Acquiring Background Knowledge to Improve Moral Value Prediction

Ying Lin
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY, USA
liny9@rpi.edu

Joe Hoover
Department of Psychology
University of Southern California
Los Angeles, CA, USA
jehoover@usc.edu

Gwenyth Portillo-Wightman
Department of Psychology
University of Southern California
Los Angeles, CA, USA
gportill@usc.edu

Christina Park
Department of Psychology
University of Southern California
Los Angeles, CA, USA
park415@usc.edu

Morteza Dehghani
Department of Psychology
and Department of Computer Science
University of Southern California
Los Angeles, CA, USA
mdehghan@usc.edu

Heng Ji
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY, USA
jih@rpi.edu

***Abstract*—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 textual features with background knowledge, our overall context-aware framework achieves performance comparable to a single human annotator. Our approach obtains 13.3% absolute F-score gains compared to our baseline model that only uses textual features.** [1]

***Index Terms*—Moral Value Prediction, Background Knowledge, Entity Linking, Natural Language Processing**

I. 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 binds society together [1]. However, moral values are not universal, and disagreements about what is moral or sacred can give rise to seemingly intractable conflicts [2], [3]. Accordingly, public demonstrations and protests often involve moral conflicts between different groups. For example, Table I shows several tweets posted during the 2015 Baltimore protests[2], a series of protests organized after Freddie Gray, black resident of Baltimore, died in police custody. Users posted their viewpoints about this event on Twitter, demonstrating divergent and even opposite moral values.

Moral Values	Tweet
Purity Degradation	*God bless* Freddie Gray. He is changing the country and making us address issues that will *make America better*. #FreddieGray #Baltimore
Fairness Cheating	During the #BaltimoreUprising there was SOME isolated "rioting," however labeling the whole thing as such is patently *dishonest*.
Authority Subversion	The mayor of Baltimore should be arrested for false imprisonment, because she busted murderers? You're nuts. #FreddieGray

TABLE I: Tweets related to 2015 Baltimore Protests.

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, [4] 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 stances in the United States. Liberals and conservatives attend to different moral intuitions [5]: Liberals focus on the notions of Harm and Fairness, while conservatives attend to ideas of Loyalty to in-group members, Authority, and Purity.

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 II) [6], 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 embeddings, emoticons, to word categories [7]–[13]. 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

[1]Our code is available at `https://github.com/limteng-rpi/mvp`
[2]`https://en.wikipedia.org/wiki/2015_Baltimore_protests`

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

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 1. 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.

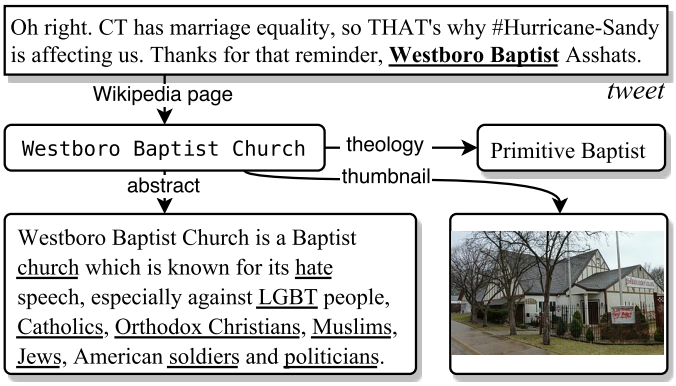

Fig. 1: Example of Westboro Baptist Church.

Overall, this paper makes the following contributions:

1. We introduce various NLP techniques, such as entity linking, 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. We use moral value prediction as a case study for this approach.

II. 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 [6] [3] aims to explain this variability as a function of five core moral factors or foundations that appear across cultures, as shown in Table II. These foundations account for various aspects of morality that serve different but related social functions. Further, degrees of sensitivity toward themvary 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 II: Moral foundation definitions.

Given the importance of human morality for social functioning [1], 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 [5], is responsive to environmental disturbances such as terrorism [14], and predicts psychologically relevant behavior [13].

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 [5], [13]–[15]. In Table III, 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 **hurts** but **loss** of people and pets is **devastating** Sending **prayers** to all who were affected by Sandy
Fairness Cheating	Complicit lap dog **biased** corrupt media is saying Obama has done good job w Sandy WHAT **LIES** Organization & Distribution get double F s
Loyalty Betrayal	Love my **fellow brothers and sisters** in New Jeersey [sic] And **fellow Americans** standing strong as a nation Sandy please donate to local shelters
Authority Subversion	I maintain a profound **respect for govchristie** new-jersey sandy AT_USER humanitarian
Purity Degradation	Everyone should unfollow AT_USER immediately Making hurricane jokes is pathetic and insensitive and **disgusting** sandy

TABLE III: Tweets reflecting each of the foundations.

III. Approach Overview

In this work, our goal is to predict the moral values expressed in social media text based on Moral Foundations Theory via a suite of Natural Language Processing techniques. For example, moral values of Care/Harm and Purity/Degradation

[3] http://moralfoundations.org

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:

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

A. Framework

As Figure 2 depicts, we propose a framework for moral value prediction that consists of two modules.

Textual Feature Extraction: The basis of the framework is a recurrent neural network that iterates through the given tweet and outputs a vector that carries textual features extracted from the input word embeddings. We will elaborate on the learning model in Section III-B. Additionally, we include word usage-related features using the Moral Foundation Dictionary and Linguistic Inquiry and Word Count.

Background Knowledge Extraction: To incorporate background knowledge, we utilize entity linking techniques to identify name mentions in the given tweet and associate them with their canonical entities in Wikipedia. For each tweet, we represent Wikipedia abstracts and properties of all linked entities in a single fixed-length vector.

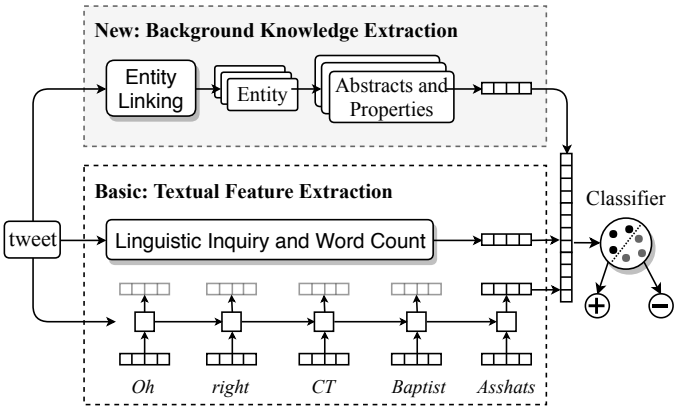

Fig. 2: Overall framework.

After that, we concatenate all feature vectors and feed them into a binary classifier. We train a separate classifier that returns y_j for each foundation f_j and merge classification results from all classifiers. A tweet will be predicted as "Non-moral" if none of the classifiers return `True`.

B. Learning Model

In previous studies on predicting attributes such as gender, personality, power, and political orientation [11], [16]–[19], 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) [20] performed the best. LSTM is a specific Recurrent Neural Network variant designed to better model long-term dependencies. LSTM cells take as input a sequence of embeddings of words $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_l\}$ in a tweet and output hidden states $\{\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_l\}$ in succession. The last output $\boldsymbol{h}_l$ of the LSTM layer is expected to encode key information of the entire tweet for moral value prediction. We concatenate it with additional feature vectors and feed them into a fully-connected network. On top of the model, we add a softmax layer to transform the neural network output into a probability distribution over target labels. To prevent overfitting, we apply Dropout [21] to outputs of the embedding, LSTM, and fully connected layers with a dropout rate of 0.5.

C. 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. Word embeddings have been shown to boost the performance on a range of NLP tasks, such as part-of-speech tagging, sentiment analysis, and semantic role labeling [22], [23].

Moral Foundation Dictionary (MFD): Linguistic Inquiry and Word Count (LIWC) [24] 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 [11], [25]. In this work, we use MFD [5], a LIWC dictionary that contains 324 foundation-supporting and foundation-violating words and word stems under 11 categories.

IV. 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.

A. Background Knowledge Acquisition

To incorporate background knowledge, we apply entity linking to associate mentions with their referent entities. We

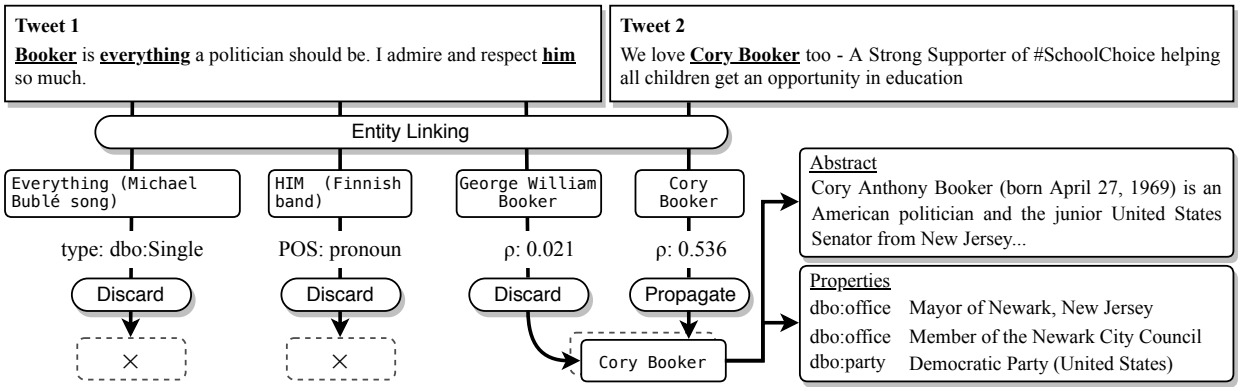

Fig. 3: Acquiring background knowledge.

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. Although many entity-related tasks only focus on limited types of entities: typically persons, locations and organizations, we consider widely defined entities, such as "`Social equality`", as they may provide crucial background information of the moral rhetoric involved in the tweet. We will elaborate each step of the background knowledge acquisition with the example illustrated in Figure 3.

Entity linking. First, we identify and link mentions to entities in the KB using TAGME [26], 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[4], which returns a JSON response including identified mentions, offsets, confidence scores, Wikipedia page titles, and Wikipedia abstract. For tweet 1 in Figure 3, the linker identifies "Booker", "everything", and "him" from the text and associates them with "`George William Booker`", "`Everything (Michael Bublé song)`", and "`HIM (Finnish band)`" respectively.

Result refinement. In order to cover more mentions in poorly composed tweets, TAGME annotates phrases regardless of letter case 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, 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 reports a confidence score (ρ) that estimates the linking quality. We remove entities with a low score (< 0.1 in our experiments[5]) such as `George William Booker` ($\rho = 0.021$) in the example.

2. Type of entity. We observe that, 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 (e.g., `Everything (Michael Bublé song)`). We collect all entity types in DBpedia 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", which acts as a pronoun in tweet 1, should not be marked as a name. We utilize a tweet-oriented part-of-speech tagger [27] 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 assumption: 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

[4]`https://sobigdata.d4science.org/web/tagme`

[5]$\rho \in [0.1, 0.3]$ is suggested by the official documentation.

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. In this work, we use abstracts returned by TAGME, which are derived from Wikipedia.

2. Entity property: structured metadata and facts of an entity, such as the title of a person. We obtain entity properties from DBpedia and select the following property types that are related to moral value: `purpose`, `office`, `background`, `meaning`, `orderInOffice`, `seniority`, `title`, and `role`.

B. Background Knowledge Incorporation

We first merge the abstract and properties of each linked entity into a single document. For example, the merged document of `Cory Booker` in Figure 3 is "*Cory Anthony Booker (Born April 27, 1969) is an American politician ... Mayor of Newark, New Jersey. Democratic Party (United States)*".

After that, we represent the background knowledge of each tweet as:

$$\boldsymbol{v} = \sum_{e \in \mathcal{E}} \frac{\rho_e}{|\mathcal{T}_e|} \sum_{t \in \mathcal{T}_e} \boldsymbol{w}_t,$$

where $\mathcal{E}$ is the set of linked entities, ρ_e is the confidence score of entity e estimated by TAGME, $\mathcal{T}_e$ is the set of tokens of the merged document of e, and $\boldsymbol{w_t}$ represents the word embedding of token t. In other words, we weight the representation of each entity using its linking confidence score.

Instead of directly concatenating the background knowledge vector $\boldsymbol{v}$ with other feature vectors, we process it with a fully-connected network. We introduce this network to extract only task-related features from the weighted average of word embeddings and reduce the dimension of the vector.

V. EXPERIMENTS

A. Data Sets

We evaluate the proposed framework two data sets:

Hurricane Sandy. 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.

Black Lives Matter. This data set contains $4,099$ tweets with hashtags related to the Black Lives Matter movement, such as "#BlackLivesMatter" and "#AllLivesMatter". Black Lives Matter is a movement against racialized violence towards black people.

To establish ground truth for our analyses, three trained annotators coded these two data sets. 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 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. In this work, we collapse virtue and vice into a single category (e.g., Care/Harm) as they are strongly related.

Hurricane Sandy			
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)

Black Lives Matter			
Foundation	**Positive**	**Negative**	**Pos:Neg**
Care/Harm	1,103	2,996	1:2.72 (0.37)
Fairness/Cheating	1,201	2,898	1:2.41 (0.41)
Loyalty/Betrayal	575	3,524	1:6.13 (0.16)
Authority/Subversion	485	3,614	1:7.45 (0.13)
Purity/Degradation	244	3,855	1:15.8 (0.06)
Non-moral	994	3,105	1:3.12 (0.32)

TABLE IV: 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 corpora revealed considerable negative bias, such that the absence of each class occurred with greater frequency than its presence (See Table IV). However, this is unsurprising, as there is no reason to expect half or even close to half of the texts in these corpora 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. In each training epoch, we randomly duplicate positive examples until both classes are balanced and shuffle the up-sampled data set.

To evaluate the annotation quality of these corpora, we measure inter-annotator agreement (IAA) using prevalence-adjusted bias-adjusted kappa (PABAK) [28], which is suitable for imbalanced data. Compared to Cohen's Kappa, PABAK is calculated only from the observed proportion of agreement between annotators. Based on the widely referenced standards for Kappa proposed in [29], IAA scores of these data sets range from *moderate* (0.41-0.60) to *almost perfect* (0.81-1.00).

For both data sets, we sample 80%, 10%, and 10% of all instances as the train, development, and test sets respectively.

B. Experimental Setup

We use pre-trained word embeddings in our experiments. For the Hurricane Sandy dataset, we train word embeddings from the complete corpus containing 7 million tweets using the word2vec package. For the Black Lives Matter dataset, we

use 100-dimensional GloVe embeddings trained from 2 billion tweets[6]. For the background knowledge representation, we use 100-dimensional word embeddings trained from English Wikipedia articles with word2vec. We optimize the model with Stochastic Gradient Descent with a learning rate of 0.005, batch size of 10, and momentum β of 0.9. We set the LSTM hidden state size to 100 and MFD and background knowledge feature dimensions to 5.

C. Overall Results

We evaluate our model with four feature sets: embedding alone (E), the combination of embedding and MFD (E+MFD), the combination of embedding and background knowledge (E+BK), and the combination of all features (E+BK+MFD). Model performance is evaluated using F-scores.

Hurricane Sandy				
Foundation	**E**	**E+MFD**	**E+BK**	**E+MFD+BK**
Care/Harm	79.6	80.4	**80.6**	78.7
Fairness/Cheating	55.6	60.3	**62.5**	60.7
Loyalty/Betrayal	65.6	66.1	**67.4**	66.9
Authority/Subversion	42.8	**50.7**	49.3	44.8
Purity/Degradation	27.3	32.7	**40.6**	36.2
Non-moral	49.5	49.7	**54.9**	52.8
Black Lives Matter				
Foundation	**E**	**E+MFD**	**E+BK**	**E+MFD+BK**
Care/Harm	74.8	74.1	**78.9**	75.0
Fairness/Cheating	86.7	86.0	87.1	**87.2**
Loyalty/Betrayal	91.9	93.1	**93.2**	93.1
Authority/Subversion	92.7	91.8	**94.9**	93.4
Purity/Degradation	84.7	85.0	**85.1**	84.5
Non-moral	77.5	78.3	**81.4**	80.5

TABLE V: Overall results (%, F-score). E and BK represent embedding and background knowledge respectively.

Table V shows that our model achieves much higher F-scores on the Black Lives Matter data set. As numbers of training examples are similar for both data sets, we think the inter-annotator agreement is a crucial factor that affects the performance. For the Hurricane Sandy data set, IAAs range from 0.45 to 0.82, while IAAs on the Black Lives Matter data set are higher than 0.80 except for the Care/Harm foundation. Generally, a higher IAA indicates better annotation consistency and less noise in the data.

Our experiment results provide evidence that integrating background knowledge into the representation of tweets improves detection of moral values. We calculate the significance of F-score differences using unpaired t-test. All differences are significant with p-values < 0.05.

The following example demonstrates this process for a tweet which contains Authority/Subversion rhetoric:

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

[6]https://nlp.stanford.edu/projects/glove

The baseline model fails to identify the moral sentiment on Authority/Subversion. 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. In their entity abstracts and properties, our model extract useful information from politics-related words such as "governor" and "party" to confirm the moral sentiment on Authority/Subversion in this tweet.

In another example:

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

Although the linker successfully links "Isaiah" and use the knowledge to correct the prediction, it fails to associate "HIM" with God, which shows the limitations of existing techniques. Humans are able to make a quick inference about the referent of "HIM" from its 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, a rock band), especially in poorly composed texts.

Unexpectedly, we observe that adding the Moral Foundation Dictionary does not further improve the performance if we have background knowledge. A possible reason is that with a limited amount of training data, increasing the number of features can lead to overfitting and thus hurt the performance on development sets.

D. Comparison with the Human Annotator

Foundation	**the 4th Human 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 VI: A comparison of performance between human and our method on the Hurricane Sandy dataset (%, F-score).

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 from the Hurricane Sandy data set and used both the coder's annotations and the predictions from our model to label moral concerns on these tweets. This enables us to compare the performance of the model to the performance of an independent human annotator. On most categories, our model performs comparably to the human annotator (see Table VI). We 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.

E. Remaining Challenges

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

Tweets are often too short to provide contextual cues sufficient for entity disambiguation. For example, in "*Willard is a Frickin Lying Hypocrite*", it is hard for the entity linker to determine which entity "Willard" refers to.

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 US government facilities in Benghazi in 2012 instead of other facts in the KB (e.g., population). To address this issue, we need to exploit more comprehensive knowledge from other sources such as news and tweets.

In this work, we manually select the entity types to remove and property types to keep in the background knowledge extraction step due to the limited 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.

Additionally, we only focus on English tweets in this paper. There are two obstacles to applying this framework to other languages: 1. Large-scale Knowledge Bases are not available for all languages. 2. Most EL systems are designed for English or a few languages. Without a mature entity linker, we are not able to identify and disambiguate named entities in text. Cross-lingual EL [30] is a possible solution to these issues.

We observe that a model trained on data in one domain usually does not work well on another domain. Therefore, we plan to incorporate cross-domain transfer techniques to reduce the annotation cost when migrating the model to new domains.

VI. RELATED WORK

Textual Content Analysis in Computational Social Science. 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 moral value prediction [13], [14], sentiment analysis [31]–[35], sarcasm detection [36]–[38], gender prediction [7], [17], hate speech detection [39], [40], personality prediction [9], [11], leadership role identification [41], expertise location [42], and social interaction analysis [43], [44]. Moral value prediction shares similarities with several NLP tasks, such as sentiment analysis and emotion detection, whereas it differs from them in the following respects: 1. We view moral value prediction as a multi-label classification problem instead of a multi-class classification problem. A tweet can hold moral rhetoric on more than one foundation at the same time. 2. Generally, moral values are conveyed in an implicit manner and closely related to attributes and events of the mentioned names. Therefore, moral value prediction has higher requirements for the background knowledge of the given tweet in order to construct the context and infer the involved moral rhetoric. 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 [45], their "background knowledge" refers to articles describing the target movies, which are different from the background knowledge we actively extract from the KB.

VII. 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 [46].

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*"). We also plan to integrate sarcasm detection to enhance text understanding due to the prevalence of irony in tweets. 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. Moreover, because users post tweets on diverse topics and in different languages on social media, we are migrating our framework to more languages and domains.

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 verification of social science theories as well. It is worth noting that the socio-psychological framework of Moral Foundations Theory is one of the many proposed frameworks aiming at explaining the full spectrum of

human moral reasoning. These frameworks are complimentary in a number of ways, but are also seen as competitors [47]. We believe computational methods, such as the method described in this paper, can help disentangle the differences between these frameworks.

VIII. Acknowledgements

This work was supported by the U.S. ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] J. Haidt, *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

[2] M. Dehghani, S. Atran, R. Iliev, S. Sachdeva, D. Medin, and J. Ginges, "Sacred values and conflict over iran's nuclear program," *Judgment and Decision Making*, 2010.

[3] J. Ginges, S. Atran, D. Medin, and K. Shikaki, "Sacred bounds on rational resolution of violent political conflict," *Proceedings of the National Academy of Sciences*, 2007.

[4] S. P. Koleva, J. Graham, R. Iyer, P. H. Ditto, and J. Haidt, "Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes," *Journal of Research in Personality*, 2012.

[5] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations," *Journal of personality and social psychology*, 2009.

[6] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," *Advances in Experimental Social Psychology*, 2013.

[7] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," in *SMUC*, 2010.

[8] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," *ICWSM*, 2010.

[9] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from Twitter," in *SocialCom/PASSAT*, 2011.

[10] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonalves, F. Menczer, and A. Flammini, "Political polarization on Twitter," in *ICWSM*, 2011.

[11] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, 2013.

[12] M. Dehghani, K. Sagae, S. Sachdeva, and J. Gratch, "Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the ground zero mosque," *Journal of Information Technology & Politics*, 2014.

[13] M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N. J. Parmar, S. Vaisey, R. Iliev, and J. Graham, "Purity homophily in social networks," *Journal of Experimental Psychology: General*, 2016.

[14] E. Sagi and M. Dehghani, "Measuring moral rhetoric in text," *Social science computer review*, 2014.

[15] J. Garten, R. Boghrati, J. Hoover, K. M. Johnson, and M. Dehghani, "Morality between the lines: Detecting moral sentiment in text," in *IJCAI*, 2016.

[16] D. Gomez, E. González-Arangüena, C. Manuel, G. Owen, M. del Pozo, and J. Tejada, "Centrality and power in social networks: a game theoretic approach," *Mathematical Social Sciences*, 2003.

[17] J. D. Burger, J. C. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," in *EMNLP*, 2011.

[18] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language." *Journal of personality and social psychology*, 2015.

[19] D. Katerenchuk and A. Rosenberg, "Hierarchy prediction in online communities," in *AAAI*, 2016.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[23] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.

[24] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, 2001.

[25] J. W. Pennebaker, *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press, 2011.

[26] P. Ferragina and U. Scaiella, "Fast and accurate annotation of short texts with wikipedia pages," *IEEE Software*, 2012.

[27] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters," in *NAACL*, 2013.

[28] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: Use, interpretation, and sample size requirements," *Physical Therapy*, 2005.

[29] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, 1977.

[30] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, "Cross-lingual name tagging and linking for 282 languages," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017.

[31] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *LREc*, 2010.

[32] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" *Icwsm*, vol. 11, no. 538-541, p. 164, 2011.

[33] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[34] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, 2014.

[35] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *NAACL HLT*, 2015.

[36] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter." in *ICWSM*, 2015.

[37] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on twitter: A behavioral modeling approach," in *WSDM*, 2015.

[38] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *ACL*, 2015.

[39] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *NAACL Student Research Workshop*, 2016.

[40] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *WWW*, 2016.

[41] Y. Tyshchuk, H. Li, H. Ji, and W. A. Wallace, "Evolution of communities on twitter and the role of their leaders during emergencies," in *ASONAM*, 2013.

[42] B. D. Horne, D. Nevo, J. Freitas, H. Ji, and S. Adali, "Expertise in social networks: How do experts differ from other users?" in *ICWSM*, 2016.

[43] T. Althoff, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "How to ask for a favor: A case study on the success of altruistic requests," in *ICWSM*, 2014.

[44] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee, "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions," in *WWW*, 2016.

[45] R. Boghrati, J. Garten, A. Litvinova, and M. Dehghani, "Incorporating background knowledge into text classification." in *CogSci*, 2015.

[46] L. R. Goldberg, "An alternative" description of personality": the big-five factor structure." *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[47] J. Graham, "Mapping the moral maps: From alternate taxonomies to competing predictions," *Personality and Social Psychology Review*, 2013.