Kenneth R. Fleischmann, Clay Templeton, **Jordan Boyd-Graber**, An-Shou Cheng, Douglas W. Oard, Emi Ishita, Jes A. Koepfler, and William A. Wallace. **Explaining Sentiment Polarity: Automatic Detection of Human Values in Texts**. In Preparation.

```
    @article{Fleischmann:Templeton:Boyd-Graber:Cheng:Oard:Ishita:Koepfler:Wallace-In Preparation,
Author = {Kenneth R. Fleischmann and Clay Templeton and Jordan Boyd-Graber and An-Shou Cheng and Douglas W. Oar
Year = {In Preparation},
Title = {Explaining Sentiment Polarity: Automatic Detection of Human Values in Texts},
}
```

Noname manuscript No.
(will be inserted by the editor)

Explaining Sentiment Polarity

Automatic Detection of Human Values in Texts

Kenneth R. Fleischmann · Clay Templeton · Jordan Boyd-Graber · An-Shou Cheng · Douglas W. Oard · Emi Ishita · Jes A. Koepfler · William A. Wallace

Received: date / Accepted: date

Abstract How can we explain and predict differences in sentiment among different groups of people? Automatic techniques can determine sentiment polarity by labeling an author's text as expressing positive, negative, or neutral opinion, but such approaches ignore motivations behind the sentiment. Human values are what determine an individual's priorities, and, as such, can help explain why someone has a particular sentiment. We present a theoretical framework and corresponding research program with the goal of scaling up social science text analysis by automatically detecting and classifying human values in texts to explain sentiment. We present two studies with different approaches to machine learning about human values. In the first, which focuses on the Net neutrality debate, content analysis by experts leads to annotations of values in texts with varying sentiment polarity toward Net neutrality. Machine learning allows leveraging these annotations for automatically detecting human values in texts.

K.R. Fleischmann
University of Texas at Austin, USA
301-661-7990
E-mail: kfleisch@ischool.utexas.edu

T.C. Templeton
University of Texas at Austin, USA

J. Boyd-Graber
University of Maryland, USA

A.-S. Cheng
National Sun Yat-Sen University, Taiwan

D.W. Oard
University of Maryland, USA

E. Ishita
Kyushu University, Japan

J.A. Koepfler
University of Maryland, USA

W.A. Wallace
Rensselear Polytechnic Institute, USA

In the second, which focuses on the Park51 controversy, crowdsourcing is leveraged to collect data about individuals' values and their agreement or disagreement with texts with varying sentiment. A radically different approach to sentiment analysis used here allows us to mimic the human property of reacting to events based on an individuals set of basic human values. Both of these approaches use the explanatory power of human values, which play a fundamental role in shaping the different sentiment polarities about topics among different individuals.

Keywords human values · sentiment analysis · crowdsourcing · annotation

1 Introduction

Great advances have been made in the automatic detection of sentiment polarity—that is, determining whether a text indicates positive, negative, or neutral sentiment of the author toward a topic, issue, individual, group, or product (Pang and Lee, 2008). However, understanding why the author expresses and presumably experiences that sentiment is nontrivial. The rich literature within social psychology and, broadly, social science provides many possible frameworks that could explain why people feel the way they do about particular topics. Here, we focus on one framework, human values, as a feature set that can explain sentiment polarity, in turn furthering our understanding of human values and their effects on human behavior.

Human values can be defined as "guiding principles of what people consider important in life" (Cheng and Fleischmann, 2010). According to Rokeach (1973), "Values are determinants of virtually all kinds of behavior that could be called social behavior or social action, attitudes and ideology, evaluations, moral judgments and justifications of self to others, and attempts to influence others" (p. 5). As such, values are useful constructs for understanding human attitudes and behavior. Schwartz (2007) has done extensive and systematic research on human values, and has demonstrated that values can be used to predict both national and international variation of attitudes and behavior. For example, he used data from the European Social Survey to demonstrate the relationship between specific values and attitudes toward immigration, as well as behaviors such as interpersonal trust, social involvement, organizational membership, and political activism.

This paper demonstrates a theoretical basis for a relationship between values and sentiment polarity, inspired by related work that integrate insights from natural language processing and social science content analysis to develop new approaches to computational social science (Hopkins and King, 2010; Crowston et al, In Press). The next section provides an in-depth discussion of human values. The following sections describe two studies that detect values in texts using fundamentally distinct approaches to computational social science. The first series of studies uses machine learning to replicate the annotations from content analysis on formal, prepared testimonies presented at Congressional and FCC Hearings about Net neutrality. The second series of studies applies machine learning based on crowdsourced data to study the relationship between what people say they value and how they express their opinions on the Park51 debate. The paper concludes with a call for additional research on the

role of human values in determining sentiment polarity, and provides examples of the potential theoretical and practical impacts of developing this capability.

2 Background

An important part of understanding human nature and what makes people different from each other is an understanding of human values. As Weber (1947) argues, "Many ultimate ends or values toward which experience shows that human action may be oriented, often cannot be understood completely, though sometimes we are able to grasp them intellectually. The more radically they differ from our own ultimate values, however, the more difficult it is for us to make them understandable by imaginatively participating in them" (p. 91). Parsons (1964) notes that Weber "developed a comprehensive analysis of the ways in which values systems influence concrete behavior" (p. 175). Fundamentally, the social sciences aim to explain human behavior. Understanding human behavior is an important step not only in explaining the motivations behind sentiment polarity, but also for practical application of sentiment analysis that can be applied to diverse populations. Human values provide an explanatory framework for investigating the factors that motivate human sentiment polarity and behavior.

Values are particularly important based on their durability and broad applicability. According to Habermas (1984), "Interest positions change, whereas generalizable values are always valid for more than merely one type of situation" (p. 172). Massey (1979), as cited in Hill (2004) argues that values form early in life, through three stages: imprinting, from ages 1–7; modeling, from ages 7–13; and socialization, from ages 14–20. Thus, values tend to be fairly solidified and fixed by the time one reaches their early 20s.

One important debate about human values is whether they are universal; that is, whether the same set of values are shared across cultures. Several scholars have sought to organize human values into coherent frameworks, and have studied the extent to which these frameworks are universal. Rokeach (1979) devised one of the first such values instruments, the Rokeach Value Survey. The Rokeach Value Survey includes 18 instrumental values (such as "ambitious") and 18 terminal values (such as "a sense of accomplishment").

Schwartz (1994) developed the best known and most widely used values instrument, the Schwartz Value Survey. The Schwartz Value Survey includes three levels of hierarchy, including two orthogonal value dimensions that form four value quadrants ("conservation" to "openness to change" and "self-enhancement" to "self-transcendence"), ten value types (such as "benevolence"), and 56 basic human values (such as "wealth"). In his original study, Schwartz compared results from 44 different countries; subsequent research has further expanded the range of countries studied. Schwartz developed a second instrument, the Portrait Value Questionnaire, which adopts an indirect approach to masuring values. Instead of asking respondents directly about how much they value various things, as in the case of the Rokeach Value Survey and the Schwartz Value Survey, the PVQ asks respondents how much they relate to various portraits of individuals expressing and acting on specific value types (Schwartz, 2007).

Many other value instruments exist. Cheng and Fleischmann (2010) review the Rokeach Value Survey, the Schwartz Value Survey, and ten additional value instruments from a wide range of fields. They then integrate these twelve value instruments into a new instrument, the meta-inventory of human values. Thus, many social science instruments can help identify human values in texts.

The burgeoning fields of sentiment analysis (Pang and Lee, 2008) and opinion mining (Wilson, 2008; Turney and Littman, 2003) attempt to expose a person's "private states" (Wilson and Wiebe, 2005), or their internal thought process, from the observed text. Much of this research has taken the form of a supervised learning problem: given pieces of text with corresponding labels of sentiment, learn a mapping from unannotated text to their opinion.

While sentiment analysis is important, for both basic research and its current commercial applications, it often treats individuals as homogenous when they are certainly heterogenous. First, individuals' opinion on a subject often hinges on their perspective; for example, how a commenter reacts to Israeli settlements in the West Bank will be informed by their prior perspective on the issue. In the Machine Learning literature, similar insights have been leveraged through both supervised (Lin et al, 2006; Hardisty et al, 2010) and unsupervised (Paul and Girju, 2010) approaches.

The approach presented here allows us to model a scaffolded perspective in which we propose to incorporate *values* as a low-dimensional representation from which we can estimate how an individual will likely react to particular issues.

3 Content Analysis of Values in the Net Neutrality Debate

How can values be detected, and how can they in turn be used to explain sentiment? A series of studies has employed a range of approaches, including content analysis, crowdsourcing, and machine learning. This section describes research to date that demonstrates connections between values and sentiment while also serving as a proof of concept for the feasibility of automatically detecting these relationships.

Our first attempt at detecting values in texts was a study of human values in computational models (Fleischmann and Wallace, 2006, 2009; Fleischmann et al, 2010, 2011b). The field sites for this study included corporate, academic, and government research laboratories heavily involved in computational modeling. Data collection for the study included surveys (including the Schwartz Value Survey), interviews, and focus groups. For the interview data, two coders independently coded text. First, they coded for the presence of values, with a process for comparing coding results and reaching consensus. Next, they coded each value-laden expression, again comparing the independent coding results and reaching consensus. The coding instrument for the second task was a list of values that included the values from the Schwartz (1994) Value Survey. One analysis performed based on this coding was a comparison of the frequency of expression of values across the three sites. For example, we found that the values of "responsible" and "wealth" most frequently occurred in data from the corporate laboratory, while the value of "influential" occurred least frequently in the corporate laboratory data (Fleischmann et al, 2011a). Thus, this study served as an initial proof-of-concept that values could be systematically detected in text, and that

the analysis of these values could yield interesting and useful results that can help to explain human behavior.

We then made an early effort to automate this process using a simple thesaurus-based approach (Zhou et al, 2010). The corpus under investigation was the Enron e-mail dataset, a collection of 252,830 unique e-mail messages sent and received by Enron employees that was publicly released after Enron's sudden bankruptcy. Our goal was to study how values related to communication within a social network. A bag of words approach was used, such that the keyword corresponding to each value, as well as all of its synonyms, constituted a bag of words $\mathbf{W}_v$ for that value, although it was necessary to remove some words that had multiple meanings. Then, the Enron e-mail dataset was clustered according to the values expressed in the e-mails sent by specific individuals m_p. Each individual was assigned to the value cluster v that accounted for the largest proportion of the words x written by an author p in their emails e,

$$\text{value}(p) \equiv \arg\max_{v} \sum_{e \in m_p} \sum_{x \in e} \mathbb{I}\left[x \in \mathbf{W}_v\right] \tag{1}$$

A total of six value-based clusters were assigned, and the communication density within those clusters was higher than the communication density between clusters for four of the six clusters. Thus, we concluded that people may be more likely to communicate with individuals with similar values. This approach was fraught with problems, including the failure to include multi-word phrases and the potential for use of words in different ways. Since no manual coding was performed, there was no way to measure the accuracy of the system. Thus, we determined that the next logical step was to develop a more rigorous and effective approach for annotating values in texts prior to automation of the analysis.

Next, we set out to develop an effective approach for annotating values in texts. We selected the issue of Net neutrality (Schwartz and Weiser, 2009), which had the advantage of being a binary issue where sentiment polarity can be determined easily. The corpus constructed for this task was a set of 102 prepared testimonies presented at public hearings organized by the U.S. House, Senate, and Federal Communications Commission (FCC). To annotate these testimonies, we first used the values from the Schwartz Value Survey (Schwartz, 1994), following the approach described above for the computational modeling study (Cheng et al, 2010). We annotated these testimonies at the sentence level, such that each sentence contained from 0 to N values. The biggest challenge was attaining high inter-annotator agreement. For this pilot study using the Schwartz Value Survey, only 17 of the 56 values were detected by both coders in a sample of four of the testimonies. Substantial agreement (Landis and Koch, 1977), or a Cohen's Kappa (Cohen, 1960) of .61–.80, was only achieved for two of the 17 values, and moderate agreement, or a Cohen's Kappa of .41–.60, for four of the 17 values. Thus, this pilot study served as a proof-of-concept, but more refinement of the annotation scheme and process was necessary.

Despite our limited success with attaining sufficiently high inter-annotator agreement, we moved forward with our parallel effort to explore ways of automating the annotation process (Ishita et al, 2009). Our first approach was to treat this as a multiple answer-multiple choice assignment task, since each value-laden sentence could be annotated with one or more human values. As such, we separated the tasks of detecting

Study	F-measure	Frequency Baseline
Shwartz Value Survey	0.30	0.37
Modified Schwartz Value Survey	0.45	0.49
Meta-Inventory of Human Values	0.70	0.47

Table 1 F_1 values for classification of net neutrality testimonies using three vales inventories with increasing inter-annotator agreement and inscreasingly sophisticated systems. (baseline=most frequent category)

value presence (i.e., whether *any* value is present) and classifying with specific values (i.e., *which* values are present), and focused initially on the latter task. We chose to initially focus on variations of the k Nearest Neighbor (kNN) algorithm, which easily accommodates multi-label classification. We first learned to rank order the labels using cross-validation, evaluating the resulting ordering using ranked retrieval measures such as Mean Average Precision (MAP). Looking ahead to our ultimate need to produce sharp decisions, we then learned a simple fixed threshold to select the top N of those values again using cross-validation. Our best result was achieved for $k = 20$, resulting in an F_1 measure of 0.30, an unimpressive result when compared to a dumb baseline that simply guessed the most frequently occurring value, which obtained an F_1 measure of 0.37, as shown in Table 1

To improve our results for automated approaches, we first had to improve our inter-annotator agreement, a common hurdle in mechanizing social science coding schemes (Krippendorff, 1980). Next, we refined the annotation scheme, and developed a coding guide. We used the Schwartz (1994) Value Survey as an initial inspiration, systematically eliminating values that did not occur within our dataset and systematically combining values that seemed difficult for human annotators to differentiate based on analysis of confusion matrices. The result was a coding scheme that contained ten values (effectiveness, human welfare, importance, independence, innovation, law and order, nature, personal welfare, power, and wealth). We applied this coding scheme to a subset of the overall corpus that included 28 testimonies. We obtained substantial agreement for five of the ten values and moderate agreement for two others. We also annotated the sentiment polarity toward Net neutrality expressed by the testimony author (pro, con, or neutral), and found that testimonies that argued in favor of Net neutrality more frequently invoked the value of innovation, while testimonies that argued against Net neutrality more frequently invoked the value of wealth (Cheng et al, 2012). Thus, this approach to annotating values was able to generate a useful social science finding.

Attempting to automate this process, we again used kNN with the more consistently appliable annotation scheme, again using cross validation after re-annotating the same corpus of 28 testimonies. Here, we were able to obtain an F_1 measure of 0.45 for a kNN classifier with a learned fixed threshold on the number of value labels to be assigned (Ishita et al, 2010). Although this yielded a substantial increase in the F_1 measure, the simple most-frequent-category baseline also increased to $F_1 = 0.49$. We therefore decided to continue to refine the annotation scheme, the coding guidelines, and the approach to automating the task.

We did this by first going back to the literature to find value concepts that transcended individual coding frames and for which broad consensus could be found. We

identified 12 relevant value surveys with similar purposes but from radically different fields, including social psychology, sociology, management, advertising, and human-computer interaction. We then conducted a thematic analysis of the values contained within the various inventories, clustering similar values from different inventories. Finally, we created a list of 16 values that were contained in at least five different value inventories, which we named the Meta-Inventory of Human Values (Cheng et al, 2010). We then refined this list of 16 values into a set of six values based on eliminating values that did not occur within the dataset and combining similar values that were easily confused. We also further refined the annotation guidelines. We applied the new inventory and guidelines to the entire corpus of 102 testimonies, and had two expert annotators independently code 20 of the testimonies. We obtained substantial agreement for four of the six values and moderate agreement for the other two values in the Meta-Inventory of Human Values. have used these annotations to explore a wide range of social science research questions and have obtained further meaningful findings (Cheng, 2012).

In addition to the improvements to the annotation scheme and annotation guide which improved inter-annotator agreement, we also adopted a different classification approach, modeling the task as a set of binary decisions, such that each sentence either reflected or did not reflect each value. We applied three machine learning approaches to this binary classification task using cross-validation: kNN, Support Vector Machine (SVM), and Naïve Bayes. We found that for this binary classification task SVM unsuprisingly outperformed kNN, and Naïve Bayes. For the four value labels that attained substantial agreement among human annotators, the average F_1 measure was 0.70. Moreover, this result substantially outperformed the most-frequent-category baseline, which obtained an F_1 measure of 0.47. All annotations produced for the 102 Net Neutrality testimonies are available online.[1]

Our next step will be to replicate the social science data analysis using completely automatic methods, both our own novel classification of values and off the shelf approaches for sentiment polarity, to attempt to replicate the social science findings obtained by human researchers. The studies that we have conducted to date indicate some promise for employing automatic annotation methods to assist social scientists with their research. The next section describes a study that built on this study, employing crowdsourcing as well as automatic methods to study a less formal debate around another contentious issue.

4 Crowdsourcing Values in the Park51 Debate

For our next series of studies, we took a different approach to detecting human values—instead of taking the approach of asking experts to annotate values, which provides the challenge of achieving sufficient inter-annotator agreement, we turned to Mechanical Turk (Snow et al, 2008) to collect crowdsourced data. Specifically, we asked Turkers to complete a value survey, the Portrait Value Questionnaire (PVQ) (Schwartz, 2007) and indicate their agreement or disagreement with statements that express different sentiment polarities toward a politically contentious development proposal, Park51.

[1] `http://stick.ischool.umd.edu/popit/`

Park51 is the official name for the Manhattan center that backers formerly called the "Cardoba House" and opponents called the "Ground Zero Mosque". Its backers sought to build a community center in a disused Burlington Coat Factory building. Opponents wanted to stop the development because of its proximity to the location of terrorist attacks on the World Trade Center in downtown Manhattan. We focused on the Park51 discussion in the fall of 2010, when it was widely covered in the media.

The PVQ expresses values as a set of portraits of people, and asks respondents to rate their identification with the portrait. For example, the value universalism is elaborated by Schwartz (1994) as including "Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature". In the PVQ, universalism is measured using three miniature portraits:

1. She thinks it is important that every person in the world be treated equally. She believes everyone should have equal opportunities in life (Schwartz, 2007).
2. It is important to her to listen to people who are different from her. Even when she disagrees with them, she still wants to understand them (Schwartz, 2007).
3. She strongly believes that people should care for nature. Looking after the environment is important to her (Schwartz, 2007).

Similarly, Schwartz (1994) elaborated security as "Safety, harmony, and stability of society, of relationships, and of self". In the PVQ, security is captured in two portrait items:

1. It is important to her to live in secure surroundings. She avoids anything that might endanger her safety (Schwartz, 2007).
2. It is important to her that the government insure her safety against all threats. She wants the state to be strong so it can defend its citizens (Schwartz, 2007).

After they completed the PVQ, we then asked individuals to respond to opinionated paragraphs on Park 51 by indicating how much each paragraph matched their point of view. We also assessed the sentiment expressed in each paragraph toward the project.

These three variables—individuals' values, their response to paragraphs, and the sentiment expressed in those paragraphs — allowed us to infer values salient to debate about Park51. In this context, we call a value salient if it is predictive of agreement with paragraphs. More specifically, salient values yield significant value and interaction parameters in an ordinary least squares regression model of the form

$$\mathbf{A} = \beta_0 + \beta_1 \mathbf{S} + \beta_2 \mathbf{V} + \beta_3 \mathbf{S} \cdot \mathbf{V} \tag{2}$$

In equation 2, A represents a vector of agreement scores, each characterizing an encounter between an individual and a paragraph. S represents sentiment scores for the paragraphs in these encounters, and V represents value level (high or low) of the individuals doing the agreeing or disagreeing. Intuitively, switching between levels of a salient value yields significantly different relationships between sentiment and agreement.

We found that the values universalism and security were most salient in the Park51 debate. For subjects scoring above the median on the value universalism, we found a strong positive correlation between paragraph sentiment polarity and agreement with paragraph. For subjects scoring below the median, we found a strong negative

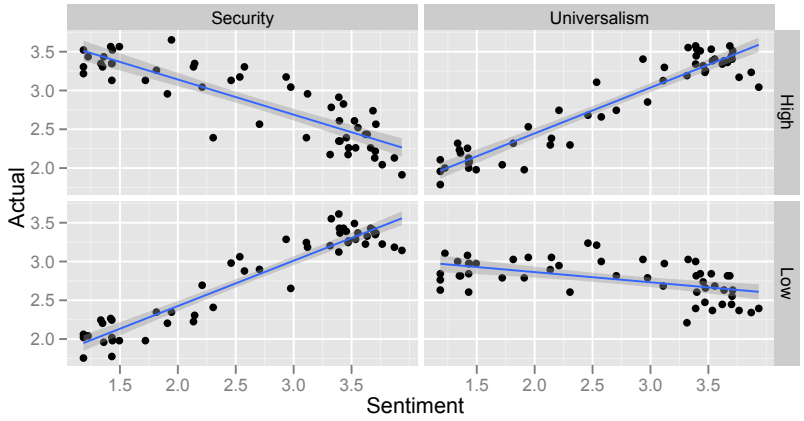

Fig. 1 Regression results using actual average agreement. This figure shows a plot of sentiment vs. average agreement for each value level using average agreement scores calculated using actual data. Results for the human value security are shown on the left, and results for universalism are shown on the right of the figure. For the subset of subjects who score low on the value of security (level = 0), agreement with paragraphs increases as the sentiment polarity expressed in the paragraph toward the Park51 project becomes more positive. For the subset who score high on security, agreement decreases as sentiment polarity becomes more positive. For the value of universalism, the entire pattern is reversed.

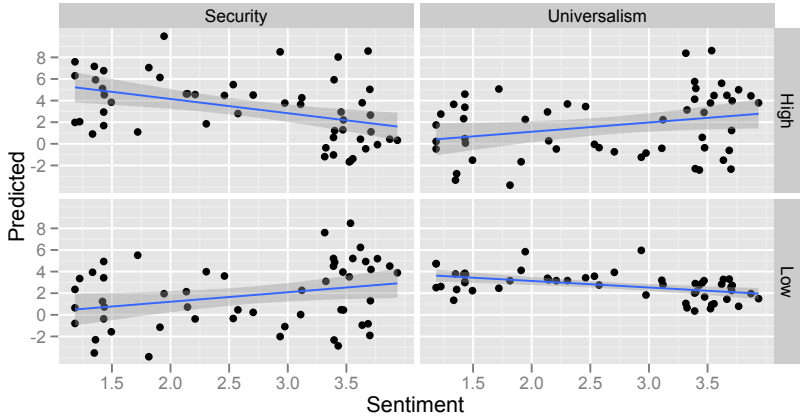

Fig. 2 Regression results using predicted average agreement. Figure 2 shows a plot of sentiment vs. average agreement for each value level using average agreement predicted by SVM regression. As is visible in the figure, the difference in slope between high and low value levels has the same sign in regressions using actual (figure 1) and predicted (figure 2) agreement scores, as does the difference in intercept.

correlation. The entire pattern was reversed for the value security (Fleischmann et al, 2011a; Templeton et al, 2011a; Templeton and Fleischmann, 2011). Thus, our crowdsourcing technique provides an alternative method of assessing the values salient to a debate.

It was a natural next step to see if we could train a model to behave as if it were a population of individuals. To explore the possibility of automatically obtaining social science results related to audience-specific behavior, we trained an SVM regression model for each audience to predict its expected agreement with paragraphs.

For each value, we split our audience of Turkers into two groups: Turkers who scored above the median, and Turkers who scored below. For each group thus obtained, we trained an SVM regression model (Joachims, 1999) to predict how much, on average, that group would agree with a given paragraph. The SVMs were evaluated using a leave one paragraph out approach.

Figure 1 and Figure 2 show results for the most salient values identified in our analysis, security and universalism. As is visible in the figures, regressions obtained by predicting agreement scores from text using SVMs produce trends that are similar to those that are obtained by using emperical agreement data. This was true for all the values we considered. The sign (positive or negative) of every parameter that was significant in the regressions obtained from actual average agreement scores was identical in the regressions obtained from predicted agreement scores. Moreover, there were no false positives in the regressions obtained from predicted agreement scores. Every statistically significant parameter in regressions obtained from predicted agreement scores was at least as significant in regressions obtained from actual agreement scores (Templeton et al, 2011b).

When we applied the crowdsourcing technique to opinionated articles published in the wake of the Fukushima nuclear disaster, the results were less distinct. We reasoned that perhaps public discussion of nuclear power was too rich and nuanced to identify salient values by simply examing the relationship between sentiment and agreement at different value levels. Perhaps the same value was being activated in support of more than one position.For example, texts in favor of nuclear power might invoke values related to the environment by contrasting nuclear power with coal, but so might anti nuclear texts in contrasting nuclear power with renewable energy. We chose issue frames (Chong and Druckman, 2007) as an organizing principle to help us more clearly identify values salient to a public policy issue (Koepfler et al, 2012).

5 Learning to be Social Scientists or to be People?

These two series of studies take fundamentally different approaches to employing machine learning for computational social science. The first series of studies, based on content analysis, involves training machines to detect human values in texts—essentially, machines learning to act like social scientists. The second series of studies, based on crowdsourcing, involves training machines to react differently to texts based on "values"—essentially, machines learning to act like the objects of study of social scientists: people.

One example of a research tool that these approaches might ultimately help to simulate is a focus group (Schwartz, 2007). Focus groups bring together stakeholders to gauge their perspective on a product, technology, message, or advertisement. Stakeholders may have different views depending, in part, on their values. Even the best-designed focus group study may, however, not be able to encompass all rel-

evant stakeholder groups and all perspectives (Krueger and Casey, 2009) because focus groups can be expensive, conducting a focus group requires significant expertise (Massey and Wallace, 1991), and because the voluntary nature of participation means that the participants are always self-selected. Moreover, if the information discussed is sensitive there may be some risk that participants would repeat what they have learned about the views of others in inappropriate social contexts. Thus, while focus groups are a tremendously valuable and effective social science research method, there will invariably be situations where other research methods are needed.

If we are able to predict how individuals with different values will interpret different texts, we might one day be able to construct a "focus group in a box" that allows for the simulation of different individuals or types of individuals reactions to texts (or, perhaps eventually, other media). This would be of interest to social scientists, who could have fast, inexpensive, and non-invasive access to a pool of simulated "individuals" without having to worry about potential harm to actual individuals. Diplomats could benefit from the ability to test messages on different audiences without risking a diplomatic crisis. Marketing analysts could cheaply and easily try out new sales pitches on diverse audiences. Political strategists could test new political campaign themes without worrying about the wrong themes going viral unexpectedly and being spread without permission or control, since simulated audiences tweet no tales (at least for now). Of course, much remains to be done before we could construct such simulated focus groups, and much more remains to be done before we will be able to characterize the degree of fidelity that they are able to achieve in comparison with actual human focus groups.

Each of our approaches has different strengths and weaknesses. Content analysis is an accepted social science research method, and one that can fairly easily be automated. However, one interesting feature of content analysis is that it requires consistent annotation among multiple annotators, which is typically achieved through a rigorous, time-consuming training process. Interestingly, the human must first be trained (to act like a machine) before the machine can be trained (to act like a human acting like a machine). People are not a uniform group that see the world the same way—rather, it is our diversity of perspectives that makes us human. Thus, it is useful to consider computational methods that see human diversity as a feature, rather than as a bug.

Our crowdsourced approach reflects human diversity as a feature that could inform downstream sentiment analysis tasks. There admittedly are some challenges with that approach, including the diversity of perspectives within and between groups, determining where to draw cutoffs between groups, and determining whether people have faithfully answered surveys in ways that truly reflect their values. Nonetheless, this approach does seems to us to have promise. Humans have diverse perspectives, and machine learning can and should embrace that diversity as an opportunity to model more human-like intelligences.

References

Cheng AS (2012) Values in the net neutrality debate: Applying content analysis to testimonies from public hearings. PhD thesis, University of Maryland

Cheng AS, Fleischmann KR (2010) Developing a meta-inventory of human values. In: Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47, American Society for Information Science, Silver Springs, MD, USA, ASIS&T '10, pp 3:1–3:10

Cheng AS, Fleischmann KR, Wang P, Ishita E, Oard D (2010) Values of stakeholders in the net neutrality debate: Applying content analysis to telecommunications policy. In: Proceedings of the 43rd Hawai'i International Conference on System Sciences

Cheng AS, Fleischmann KR, Wang P, Ishita E, Oard DW (2012) The role of innovation and wealth in the net neutrality debate: A content analysis of human values in congressional and fcc hearings. Journal of the American Society for Information Science and Technology 63:1360–1373

Chong D, Druckman JN (2007) A theory of framing and opinion formation in competitive elite environments. Journal of Communication 57(1):99–118

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:3746

Crowston K, Allen EE, Heckman R (In Press) Using natural language processing for qualitative data analysis. International Journal of Social Research Methodology

Fleischmann KR, Wallace WA (2006) Ethical implications of values embedded in computational models: An exploratory study. In: Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology

Fleischmann KR, Wallace WA (2009) Ensuring transparency in computational modeling. Communications of the ACM 52(3):131–134

Fleischmann KR, Wallace WA (2010) Value conflicts in computational modeling. Computer 43(7):57–63

Fleischmann KR, Wallace WA, Grimes J (2010) The values of computational modelers and professional codes of ethics: Results from a field study. In: Hawaii International Conference on System Sciences

Fleischmann KR, Templeton TC, Boyd-Graber J (2011a) Modeling diverse standpoints in text classification: Learning to be human by modeling human values. In: iConference

Fleischmann KR, Wallace WA, Grimes J (2011b) Computational modeling and human values: A comparative study of corporate, academic and government research labs. In: Hawaii International Conference on System Sciences

Fleischmann KR, Wallace WA, Grimes J (2011c) How values can reduce conflicts in the design process: Results from a multi-site mixed-method field study. In: Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology

Habermas J (1984) Theory of Communicative Action. Beacon Press

Hardisty E, Boyd-Graber J, Resnik P (2010) Modeling perspective using adaptor grammars. In: Proceedings of Emperical Methods in Natural Language Processing

Hill KS (2004) Defy the decades with multigenerational teams. Nursing Management 35:32–35

Hopkins DJ, King G (2010) A method of automated nonparametric content analysis for social science. American Journal of Political Science 54:229–247
Ishita E, Cheng AS, Oard DW, Fleischmann KR (2009) Multi-label classification for human values. In: Proceedings of the Annual Conference of the Japan Society of Library and Information Science, Tokyo, Japan
Ishita E, Oard DW, Fleischmann KR, Cheng AS, Templeton TC (2010) Investigating multi-label classification for human values. Proceedings of the American Society for Information Science and Technology 47(1):1–4, DOI 10.1002/meet.14504701116
Joachims T (1999) Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A (eds) Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, MA, chap 11, pp 169–184
Koepfler J, Templeton TC, Fleischmann KR (2012) Exploration of values and frames in social media texts related to the homeless hotspots debate. In: Proceedings of the 75th Annual Meeting of the American Society for Information Science and Technology (ASIS&T), Baltimore, MD
Krippendorff K (1980) Content Analysis: An Introduction to Its Methodology. Sage
Krueger RA, Casey MA (2009) Focus groups: A practical guide for applied research. Sage
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
Lin WH, Wilson T, Wiebe J, Hauptmann A (2006) Which side are you on? identifying perspectives at the document and sentence levels. In: Proceedings of the Conference on Natural Language Learning (CoNLL)
Massey AP, Wallace WA (1991) Focus groups as a knowledge elicitation technique: An exploratory study. IEEE Transactions on Knowledge and Data Engineering 3:193–200
Massey M (1979) The People Puzzle: Understanding Yourself and Others. Reston Publishing
Pang B, Lee L (2008) Opinion Mining and Sentiment Analysis. Now Publishers Inc
Parsons T (1964) Social structure and personality. Free Press of Glencoe
Paul M, Girju R (2010) A two-dimensional topic-aspect model for discovering multi-faceted topics. In: Association for the Advancement of Artificial Intelligence
Rokeach M (1973) The nature of human values. Free Press
Rokeach M (1979) Understanding Human Values. Free Press
Schwartz M, Weiser PJ (2009) Introduction to a special issue on network neutrality. Review of Network Economics 8(1):1
Schwartz SH (1994) Are There Universal Aspects in the Structure and Contents of Human Values? Journal of Social Issues 50(4):19–45, DOI 10.1111/j.1540-4560.1994.tb01196.x
Schwartz SH (2007) Universalism values and the inclusiveness of our moral universe. Journal of Cross-Cultural Psychology 38(6):711–728
Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of Emperical Methods in Natural Language Processing
Templeton TC, Fleischmann KR (2011) The relationship between human values and attitudes toward the park51 and nuclear power controversies. In: Proceedings of

the 74th Annual Meeting of the American Society for Information Science and Technology (ASIS&T), New Orleans, LA

Templeton TC, Fleischmann KR, Boyd-Graber J (2011a) Comparing values and sentiment using mechanical turk. In: Proceedings of the 2011 iConference, ACM, pp 783–784

Templeton TC, Fleischmann KR, Boyd-Graber J (2011b) Simulating audiences: Automating analysis of values, attitudes, and sentiment. In: IEEE International Conference on Social Computing

Turney PD, Littman ML (2003) Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) 21(4):315–346

Weber M (1947) The Theory of Social and Economic Organization. Free Press

Wilson T, Wiebe J (2005) Annotating attributions and private states. In: CorpusAnno '05: Proceedings of the Workshop on Frontiers in Corpus Annotations II, Association for Computational Linguistics, Morristown, NJ, USA

Wilson TA (2008) Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity and attitudes of private states. PhD thesis, University of Pittsburgh

Zhou Y, Fleischmann K, Wallace W (2010) Automatic text analysis of values in the enron email dataset: Clustering a social network using the value patterns of actors. In: System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pp 1 –10, DOI 10.1109/HICSS.2010.77