53	25	28	10	4	9	6	3	22
Authority	192	62	130	24	44	50	38	10	34
Subversion	419	251	168	34	169	75	73	25	60
Purity	174	86	88	24	3	102	5	24	41
Degradation	66	34	32	5	0	31	0	4	31
Non-moral	334	198	136	17	143	28	47	7	96

Table 2: Distributions of Moral Foundations. Overall is across the entire dataset. Party is the Republican (Rep) or Democrat (Dem) specific distributions. Issue lists the six issue-specific distributions (Abortion, ACA, Guns, Immigration, LGBTQ, Terrorism).

Moral Foundation and Co-occuring Frames
Care: Capacity & Resources, Security & Defense, Health & Safety, Quality of Life, Public Sentiment, External Regulation & Reputation
Harm: Economic, Crime & Punishment
Fairness: Fairness & Equality
Loyalty: Cultural Identity
Subversion: Legality, Constitutionality, & Jurisdiction, Political Factors & Implications, Policy Description, Prescription, & Evaluation
Purity: Morality & Ethics, Personal Sympathy & Support
Non-moral: Factual, (Self) Promotion

Table 3: Foundations and Co-occuring Frames. *Cheating, Betrayal, Authority,* and *Degradation* did not co-occur frequently with any frames.

prayers or the fight against ISIL/ISIS. (2) *Loyalty* is for tweets that discuss "stand(ing) with" others, American values, troops, or allies, or reference a demographic that the politician belongs to, e.g. if the politician tweeting is a woman and she discusses an issue in terms of its effects on women. (3) At the time the dataset was collected, the President was Barack Obama and the Republican party controlled Congress. Therefore, any tweets specifically attacking Obama or Republicans (the controlling 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*.

4 Feature Extraction for PSL Models

For this work, we designed extraction models and PSL models that were capable of adapting to the dynamic language used on Twitter and predicting the moral foundation of a given tweet. Our approach uses weakly supervised extraction models, whose only initial supervision is a set of unigrams and the political party of the tweet's author, to extract features for each PSL model. These features are represented as PSL predicates and combined into the probabilistic rules of each model, as shown in Table 4, which successively build upon the rules of the previous model.

4.1 Global Modeling Using PSL

PSL is a declarative modeling language which can be used to specify weighted, first-order logic rules that are compiled into a hinge-loss Markov random field. This field defines a probability distribution over possible continuous value assignments to the random variables of the model (Bach et al., 2015) and is represented as:

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z} \exp\left(-\sum_{r=1}^{M} \lambda_r \phi_r(\mathbf{Y}\,, \mathbf{X})\right)$$

where Z is a normalization constant, λ is the weight vector, and

$$\phi_r(\mathbf{Y}, \mathbf{X}) = (\max\{l_r(\mathbf{Y}, \mathbf{X}), 0\})^{\rho_r}$$

is the hinge-loss potential specified by a linear function l_r. The exponent $\rho_r \in 1, 2$ is optional. Each potential represents the instantiation of a rule, which takes the following form:

$$\lambda_1 : P_1(x) \wedge P_2(x, y) \rightarrow P_3(y)$$
$$\lambda_2 : P_1(x) \wedge P_4(x, y) \rightarrow \neg P_3(y)$$

P_1, P_2, P_3, and P_4 are predicates (e.g., party, issue, and frame) and x, y are variables. Each rule has a weight λ to reflect its importance to the model. Using concrete constants *a, b* (e.g., tweets) which instantiate the variables x, y, model atoms are mapped to continuous [0,1] assignments.

Mod.	Information Used	Example of PSL Rule
M1	Unigrams (MFD or AR)	Unigram$_M$(T, U) → Moral(T, M)
M2	M1 + Party	Unigram$_M$(T, U) ∧ Party(T, P) → Moral(T, M)
M3	M2 + Issue	Unigram$_M$(T, U) ∧ Party(T, P) ∧ Issue(T, I) → Moral(T, M)
M4	M3 + Phrase	Unigram$_M$(T, U) ∧ Party(T, P) ∧ Phrase(T, PH) → Moral(T, M)
M5	M4 + Frame	Unigram$_M$(T, U) ∧ Phrase(T, PH) ∧ Frame(T, F) → Moral(T, M)
M6	M5 + Party-Bigrams	Unigram$_M$(T, U) ∧ Party(T, P) ∧ Bigram$_P$(T, B) → Moral(T, M)
M7	M6 + Party-Issue-Bigrams	Unigram$_M$(T, U) ∧ Party(T, P) ∧ Bigram$_{PI}$(T, B) → Moral(T, M)
M8	M7 + Phrase	Bigram$_{PI}$(T, B) ∧ Phrase(T, PH) → Moral(T, M)
M9	M8 + Frame	Bigram$_{PI}$(T, B) ∧ Frame(T, F) → Moral(T, M)
M10	M9 + Party-Trigrams	Unigram$_M$(T, U) ∧ Party(T, P) ∧ Trigram$_P$(T, TG) → Moral(T, M)
M11	M10 + Party-Issue-Trigrams	Unigram$_M$(T, U) ∧ Party(T, P) ∧ Trigram$_{PI}$(T, TG) → Moral(T, M)
M12	M11 + Phrase	Trigram$_{PI}$(T, TG) ∧ Phrase(T, PH) → Moral(T, M)
M13	M12 + Frame	Trigram$_{PI}$(T, TG) ∧ Frame(T, F) → Moral(T, M)

Table 4: Examples of PSL Moral Model Rules Using Gold Standard Frames. For these rules, the Frame predicate is initialized with the known frame labels of the tweet. Each model builds successively on the rules of the previous model.

M2: Unigrams + Party	Unigram$_M$(T, U) ∧ Party(T, P) ∧ Frame(T, F) → Moral(T, M) Unigram$_M$(T, U) ∧ Party(T, P) ∧ Moral(T, M) → Frame(T, F)
M13: All Features	Trigram$_{PI}$(T, TG) ∧ Phrase(T, PH) ∧ Frame(T, F) → Moral(T, M) Trigram$_{PI}$(T, TG) ∧ Unigram$_M$(T, U) ∧ Moral(T, M) → Frame(T, F)

Table 5: Examples of PSL Joint Moral and Frame Model Rules. For these models, the Frame predicate is *not initialized with known values*, but is predicted jointly with the Moral predicate.

4.2 Feature Extraction Models

For each aspect of the tweets that composes the PSL models, scripts are written to first identify and then extract the correct information from the tweets. Once extracted, this information is formatted into PSL predicate notation and input to the PSL models. Table 4 presents the information that composes each PSL model, as well as an example of how rules in the PSL model are constructed.

Language: Works studying the Moral Foundations Theory typically assign a foundation to a body of text based on a majority match of the words in the text to the Moral Foundations Dictionary (MFD), a predefined list of unigrams associated with each foundation. These unigrams capture the conceptual idea behind each foundation. Annotators noted, however, that when choosing a foundation they typically used a small phrase or the entire tweet, not a single unigram. Based on this, we compiled all of the annotators' phrases per foundation into a unique set to create a new list of unigrams for each foundation. These unigrams are referred to as "Annotator's Rationale (AR)" throughout the remainder of this paper. The PSL predicate Unigram$_M$(T, U) is used to input any unigram U from tweet T that matches the M list of unigrams (either from the MFD or AR lists) into the PSL models. An example of a rule using this predicate is shown in the first row of Table 4.

During annotation, we observed that often a tweet has only one match to a unigram, if any, and therefore a majority count approach may fail. Further, as shown in Figure 2, many tweets have one unigram that matches one foundation and another unigram that matches a different foundation. In such cases, the correct foundation cannot be determined from unigram counts alone. Based on these observations and the annotators' preference for using phrases, we incorporate the most frequent bigrams and trigrams for each political party (Bigram$_P$(T, B) and Trigram$_P$(T, TG)) and for each party on each issue (Bigram$_{PI}$(T, B) and Trigram$_{PI}$(T, TG)). These top 20 bigrams and trigrams contribute to a more accurate prediction than unigrams alone (Johnson et al., 2017).

Ideological Information: Previous works have shown a strong correlation between ideology and the moral foundations (Haidt and Graham, 2007), as well as between ideology and policy issues (Boydstun et al., 2014). Annotators were able to agree on labels when instructed to label from the ideological point of view of the tweet's author, even if it opposed their own views. Based on these positive correlations, we incorporate both the issue of the tweet (ISSUE(T, I)) and the political party of the author of the tweet (PARTY(T, P)) into the PSL models. Examples of how this information is represented in the PSL models are shown in rows two and three of Table 4.

Abstract Phrases: As described previously, annotators reported that phrases were more useful than unigrams in determining the moral foundation of the tweet. Due to the dynamic nature of language and trending issues on Twitter, it is impracticable to construct a list of all possible phrases one can expect to appear in tweets. However, because politicians are known for sticking to certain talking points, these phrases can be *abstracted* into higher-level phrases that are more stable and thus easier to identify and extract.

For example, a tweet discussing "President Obama's signing a bill" has two possible concrete phrases: *President Obama's signing* and *signing a bill*. Each phrase falls under two possible abstractions: political maneuvering (Obama's actions) and mentions legislation (signing of a bill). In this paper we use the following high-level abstractions: `legislation or voting`, `rights and equality`, `emotion`, `sources of danger or harm`, `positive benefits or effects`, `solidarity`, `political maneuvering`, `protection and prevention`, `American values or traditions`, `religion`, and `promotion`. For example, if a tweet mentions "civil rights" or "equal pay", then these phrases indicate that the `rights and equality` abstraction is being used to express morality. Some of these abstractions correlate with the corresponding MF or frame, e.g., the `religion` abstraction is highly correlated with the *Purity* foundation and `political maneuvering` is correlated with the Political Factors & Implications Frame.

To match phrases in tweets to these abstractions, we use the embedding-based model of Lee et al. (2017). This phrase similarity model was trained on the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) and incorporates a Convolutional Neural Network (CNN) to capture sentence structures. This model generates the embeddings of our abstract phrases and computes the cosine similarities between phrases and tweets as the scores. The input tweets and phrases are represented as the average word embeddings in the input layer, which are then projected into a convolutional layer, a max-pooling layer, and finally two fully-connected layers. The embeddings are thus represented in the final layer. The learning objective of this model is:

$$\min_{W_c,W_w} \Bigg(\sum_{<x_1,x_2>\in X} max(0, \delta - cos(g(x_1), g(x_2)) + cos(g(x_1), g(t_1))) + max(0, \delta - cos(g(x_1), g(x_2))) + cos(g(x_2), g(t_2))) \Bigg) + \lambda_c ||W_c||^2 + \lambda_w ||W_{init} - W_w||^2,$$

where X is all the positive input pairs, δ is the margin, $g(\cdot)$ represents the network, λ_c and λ_w are the weights for L2-regularization, W_c is the network parameters, W_w is the word embeddings, W_{init} is the initial word embeddings, and t_1 and t_2 are negative examples that are randomly selected.

All tweet-phrase pairs with a cosine similarity over a given threshold are used as input to the PSL model via the predicate PHRASE(T, PH), which indicates that tweet T contains a phrase that is similar to an abstracted phrase (PH). [3] Rows four, eight, and twelve of Table 4 show examples of the phrase rules as used in our modeling procedure.

Nuanced Framing: Framing is a political strategy in which politicians carefully word their statements in order to bias public opinion towards their stance on an issue. This technique is a fine-grained view of how issues are expressed. Frames are associated with issue, political party, and ideologies. For example, if a politician emphasizes the economic burden a new bill would place on the public, then they are using the *Economic* frame. Different from this, if they emphasize how people's lives will improve because of this bill, then they are using the *Quality of Life* frame.

In this work, we explore frames in two settings: (1) where the actual frames of tweets are known and used to predict the moral foundation of the tweets and (2) when the frames are unknown and predicted jointly with the moral foundations. Using the Congressional Tweets Dataset as the true labels for 17 policy frames, this information is input to the PSL models using the FRAME(T, F) predicate as shown in Table 4. Conversely, the

[3] A threshold score of 0.45 provided the most accurate matches while minimizing noise.

same predicate can be used as a joint prediction target predicate, with no initialization, as shown in Table 5.

5 Experimental Results

In this section, we present an analysis of the results of our modeling approach. Table 6 summarizes our overall results and compares the traditional BoW SVM classifier[4] to several variations of our model. We provide an in-depth analysis, broken down by the different types of moral foundations, in Tables 7 and 8.

We also study the relationship between moral foundations, policy framing, and political ideology. Table 9 describes the results of a joint model for predicting moral foundations and policy frames. Finally, in Section 6 we discuss how moral foundations can be used for the downstream prediction of political party affiliation.

MODEL	MFD	AR
SVM BOW	18.70	—
PSL BOW	21.88	—
MAJORITY VOTE	12.50	10.86
M1 (UNIGRAMS)	7.17	8.68
M3 (+ POLITICAL INFO)	22.01	30.45
M5 (+ FRAMES)	28.94	37.44
M9 (+ BIGRAMS)	67.93	66.50
M13 (ALL FEATURES)	72.49	69.38

Table 6: Overview of Macro-weighted Average F_1 Scores of SVM and PSL Models. The top portion of the table shows the results of the three baselines. The bottom portion shows a subset of the PSL models (parentheses indicate features added onto the previous models).

Evaluation Metrics: Since each tweet can have more than one moral foundation, our prediction task is a multilabel classification task. The precision of a multilabel model is the ratio of how many predicted labels are correct:

$$Precision = \frac{1}{T}\sum_{t=1}^{T}\frac{|Y_t \cap h(x_t)|}{|h(x_t)|} \qquad (1)$$

The recall of this model is the ratio of how many of the actual labels were predicted:

$$Recall = \frac{1}{T}\sum_{t=1}^{T}\frac{|Y_t \cap h(x_t)|}{|Y_t|} \qquad (2)$$

In both formulas, T is the number of tweets, $\mathrm{Y_t}$ is the true label for tweet *t*, $\mathrm{x_t}$ is a tweet example, and $\mathrm{h(x_t)}$ are the predicted labels for that tweet. The $\mathrm{F_1}$ score is computed as the harmonic mean of the precision and recall. Additionally, the last lines of Tables 7 and 8 provide the macro-weighted average $\mathrm{F_1}$ score over all moral foundations.

Analysis of Supervised Experiments: We conducted supervised experiments using five-fold cross validation with randomly chosen splits. Table 6 shows an overview of the average results of our supervised experiments for five of the PSL models. The first column lists the SVM or PSL model. The second column presents the results of a given model when using the MFD as the source of the unigrams for the initial model (M1). The final column shows the results when the AR unigrams are used as the initial source of supervision. The first two rows show the results of predicting the morals present in tweets using a bag-of-words (BoW) approach. Both the SVM and PSL models perform poorly due to the eleven predictive classes and noisy input features. The third row shows the results when taking a majority vote over the presence of MFD unigrams, similar to previous works. This approach is simpler and less noisy than M1, the PSL model closest to this approach.

The last five lines of this table also show the overall trends of the full results shown in Tables 7 and 8. As can be seen in all three tables, as we add more information with each PSL model, the overall results continue to improve, with the final model (M13) achieving the highest $\mathrm{F_1}$ score for both sources of unigrams.

An interesting trend to note is that the AR unigrams based models result in better average performance for most of the models until M9. Models M9 and above incorporate the most powerful features: bigrams and trigrams with phrases and frames. This suggests that the AR unigrams, designed specifically for the political Twitter domain, are more useful than the MFD unigrams, *when only unigrams are available*. Conversely, the MFD unigrams are designed to *conceptually* capture morality, and therefore have weaker performance in the unigram-based models, but achieve higher performance when combined with the more powerful features of the higher models. For all models, incorporating phrases and frames results in a more accurate prediction than when using unigrams alone.

[4] For this work, we used the SVM implementation provided by scikit-learn.

Moral Fdn.	Results of Non-Joint PSL Model Predictions												
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Care	16.61	52.51	43.34	53.24	53.38	53.59	55.64	62.40	66.00	66.48	67.32	67.59	**67.78**
Harm	12.57	47.62	42.58	50.39	57.24	55.29	60.06	67.06	71.58	71.58	72.39	**73.68**	73.54
Fairness	24.68	52.22	45.16	50.22	51.50	50.86	61.54	71.13	74.00	74.50	75.32	**75.48**	**75.48**
Cheating	0.00	0.00	0.00	0.00	0.00	0.00	0.00	21.05	51.85	51.85	56.14	**60.00**	**60.00**
Loyalty	18.29	44.53	41.49	43.87	43.59	44.22	47.65	59.15	62.82	63.75	63.75	63.95	**64.20**
Betrayal	0.00	0.00	10.00	20.00	20.00	20.00	18.18	34.78	66.67	66.67	68.42	**70.00**	**70.00**
Authority	0.00	30.93	30.19	33.10	35.53	33.96	45.52	55.29	62.50	65.91	67.78	69.23	**69.61**
Subversion	3.77	32.69	13.39	25.90	24.66	42.36	59.29	72.66	77.29	78.08	78.41	79.22	**79.61**
Purity	0.00	8.89	4.88	9.88	9.76	56.12	63.86	70.86	72.13	74.16	76.09	79.14	**80.41**
Degradation	2.99	15.38	9.52	10.00	10.00	8.00	20.69	52.94	61.54	61.54	68.09	**73.47**	**73.47**
Non-moral	0.00	0.00	1.60	3.51	12.70	12.31	54.55	71.14	80.90	81.82	82.35	82.54	**83.33**
Average	7.17	25.89	22.01	27.28	28.94	34.25	44.27	58.04	67.93	68.76	70.55	72.21	**72.49**

Table 7: F_1 Scores of PSL Models Using the Moral Foundations Dictionary (MFD). The highest prediction per moral foundation is marked in bold.

Moral Fdn.	Results of Non-Joint PSL Model Predictions												
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Care	7.29	29.72	30.51	30.86	30.62	35.66	46.41	54.17	61.77	62.16	62.91	64.79	**64.91**
Harm	2.25	8.89	19.31	21.89	26.18	26.09	37.28	52.40	62.18	62.18	63.74	64.67	**64.86**
Fairness	9.15	26.43	27.12	28.70	30.43	31.92	53.56	69.88	72.52	72.52	74.26	**74.63**	**74.63**
Cheating	4.76	13.33	25.45	25.45	38.71	39.34	40.68	51.61	62.16	62.16	64.94	**65.82**	**65.82**
Loyalty	2.61	19.66	23.85	25.10	27.31	29.57	38.06	47.73	54.30	55.22	55.59	57.34	**57.91**
Betrayal	0.00	0.00	0.00	6.25	12.12	11.76	18.18	28.57	60.47	60.47	62.22	**65.22**	**65.22**
Authority	13.59	40.19	48.40	51.82	56.25	56.14	57.04	63.30	66.45	66.67	67.32	**67.53**	**67.53**
Subversion	4.79	40.69	42.34	43.21	43.93	44.03	47.20	55.12	56.47	56.47	57.07	57.53	**57.65**
Purity	5.62	13.64	19.78	23.16	30.00	60.38	69.66	76.67	79.35	79.35	80.21	81.82	**82.52**
Degradation	16.66	31.37	37.74	44.83	51.61	51.61	57.14	68.75	73.53	73.53	77.33	**78.95**	**78.95**
Non-moral	28.78	52.99	60.48	61.33	64.72	66.00	73.62	79.41	82.25	82.25	82.55	82.78	**83.20**
Average	8.68	25.17	30.45	32.96	37.44	41.14	48.98	58.87	66.50	66.63	68.01	69.19	**69.38**

Table 8: F_1 Scores of PSL Models Using Annotator's Rationale (AR). The highest prediction per moral foundation is marked in bold.

Analysis of Joint Experiments: In addition to studying the effects of each feature on the models' ability to predict moral foundations, we also explored jointly predicting both policy frames and moral foundations. These tasks are highly related as shown by the large increase in score between the baseline and skyline measurements in Table 9 once frames are incorporated into the models.

Both moral foundations and frame classification are challenging multilabel classification tasks, the former using 11 possible foundations and the latter consisting of 17 possible frames. Furthermore, joint learning problems are harder to learn due to a larger numbers of parameters, which in turn also affects learning and inference.

Table 9 shows the macro-weighted average F_1 scores for three different models. The Baseline model shows the results of predicting only the Moral of the tweet using the non-joint model M13, which uses all features with frames initialized. The Joint model is designed to predict both the moral foundation and frame of a tweet simultaneously (as shown in Table 5), with no frame initialization. Finally, the Skyline model is M13 with all features, where the frames are initialized with their known values.

The joint model using AR unigrams outperforms the baseline, showing that there is some benefit to modeling moral foundations and frames together, as well as using domain-specific unigrams. However, it is unable to beat the MFD-based unigrams model. This is likely due to the large amount of noise introduced by incorrect frame predictions into the joint model. As expected, the joint model does not outperform the skyline model which is able to use the known values of the frames in order to accurately classify the moral foundations associated with the tweets.

Finally, the predictions for the frames in the joint model were quite low, going from an average F_1 score of 26.09 in M1 to an average F_1 score of 27.99 in M13. This likely has two causes: (1) frame prediction is a challenging 17-label classification task, with a random baseline of 6% (which our approach is able to exceed) and (2) the lower performance is because the frames are predicted with *no initialization*. In previous works, the frame prediction models are initialized with a set of unigrams expected to occur for each frame. Different from this approach, the only information our models provide to the frames are political party, issue, associated bigrams and trigrams, and the *predicted values for the moral foundations* from using this information. The F_1 score of 27.99 with such minimal initialization indicates that there is indeed a relationship between policy frames and the moral foundations expressed in tweets worth exploring in future work.

PSL MODEL	MFD	AR
BASELINE	55.49	55.88
JOINT	51.22	58.75
SKYLINE	72.49	69.38

Table 9: Overview of Macro-weighted Average F_1 Scores of Joint PSL Model M13. BASELINE is the MORAL prediction result. JOINT is the result of jointly predicting the MORAL and uninitialized FRAME predicates. SKYLINE shows the results when using all features with initialized frames.

6 Qualitative Results

Previous works (Makazhanov and Rafiei, 2013; Preoţiuc-Pietro et al., 2017) have shown the usefulness of moral foundations for the prediction of political party preference and the political ideologies of Twitter users. The moral foundation information used in these tasks is typically represented as word-level features extracted from the MFD. Unfortunately, these dictionary-based features are often too noisy to contribute to highly accurate predictions.

Recall the example tweets shown in Figures 1 and 2. Both figures are examples of tweets that are mislabeled by the traditional MFD-based approach, but correctly labeled using PSL Model M13. Using the MFD, Figure 1 is labeled as *Authority* due to "permit", the only matching unigram, while Figure 2 is incorrectly labeled as *Care*, even though there is one matching unigram for *Harm* and one for *Care*. To further demonstrate this point we compare the dictionary features to features extracted from the MORAL predictions of our PSL model.

Table 10 shows the results of using the different feature sets for the prediction of political affiliation of the author of a given tweet. All three models use moral information for prediction, but this information is represented differently in each of the models. The MFD model (line 1) uses the MFD unigrams to directly predict the political party of the author. The PSL model (line 2) uses the MF *prediction* made by the best performing model (M13) as features. Finally, the GOLD model (line 3) uses the actual MF annotations.

The difference in performance between the GOLD and MFD results shows that directly mapping the expected MFD unigrams to politicians' tweets is not informative enough for party affiliation prediction. However, by using abstract representations of language, the PSL model is able to achieve results closer to that which can be attained when using the *actual annotations* as features.

PSL MODEL	REP	DEM
MFD	48.72	51.28
PSL	61.25	66.92
GOLD	68.57	71.43

Table 10: Accuracy of Author Political Party Prediction. REP represents Republican and DEM represents Democrat.

7 Conclusion

Moral foundations and policy frames are employed as political strategies by politicians to garner support from the public. Politicians carefully word their statements to express their moral and social positions on issues, while maximizing their base's response to their message. In this paper we present PSL models for the classification of moral foundations expressed in political discourse on the microblog, Twitter. We show the benefits and drawbacks of traditionally used MFD unigrams and domain-specific unigrams for initialization of the models. We also provide an initial approach to the joint modeling of frames and moral foundations. In future works, we will exploit the interesting connections between moral foundations and frames for the analysis of more detailed ideological leanings and stance prediction.

Acknowledgments

We thank Lyle Ungar for his insight and discussion which inspired this work. We also thank the anonymous reviewers for their thoughtful comments and suggestions. This research was partly funded by a Google Focused Research Award.

References

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2015. Hinge-loss markov random fields and probabilistic soft logic. *arXiv preprint arXiv:1505.04406*.

Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor. 2013. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Proc. of UAI*.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proc. of ACL*.

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*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. The paraphrase database. In *Proc. of NAACL-HLT*.

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *IJCAI workshops*.

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 Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

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 and Dan Goldwasser. 2016. “all i know about politics is what i read in twitter”: Weakly supervised models for extracting politicians stances from twitter. In *Proceedings of COLING*.

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*.

I-Ta Lee, Mahak Goindani, Chang Li, Di Jin, Kristen Johnson, Xiao Zhang, Maria Pacheco, and Dan Goldwasser. 2017. Purduenlp at semeval-2017 task 1: Predicting semantic textual similarity with paraphrase and event embeddings. In *Proc. of SemEval*.

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*.

Aibek Makazhanov and Davood Rafiei. 2013. Predicting political preference of twitter users. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 298–305, New York, NY, USA. ACM.

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*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Ferran Pla and Lluís F Hurtado. 2014. Political tendency identification in twitter using sentiment analysis techniques. In *Proc. of COLING*.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 729–740.

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

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

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*.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 647–653.