Detecting and Understanding Moral Biases in News

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, Elena Zheleva
University of Illinois at Chicago
Chicago, IL
{hshahi6,bdieugen,arojecki,ezheleva}@uic.edu

Abstract

We describe work in progress on detecting and understanding the moral biases of news sources by combining framing theory with natural language processing. First we draw connections between issue-specific frames and moral frames that apply to all issues. Then we analyze the connection between moral frame presence and news source political leaning. We develop and test a simple classification model for detecting the presence of a moral frame, highlighting the need for more sophisticated models. We also discuss some of the annotation and frame detection challenges that can inform future research in this area.

1 Introduction

While much attention has focused on the role of fake news in political discourse, comparatively little attention has been paid to the dissemination of news frames. Framing in news coverage–highlighting certain aspects of an issue or event–can have a significant impact on public opinion formation (Callaghan, 2014). Framing theory posits that preference formation depends on which subset of relevant considerations or beliefs–"frame in mind"–are activated by a particular message–"frame in communication." Scholars refer to the power of such a frame as a framing effect, a phenomenon widely reported in academic scholarship on domestic and foreign issues alike (Jacob, 2000; Grant and Rudolph, 2003; Nicholson and Howard, 2003; Baumgartner and Boydstun, 2008; Perla, 2011). If one-sided and morally charged, it can exacerbate polarization and post-truth politics.

According to the most widely cited model used by social scientists (Entman, 1993), the essential components of a frame include problem definition, diagnosis of cause, moral judgment, and prescribed remedy. For example, obesity may be defined as a significant national health problem, diagnosed as the result of increasingly passive lifestyles judged as detrimental to the strength of society and individuals, and effectively treated by increased physical activity. Such an emphasis on individual choice redirects attention from other possible causes such as genetic disposition or advertising campaigns for caloric rich foods.

Even though moral judgment is central to frame analysis, much of the frame analysis research neglects the moral dimension. Our work responds to this gap by adapting Moral Foundations Theory (MFT) which proposes a set of five modalities–each with a virtue and vice binary partner–that underlie moral thinking (Graham et al., 2013). Morally-inflected frames follow the contours of political ideology (Graham et al., 2009), are more likely to be shared on social media (Valenzuela et al., 2017), and, most importantly, reinforce attitudes, making compromise more difficult (Koleva et al., 2012).

Technology offers little solution to mitigate these framing effects at the scale and speed of modern information networks. Our work responds to the need for cross-disciplinary frameworks that enable the early detection, propagation, and influence of moral frames in such networks. In this paper, we offer an initial analysis on the steps necessary for detecting and understanding the prominence of moral frames in news. We annotate a small corpus of news articles with moral frames and look into their connection to issue-specific frames and news source leaning, together with models for detecting them.

2 Related work

In the last few years, a number of NLP approaches have been devised for frame identification in text: most focus on coarser-grained primary frame identification (Card et al., 2016; Ji and Smith, 2017; Johnson et al., 2017a), possibly based on a probabilistic distribution (Burscher et al., 2014); few

Statistics	Values
Sentences with at least one moral frame	2.81 %
Articles with at least one moral frame	20.61 %
Article frame presence agreement alpha	0.0485
Sentence frame presence agreement alpha	-0.0264
Article frame type agreement alpha	0.8435
Sentence frame type agreement alpha	0.8525

Table 1: Dataset annotation statistics.

address finer-grained frame tagging at the paragraph level (Tsur et al., 2015). Most research relies on word-based approaches, from direct keyword matching to latent representations (Boydstun et al., 2013; Burscher et al., 2014; Baumer et al., 2015; Tsur et al., 2015; Johnson et al., 2017a,b). Few studies use rhetorical information such as discourse structure (Ji and Smith, 2017).

Additionally, work has been done on general frames, such as *economy* or *law and order* (Card et al., 2016; Burscher et al., 2014) that can apply across issues, or on issue-specific (also called topical) frames, such as *innocence* as concerns capital punishment; than on identifying moral foundations. Approaches to the latter mostly rely on moral foundation keyword dictionaries, again directly (Fulgoni et al., 2016) or via latent representations (Kaur and Sasahara, 2016; Garten et al., 2016).

3 Datasets and annotation

Since there is no existing corpus with moral frame annotations for news articles, we put together a small initial dataset to help us understand the intricacies of moral frame annotation and analysis. Our dataset contains 400 articles on four different issues. 300 articles are from a previously collected corpus (Card et al., [n. d.]), 100 articles for each of immigration, smoking and same-sex marriage issues from 13 news sources. Another set of 100 articles was collected on the racial unrest in Baltimore from 16 national and local newspaper sources.

Three undergraduate student annotators were hired as summer interns for this project. Each article was independently annotated with sentence-level and article-level moral frames by all three annotators based on the 10 moral foundations (Graham et al., 2013) – Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/ Subversion, Sanctity/Degradation – or with NA. The annotators used the BRAT software to perform the annotations (https://brat.nlplab.org/).

The annotation process proceeded in two stages, each of which involved a detailed annotation manual, that was modified in the second stage [1].

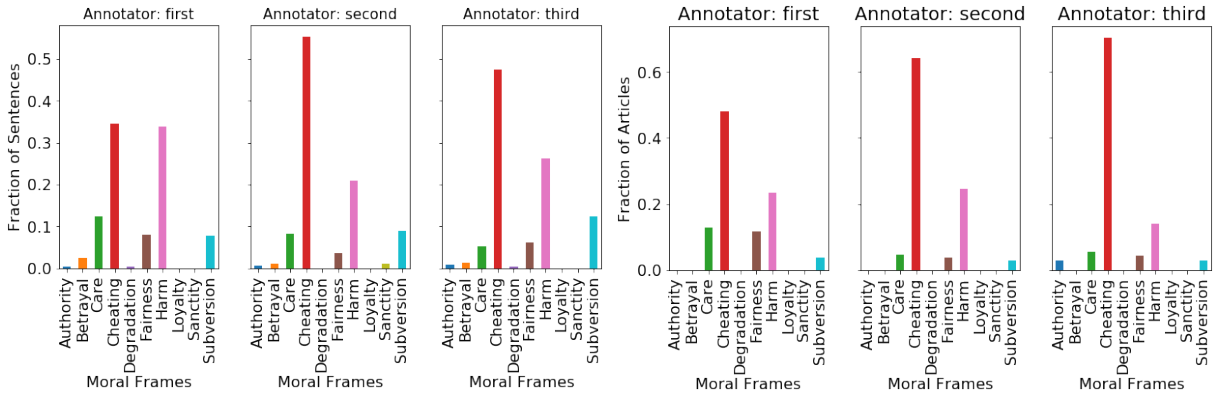
Stage 1. The annotation manual instructed the annotators to proceed with coding in 3 ordered steps: (1) to identify the moral frame type; (2) to decide whether the author supports or rejects the frame; and (3) to decide whether the author explicitly favors or opposes the specific issue the article is about. The annotators were also instructed to do so for both sentences and the whole article; at the article-level, the annotators were asked to evaluate the entire article and specify what they regarded as its main moral frame. The annotators were told to first annotate the sentences in an article and then the article as a whole, but no explicit written guidelines were provided in this regard.

Stage 2. After the initial set of annotations from Stage 1 (which were discarded), the protocol was adjusted based on annotator feedback, with the goal of making the annotation process less ambiguous. First, a preliminary step was added to the three annotation steps, a.k.a step 0: annotators were instructed to identify the presence or absence of any moral frame before embarking in the subsequent three steps. Second, the sentence and article annotation were clearly separated, and for the article annotation specific guidelines were provided: *Evaluate the entire article and specify what you regard is its main moral claim. Keep in mind that it may or may not be the most frequent one (based on counting sentences with moral claims).*

Another main adjustment was providing the annotators with a list of keywords associated with each moral foundation developed by Graham et al. (Graham et al., 2013). The annotators were instructed to use such sets as keywords as guidance, but were warned that (a) a moral frame may contain none of the keywords listed in the codebook, and that (b) the presence of a keyword does not necessarily indicate the presence of a moral frame. The annotators were provided with examples of both (a) and (b).

We refer to the dataset with sentence-level annotations as ***mf-sent*** and with article-level annotations as ***mf-art***. 116 articles have both an article-level frame and at least one sentence-level frame, as annotated by at least one annotator. The percentage of articles whose moral frame is different from the most frequent moral frame among the article's sentences is 14.3%.

[1] Annotation manual: https://bit.ly/2LXiiR5



(a) Sentence level

(b) Article level

Figure 1: Moral frame distribution for each annotator in moral frames datasets

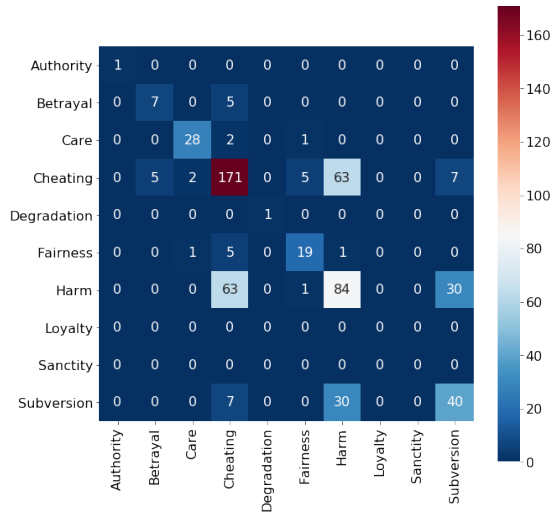


Figure 2: Coincidence matrix showing number of times annotators (dis)agree on sentence-level frame types.

To study the connection between issue-specific frames and more general, moral frames, a domain expert annotated a subset of 48 Baltimore unrest articles from *mf-sent* with issue-specific frames at the sentence level following (Rojecki, 2017): Black Criminality, Police Racism, Rogue Cops, and Structural Inequality. We refer to this dataset as ***ol-sent***.

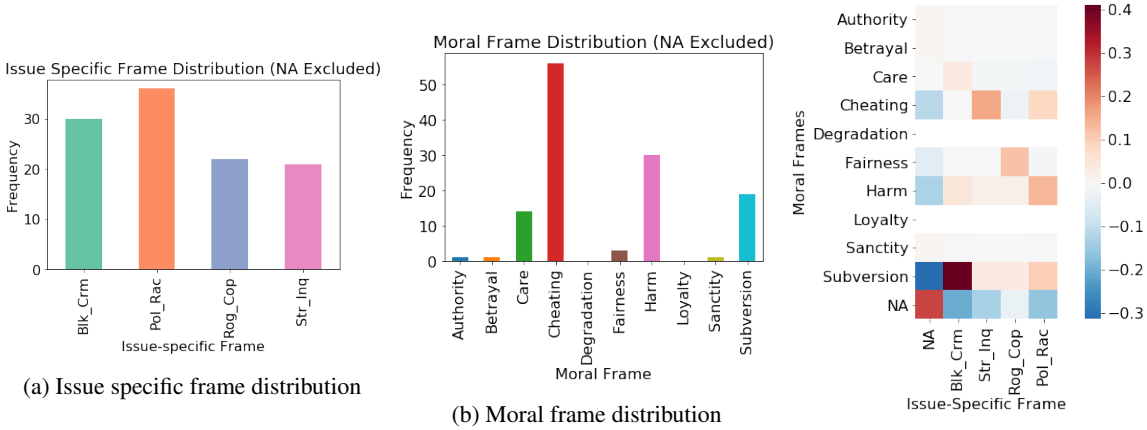
Subjectivity in Moral Frame Annotation. Table 1 shows dataset annotation statistics including Krippendorff's alpha for inter-annotator agreement. Despite the protocol iterations, the annotators had a fairly low level of agreement on the presence/absence of a moral frame both at the article level (alpha=0.0485) and the sentence level (alpha=-0.0264). However, when at least two annotators agreed that a moral frame is present, the frame type agreement was relatively high both at the article level (alpha=0.8435) and at the sentence level (alpha=0.8525). Figure 1 shows the distribution of frames for each annotator at article and sentence level. While some frames like Cheating and Harm are prevalent across annotators, the actual distributions are different. Figure 2 shows the frame confusion matrix at the sentence level. Each box represents the number of times a moral frame disagreement occurred at the sentence level. The figure shows that annotators often disagree on the most frequent frames, Harm, Subversion and Cheating.

These results reflect the challenges in using non-experts for moral frame annotation. A number of annotation studies have analyzed the reliability of non-expert annotations, and investigated whether corrections need to be applied to the annotation process and / or to the models derived from the non-expert annotated datasets (Snow et al., 2008; Welinder and Perona, 2010; Patton et al., 2019; Lavee et al., 2019). However, many of these studies can actually compare the performance of non-expert and expert annotation, since datasets annotated by experts for the phenomenon of interest did exist; this was not the case for us. In fact, this initial effort of ours at annotation can be taken as an indication of how difficult annotating for moral frames is for non-experts; it remains to be seen how expert annotators would fare on this task. This is part of the future research we will undertake to understand whether this task can be crowdsourced successfully at scale or whether it requires expert annotators.

4 Moral frame analysis

4.1 Issue-specific vs. moral frames

We analyze the connection between issue-specific and moral frames in the Baltimore unrest articles (*ol-sent* dataset). When a sentence is annotated with multiple frames, we consider the one with the

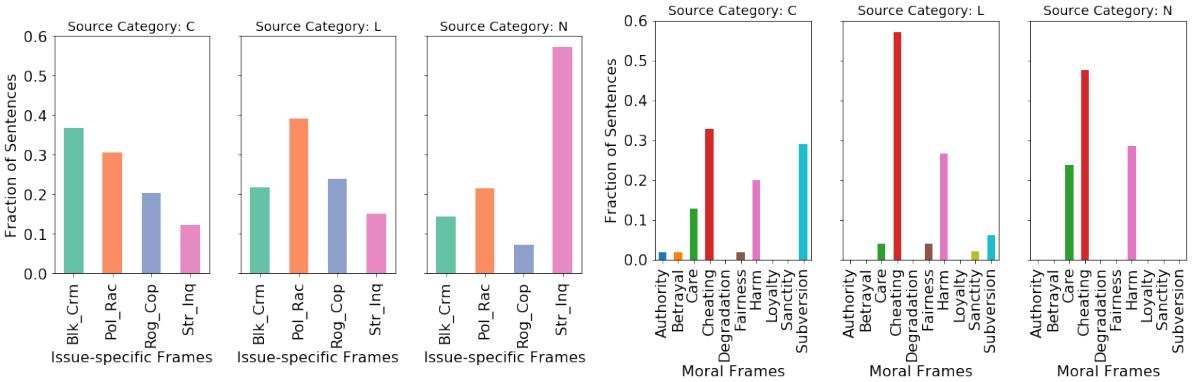


(a) Issue specific frame distribution

(b) Moral frame distribution

(c) Frame type Pearson correlation.

Figure 3: Frame distributions for different types of frames in *ol-sent* dataset.



(a) Issue specific frame distribution

(b) Moral frame distribution

Figure 4: Frame type distributions in *ol-sent* dataset based on news source type (**C**onservative, **L**iberal, **N**eutral).

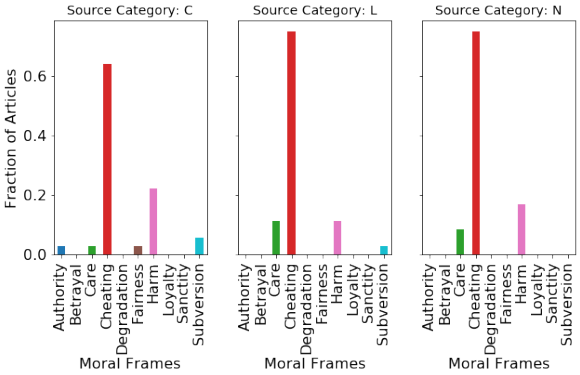


Figure 5: Moral frame distributions for different types of sources in Baltimore articles from *mf-art* dataset.

highest agreement. The sentence-level distribution of issue-specific and moral frames is given in Figure 3. While issue-specific frames are more evenly distributed, moral frames have a skewed distribution with Cheating and Harm being the dominant frames. The most likely reason for this is that, the Care/Harm and Fairness/Cheating foundations are valued by conservatives and liberals alike and is therefore more likely to be present in news frames.

We computed the Pearson correlation between different frame labels, and their heat-map representation can be found in Figure 3c. It is interesting to note that for most moral frames, there is a dominant corresponding issue-specific frame. For Subversion, it is Black Criminality (Blk_Crm), for Fairness, it is Rogue Cops (Rog_Cop), for Cheating/Injustice it is Structural Inequality (Str_Inq).

4.2 Moral frames and news source leaning

In order to understand whether moral frames can explain the political leanings of news sources, a domain expert labeled each news source based on the history of their support for a liberal/conservative candidate. [2] We use the *ol-sent* dataset for this purpose since it has both issue-specific and moral frame labels. The sentence-level distributions for liberal/conservative/neutral news sources can be seen in Figure 4. Issue-specific frame distributions (Fig. 4(a)) are very revealing and consistent

[2]Other possible news-source leaning annotations (e.g., (Wihbey et al., 2017)) can be considered in future work.

Model	Precision	Recall	F-score
Keyword match	0.08	0.65	0.14
SVM	0.20	0.41	0.27

Table 2: Moral frame presence classification results.

Model	Precision	Recall	F-score
Keyword match	0.69	0.29	0.31
SVM	0.68	0.70	0.68

Table 3: Multi-class moral frame classification results.

with previous work (Rojecki, 2017): conservative sources tend to criminalize the protesters while liberal sources focus more on police racism and rogue cops. Neutral sources are harder to explain, however, according to the domain expert, the structural inequality in issue-specific frames can be explained by the fact that these sources are more likely to be aligned with the liberal sources.

Since Black Criminality strongly correlates with Subversion as shown in Figure 3c, we can see in Figure 4(b) that subversion frame is heavily used by conservatives as opposed to liberals. Cheating or Injustice is heavily used by liberal sources as opposed to conservative. Authority and Betrayal are present in conservative sources and absent in the liberal ones. Liberal sources have some sentences labeled as Sanctity which is missing from conservative sources.

Figure 5 shows the distribution of article-level moral frames for the 100 Baltimore articles in the *mf-art* dataset. It shows similar patterns as the sentence-level annotations, except that the differences between news source categories are not as pronounced.

4.3 Moral frame detection

We train a binary classifier to detect whether a moral frame is present or absent in a sentence using all articles in *mf-sent*. Each sentence is represented by a normalized sum of its $word2vec$ word vectors. A balanced SVM classifier is tuned and trained using 5-fold stratified cross-validation. Its accuracy is reported in Table 2. It is compared to a baseline *Keyword match* which reports a frame present if at least one of the MFT keywords (Graham et al., 2013) is present in the sentence. The relatively poor results partially reflect the class skew (95% of sentences do not have a moral frame present). For the subset of sentences with moral frames present (396 in total), we used the same methods and evaluation mechanism as above to classify sentences in specific moral frame categories. The only exception is that we used frame-specific MFT keywords and SVM is trained using a one-versus-one multi-class classification setup. The weighted-average results are reported in Table 3.

5 Conclusion

We presented a small-scale study of moral frames in news showing that moral frames have the potential to explain issue-specific frames and the biases of their news sources. In order to increase the analysis scale and reliability, we need to collect a larger dataset covering more issues and news sources. We also need to improve the annotation protocol and overcome the challenges associated with annotator subjectivity. Future directions include improving on the machine learning models for predicting moral frames in news articles and studying their impact on opinions expressed in social media.

Acknowledgments

The authors would like to thank Sumayya Siddiqui, Navya Reddy and Hasan Sehwail for their help with annotating the data.

References

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *NAACL*. 1472–1482.

De Boef Suzanna L. Baumgartner, Frank R. and Amber E. Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press, Cambridge ; New York.

Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2013. Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop, London*.

Björn Burscher, Daan Odijk, Rens Vliegenthart, Maarten De Rijke, and Claes H De Vreese. 2014. Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures* 8, 3 (2014), 190–206.

Karen Callaghan. 2014. *Framing American Politics*. University of Pittsburgh Press, Pittsburgh PA.

Dallas Card, Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. [n. d.]. The media frames corpus: Annotations of frames across issues. In *ACL*.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *EMNLP*. 1410–1420.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43, 4 (1993), 51–58.

Dean Fulgoni, Jordan Carpenter, Lyle H Ungar, and Daniel Preoţiuc-Pietro. 2016. An Empirical Exploration of Moral Foundations Theory in Partisan News Sources. In *LREC*.

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

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*. Vol. 47. Elsevier, 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 (2009), 1029.

J. Tobin Grant and Thomas J. Rudolph. 2003. Value Conflict, Group Affect, and the Issue of Campaign Finance. *American Journal of Political Science* 47, 3 (2003), 453.

William G. Jacob. 2000. Issue Framing and Public Opinion on Government Spending. *American Journal of Political Science* 44, 4 (2000), 750.

Yangfeng Ji and Noah Smith. 2017. Neural discourse structure for text categorization. In *ACL*.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017a. Leveraging Behavioral and Social Information for Weakly Supervised Collective Classification of Political Discourse on Twitter. In *ACL (Volume 1: Long Papers)*, Vol. 1. 741–752.

Kristen Johnson, I-Ta Lee, and Dan Goldwasser. 2017b. Ideological Phrase Indicators for Classification of Political Discourse Framing on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 90–99.

Rishemjit Kaur and Kazutoshi Sasahara. 2016. Quantifying moral foundations from various topics on Twitter conversations. In *IEEE BigData*. 2505–2512.

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

Tamar Lavee, Lili Kotlerman, Matan Orbach, Yonatan Bilu, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*. Association for Computational Linguistics, Hong Kong, 29–38. https://doi.org/10.18653/v1/D19-5905

Stephen P. Nicholson and Robert M. Howard. 2003. Framing Support for the Supreme Court in the Aftermath of Bush v. Gore. *The Journal of Politics* 65, 3 (2003), 676–695.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Héctor Perla. 2011. Explaining Public Support for the Use of Military Force: The Impact of Reference Point Framing and Prospective Decision Making. *International Organization* 65, 1 (2011), 139–167.

Andrew Rojecki. 2017. Racial Threat and Local Framing of Baltimores Unrest. In *News of Baltimore: race, rage and the city*, Linda Steiner and Silvio R. Waisbord (Eds.). Routledge, New York.

Rion Snow, Brendan OConnor, Dan Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast–but is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *ACL (Volume 1: Long Papers)*. 1629–1638.

Sebastin Valenzuela, Josefina Ramrez, and Martina Pia. 2017. Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing. *Journal of Communication* 67, 5 (8 2017), 803–826.

P. Welinder and P. Perona. 2010. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 25–32.

John Wihbey, Thalita Dias Coleman, Kenneth Joseph, and David Lazer. 2017. Exploring the Ideological Nature of Journalists' Social Networks on Twitter and Associations with News Story Content. *KDD Workshop on Data Science + Journalism* (2017).