Morality Between the Lines: Detecting Moral Sentiment In Text

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M. Johnson, Morteza Dehghani
University of Southern California, Los Angeles, CA
jgarten, boghrati, jehoover, katejohn, mdehghan@usc.edu

Abstract

Expressions of moral sentiment play a fundamental role in political framing, social solidarity, and basic human motivation. Moral rhetoric helps us communicate the reasoning behind our choices, how we feel we should govern, and the communities to which we belong. In this paper, we use short-post social media to compare the accuracy of text analysis methods for detecting moral rhetoric and longer form political speeches to explore detecting shifts in that rhetoric over time. Building on previous work using word count methods and the Moral Foundations Dictionary [Graham *et al.*, 2009], we make use of pre-trained distributed representations for words to extend this dictionary. We show that combining the MFD with distributed representations allows us to capture a cleaner signal when detecting moral rhetoric, particularly with short-form text. We further demonstrate how the addition of distributed representations can simplify dictionary creation. Finally, we demonstrate how capturing moral rhetoric in text over time opens up new avenues for research such as assessing when and how arguments become moralized and how moral rhetoric impacts subsequent behavior.

1 Introduction

Communication plays a crucial role as the primary window onto a range of internal psychological processes. From self-report surveys and interviews to observations of public discourse, most psychological studies rely on language to capture the underlying attitudes and beliefs which drive human action.

However, the flexibility and range of meanings underlying even the most basic of human communication complicates the clear interpretation of language for research purposes. Sometimes, social desirability concerns lead people to specifically try to obscure or deflect others' perceptions of their underlying attitudes. Even when attempting to answer truthfully, survey and interview question wording, context, and framing all strongly influence how people interpret researchers' meaning and how they answer [Schwarz, 1999]. Research using naturally occurring communication is able to avoid concerns of researcher language influence, however, the lack of restricted and detailed contextual information introduces additional interpretation difficulties. The same utterance can be used to convey a fact, a promise, an order, or signal connection depending on who is speaking to whom [John, 1962].

While much of social scientific research has historically focused on understanding language within survey and experimental methodologies, the explosion of naturally-occurring communication available to researchers through the Internet and social media has led to renewed interest in understanding how to assess text for capturing social and psychological phenomena and modeling the processes which drive them.

However, the choice of topics has been influenced by the varying accessibility of annotated data. Sentiment analysis has been driven by the ease of combining raw reviews with discrete signals such as star ratings [Pang and Lee, 2005]. Other work has focused around contexts where users have filled out questionnaires on topics such as personality [Park *et al.*, 2015].

However, in natural language, humans rarely stop to annotate their utterances. We don't end an email by rating how we're feeling or what beliefs contributed to our choice of words. Often times we aren't even aware of the latent factors which drive our choices. This lack of targets to optimize against has limited the applicability of many of the supervised learning techniques which have become commonplace in areas like sentiment analysis. Further, the theory-blind aspect of these methods has made it difficult to derive even correlational interactions with underlying theories.

In contrast, theory-driven studies have made extensive use of psychological dictionaries [Stone *et al.*, 1968; Pennebaker *et al.*, 2001] which are often applied through word count methods. While some of these have been developed for specific domains, a great deal of this work has been driven by application of the pre-existing dictionaries, in particular those available through the LIWC package [Tausczik and Pennebaker, 2010].

Recent work in the natural language processing (NLP) community has increasingly focused on data-driven rather than theory-driven methods. Explicitly chosen features have largely been replaced with feature learning methods and bottom-up processing. However, a strong line of theory-driven work has continued, driven in large part by increasing social science interest in applications of big data [Dehghani *et al.*, 2016a].

Increasingly, NLP techniques such as topic modeling [Blei *et al.*, 2003] have been incorporated into the analysis of social behavior [Mitchell *et al.*, 2015]. Of particular interest has been work which has combined textual measures with longstanding cognitive theories [Dehghani *et al.*, 2013]. Such work has been applied to domain areas including personal values [Boyd *et al.*, 2015], social characteristics [Iliev and Smirnova, 2014], general personality [Park *et al.*, 2015], and dark triad personality [Garcia and Sikström, 2014]. In addition to direct prediction of underlying states, such work has also explored outcomes such as heart disease [Eichstaedt *et al.*, 2015], black mortality [Chae *et al.*, 2015], depression [Schwartz *et al.*, 2014], and suicide [Kumar *et al.*, 2015].

In this paper, we consider the task of quantifying the presence of one such latent psychological factor, the presence of moral sentiment around a given topic. Moral values and judgments are central to social decision making and cultural cohesion [Dehghani *et al.*, 2016a; Kaplan *et al.*, 2016]. Culture wars over social issues and intergroup violence can be traced to differences in moral concerns [Koleva *et al.*, 2012], and perceptions of moral value fit affect communal participation [Johnson *et al.*, 2014; Motyl *et al.*, 2014].

Moral rhetoric is often used to communicate normative prescriptions for behaviors in which we *should* engage or normative proscriptions for behaviors in which we *should not*. While not all calls to action include a moral argument ("Go clean your room" may be associated with purity concerns, but those concerns aren't part of the argument itself), when present, expressions of moral relevance can be a powerful motivator for personal and collective action.

2 Detecting Moral Rhetoric

Detecting moral rhetoric is a critical step to understanding why it is being used and what effects those uses have on recipients. To date, much of the research on moral rhetoric has relied on word count methods using the Moral Foundations Dictionary (MFD) [Haidt *et al.*, 2009] which was created to capture the five foundations laid out in Moral Foundations Theory (MFT) [Graham *et al.*, 2009]. For example, this approach has been used to compare the content of sermons delivered in liberal and conservative churches [Graham *et al.*, 2009].

MFT posits that, cross-culturally, all moral concerns are rooted in five general content domains: Care/harm (sensitivity to the suffering of others), Fairness/cheating, (reciprocal social interactions and the motivations to be fair and just when working together), Loyalty/betrayal (promoting ingroup cooperation, sacrifice, and trust), Authority/subversion (endorsing social hierarchy), and Purity/degradation (promoting cleanliness of the body and the soul over hedonism). Each of these domains is divided into positive and negative aspects, yielding a total of 10 domains that can be captured using the MFD.

While the MFD provides a valuable tool for theory-driven analysis of naturally occurring moral rhetoric, it also highlights the difficulty of creating a dictionary of moral terms. While certain words may generally indicate a given moral concern, there is a huge range in both the issues that are moralized and the language around those moralizations across age cohorts, regions, religious groups, cultural groups, socioeconomic classes, and contexts. These differences are in fact one of the key findings of research in this area [Haidt *et al.*, 2009].

A second limitation comes in the way that these dictionaries are applied. In word count methods, documents are scanned for words appearing in the dictionaries. Not only does this limit applicability since these methods require enough text to find words in the limited dictionary set, it also distorts the process of dictionary creation in that authors are forced to look for low frequency terms at the margin of a concept in order to get sufficient coverage for the dictionary to be applied.

In response to these difficulties, we introduce a method which addresses some of these challenges while leveraging the structure of the existing dictionaries.

2.1 Generating Moral Concept Representations

Distributed representations offer one way past some of these difficulties. In these, words are represented as points in a low-dimensional space (generally 10-2000 dimensions). Similarity can be extracted based on nearness in the space while certain relationship types can be observed as linear transforms [Mikolov *et al.*, 2013b].

Here, we make use of the regularities in these spaces to answer some of the limitations of the MFD. To generate concept representations from the MFD, we average the vector representations of the words in each of the MFD categories based on a chosen distributed representation. This extends a method of generating sentence level representations by adding and averaging the vector representations of the individual words [Foltz *et al.*, 1998; Mitchell and Lapata, 2008]. With this, we can compare the similarity of the concept to other words, phrases, or documents. This combination of dictionaries and distributed representations has a number of advantages.

On the dictionary generation side, it allows for simplified and improved authorship by allowing a focus on psychological validity rather than linguistic range. Rather than worrying about finding all words which might indicate a concept (for concepts lacking annotated data), dictionary authors can focus on the core of the concept and make use of distributional similarity to capture other relevant terms. This has the added benefit of opening the possibility of applying text analysis to concepts where resources might not have been available for the development of broad coverage dictionaries.

Additionally, by measuring continuous similarity to a concept rather than a discrete count of words, we generate a smoother measure than prior word count methods. Further, the structure of the measure allows it to be applied to any size piece of text (down to the individual word level), allowing dictionary methods to be applied in contexts where they were previously limited. For example, this allows the application of dictionaries to social media posts which are often only a few words long and so difficult to measure with word count methods.

Formally, we assume the presences of a non-empty dictionary D of m words $\{w_1, w_2, \ldots, w_m\}$ and a pre-trained n

dimensional distributed representation R. R can be treated as a map defined over the words in its vocabulary V such that, for each word in that vocabulary, R maps the word to an n-dimensional real-valued vector:

$$R(w) = [d_1, d_2, \cdots, d_n], \forall w \in V$$

The next step is to generate the representation of the concept dictionary C_R in the chosen distributed representation R. This creates a concept representation within that particular representation, only applicable in that space.

As we can only consider dictionary words which are in the representation's vocabulary, we consider the subset:

$$D_R = D \cap V$$

Finally, we add the representations of the words in this intersection together and normalize to generate a concept representation:

$$C_R = \frac{\sum_{w \in D_R} R(w)}{\| \sum_{w \in D_R} R(w) \|}$$

Given C_R, we can calculate its cosine similarity to any word in the space by:

$$D(C_R, w) = \frac{C_R \cdot R(w)}{\|C_R\| \|R(w)\|}$$

To compare a document or phrase, we simply make use of the additive composition method for the words in the span of text and find the cosine similarity of the resulting phrase or document vector and C_R.

2.2 The Current Research

We first validate this method of generating moral concept representations in the context of a new dataset of Twitter posts focused on Hurricane Sandy. We compare applying the MFD through word-count and distributed concept representations and explore the potential of smaller dictionaries by using a subset of the MFD to generate moral concept representations. Next, we demonstrate how these same methods can be used to explore changes in moral rhetoric over time, providing one way to link the study of individual attitudes with larger psychosocial trends.

3 Study 1: Morality in 140 Characters

In this study, our aim is to compare the amount of signal detected by word count and distributed concept representations. To do this, we compare the performance of features generated by each of these methods when applied to the task of identifying moral rhetoric in Twitter posts calling for donations to help the victims of Hurricane Sandy.

This is challenging for a number of reasons. First, the short format of Twitter posts, only 140 characters, poses significant challenges for traditional language processing techniques [Kouloumpis *et al.*, 2011]. Second, moral rhetoric is itself difficult to precisely classify even for human annotators. This is particularly true in the context of Twitter posts which are characterized by allusiveness and minimal regard for traditional spelling and grammar. Third, in contrast to standard sentiment analysis, this is a multi-class problem where a single post can contain multiple moral domains. Given a lack of external ratings, we required human annotation which allowed us to further compare system performance versus variations in human coding.

3.1 Method

To generate this corpus, we selected 3000 tweets from a total set of approximately 7 million posts on the topic of Hurricane Sandy. The raw set was first filtered to exclude retweets and those lacking location information, then limited to tweets discussing donation. Three trained coders each coded 2000 Tweets on eleven factors, the positive and negative aspect for each of the five moral dimensions and an additional "non-moral" class for tweets without moral content. Coders were trained over multiple sessions by first being introduced to the overall MFT framework with subsequent sessions detailing the domains and covering potential ambiguities. They were not specifically trained on the MFD. For each tweet, coders were able to mark as present an arbitrary number of types of moral rhetoric. So, a single tweet such as "Thank God for our FirstResponders they endanger their lives for ours show RESPECT to them God keep em safe during Hurricane-Sandy" was coded by one annotator as displaying "care-pos", "loyalty-pos", and "authority-pos".

We compared three automated methods of identifying moral rhetoric. First, we considered classic word counts using the MFD categories. For each tweet in our set, we count the number of words from each of the categories, convert these to percentage scores given the total number of words in the tweet, and use these as features for classification.

Second, we made use of concept representations generated from the full MFD dictionaries. These were generated using two publicly available distributed representations. One was trained using Word2Vec [Mikolov *et al.*, 2013a] on a corpus of approximately 100 billion words of articles from Google News[1] with a total vocabulary of approximately 3 million words. The other was trained using GloVe [Pennington *et al.*, 2014] on 2 billion tweets with a resulting vocabulary size of approximately 1.2 million words[2]. For both representations, we created separate concept representations for each of the 10 MFD categories. Then, for each tweet, we calculated the distance between the tweet and the 10 concept representations to yield features for use in classification.

Third, we consider concept representations generated using simpler dictionaries, potentially allowing applications in domains where larger dictionaries are unavailable. Here, in consultation with the original MFD authors, we selected representative subsets of the complete MFD categories (four words from each category). These were then applied as in the previous step.

Classification was done on a per-class basis using logistic regression with 10-fold cross-validation. Due to the imbalance in the complete set of positive and negative cases for any particular class, upsampling was employed during training, selecting with replacement cases from the lower-frequency class.

[1] available at https://code.google.com/p/word2vec/

[2] available at http://nlp.stanford.edu/projects/glove/

Model	Precision	Recall	F1
MFD - word count	0.181	0.457	0.275
Full MFD - Google News	0.363	0.837	0.485
Seed MFD - Google News	0.372	0.840	0.496
Full MFD - Twitter	0.312	0.764	0.421
Seed MFD - Twitter	0.305	0.763	0.415

Table 1: Results for experiment 2 method performance averaged across coders and dimensions.

3.2 Results

As we made use of human-annotated data as the gold standard for this task, evaluation of inter annotator agreement was critical. Agreement was measured using Prevalence and Bias adjusted Kappa (PABAK) [Byrt *et al.*, 1993; Sim and Wright, 2005], an extension of Cohen's Kappa that is robust to unbalanced data sets. PABAK, which can be evaluated using the same rough guidelines as Kappa, was reasonably high for all dimensions (for the moral dimensions averaged across coder pairs, M = 0.81, SD = 0.07).

All classifiers were evaluated on precision, recall, and F1 for each of the 10 MFD categories. Significance of F1 differences across methods was calculated using permutation testing with 10,000 iterations. All differences in F1 scores were significant. As can be seen in Table 1, concept representations significantly outperformed traditional word count on these tests for all combinations of dictionaries and distributed representation. This effect held across coders and criterion dimensions, providing strong evidence that this is not unique to a particular coder's response style or criterion dimension. Concept representations from the Google News corpus significantly outperformed those trained on Twitter data.

3.3 Discussion

The use of distributed concept representations showed consistently better performance on these tests. One reason for that is, given the short length of the tweets, many of them included no words from *any* of the MFD categories. For that portion of the data, the word-count classifier could do no better than chance. Not only does this mean that using distributed concept representations allows dictionaries to be applied to previously inaccessible contexts, it also allows for more subtle analyses of longer form content. For example, separate measures can be taken at paragraph, sentence, or subsentential levels allowing for consideration in shifts over a single document.

The other important aspect of these results is the demonstration of the applicability of much shorter lists of words than found in traditional psychological dictionaries. While the difference in results between Google News and Twitter embeddings show that smaller dictionaries do not always improve performance, the fact that they showed comparable and, in one case, superior performance suggests that they could be viable for many applications. This matters for several reasons. First, it allows for much more rapid iteration as dictionary authors explore word choices. Second, it allows for psychological concepts which may have lacked the funding for full dictionary development to be operationalized and deployed in large-scale text analysis. Finally, as seen here, it may allow for better downstream task performance.

In terms of the differences between the performance using Twitter versus Google News representations we don't believe there is sufficient data to justify strong conclusions. However, for a domain such as morality, the use of vectors trained in a wide range of contexts as in the Google News case may better capture some of the subtle interactions among expressions of moral rhetoric.

4 Study 2: Moral Rhetoric over Time

In this study, we examine how the use of moral rhetoric has shifted in speeches in the United States Senate between 1988 and 2012. In particular, we consider shifts in the moral rhetoric surrounding the word "gay". While looking at how language around certain words and concepts evolves over time doesn't directly tell us about shifts in underlying attitudes, it does let us look one aspect of the social expression of those attitudes.

4.1 Method

We downloaded a corpus of approximately 375,000 speeches from the United States Senate from the Library of Congress THOMAS website[3] making use of the TACIT web crawler[4] [Dehghani *et al.*, 2016b]. For this analysis, we divided speeches between Republican and Democratic parties, dropping independents.

We iterated through these speeches, looking for instances of the word "gay", capturing a window of ±10 words surrounding each instance. These windows were aggregated by senator on a monthly basis (dropping months with no data). We measured the distance of the vector average of each month and the concept representation for each of the MFD categories in terms of the Google News distributed representations generated in Study 1.

4.2 Results

Overall Republicans tend to use more purity rhetoric towards issues associated with the term "gay" compared to Democrats t(1351) = 2.463, p = 0.014. Specifically, even though there is no difference between use of purity rhetoric before 1996 (1988 to 1996) between Republicans and Democrats t(396) = 0.937, p = 0.350, this difference becomes significant for the period after 1996 (1996 to 2004) t(586) = 1.947, p = 0.052. Moreover, the increase in use of purity language is significant between the two periods for the Republicans t(395) = 4.723, $p < 0.001$.

We found no significant difference between Republicans and Democrats for the other MFD categories.

4.3 Discussion

This result fits with existing work which has shown the purity dimension to be the key differentiator between Republicans and Democrats in terms of moral rhetoric on social media [Dehghani *et al.*, 2016a]. However, post hoc explanations

[3] http://thomas.loc.gov/home/thomas.php

[4] http://tacit.usc.edu/

can't be treated as confirmed hypotheses. Rather, we feel that this type of theory-driven exploratory analysis has the ability to provide a direction for future work.

In the present case, the finding that the divergence on purity rhetoric occurred in 1996 raises a number of questions for further consideration. For example, applying a similar analysis to text from other sources such as news articles, talk radio transcripts, or popular media sources would provide further information about the origins of this shift. Generally: are politicians driving changes in political discourse and, if not, where are these shifts originating?

For analyses covering longer historical windows, the use of distributed concept representations has a particularly large advantage over word count methods. While we can often identify a few words that remain characteristic of a rhetorical domain over time, the less strongly associated words will generally be far less stable. Applying those stable words to representations trained on historically relevant documents provides one possible way of studying concepts over time. And, while rhetoric does not provide a direct measure of attitudes, the analysis in those shifts at least provides a suggestion as to where to focus deeper analyses.

5 Conclusions and Future Work

While moral rhetoric is an important aspect of a variety of areas of discourse, it remains a somewhat elusive concept. Theories of morality remain divided. Nonetheless, given its importance to a range of social phenomena, the ability to detect moral rhetoric in large sources of text is extremely valuable. In these experiments, we have shown that this is not only possible, but even viable in short-form social networking posts where prior methods have struggled.

Our results further suggest the value of blending theory-driven and data-driven methods. Each have strengths and weaknesses, and provide key pieces for both computational and social scientific research. In particular, our use of concept representations demonstrate the blending of dictionary based methods with modern distributional semantics in the form of distributed representations. For dictionaries, it allows their development to be driven by psychological validity rather than linguistic coverage for use in downstream applications. Further, it allows those dictionaries to be applied in contexts which were previously inaccessible. For distributed representations, it allows the inclusion of prior knowledge and one method for including concepts in the dictionary structure.

Of course, this study does not address questions around the relationship between the presence of moral rhetoric and underlying moral attitudes. However, prior work has demonstrated the process of applying lab-based experimental follow-ups to patterns observed in moral rhetoric in large scale text analysis [Dehghani *et al.*, 2016a]. Thus, extending the ability to apply this type of text analysis to both larger and smaller contexts opens the door to future work in this direction. While surface representation does not take us directly to attitude, it is one valuable step on the way. And, the study of expressions of moral rhetoric give us a valuable window into larger psychosocial processes as we continue to understand how moral sentiment shapes our choices and evolves over time.

Acknowledgments

This work has been funded in part NSF IBSS #1520031.

References

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Boyd *et al.*, 2015] Ryan L Boyd, Steven R Wilson, James W Pennebaker, Michal Kosinski, David J Stillwell, and Rada Mihalcea. Values in words: Using language to evaluate and understand personal values. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[Byrt *et al.*, 1993] Ted Byrt, Janet Bishop, and John B Carlin. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429, 1993.

[Chae *et al.*, 2015] David H Chae, Sean Clouston, Mark L Hatzenbuehler, Michael R Kramer, Hannah LF Cooper, Sacoby M Wilson, Seth I Stephens-Davidowitz, Robert S Gold, and Bruce G Link. Association between an internet-based measure of area racism and black mortality. *PloS one*, 10(4):e0122963, 2015.

[Dehghani *et al.*, 2013] Morteza Dehghani, Megan Bang, Douglas Medin, Ananda Marin, Erin Leddon, and Sandra Waxman. Epistemologies in the text of children's books: Native-and non-native-authored books. *International Journal of Science Education*, 35(13):2133–2151, 2013.

[Dehghani *et al.*, 2016a] Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. Purity homophily in social networks. *Journal of Experimental Psychology: General*, 2016.

[Dehghani *et al.*, 2016b] Morteza Dehghani, Kate M. Johnson, Justin Garten, Reihane Boghrati, Joe Hoover, Vijayan Balasubramanian, Anurag Singh, Yuvarani Shankar, Linda Pulickal, Aswin Rajkumar, et al. Tacit: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, pages 1–10, 2016.

[Eichstaedt *et al.*, 2015] Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.

[Foltz *et al.*, 1998] Peter W Foltz, Walter Kintsch, and Thomas K Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307, 1998.

[Garcia and Sikström, 2014] Danilo Garcia and Sverker Sikström. The dark side of facebook: Semantic representations of status updates predict the dark triad of personality. *Personality and Individual Differences*, 67:92–96, 2014.

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

[Haidt *et al.*, 2009] Jonathan Haidt, Jesse Graham, and Craig Joseph. Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, 20(2-3):110–119, 2009.

[Iliev and Smirnova, 2014] Rumen Iliev and Anastasia Smirnova. Revealing word order: Using serial position in binomials to predict properties of the speaker. *Journal of psycholinguistic research*, pages 1–31, 2014.

[John, 1962] L John. Austin. how to do things with words, 1962.

[Johnson *et al.*, 2014] Kate M Johnson, Ravi Iyer, Sean P Wojcik, Stephen Vaisey, Andrew Miles, Veronica Chu, and Jesse Graham. Ideology-specific patterns of moral indifference predict intentions not to vote. *Analyses of Social Issues and Public Policy*, 14(1):61–77, 2014.

[Kaplan *et al.*, 2016] Jonas T Kaplan, Sarah I Gimbel, Morteza Dehghani, Mary Helen Immordino-Yang, Kenji Sagae, Jennifer D Wong, Christine M Tipper, Hanna Damasio, Andrew S Gordon, and Antonio Damasio. Processing narratives concerning protected values: A cross-cultural investigation of neural correlates. *Cerebral Cortex*, page bhv325, 2016.

[Koleva *et al.*, 2012] Spassena P Koleva, Jesse Graham, Ravi Iyer, Peter H Ditto, and Jonathan Haidt. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality*, 46(2):184–194, 2012.

[Kouloumpis *et al.*, 2011] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11:538–541, 2011.

[Kumar *et al.*, 2015] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 85–94. ACM, 2015.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013.

[Mitchell and Lapata, 2008] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *ACL*, pages 236–244, 2008.

[Mitchell *et al.*, 2015] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. *NAACL HLT 2015*, page 11, 2015.

[Motyl *et al.*, 2014] Matt Motyl, Ravi Iyer, Shigehiro Oishi, Sophie Trawalter, and Brian A Nosek. How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology*, 51:1–14, 2014.

[Pang and Lee, 2005] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[Park *et al.*, 2015] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934, 2015.

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

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014.

[Schwartz *et al.*, 2014] H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125. Citeseer, 2014.

[Schwarz, 1999] Norbert Schwarz. Self-reports: how the questions shape the answers. *American psychologist*, 54(2):93–105, 1999.

[Sim and Wright, 2005] Julius Sim and Chris C Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.

[Stone *et al.*, 1968] Philip Stone, Dexter C Dunphy, Marshall S Smith, and DM Ogilvie. The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1):113–116, 1968.

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