

# Acquiring Background Knowledge to Improve Moral Value Prediction

Ying Lin<sup>1</sup>, Joe Hoover<sup>2</sup>, Morteza Dehghani<sup>2,3</sup>, Marlon Mooijman<sup>2</sup>, Heng Ji<sup>1</sup>

<sup>1</sup> Computer Science Department,  
Rensselaer Polytechnic Institute, Troy, NY, USA  
{liny9, jih}@rpi.edu

<sup>2</sup> Department of Psychology,  
University of Southern California, Los Angeles, CA, USA  
{jehoover, mdehghan, mooijman}@usc.edu

<sup>3</sup> Department of Computer Science,  
University of Southern California, Los Angeles, CA, USA

## Abstract

In this paper, we address the problem of detecting expressions of moral values in tweets using content analysis. This is a particularly challenging problem because moral values are often only implicitly signaled in language, and tweets contain little contextual information due to length constraints. To address these obstacles, we present a novel approach to automatically acquire background knowledge from an external knowledge base to enrich input texts and thus improve moral value prediction. By combining basic text features with background knowledge, our overall context-aware framework achieves performance comparable to a single human annotator. To the best of our knowledge, this is the first attempt to incorporate background knowledge for the prediction of implicit psychological variables in the area of computational social science.

## 1 Introduction

Moral values are principles that define right and wrong for a given individual. They influence decision making, social judgments, motivation, and behavior and are thought of as the glue that holds society together (Haidt, 2012). However, moral values are not universal, and disagreements about what is moral or sacred can give rise to seemingly intractable conflicts (Dehghani et al., 2010; Ginges et al., 2007). Accordingly, public demonstrations and protests often involve moral conflicts between different groups. For example, as Figure 1 shows, during the 2015 Baltimore protests<sup>1</sup>, users posted their viewpoints about this event on Twitter,

demonstrating divergent and even opposite moral values.

Detecting moral values in user-generated content not only can provide insight into these conflicts but also inform applications that aim to model social phenomena such as voting behavior and public opinions. For example, (Koleva et al., 2012) shows that moral concerns play an important role in one’s attitude and ideological position across a wide range of issues, such as abortion and same-sex marriage. Moral values have also been used to investigate various political attitudes in the United States. Liberals and conservatives attend to different moral intuitions (Graham et al., 2009): Liberals focus on the notions of Harm and Fairness, while conservatives attend to ideas of Loyalty to in-group members, Authority, and Purity.

God bless Freddie Gray. He is changing the country and making us address issues that will make America better. #FreddieGray #baltimore

During the #BaltimoreUprising there was SOME isolated "rioting," however labeling the whole thing as such is patently dishonest.

The mayor of Baltimore should be arrested for false imprisonment, because she busted murderers? You're nuts. #FreddieGray

Figure 1: Tweets related to 2015 Baltimore Protest.

In this work, we predict the moral values expressed in social media text via a suite of Natural Language Processing (NLP) techniques. A given text can contain any one or more moral values, as defined by Moral Foundation Theory (MFT, elaborated in Section 2) (Graham et al., 2013), or it can be *non-moral*. In previous work, computational linguistic measurements of latent attributes such as moral values, personality, and political orientation have primarily relied on textual features directly derived from target texts; these features have ranged from  $n$ -grams, word embed-

<sup>1</sup>[https://en.wikipedia.org/wiki/2015\\_Baltimore\\_protests](https://en.wikipedia.org/wiki/2015_Baltimore_protests)

dings, emoticons, to word categories (Rao et al., 2010; Tumasjan et al., 2010; Golbeck et al., 2011; Conover et al., 2011; Schwartz et al., 2013; Dehghani et al., 2014, 2016). While such approaches can yield powerful representations of text, they fall far short of human representation, which is greatly enhanced by the capacity to actively acquire background knowledge for reasoning and prediction. In the domain of moral value detection, the capacity for external knowledge integration is particularly important. For example, consider the tweet shown in Figure 2. A reader who has no knowledge of “Westboro Baptist” could look it up and learn that it is a church known for anti-LGBT and racist hate speech. This reader might then infer that this tweet conveys moral values concerning Purity/Degradation and Fairness/Cheating. Conversely, an algorithm that lacks access to background knowledge would be unable to exploit this information-rich indicator. Accordingly, we apply Entity Linking (EL) to identify entities in tweets, link them to an external knowledge-base (KB; Wikipedia in this work), and acquire their abstract descriptions and properties. From the background knowledge, we extract words showing a strong correlation with each moral foundation as additional discriminative features to improve the prediction.

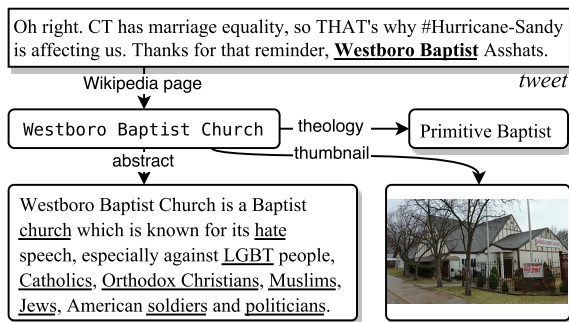


Figure 2: Example of Westboro Baptist Church.

Overall, this paper makes the following contributions:

1. We introduce various NLP techniques, such as entity linking and part-of-speech tagging, to tackle the problem of moral value prediction, which provides a new insight into the inference of latent semantic attributes in social media.
2. In the area of computational social science, most previous work involving applications of NLP to psychological measurement has relied exclusively on features derived directly from input text.

Due to the brevity and informality of tweets, however, textual features alone may not be sufficient for high-quality prediction. To address this issue, we acquire and incorporate background knowledge into our language models in order to better represent tweets and we use moral value prediction as case study for this approach. To the best of our knowledge, this is the first work to actively acquire background knowledge to enrich contextual information for more precise prediction in computational social science applications.

## 2 Moral Foundation Theory

What a given person holds to be moral or immoral can vary widely as a function of individual differences, and contextual and cultural factors. Moral Foundations Theory (Graham et al., 2013)<sup>2</sup> aims to explain this variability as a function of five core moral factors or foundations that appear across cultures, as shown in Table 1. These foundations account for various aspects of morality that serve different but related social functions. Degree of sensitivity towards them vary across different cultures and can change over time.

Foundation	Definition
Care Harm	Prescriptive moral values such as caring for others, generosity and compassion and moral values prohibiting actions that harm others.
Fairness Cheating	Prescriptive moral values such as fairness, justice, and reciprocity and moral values prohibiting cheating.
Loyalty Betrayal	Prescriptive moral values associated with group affiliation and solidarity and moral values prohibiting betrayal of one's group.
Authority Subversion	Prescriptive moral values associated with fulfilling social roles and submitting to authority and moral values prohibiting rebellion against authority.
Purity Degradation	Prescriptive moral values associated with the sacred and holy and moral values prohibiting violating the sacred.

Table 1: Moral foundation definitions.

Given the importance of human morality for social functioning (Haidt, 2012), it is perhaps unsurprising that our moral values leave residue in cultural artifacts such as texts. Indeed, research indicates that variation in moral rhetoric can reliably distinguish between cultural groups (Graham et al., 2009), is responsive to environmental disturbances such as terrorism (Sagi and Dehghani, 2014), and predicts psychologically relevant behavior (Dehghani et al., 2016).

<sup>2</sup><http://moralfoundations.org/>

While classifying the ground-truth moral content of a text is ultimately subjective and imperfect, general sentiment associated with the foundations above has been shown to be a sufficient proxy for models making secondary predictions (Graham et al., 2009; Sagi and Dehghani, 2014; Garten et al., 2016; Dehghani et al., 2016).

In Table 2, we list real tweets on the topic of Hurricane Sandy extracted from our data set that reflect each of the five foundations.

Foundation	Example
Care Harm	Loss of material things <b>hurts</b> but <b>loss</b> of people and pets is <b>devastating</b> Sending <b>prayers</b> to all who were affected by Sandy
Fairness Cheating	Complicit lap dog <b>biased</b> corrupt media is saying Obama has done good job w Sandy WHAT <b>LIES</b> Organization & Distribution get double F s
Loyalty Betrayal	Love my <b>fellow brothers and sisters</b> in New Jersey [sic] And <b>fellow Americans</b> standing strong as a nation Sandy please donate to local shelters
Authority Subversion	I maintain a profound <b>respect for govchristie</b> newjersey sandy AT.USER humanitarian
Purity Degradation	<b>God bless</b> these men Truly touched by their dedication AT.USER genTN Tomb guards Incredible Sandy

Table 2: Tweets reflecting each of the foundations.

### 3 Approach Overview

In this work, our goal is to predict the moral values expressed in social media text based on the Moral Foundation Theory via a suite of Natural Language Processing techniques. For example, moral values on Care/Harm and Purity/Degradation are expected to be detected from the following tweet – “*The Lord our Shepherd will keep & protect everyone on the East Coast Apply wisdom & be safe. Listen to the Spirits nudge. Love you. #Sandy*”. Thus, we define the Moral Value Prediction problem as follows:

DEFINITION: Given a set of documents  $\mathcal{X} = \{x_1, \dots, x_n\}$  regarding a certain topic and a set of moral foundations  $\mathcal{F} = \{f_1, \dots, f_m\}$ , for each  $x \in \mathcal{X}$ , return a binary vector  $\mathbf{y} = \{y_1, \dots, y_m\}$ , where  $y_j$  indicates whether  $x$  reflects concern on  $f_j$ .

#### 3.1 Framework

Figure 3 depicts the overall framework. In the *Textual Feature Extraction* module, textual features are extracted from the tweet and encoded into separate vectors. At the beginning of the *Background*

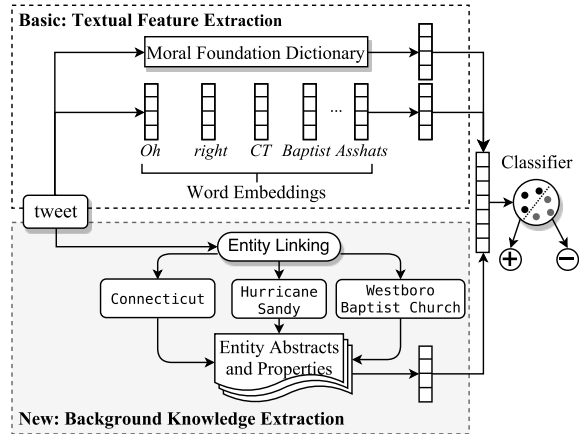


Figure 3: Overall framework.

*Knowledge Extraction* module, we apply entity linking to each tweet and acquire abstract descriptions and properties of linked entities from the KB. Next, we select information that is relevant to the target moral foundation from the prior knowledge and represent it as a vector. Lastly, all vectors are concatenated as input to a binary classifier.

We train a separate classifier that returns  $y_j$  for each foundation  $f_j$  and merge classification results from all classifiers as  $\mathbf{y}$ . A tweet will be predicted as “Non-moral” if all classifiers return False.

#### 3.2 Learning Model

In previous studies on predicting attributes such as gender, personality, power, and political orientation (Gomez et al., 2003; Burger et al., 2011; Schwartz et al., 2013; Park et al., 2015; Katerenchuk and Rosenberg, 2016), a document is usually modeled as a bag of words and represented by counting the frequency of each feature or aggregating embeddings of words. A major drawback to this approach is that bag-of-words models disregard word order and relationships between words that may serve as important information for classification. Consider the following tweets that mention “governor”:

- \* [AUTHORITY] Love our governor’s honesty #njsandy
- \* [FAIRNESS] Only 14 months till marriage #Equality comes to NJ, when @CoryBooker is sworn in as next governor.

In the first tweet, two positive words “love” and “honesty” around “governor” obviously reflect the user’s attitude towards him. In the second one, however, “governor” is not closely intertwined with other words and only modified by a

neutral word “next”. Because bag-of-words features ignore such context, the classifier may mistakenly assign Authority/Subversion to tweet 2 if “governor” is selected as a feature.

To address this issue, we experimented with various supervised learning models and found that the Recurrent Neural Network-based classifier with long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) performed the best. LSTM is a specific Recurrent NN variant designed to better model long-term dependencies. LSTM cells take as input a sequence of embeddings of words  $\{w_1, w_2, \dots, w_l\}$  in a tweet and output hidden states  $\{h_1, h_2, \dots, h_l\}$  in succession. We use the last output  $h_l$ , which is expected to encode information of the entire tweet, and handle it with a fully connected layer. Extra features including background knowledge are represented as separate vectors and processed using fully connected layers as well. Finally, we concatenate processed vectors and add a softmax layer on top for classification. To prevent overfitting, we apply dropout to outputs of the embedding, LSTM, and fully connected layers, and L2 regularization to the weight of the softmax layer. We train a separate classifier for each foundation and merge results from all classifiers.

### 3.3 Textual Features

In this paper, we use the following textual features.

**Word Embedding:** Word embedding is a dense distributed representation which embeds words to a low-dimensional space to encode their semantic and syntactic information. We use 300-dimensional Word2Vec embeddings trained on Google News<sup>3</sup>. We also carried out experiments using unigram and/or bigram as features. As they didn’t outperform word embedding and the latter is a common input to neural networks, we chose embeddings as the basic feature.

**Moral Foundation Dictionary:** Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) is a program that counts the proportion of words in different psychologically meaningful categories. Researchers have reported success applying LIWC to a range of social psychology problems (Pennebaker, 2011; Schwartz et al., 2013). In this work, we use Moral Foundation Dictionary (Graham et al., 2009), a LIWC dictionary that contains 324 foundation-supporting and foundation-violating words and word stems under

<sup>3</sup><https://code.google.com/archive/p/word2vec>

11 categories. It can be regarded as a kind of moral-oriented prior knowledge, while it is not as rich as the knowledge we propose to utilize.

## 4 Background Knowledge

Prior knowledge plays a critical role in how a human reader comprehends texts. As discussed above, background knowledge is also important in understanding expressions of moral concerns. For example, to perceive the Fairness/Cheating and Purity/Degradation-related moral concerns in the sentence “*we would also like to ban KKK*”, we need to know that “KKK” refers to Ku Klux Klan, hate groups opposing the Civil Rights Movement and same-sex marriage.

### 4.1 Background Knowledge Acquisition

To incorporate background knowledge, we apply entity linking to associate mentions with their referent entities. Next, we develop a set of criteria to automatically remove or correct erroneous linking results based on their types, linking confidence scores, or part-of-speech tags. From the KB, we extract structured and unstructured information of remaining entities. We will elaborate each step with the example illustrated in Figure 4.

**Entity linking.** First, we identify and link mentions to entities in the KB using TAGME (Fragina and Scaiella, 2012), a system developed to link mentions to pertinent Wikipedia pages. We choose this tool because most entity linkers are designed for formal texts such as news articles, while TAGME is intended for short texts and includes a special mode to handle hashtags, usernames, and URLs in tweets. TAGME provides an open API<sup>4</sup>, which returns a response including identified mentions, offsets, confidence scores, and Wikipedia titles. For tweet 1 in Figure 4, the linker identifies “Booker”, “everything”, and “him” and associates them with “George William Booker”, “Everything (Michael Bubl  song)”, and “HIM (Finnish band)”.

**Result refinement.** TAGME not only annotates capitalized phrases, thereby covering more mentions in poorly composed tweets at the cost of aggressively identifying some non-name words as mentions, such as “everything” and “him” in this example. Additionally, the lack of information that contextualizes the mention stands an obstacle to entity disambiguation. For example,

<sup>4</sup><https://sobigdata.d4science.org/web/tagme/>



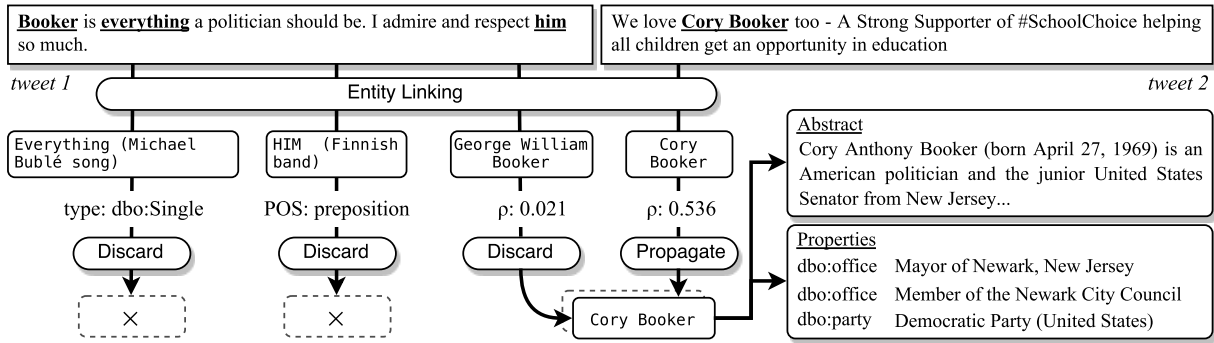


Figure 4: Acquiring background knowledge.

with the only clue - “politician” - in tweet 1, it is difficult to determine whether “Booker” refers to George William Booker or Cory Booker as both of them are politicians. To reduce these two types of errors, namely spurious annotations (“everything” and “him”) and linking errors (“Booker”→George William Booker), we refine the results based on the following attributes:

1. Linking confidence score. For each annotation, TAGME returns a confidence score ( $\rho$ ) that estimates the linking quality. We remove entities with a low score ( $< 0.1$ ) such as George William Booker ( $\rho = 0.021$ ) in the example.

2. Type of entity. Under most circumstances, concepts incorrectly linked to non-name words are literary or musical work entities, such as songs and books, which are more possibly titled using common words. We collect all entity types in DBpedia<sup>5</sup> and manually discard 113 types.

3. Part-of-speech. In general, a single verb, adjective, adverb, pronoun, determiner, or preposition is unlikely to be a name. Thus, “him” acting as a pronoun in tweet 1 should not be marked as a name. We utilize a tweet-oriented part-of-speech tagger (Owoputi et al., 2013) to annotate the part-of-speech of each word. If no word in a mention matches any nominal tag, we will remove the associated entity from the results.

**Cross-document propagation.** In the previous step, we present rules to reduce spurious annotations and linking errors. For the latter case, however, our goal is to leverage prior knowledge rather than merely eliminating incorrect entities. Therefore, if the linker returns an annotation with a low score, we reject it and try to retrieve the referent entity from annotations of the same mention in other documents. We make the following as-

sumption: within a topic, when people mention the same name, they usually refer to the same entity. For example, in tweets regarding architecture, it is very likely that all mentions of “Zaha” refer to Zaha Hadid, an architect, instead of Wilfried Zaha, a footballer. Analogously, as it is difficult to determine the referent entity of “Booker” in tweet 1, we check annotations of other “Booker”s in the entire corpus, find the most confident one (“Cory Booker” in tweet 2→Cory Booker,  $\rho = 0.536$ ), and use it as the entity of “Booker” in tweet 1.

**Knowledge extraction.** Unlike human beings, machines still lack the ability to process and comprehend complicated information (e.g., a man carries an American flag upside down in an image in the Westboro Baptist Church Wikipedia page, which can be viewed as a political statement or an act of desecration and disrespect) or disregard information contributing little to moral value prediction (e.g., population of New York State). For this reason, we only derive two types of constructive knowledge that can be processed and utilized by existing techniques and are applicable to most entities from the KB as follows.

1. Entity abstract: a summary of an entity, which usually contains useful facts such as definition, office, party, and purpose.

2. Entity property: structured metadata and facts of an entity. We obtain entity properties from DBpedia and keep the following: purpose, office, background, meaning, orderInOffice, seniority, title, and role.

## 4.2 Background Knowledge Incorporation

Many facts in the retrieved background knowledge, however, are irrelevant to the prediction of moral values (e.g., term of office and education information of Cory Booker). In addition, unlike tweets that are limited to 140 characters, an ab-

<sup>5</sup><http://wiki.dbpedia.org/>

stract can either be very concise or contain up to hundreds of words. In order to extract discriminative information from the background knowledge, we design a pointwise mutual information (PMI)-based approach as follows.

We first merge the abstract and property values of an entity into a document. For example, the merged document of Cory Booker is “*Cory Anthony Booker (Born April 27, 1969) is an American politician ... Mayor of Newark, New Jersey. Democratic Party (United States)*”. After that, we calculate the document-based PMI with corpus level significance (cPMId) (Damani, 2013) between each word in the document and the target foundation. We rank all words with respect to their cPMId’s and choose the top  $k$  ( $k = 100$  in our experiments). Thus, we extract words strongly related to each moral foundation as features.

Lastly, we encode the background knowledge as a vector consisting of  $k$  elements, each of which represents the term frequency of a selected word. However, consider this case where “hurt” is a feature, while “injury” is the word used in a document. We will ignore “injury” although it is a synonym for “hurt”. To encode the background knowledge in a “softer” and more generalizable way, we calculate the cosine similarity  $\text{sim}(u, w)$  between embeddings of feature  $u$  and each word  $w$ . If  $\text{sim}(u, w)$  exceeds a chosen threshold (0.6 in this work), we regard  $w$  as an occurrence of  $u$ .

## 5 Experiments

### 5.1 Data Set

For this work, we use a corpus of 4,191 tweets randomly sampled from a larger corpus of 7 million Tweets containing hashtags relevant to Hurricane Sandy, a hurricane that caused major damage to the Eastern seaboard of the United States in 2012. All tweets included in these analyses were processed to strip user mentions, URLs, and punctuation.

To establish ground truth for our analyses, three trained annotators coded the 4,191 sampled tweets<sup>6</sup>. Coder training consisted of multiple rounds of annotation and discussion. After completing training, annotators coded for the presence or absence of each moral foundation dimension. Additionally, tweets that contained no

<sup>6</sup>The data set and annotation guideline designed based on the Moral Foundation Theory will be published in a separate paper.

moral rhetoric were coded as “Non-moral”. Gold-standard classes for each tweet were then generated by taking the majority vote for each class across all three coders. Each tweet can be annotated with more than one moral concern at the same time.

Foundation	Positive	Negative	Pos:Neg
Care/Harm	1,802	2,389	1:1.33 (0.75)
Fairness/Cheating	667	3,524	1:5.28 (0.19)
Loyalty/Betrayal	574	3,617	1:6.30 (0.16)
Authority/Subversion	935	3,246	1:3.47 (0.29)
Purity/Degradation	159	4,032	1:25.4 (0.04)
Non-moral	713	3,478	1:4.88 (0.21)

Table 3: Data set statistics. Note that “positive” and “negative” do not refer to virtue and vice of a foundation. Rather, they indicate whether moral concern on a foundation (e.g., Fairness/Cheating) is reflected in a tweet or not.

Class frequency analyses of the coded corpus revealed considerable negative bias, such that the absence of each class occurred with greater frequency than its presence (See Table 3). However, this is unsurprising, as there is no reason to expect half or even close to half of the texts in this corpus to evoke a given moral domain. Nonetheless, extreme imbalance like this can inhibit classifier performance by inducing classification bias and failing to sufficiently represent the population of the infrequent class. To account for this in our experiments, we up-sample positive classes to prevent bias toward the majority class.

To evaluate the annotation quality of this corpus, we measure inter-annotator agreement (IAA) using prevalence-adjusted bias-adjusted kappa (PABAK) (Sim and Wright, 2005), which is suitable for imbalanced data. Based on the widely referenced standards for Kappa proposed in (Landis and Koch, 1977), IAA scores of this data set range from *moderate* (0.41 – 0.60) to *almost perfect* (0.81 – 1.00).

### 5.2 Overall Results

We evaluate our model with three feature sets: word embedding alone (E), the combination of word embedding and background knowledge (E+BK), and the combination of all features (E+BK+MFD). Model performance is evaluated using F-scores generated from 10-fold cross-validation in Figure 4.

Our experiment results provide evidence that integrating background knowledge into the repre-

Foundation	E	E+BK	E+BK+MFD
Care/Harm	81.2	<b>82.3</b>	81.9
Fairness/Cheating	66.1	70.7	<b>70.8</b>
Loyalty/Betrayal	47.2	<b>50.3</b>	<b>50.3</b>
Authority/Subversion	68.3	69.3	<b>69.9</b>
Purity/Degradation	34.7	<b>37.4</b>	37.0
Non-moral	61.7	<b>64.2</b>	63.5

Table 4: Overall results (% F-score). E, BK, and MFD represent embedding, background knowledge, and Moral Foundation Dictionary, respectively.

sentation of tweets improves detection of moral values. The following example demonstrates this process for a tweet which contains Authority/Subversion rhetoric:

\* [AUTHORITY] *Holy shit Chris Christie is asking for federal funds Sounds like a self hating republican to me hurricanesandy*

After linking “Chris Christie” and “republican” to Chris Christie and Republican Party (United States), we know the former is the 55th Governor of New Jersey and the latter a major political party in the United States. As our automatic approach selects politics-related words including “governor” and “party” as features, such background knowledge effectively confirms the moral sentiment on Authority/Subversion in this tweet.

In another example:

\* [PURITY] *Hurricane Sandy is an opportunity for believers to embody the perfect peace Isaiah 26 3 talks about as we trust in HIM hurricanesandy*

Although we successfully link “Isaiah” and use the prior knowledge to correct the prediction, the linker fails to associate “HIM” with God, which illustrates the limitations of existing techniques. Humans are able to make a quick inference about the referent of “HIM” from its distinct uppercase form because pronouns referring to God are often capitalized or uppercased. In contrast, it is difficult for machines to distinguish different “HIM”s (e.g., a common yet uppercased pronoun, a pronoun referring to God, the Finnish rock band, etc.), especially in poorly composed texts such as tweets.

It should also be noted that there seems to be a relationship between the prevalence of the positive class for a given dimension and performance for that dimension. For example, we observe that all models perform well on Care/Harm, for which the data is relatively balanced (See Table 3 and 4), while they produce particularly low scores on the

most imbalanced foundation, Purity/Degradation.

In addition, we observe that adding the Moral Foundation Dictionary does not further improve the performance if we have background knowledge. Our framework automates the extraction of the latter, hence saving much manual effort.

### 5.3 Comparison with the Human Annotator

While we have demonstrated the viability of our approach for classifying moral rhetoric, to truly evaluate the performance of these models it is necessary to compare them to human coder performance. To do this, we had a minimally trained fourth coder annotate a sample of 300 tweets and used both the coder’s annotations and the predictions from our model to predict moral concerns on these tweets. This enables us to compare the performance of the model to the performance of an independent human annotator.

Foundation	4th Coder	Our Model
Care/Harm	76.0	76.3
Fairness/Cheating	76.6	72.3
Loyalty/Betrayal	62.2	69.5
Authority/Subversion	68.5	67.8
Purity/Degradation	61.8	54.8
Non-moral	77.9	69.2

Table 5: A comparison of performance between human and our method (% F-score).

On most categories, our model performs comparably to the human annotator (see Table 5). Though, notably, the model again obtains a low score on Purity/Degradation. We also observe a large gap in the prediction of Non-moral, which may indicate that humans have a stronger ability to recognize tweets without moral content.

We also observe that although our model achieves comparable performance to the human annotator, the latter is superior in understanding deeper information in text to make inference. For example, in the following tweet:

\* [LOYALTY] *There needs to be a proper balance between individual responsibility and collective obligation Superstorm Sandy has shown us that*

Although “individual responsibility” and “collective obligation” are not typical words for Loyalty/Betrayal, a human reader is able to understand that the author’s concern on this foundation is reflected when discussing the balance between “individual responsibility” and “collective obligation”. The model, however, is unable to capture their relationship to make the correct prediction.

## 5.4 Remaining Challenges

Despite the effectiveness of our proposed model, we encounter some unsolved problems over the study period. We summarize the main remaining challenges as follows.

Tweets are often too short to provide contextual cues sufficient for entity disambiguation. For example, for the tweet “*Willard is a Frickin Lying Hypocrite*”, it is hard for the entity linking system to determine which entity “Willard” refers to. Additionally, tweets are often poorly composed and need to be normalized. People extensively use elements such as hashtags, abbreviations, slangs, and emoticons in tweets, which affects the performance of the entity linker and classifiers.

Further, knowledge from KBs is relatively static and limited. Consider the following tweet, “*Sandy could be God’s answer to Obama letting his countrymen die in Benghazi and then lying about it*”. The entity linker can easily link “Benghazi” to the Benghazi city. However, the real concerned knowledge is the attack against United States government facilities in Benghazi in 2012 instead of other facts, such as the population of the city, in the KB. To address this issue, we need to exploit more comprehensive knowledge of other types or from other sources, such as news and tweets.

Additionally, in this work, entity types to remove and property types to keep in the background knowledge extraction step are manually selected due to the limitation of data size. Manual selection may introduce individual biases and weaken generalization ability of the model on other corpora and domains. With enough occurrences of entity and property types, a number of automatic feature selection methods are applicable, such as mutual information, chi square, and information gain.

## 6 Related Work

Recently NLP techniques have been successfully applied to computational social science. Combined with social network analysis, textual content analysis has shown promise in applications such as prediction of moral value (Sagi and Dehghani, 2014; Dehghani et al., 2016), power (Gomez et al., 2003; Prabhakaran and Rambow, 2013; Katerenchuk and Rosenberg, 2016), expertise (Horne et al., 2016), leadership role (Tyshchuk et al., 2013), personality (Golbeck et al., 2011; Schwartz et al., 2013), gen-

der (Burger et al., 2011; Rao et al., 2010), hate speech (Waseem and Hovy, 2016; Nobata et al., 2016), and social interaction (Althoff et al., 2014; Tan et al., 2016). This work has extensively studied textual (e.g.,  $n$ -gram and LIWC) and structural features (e.g., Twitter relationships) on a variety of online platforms.

Nevertheless, to the best of our knowledge, our work is the first attempt to incorporate background knowledge through entity linking to enhance implicit content analysis in the area of computational social science. It should be noted that although there is a study on incorporating background knowledge into movie reviews classification by (Boghrati et al., 2015), their “background knowledge” refers to articles describing the target movies, which act like the Moral Foundation Dictionary whereas are completely different from the background knowledge we actively extract from the knowledge base.

## 7 Conclusions and Future Work

Moral value prediction is a critical task for predicting psychological variables and events. Using it as a case study, we demonstrate the importance of acquiring background knowledge for extracting implicit information through our new framework. Our framework can also be adapted for other implicit sentiment prediction tasks that are convertible to a multi-label classification problem, such as detecting personality types through text analysis (Goldberg, 1990).

In the future, we will exploit more up-to-date background knowledge from wider sources such as news articles. We also will detect specific moral value holders and target issues associated with each moral concern (e.g., women’s rights is the issue of the moral concern on Fairness/Cheating in “*oppression of women must be tackled*”). Moreover, we are interested in uniting moral value prediction with a variety of applications such as implicit community membership and leadership roles detection in social networks and event prediction.

We believe computational social science research can establish a bridge between NLP techniques and social science theories. We apply computational methods to analyze social phenomena supported by social theories, while more complex and accurate models can help theory adjudication in social science as well.



## References

- Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *ICWSM*.
- Reihane Boghrati, Justin Garten, Aleksandra Litvinova, and Morteza Dehghani. 2015. Incorporating background knowledge into text classification. In *CogSci*.
- John D. Burger, John C. Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *EMNLP*.
- Michael Conover, Jacob Ratkiewicz, Matthew R. Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *ICWSM*.
- Om P Damani. 2013. Improving pointwise mutual information (pmi) by incorporating significant co-occurrence. In *CoNLL*.
- Morteza Dehghani, Scott Atran, Rumen Iliev, Sonya Sachdeva, Douglas Medin, and Jeremy Ginges. 2010. Sacred values and conflict over iran’s nuclear program. *Judgment and Decision Making* 5(7):540.
- Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General* 145(3).
- Morteza Dehghani, Kenji Sagae, Sonya Sachdeva, and Jonathan Gratch. 2014. Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the “Ground Zero Mosque”. *Journal of Information Technology & Politics* 11(1):1–14.
- Paolo Ferragina and Ugo Scaiella. 2012. Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software* 29(1):70–75.
- Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *IJCAI*.
- Jeremy Ginges, Scott Atran, Douglas Medin, and Khalil Shikaki. 2007. Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Sciences* 104(18):7357–7360.
- Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from Twitter. In *SocialCom/PASSAT*.
- Lewis R Goldberg. 1990. An alternative “description of personality”: the big-five factor structure. *Journal of personality and social psychology* 59(6):1216.
- Daniel Gomez, Enrique González-Arangüena, Conrado Manuel, Guillermo Owen, Monica del Pozo, and Juan Tejada. 2003. Centrality and power in social networks: a game theoretic approach. *Mathematical Social Sciences* 46(1):27–54.
- Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology* 47:55–130.
- 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.
- Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- Benjamin D. Horne, Dorit Nevo, Jesse Freitas, Heng Ji, and Sibel Adali. 2016. Expertise in social networks: How do experts differ from other users? In *ICWSM*.
- Denys Katerenchuk and Andrew Rosenberg. 2016. Hierarchy prediction in online communities. In *AAAI*.
- Spassena P Koleva, Jesse Graham, Ravi Iyer, Peter H Ditto, and Jonathan Haidt. 2012. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality* 46(2):184–194.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *WWW*.
- Olutobi Owoputi, Brendan O’Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*.
- Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology* 108(6):934.
- James W Pennebaker. 2011. *The Secret Life of Pro-nouns: What Our Words Say About Us*. Bloomsbury Press.
- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001 .

- Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *IJC-NLP*.
- Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *SMUC*.
- Eyal Sagi and Morteza Dehghani. 2014. Measuring moral rhetoric in text. *Social science computer review* 32(2):132–144.
- H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE* 8(9):e73791.
- Julius Sim and Chris C. Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85(3):257–268.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *WWW*.
- Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welp. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM* 10(1):178–185.
- Yulia Tyshchuk, Hao Li, Heng Ji, and William A Wallace. 2013. Evolution of communities on twitter and the role of their leaders during emergencies. In *ASONAM*.
- Zeeraq Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL Student Research Workshop*.