Text-based inference of moral sentiment change

Jing Yi Xie*, **Renato Ferreira Pinto, Jr.***, **Graeme Hirst , and Yang Xu**

Department of Computer Science, University of Toronto, Toronto, Canada
jingyi.xie@mail.utoronto.ca, {renato,gh,yangxu}@cs.toronto.edu

Abstract

We present a text-based framework for investigating moral sentiment change of the public via longitudinal corpora. Our framework is based on the premise that language use can inform people's moral perception toward right or wrong, and we build our methodology by exploring moral biases learned from diachronic word embeddings. We demonstrate how a parameter-free model supports inference of historical shifts in moral sentiment toward concepts such as slavery and democracy over centuries at three incremental levels: moral relevance, moral polarity, and fine-grained moral dimensions. We apply this methodology to visualizing moral time courses of individual concepts and analyzing the relations between psycholinguistic variables and rates of moral sentiment change at scale. Our work offers opportunities for applying natural language processing toward characterizing moral sentiment change in society.

1 Moral sentiment change and language

People's moral sentiment—our feelings toward right or wrong—can change over time. For instance, the public's views toward *slavery* have shifted substantially over the past centuries (Oldfield, 2012). How society's moral views evolve has been a long-standing issue and a constant source of controversy subject to interpretations from social scientists, historians, philosophers, among others. Here we ask whether natural language processing has the potential to inform moral sentiment change in society at scale, involving minimal human labour or intervention.

The topic of moral sentiment has been thus far considered a traditional inquiry in philosophy (Hume, 1739; Smith, 1759; Kant, 1785), with contemporary development of this topic represented in social psychology (Piaget, 1932; Kohlberg, 1969; Stigler et al., 1990; Fiske and Taylor, 1991; Pizarro and Bloom, 2003), cognitive linguistics (Lakoff, 1996), and more recently, the advent of Moral Foundations Theory (Haidt and Joseph, 2004; Haidt et al., 2007; Graham et al., 2013). Despite the fundamental importance and interdisciplinarity of this topic, large-scale formal treatment of moral sentiment, particularly its evolution, is still in infancy from the natural language processing (NLP) community (see overview in Section 2).

We believe that there is a tremendous potential to bring NLP methodologies to bear on the problem of moral sentiment change. We build on extensive recent work showing that word embeddings reveal implicit human biases (Bolukbasi et al., 2016; Caliskan et al., 2017) and social stereotypes (Garg et al., 2018). Differing from this existing work, we demonstrate that moral sentiment change can be revealed by moral biases implicitly learned from diachronic text corpora. Accordingly, we present to our knowledge the first text-based framework for probing moral sentiment change at a large scale with support for different levels of analysis concerning *moral relevance*, *moral polarity*, and *fine-grained moral dimensions*. As such, for any query item such as *slavery*, our goal is to automatically infer its moral trajectories from sentiments at each of these levels over a long period of time.

Our approach is based on the premise that people's moral sentiments are reflected in natural language, and more specifically, in text (Bloom, 2010). In particular, we know that books are highly effective tools for conveying moral views to the public. For example, *Uncle Tom's Cabin* (Stowe, 1852) was central to the anti-slavery movement in the United States. The framework that we develop builds on this premise to explore

*Equal contribution.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4654–4663, Hong Kong, China, November 3–7, 2019.

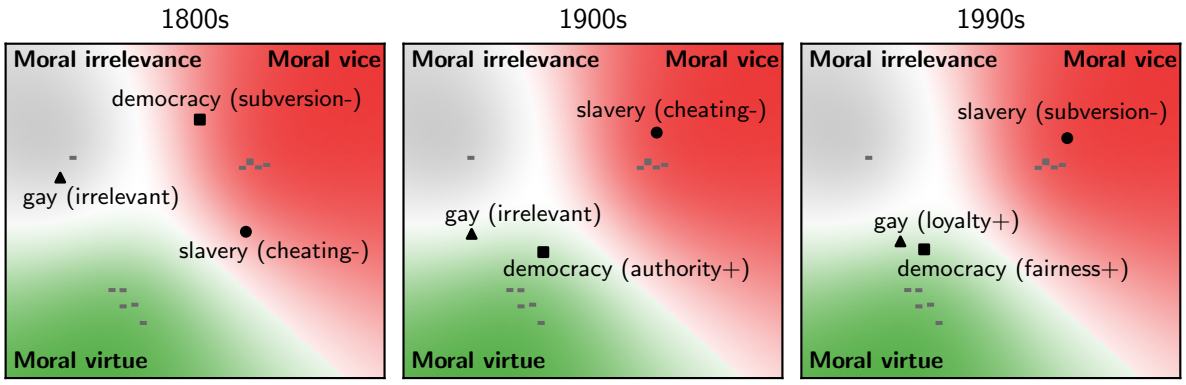

Figure 1: Illustration of moral sentiment change over the past two centuries. Moral sentiment trajectories of three probe concepts, *slavery*, *democracy*, and *gay*, are shown in moral sentiment embedding space through 2D projection from Fisher's discriminant analysis with respect to seed words from the classes of *moral virtue*, *moral vice*, and *moral irrelevance*. Parenthesized items represent moral categories predicted to be most strongly associated with the probe concepts. Gray markers represent the fine-grained centroids (or anchors) of these moral classes.

changes in moral sentiment reflected in longitudinal or historical text.

Figure 1 offers a preview of our framework by visualizing the evolution trajectories of the public's moral sentiment toward concepts signified by the probe words *slavery*, *democracy*, and *gay*. Each of these concepts illustrates a piece of "moral history" tracked through a period of 200 years (1800 to 2000), and our framework is able to capture nuanced moral changes. For instance, *slavery* initially lies at the border of moral virtue (positive sentiment) and vice (negative sentiment) in the 1800s yet gradually moves toward the center of moral vice over the 200-year period; in contrast, *democracy* considered morally negative (e.g., subversion and anti-authority under monarchy) in the 1800s is now perceived as morally positive, as a mechanism for fairness; *gay*, which came to denote homosexuality only in the 1930s (Kay et al., 2019), is inferred to be morally irrelevant until the modern day. We will describe systematic evaluations and applications of our framework that extend beyond these anecdotal cases of moral sentiment change.

The general text-based framework that we propose consists of a parameter-free approach that facilitates the prediction of public moral sentiment toward individual concepts, automated retrieval of morally changing concepts, and broad-scale psycholinguistic analyses of historical rates of moral sentiment change. We provide a description of the probabilistic models and data used, followed by comprehensive evaluations of our methodology.

2 Emerging NLP research on morality

An emerging body of work in natural language processing and computational social science has investigated how NLP systems can detect moral sentiment in online text. For example, moral rhetoric in social media and political discourse (Garten et al., 2016; Johnson and Goldwasser, 2018; Lin et al., 2018), the relation between moralization in social media and violent protests (Mooijman et al., 2018), and bias toward refugees in talk radio shows (Gillani and Levy, 2019) have been some of the topics explored in this line of inquiry. In contrast to this line of research, the development of a formal framework for moral sentiment change is still under-explored, with no existing systematic and formal treatment of this topic (Bloom, 2010).

While there is emerging awareness of ethical issues in NLP (Hovy et al., 2017; Alfano et al., 2018), work exploiting NLP techniques to study principles of moral sentiment change is scarce. Moreover, since morality is variable across cultures and time (Graham et al., 2013; Bloom, 2010), developing systems that capture the diachronic nature of moral sentiment will be a pivotal research direction. Our work leverages and complements existing research that finds implicit human biases from word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Garten et al., 2016) by developing a novel perspective on using NLP methodology to discover principles of moral sentiment change in human society.

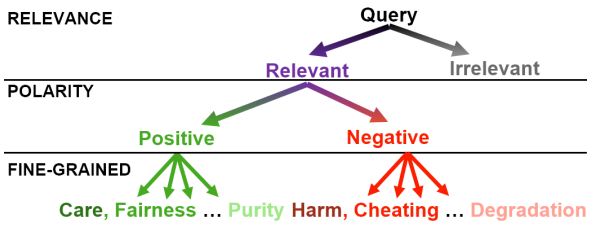

Figure 2: Illustration of the three-tier framework that supports moral sentiment inference at different levels.

3 A three-tier modelling framework

Our framework treats the moral sentiment toward a concept at three incremental levels, as illustrated in Figure 2. First, we consider moral relevance, distinguishing between morally irrelevant and morally relevant concepts. At the second tier, moral polarity, we further split morally relevant concepts into those that are positively or negatively perceived in the moral domain. Finally, a third tier classifies these concepts into fine-grained categories of human morality.

We draw from research in social psychology to inform our methodology, most prominently Moral Foundations Theory (MFT; Graham et al., 2013). MFT seeks to explain the structure and variation of human morality across cultures, and proposes five moral foundations: Care / Harm, Fairness / Cheating, Loyalty / Betrayal, Authority / Subversion, and Sanctity / Degradation. Each foundation is summarized by a positive and a negative pole, resulting in ten fine-grained moral categories.

3.1 Lexical data for moral sentiment

To ground moral sentiment in text, we leverage the Moral Foundations Dictionary (MFD; Graham et al., 2009). The MFD is a psycholinguistic resource that associates each MFT category with a set of *seed words*, which are words that provide evidence for the corresponding moral category in text. We use the MFD for moral polarity classification by dividing seed words into positive and negative sets, and for fine-grained categorization by splitting them into the 10 MFT categories.

To implement the first tier of our framework and detect moral relevance, we complement our morally relevant seed words with a corresponding set of seed words approximating moral irrelevance based on the notion of valence, i.e., the degree of pleasantness or unpleasantness of a stimulus. We refer to the emotional valence ratings collected by Warriner et al. (2013) for approximately 14,000 English words, and choose the words with most neutral valence rating that do not occur in the MFD as our set of morally irrelevant seed words, for an equal total number of morally relevant and morally irrelevant words.

3.2 Models

We propose and evaluate a set of probabilistic models to classify concepts in the three tiers of morality specified above. Our models exploit the semantic structure of word embeddings (Mikolov et al., 2013) to perform tiered moral classification of query concepts. In each tier, the model receives a query word embedding vector $\mathbf{q}$ and a set of seed words for each class in that tier, and infers the posterior probabilities over the set of classes c to which the query concept is associated with.

The seed words function as "labelled examples" that guide the moral classification of novel concepts, and are organized per classification tier as follows. In moral relevance classification, sets $\mathbf{S}_0$ and $\mathbf{S}_1$ contain the morally irrelevant and morally relevant seed words, respectively; for moral polarity, $\mathbf{S}_+$ and $\mathbf{S}_-$ contain the positive and negative seed words; and for fine-grained moral categories, $\mathbf{S}_1, \ldots, \mathbf{S}_{10}$ contain the seed words for the 10 categories of MFT. Then our general problem is to estimate $p(c \mid \mathbf{q})$, where $\mathbf{q}$ is a query vector and c is a moral category in the desired tier.

We evaluate the following four models:

- A Centroid model summarizes each set of seed words by its expected vector in embedding space, and classifies concepts into the class of closest expected embedding in Euclidean distance following a softmax rule;
- A Naïve Bayes model considers both mean and variance, under the assumption of independence among embedding dimensions, by fitting a normal distribution with mean vector and diagonal covariance matrix to the set of seed words of each class;
- A k-Nearest Neighbors (kNN) model exploits local density estimation and classifies concepts according to the majority vote of the k seed words closest to the query vector;
- A Kernel Density Estimation (KDE) model performs density estimation at a broader scale by considering the contribution of each seed word toward the total likelihood of each

Model	Parameter	Posterior Inference
Centroid	–	$p(c \mid \mathbf{q}) \propto \exp\left(-\left\|\mathbf{q} - \mathrm{E}\left[\mathbf{S}_c\right]\right\|\right)$
Naïve Bayes	–	$p(c \mid \mathbf{q}) \propto \prod_{j=1}^{d} f_N\left(\mathbf{q}_j; \mu = \mathrm{E}\left[\mathbf{S}_{c,j}\right], \sigma^2 = \mathrm{Var}\left[\mathbf{S}_{c,j}\right]\right)$
k-Nearest Neighbors (kNN)	k	$p(c \mid \mathbf{q}) \propto \left\lvert \{k \text{ nearest seed words to } \mathbf{q}\} \cap \mathbf{S}_c \right\rvert$
Kernel Density Estimation (KDE)	h	$p(c \mid \mathbf{q}) \propto \frac{1}{\lvert\mathbf{S}_c\rvert} \sum_{\mathbf{w} \in \mathbf{S}_c} f_{MN}\left(\mathbf{q}; \mu = \mathbf{w}, \Sigma = \mathrm{diag}(h)\right)$

Table 1: Summary of models for moral sentiment classification. Each model infers moral sentiment of a query word vector $\mathbf{q}$ based on moral classes c (at any of the three levels) represented by moral seed words $\mathbf{S}_c$. $\mathrm{E}\left[\mathbf{S}_c\right]$ is the mean vector of $\mathbf{S}_c$; $\mathrm{E}\left[\mathbf{S}_{c,j}\right], \mathrm{Var}\left[\mathbf{S}_{c,j}\right]$ refer to the mean and variance of $\mathbf{S}_c$ along the j-th dimension in embedding space. d is the number of embedding dimensions; and f_N, f_{MN} refer to the density functions of univariate and multivariate normal distributions, respectively.

class, regulated by a bandwidth parameter h that controls the sensitivity of the model to distance in embedding space.

Table 1 specifies the formulation of each model. Note that we adopt a parsimonious design principle in our modelling: both Centroid and Naïve Bayes are parameter-free models, kNN only depends on the choice of k, and KDE uses a single bandwidth parameter h.

4 Historical corpus data

To apply our models diachronically, we require a word embedding space that captures the meanings of words at different points in time and reflects changes pertaining to a particular word as diachronic shifts in a common embedding space.

Following Hamilton et al. (2016), we combine skip-gram word embeddings (Mikolov et al., 2013) trained on longitudinal corpora of English with rotational alignments of embedding spaces to obtain diachronic word embeddings that are aligned through time.

We divide historical time into decade-long bins, and use two sets of embeddings provided by Hamilton et al. (2016), each trained on a different historical corpus of English:

- Google N-grams (Lin et al., 2012): a corpus of 8.5×10^{11} tokens collected from the English literature (Google Books, all-genres) spanning the period 1800–1999.
- COHA (Davies, 2010): a smaller corpus of 4.1×10^{8} tokens from works selected so as to be genre-balanced and representative of American English in the period 1810–2009.

5 Model evaluations

We evaluated our models in two ways: classification of moral seed words on all three tiers (moral relevance, polarity, and fine-grained categories), and correlation of model predictions with human judgments.

5.1 Moral sentiment inference of seed words

In this evaluation, we assessed the ability of our models to classify the seed words that compose our moral environment in a leave-one-out classification task. We performed the evaluation for all three classification tiers: 1) moral relevance, where seed words are split into morally relevant and morally irrelevant; 2) moral polarity, where moral seed words are split into positive and negative; 3) fine-grained categories, where moral seed words are split into the 10 MFT categories. In each test, we removed one seed word from the training set at a time to obtain cross-validated model predictions.

Table 2 shows classification accuracy for all models and corpora on each tier for the 1990–1999 period.[1] We observe that all models perform substantially better than chance, confirming the efficacy of our methodology in capturing moral dimensions of words. We also observe that models using word embeddings trained on Google N-grams perform better than those trained on COHA, which could be expected given the larger corpus size of the former.

In the remaining analyses, we employ the Centroid model, which offers competitive accuracy and a simple, parameter-free specification.

5.2 Alignment with human valence ratings

We evaluated the approximate agreement between our methodology and human judgments using valence ratings, i.e., the degree of pleasantness or un-

[1] We also computed average accuracy over all decades using Google N-grams, the only corpus covering all moral categories through time. See Supplementary Material.

	Google N-grams			COHA		
Model	**Relevance**	**Polarity**	**Category**	**Relevance**	**Polarity**	**Category**
Random	0.50	0.50	0.10	0.50	0.50	0.10
Centroid	**0.84**	0.90	**0.59**	0.78	**0.80**	**0.40**
Naïve Bayes	**0.84**	0.89	0.53	0.76	0.78	0.39
1-NN	0.80	0.88	0.53	0.74	0.76	0.32
5-NN	0.83	**0.93**	0.57	0.74	0.75	0.33
KDE	0.82	0.90	0.57	**0.80**	0.76	0.33

Table 2: Classification accuracy of moral seed words for moral relevance, moral polarity, and fine-grained moral categories based on 1990–1999 word embeddings for two independent corpora, Google N-grams and COHA.

Corpus	**Correlation**
Google N-grams	0.43 ($n = 12293; p < 0.0001$)
COHA	0.38 ($n = 7141; p < 0.0001$)

Table 3: Pearson correlations between model predicted moral sentiment polarities and human valence ratings.

pleasantness of a stimulus. Our assumption is that the valence of a concept should correlate with its perceived moral polarity, e.g., morally repulsive ideas should evoke an unpleasant feeling. However, we do not expect this correspondence to be perfect; for example, the concept of *dessert* evokes a pleasant reaction without being morally relevant.

In this analysis, we took the valence ratings for the nearly 14,000 English nouns collected by Warriner et al. (2013) and, for each query word q, we generated a corresponding prediction of positive moral polarity from our model, $P(c_+ \mid \mathbf{q})$. Table 3 shows the correlations between human valence ratings and predictions of positive moral polarity generated by models trained on each of our corpora. We observe that the correlations are significant, suggesting the ability of our methodology to capture relevant features of moral sentiment from text.

In the remaining applications, we use the diachronic embeddings trained on the Google N-grams corpus, which enabled superior model performance throughout our evaluations.

6 Applications to diachronic morality

We applied our framework in three ways: 1) evaluation of selected concepts in historical time courses and prediction of human judgments; 2) automatic detection of moral sentiment change; and 3) broad-scale study of the relations between psycholinguistic variables and historical change of moral sentiment toward concepts.

6.1 Moral change in individual concepts

Historical time courses. We applied our models diachronically to predict time courses of moral relevance, moral polarity, and fine-grained moral categories toward two historically relevant topics: slavery and democracy. By grounding our model in word embeddings for each decade and querying concepts at the three tiers of classification, we obtained the time courses shown in Figure 3.

We note that these trajectories illustrate actual historical trends. Predictions for democracy show a trend toward morally positive sentiment, consistent with the adoption of democratic regimes in Western societies. On the other hand, predictions for slavery trend down and suggest a drop around the 1860s, coinciding with the American Civil War. We also observe changes in the dominant fine-grained moral categories, such as the perception of democracy as a fair concept, suggesting potential mechanisms behind the polarity changes and providing further insight into the public sentiment toward these concepts as evidenced by text.

Prediction of human judgments. We explored the predictive potential of our framework by comparing model predictions with human judgments of moral relevance and acceptability. We used data from the Pew Research Center's 2013 Global Attitudes survey (Pew Research Center, 2013), in which participants from 40 countries judged 8 topics such as *abortion* and *homosexuality* as one of "acceptable", "unacceptable", and "not a moral issue".

We compared human ratings with model predictions at two tiers: for moral relevance, we paired the proportion of "not a moral issue" human responses with irrelevance predictions $p(c_0 \mid \mathbf{q})$ for

Concept	Rate (e−03/decade)	Relevant Moral Category	Switching Period
abortion	4.24**	cheating−	1890
propriety	4.06***	fairness+	1870
commandment	3.68***	sanctity+	1880
righteousness	3.56***	sanctity+	1800
authorities	3.42***	authority+	1890
apostle	3.41***	authority+	1900
intervention	3.33**	authority+	1860
jew	3.32***	degradation−	1870
foreigner	3.30***	authority+	1800
individuality	3.26***	authority+	1860

Table 4: Top 10 changing words towards moral relevance during 1800–2000, with model-inferred moral category and switching period. *, **, and *** denote $p < 0.05$, $p < 0.001$, and $p < 0.0001$, all Bonferroni-corrected.

Concept	Rate (e−03/decade)	Early Category	Modern Category	Switching Period
wage	4.38**	subversion−	fairness+	1810
commitment	3.78	harm−	authority+	1830
innovation	3.21	authority+	authority+	1800
help	3.20**	sanctity+	care+	1880
mandate	3.17*	authority+	authority+	1800
guidance	3.04***	authority+	authority+	1800
licence	3.01*	authority+	authority+	1800
abortion	2.95	degradation−	cheating−	1890
democracy	2.93**	authority+	fairness+	1800
disclosure	2.89***	fairness+	fairness+	1840
propaganda	−7.75 *	authority+	subversion−	1910
humiliation	−4.05 ***	authority+	harm−	1800
seriousness	−4.02 **	authority+	harm−	1800
legacy	−3.76 ***	authority+	subversion−	1800
behavior	−3.73 **	harm−	authority+	1830
cheerfulness	−3.66 ***	authority+	sanctity+	1800
candour	−3.37 *	authority+	degradation−	1800
offense	−3.37	fairness+	subversion−	1840
indulgence	−3.37 ***	authority+	degradation−	1800
exertion	−3.26 ***	authority+	degradation−	1800

Table 5: Top 10 changing words towards moral positive (upper panel) and negative (lower panel) polarities, with model-inferred most representative moral categories during historical and modern periods and the switching periods. *, **, and *** denote $p < 0.05$, $p < 0.001$, and $p < 0.0001$, all Bonferroni-corrected for multiple tests.

each topic, and for moral acceptability, we paired the proportion of "acceptable" responses with positive predictions $p(c_+ \mid \mathbf{q})$. We used 1990s word embeddings, and obtained predictions for two-word topics by querying the model with their averaged embeddings.

Figure 4 shows plots of relevance and polarity predictions against survey proportions, and we observe a visible correspondence between model predictions and human judgments despite the difficulty of this task and limited number of topics.

6.2 Retrieval of morally changing concepts

Beyond analyzing selected concepts, we applied our framework predictively on a large repertoire of words to automatically discover the concepts that have exhibited the greatest change in moral sentiment at two tiers, moral relevance and moral polarity.

We selected the 10,000 nouns with highest total frequency in the 1800–1999 period according to data from Hamilton et al. (2016), restricted to words labelled as nouns in WordNet (Miller,

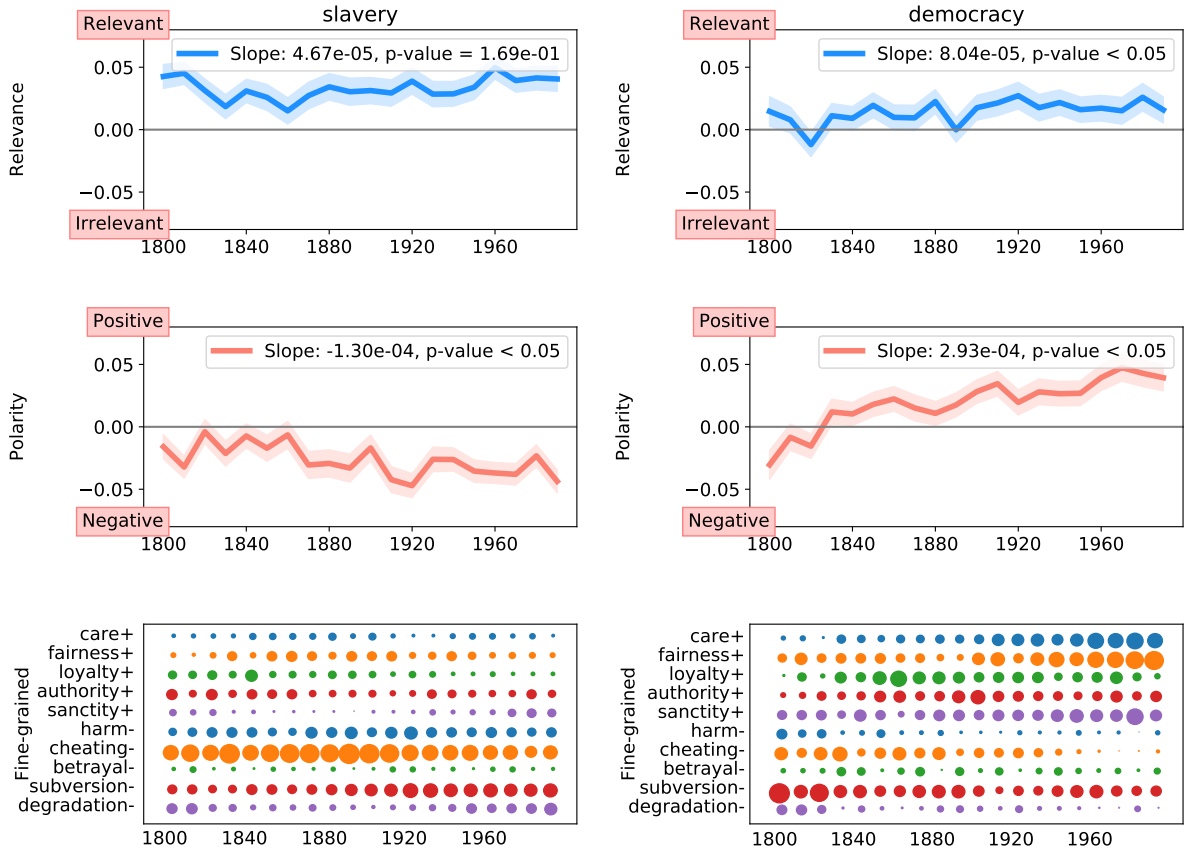

Figure 3: Moral sentiment time courses of *slavery* (left) and *democracy* (right) at each of the three levels, inferred by the Centroid model. Time courses at the moral relevance and polarity levels are in log odds ratios, and those for the fine-grained moral categories are represented by circles with sizes proportional to category probabilities.

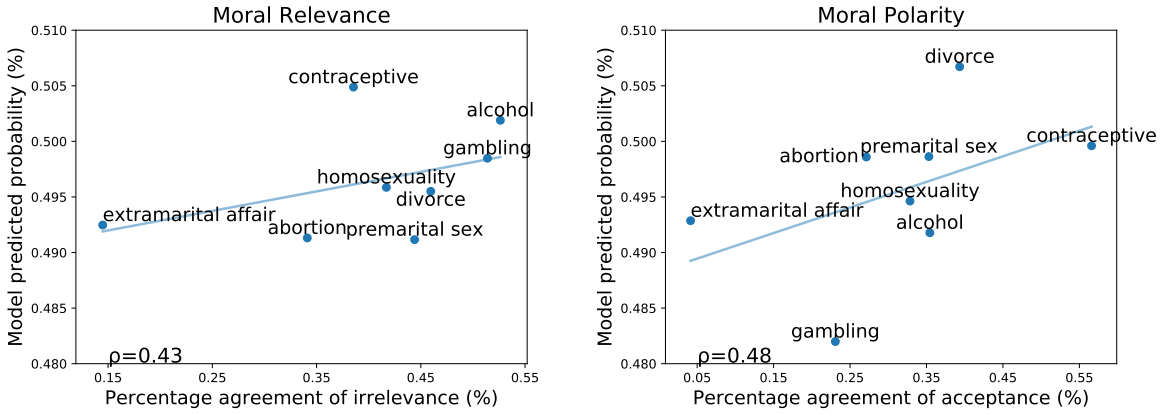

Figure 4: Model predictions against percentage of Pew respondents who selected "Not a moral concern" (left) or "Acceptable" (right), with lines of best fit and Pearson correlation coefficients ρ shown in the background.

1995) for validation. For each such word $\mathbf{q}$, we computed diachronic moral relevance scores $R_i = p(c_1 \mid \mathbf{q}), i = 1, \ldots, 20$ for the 20 decades in our time span. Then, we performed a linear regression of R on $T = 1, \ldots, n$ and took the fitted slope as a measure of moral relevance change. We repeated the same procedure for moral polarity. Finally, we removed words with average relevance score below 0.5 to focus on morally relevant retrievals.

Table 4 shows the words with steepest predicted change toward moral relevance, along with their predicted fine-grained moral categories in mod-

ern times (i.e., 1900–1999). Table 5 shows the words with steepest predicted change toward the positive and negative moral poles. To further investigate the moral sentiment that may have led to such polarity shifts, we also show the predicted fine-grained moral categories of each word at its earliest time of predicted moral relevance and in modern times. Although we do not have access to ground truth for this application, these results offer initial insight into the historical moral landscape of the English language at scale.

6.3 Broad-scale investigation of moral change

In this application, we investigated the hypothesis that concept concreteness is inversely related to change in moral relevance, i.e., that concepts considered more abstract might become morally relevant at a higher rate than concepts considered more concrete. To test this hypothesis, we performed a multiple linear regression analysis on rate of change toward moral relevance of a large repertoire of words against concept concreteness ratings, word frequency (a correlate of semantic change; see Hamilton et al., 2016), and word length (as a proxy for concept complexity; see Lewis and Frank, 2016).

We obtained norms of concreteness ratings from Warriner et al. (2013). We collected the same set of high-frequency nouns as in the previous analysis, along with their fitted slopes of moral relevance change. Since we were interested in moral relevance change within this large set of words, we restricted our analysis to those words whose model predictions indicate change in moral relevance, in either direction, from the 1800s to the 1990s.

We performed a multiple linear regression under the following model:

$$\tilde{\rho}(w) = \beta_f \log(f(w)) + \beta_l l(w) + \beta_c c(w) + \beta_0 + \tilde{\varepsilon} \quad (1)$$

Here $\rho(w)$ is the slope of moral relevance change for word w; $f(w)$ is its average frequency; $l(w)$ is its character length; $c(w)$ is its concreteness rating; β_f, β_l, β_c, and β_0 are the corresponding factor weights and intercept, respectively; and $\varepsilon \sim \mathcal{N}(0, \sigma)$ is the regression error term.

Table 6 shows the results of multiple linear regression. We observe that concreteness is a significant negative predictor of change toward moral relevance, suggesting that abstract concepts are more strongly associated with increasing moral relevance over time than concrete concepts. This

Factor	Coeff. (e-03)	Significance
Frequency	0.1	$p < 0.001$
Length	-0.03	n.s. ($\alpha = 0.05$)
Concreteness	-0.2	$p < 0.002$

Table 6: Results from multiple regression that regresses rate of change in moral relevance against the factors of word frequency, length, and concreteness ($n = 606$).

significance persists under partial correlation test against the control factors ($p < 0.01$).

We further verified the diachronic component of this effect in a random permutation analysis. We generated 1,000 control time courses by randomly shuffling the 20 decades in our data, and repeated the regression analysis to obtain a control distribution for each regression coefficient. All effects became non-significant under the shuffled condition, suggesting the relevance of concept concreteness for diachronic change in moral sentiment (see Supplementary Material).

7 Discussion and conclusion

We presented a text-based framework for exploring the socio-scientific problem of moral sentiment change. Our methodology uses minimal parameters and exploits implicit moral biases learned from diachronic word embeddings to reveal the public's moral perception toward a large concept repertoire over a long historical period.

Differing from existing work in NLP that treats moral sentiment as a flat classification problem (Garten et al., 2016; Johnson and Goldwasser, 2018), our framework probes moral sentiment change at multiple levels and captures moral dynamics concerning relevance, polarity, and fine-grained categories informed by Moral Foundations Theory (Graham et al., 2013). We applied our methodology to the automated analyses of moral change both in individual concepts and at a broad scale, thus providing insights into psycholinguistic variables that associate with rates of moral change in the public.

Our current work focuses on exploring moral sentiment change in English-speaking cultures. Future research should evaluate the appropriateness of the framework to probing moral change from a diverse range of cultures and linguistic backgrounds, and the extent to which moral sentiment change interacts and crisscrosses with linguistic meaning change and lexical coinage. Our

work creates opportunities for applying natural language processing toward characterizing moral sentiment change in society.

Acknowledgments

We would like to thank Nina Wang, Nicola Lacerata, Dan Jurafsky, Paul Bloom, Dzmitry Bahdanau, and the Computational Linguistics Group at the University of Toronto for helpful discussion. We would also like to thank Ben Prystawski for his feedback on the manuscript. JX is supported by an NSERC USRA Fellowship and YX is funded through a SSHRC Insight Grant, an NSERC Discovery Grant, and a Connaught New Researcher Award.

References

Mark Alfano, Dirk Hovy, Margaret Mitchell, and Michael Strube. 2018. *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics.

Paul Bloom. 2010. How do morals change? *Nature*, 464(7288):490.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Mark Davies. 2010. The Corpus of Historical American English: 400 million words, 1810–2009.

Susan T Fiske and Shelley E Taylor. 1991. *Social Cognition*. McGraw-Hill Book Company.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*.

Nabeel Gillani and Roger Levy. 2019. Simple dynamic word embeddings for mapping perceptions in the public sphere. *Computing Research Repository*, arXiv:1904.03352.

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. In *Advances in Experimental Social Psychology*, volume 47, pages 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 and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Jonathan Haidt, Craig Joseph, et al. 2007. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The Innate Mind*, 3:367–391.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the ACL*, pages 1489–1501. Association for Computational Linguistics.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach. 2017. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics.

David Hume. 1739. *A Treatise of Human Nature*. Clarendon Press.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Immanuel Kant. 1785. *Groundwork for the Metaphysics of Morals*. Oxford University Press.

Christian Kay, Jane Roberts, Michael Samuels, Iren Wotherspoon, and Marc Alexander (eds.). 2019. *The Historical Thesaurus of English*. Glasgow: University of Glasgow. Version 4.21.

Lawrence Kohlberg. 1969. *Stage and Sequence: The Cognitive-developmental Approach to Socialization*. Rand McNally.

George Lakoff. 1996. *Moral politics: What conservatives know that liberals don't*. University of Chicago Press, Chicago.

Molly L Lewis and Michael C Frank. 2016. The length of words reflects their conceptual complexity. *Cognition*, 153:182–195.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve

moral value prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559. IEEE.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, 2(6):389.

John R Oldfield. 2012. *Popular Politics and British Anti-Slavery: The mobilisation of public opinion against the slave trade 1787-1807*. Routledge.

Pew Research Center. 2013. Global Views on Morality. Pew Research Center's Global Attitudes Project.

Jean Piaget. 1932. *The Moral Judgment of the Child*. London: Routledge and Kegan Paul.

David A. Pizarro and Paul Bloom. 2003. The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review*, 110(1):193–196.

Adam Smith. 1759. *The Theory of Moral Sentiments*. History of Economic Thought Books. McMaster University Archive for the History of Economic Thought.

James W Stigler, Richard A Schweder, and Gilbert Herdt. 1990. *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press.

Harriet Beecher Stowe. 1852. *Uncle Tom's Cabin*. Tauchnitz.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.