九州大学学術情報リポジトリ
Kyushu University Institutional Repository

A Word-Scale Probabilistic Latent Variable Model for Detecting Human Values

Takayama, Yasuhiro
Nat’ l Institute of Technology, Tokuyama College

Tomiura, Yoichi
Department of Informatics, Faculty of Information Science and Electrical Engineering, Kyushu University

Ishita, Emi
Kyushu University Library

Oard, Douglas W.
iSchool/UMIACS University of Maryland

他

https://hdl.handle.net/2324/1833092

出版情報：ACM International Conference on Information and Knowledge Management (CIKM). 2014, pp.1-, 2014-12-06. ACM
バージョン：
権利関係：

A Word-Scale Probabilistic Latent Variable Model for Detecting Human Values

Yasuhiro Takayama
Nat'l Institute of Technology, Tokuyama College
3538 Gakuendai, Shunan, Yamaguchi 745-8585 Japan
takayama@tokuyama.ac.jp

Yoichi Tomiura
Kyushu University
744 Motooka, Nishi-ku, Fukuoka 812-0395 Japan
tom@inf.kyushu-u.ac.jp

Emi Ishita
Kyushu University
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581 Japan
ishita.emi.982@m.kyushu-u.ac.jp

Douglas W. Oard
iSchool/UMIACS
University of Maryland
College Park, MD 20742 USA
oard@umd.edu

Kenneth R. Fleischmann
University of Texas at Austin
1616 Guadalupe Suite #5.202
Austin, TX 78701 USA
kfleisch@ischool.utexas.edu

An-Shou Cheng
Nat'l Sun Yat-sen University
70 Lien-hai Rd.
Kaohsiung City 80424 Taiwan
ascheng@mail.nsysu.edu.tw

ABSTRACT

This paper describes a probabilistic latent variable model that is designed to detect human values such as *justice* or *freedom* that a writer has sought to reflect or appeal to when participating in a public debate. The proposed model treats the words in a sentence as having been chosen based on specific values; values reflected by each sentence are then estimated by aggregating values associated with each word. The model can determine the human values for the word in light of the influence of the previous word. This design choice was motivated by syntactic structures such as noun+noun, adjective+noun, and verb+adjective. The classifier based on the model was evaluated on a test collection containing 102 manually annotated documents focusing on one contentious political issue — *Net neutrality*, achieving the highest reported classification effectiveness for this task. We also compared our proposed classifier with human second annotator. As a result, the proposed classifier effectiveness is statistically comparable with human annotators.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing - text analysis.

General Terms

Algorithms, Experimentation.

Keywords

Computational social science, computational linguistics, human values, probabilistic model.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661966.

1. INTRODUCTION

Social scientists have long found it useful to consider human values as latent variables that have explanatory value for the choices that people make [35]. For example, someone who values *innovation* over *wealth* might advocate open-source over proprietary software, while someone who values *freedom* over *social order* might resist efforts for gun registration. We can think of values as influencing not only how people form their own opinions, but also as undergirding how people seek to influence the opinions of others. In this paper, we focus on automatic detection of human values reflected in texts written by advocates of specific policy positions. We take a step in that direction by evaluating automated classification of human values.

Several inventories of human values are used in social science research (e.g., Friedman et al. [13]; Kahle et al. [22]; Kluckhohn [23]; Rokeach [29]; Schwartz [32]). Integrating key components of these studies, we adopted Cheng and Fleischmann's [6] human value definition, that is, "values serve as guiding principles of what people consider important in life." We also base our work on the Meta-Inventory of Human Values (MIHV), which was developed by Cheng and Fleischmann specifically for the test collection that we use by selecting values specific to the debate at issue and by iteratively refining annotation guidelines [7, 12, 10]. Our results, generated using a redistributable collection containing 102 documents with zero or more of six sentence-level human values annotations, indicate that high precision (near 0.8) can be reliably achieved for frequently invoked values with a useful degree of recall (0.55–0.82).

We achieved statistically significant classification effectiveness over existing baselines for this task using a new probabilistic latent variable model in which we first infer the association between human values and individual word-level human values as latent variables, and then we aggregate those results over all words in a sentence. The structure of our model allows us to model the potential effect of the preceding word, which proves to be useful. Moreover, analysis of 20 dual-annotated documents indicate that with about 80 training documents our automated technique is able to achieve results that are nearly as accurate as those obtained by an independent human annotator as a pseudo-classifier.

The remainder of this paper is organized as follows. In Section 2, we describe related work on human value research and on classification methods. Section 3 then introduces the test collection that we have used. Section 4 describes our approach to detect human values and Section 5 describes our proposed latent value model. Section 6 presents our results and Section 7 concludes the paper.

2. RELATED WORK

Content analysis is one of the approaches to detect human values [11]. The key idea in content analysis is for the social science researchers to personally examine naturally occurring content and to assign codes to that content that reflect their interpretation of that content using some pre-existing coding scheme. Subsequent statistical analysis is then done on the assigned codes rather than on the content. Hsieh and Shannon [17] refer to this combination of human interpretation and an existing coding scheme as a "directed approach". One of the limiting factors the directed approach is that the annotation costs scale linearly with the size of the collection. Early in the annotation process, personal involvement of the researcher is important because the theory on which any pre-existing coding scheme is built may need to be adapted for reflecting the unique characteristics of a collection on which social scientists wish to focus. Our automated techniques are intended only for the part of the process when coding guidelines have stabilized and a substantial amount of annotated data is available.

After we obtained sufficient annotated data, we could automate the annotation process using text classifiers [33] trained with that data. We are not the first to explore the automated annotation of human values for social science research. For example, Bengston et al. [2] used dictionary-based computer aided content analysis to identify how values about forestry have shifted from anthropocentric values to biocentric values over the period 1980 through 2002.

We first compared the effectiveness of a wide range of classifiers available within Weka [16], and we found that Support Vector Machines (SVMs) [20, 21] performed best. Therefore, we compare our proposed method to SVMs using bag-of-words and bigram features in Section 6. Because we introduce a latent variable model, supervised Latent Dirichlet Allocation (sLDA) offers another appropriate baseline [3]. Essentially, sLDA is an extension of LDA [4] in which the process of constructing the probabilistic latent variable model is influenced by the known association of words with labels in a set of training documents. sLDA based on generalized linear models is a general framework to model the documents and the responses. Our proposed probabilistic latent variable model also captures the relationships between the sentences and values. Thus, we compare our method with sLDA in Section 6.

Griffiths et al. [15] found modeling sequential dependencies between word classes to be helpful. Sequential dependencies between the words themselves can also be useful, but sparsity risks must be managed. With this in mind, we model sequential word dependencies with the label(s) assigned to one word stem depending in part on the label(s) assigned to only the previous word stem.

The structure of our problem resembles that of sentiment classification, which has been extensively researched [25, 27]. An important difference is that our classification of human values is most naturally cast as a multi-category multi-label classification, whereas sentiment analysis is typically modeled as single-category classification. Importantly, human values can help to explain sentiment, given their explanatory power in relation to attitudes and behavior [34]. What distinguishes our work is our focus on human values with a redistributable test collection and our modeling of relation between sentence-level and word-level values, and sequential dependencies among words in a sentence.

3. HUMAN VALUES TEST COLLECTION

The test collection we used in this paper was originally developed by Cheng et al. [7]. The collection includes 102 written prepared statements ("testimonies") from public hearings held by the U.S. Senate, House, and Federal Communications Commission (FCC) in which representatives of stakeholder groups offered advice to legislative and regulatory bodies on *Net neutrality*. The key question in the Net neutrality debate is whether the public interest is better served by nondiscriminatory access for all Internet traffic or by some set of reasonable network traffic management practices for certain types of content or services. Their annotation task focused on the relationship between advocacy positions and detectable human values reflected by (or appealed by) written prepared statements.

Traditional paper-based annotations for values posed two challenges: (1) annotated passages could be of any length, and indeed both short (clause-scale) and long (paragraph-scale) passages were annotated; and (2) annotated passages often did overlap, indicating that evidence for multiple values was present in some places. Cheng et al. [7] therefore elected to constrain the scope of each annotation to be a single sentence, but to allow more than one value per sentence. Clause annotations were extended to sentences, and passages that spanned sentences were accommodated by annotating several consecutive sentences. This set up a well-structured sentence annotation task for supervised machine learning. Their initial experience with sentence annotation revealed poor inter-annotator agreement. After some iteration of annotation guidelines, they concluded that the Schwartz Values Inventory [32], which was developed through and for surveys, was not necessarily transferable to (manual or automatic) content analysis. To address this concern, Cheng and Fleischmann [6] developed the Meta-Inventory of Human Values (MIHV) by looking for commonalities among the full range of values inventories proposed towards categories that could be reliably inferred during annotation for content analysis. They selected a subset of their MIHV appropriate to our collection, iteratively coding a subset of our collection and iteratively refined annotation guidelines using two annotators until inter-annotator agreement stabilized. The resulting collection is annotated for six human values, the definitions of which are in Appendix A.

Sentence splitting for the test collection had been performed manually, and all 9,890 sentences in 102 documents were manually annotated. Table 1 shows examples of annotated sentences. We subsequently removed sentences whose boundaries disagreed with those of TreeTagger [31], sentences that after removing words in the SMART stopword list [30] contained more than 40 words, and sentences that (after stopword removal) contained no words. The remaining 8,660 sentences were then stemmed by the Porter stemmer [28]. Table 2 shows the distribution across the six val-

Values	Sentence
freedom, *s-order*	Consumers are entitled to access the lawful Internet content of their choice
honor	I am one of the network engineers involved for many years in designing, implementing and standardizing the software protocols that underpin the Internet
innov., *freedom*	Part of the reason why the Internet is such a creative forumfor new ideas is that there are very few barriers to using the Internet to deliver products, information and services.
justice	Under these circumstances, requiring those most responsible for congestion to bear a greater percentage of the costs would be both good network management and fair from a consumer standpoint.
s-order	The Commission, under Title I of the Communications Act, has the ability to adopt and enforce the net neutrality principles it announced in the Internet Policy Statement.
wealth	Private investors will fund the construction of a broadband network only if there is a reasonable expectation that the company making that investment will recover the cost of its investment, including acompetitive return on capital.

Table 1: Examples of human values annotation.

Value	κ	# doc	# sentences
wealth	0.629	102	3,563
social order	0.683	102	2,859
justice	0.420	99	2,641
freedom	0.620	101	2,431
innovation	0.715	94	1,147
honor	0.430	80	352

Table 2: Inter-annotator agreement and prevalence.

ues. A total of 1,545 sentences were annotated as containing *no value*.

A second annotator independently had annotated 20 of the prepared statements (containing 2,430 sentences, after the same filtering process was applied). Table 2 also shows Cohen's kappa, a chance-corrected measure of inter-annotator agreement [1, 8, 24] for those 20 documents.

4. APPROACH FOR DETECTING VALUES

In order to detect human values, we have to take into account how the values are reflected in text. Surface language expressions for human values are different from those for most other subject classification problems. In using subject classification to classify a theme of the document, the themes are often directly represented by language expressions, typically by words that occur in the documents [33]. In the case of the human value classification, while a value may be indicated by a specific word in some cases, in many cases the value may be invoked somewhat more indirectly using situation-specific terminology. In a preliminary analysis of the corpus that we use in this paper, we found the following cases:

(1) A word represents value(s).
The word in a sentence represents the certain values. For example, the word "freedom" in sentence (a) expresses the value *freedom*, the word "protect" in sentence (b) expressed the value *social order*, and the word "winner" in sentence (c) expressed the both of *justice* and *wealth*. As shown in (a), value names themselves are usually good cue words for the values.

(a) *"This preserves consumers's freedom to go where they want, use the lawful services they want, and read and say what they want online."*

(b) *"Protecting customers and delivering a good Internet experience is not limited to curtailing spam or thwarting identity theft, for example."*

(c) *"Consumers, not network operators, must be allowed to continue to choose winners and losers in the content and applications marketplace."*

(2) A pair of words represents value(s).
The following sentence (d) has the value *innovation*, but the sentence (e) does not. The word pair "good idea" (adjective+noun) provides the value *innovation* ("idea" means a suggestion for possible course of action), but the word "idea" in the sentence (e) does not specify any value ("idea" means just a thought).

(d) *"Make sure there is always a fertile place for all of our good ideas to flourish."*

(e) *"That was, I believe, the first time that idea had been presented to this Committee."*

(3) A whole sentence represents value(s).
The annotator determined that the following sentence (f) invokes the value of *honor* based on its statement.

(f) *"I am one of the network engineers involved for many years in designing, implementing and standardizing the software protocols that underpin the Internet."*

(4) Contextual information is required to infer value(s).
The following sentences (h) and (i) are annotated based on context. Sentence (h) has *honor* because of the previous sentence (g) which has the values *honor*, *innovation*, and *wealth*. Sentence (i) has *freedom* and *wealth* with influence by the next sentence (j) (which also has the values *freedom* and *wealth*).

(g) *"This is an extraordinarily positive development for the nation's economy ... for our global competitiveness ... and for the next wave of broadband-driven investment and innovation."*

(h) *"How do we continue this progress?"*

(i) *"First and foremost, by recognizing that this market is contestable to all who wish to invest."*

(j) *"This is plainly evidenced by the growing array of companies doing just that in the marketplace today ... cable ... phone ... satellite power ... municipality ... WiFi ... WiMax ... Google and more ... all investing in what is increasingly a free-for-all for consumers' broadband business."*

From the above actual examples in the corpus, we can see that the human values are expressed in variety of forms and multiple values are assigned to a sentence.

Among several approaches to estimate the presence of a category from text, typical basic methods are naive Bayes, k nearest neighbors (kNN), and SVMs. Ishita et al. [18] adopted these methods to detect human values, however, the results showed that these methods alone are not sufficient. One reason is that human values cannot be represented by simple functions such that summation of factors of words in a sentence, contributing to each human value. These function cannot be capture that some specific words play a determining role to detect certain values. Based on the above considerations, we design our model to first infer the word-level human values corresponding to each word in a sentence as latent variables, and then aggregate them by logical bitwise OR (see Section 5.1) to estimate the sentence-level human values.

Another characteristic of language expressing human values is that multiple values can be expressed by a single sentence. There are the cases in which one word reflects multiple values, as example (c) above illustrates, and multiple words with values can appear in a sentence. As an example of that, the sentence *"Part of the reason why the Internet is such a creative forum for new ideas is that there are very few barriers to using the Internet to deliver products, information and services."* has the value *innovation* based on the word "creative" and the word pair "new ideas" (adjective+noun); and *freedom* based on the word "barrier." The above examples (d) and (e) in the case (2) suggest that word sense disambiguation directed by syntax is required to detect correct human values for word-level. Among several syntax patterns, we focus in our work on two-word collocations, modeling the value of the word in a way that can be influenced by the previous word because this covers many typical and frequent syntax patterns, as the above examples show.

In this paper, we model cases (1) and (2) above in the next section, that provide an adequate coverage of major cases, in anticipating that the above cases (3) and (4) are minor. We expect our design choice is effective for these major cases, and in our future research, we could perhaps further extend our model to represent whole-sentence meanings and long-distance context in more nuanced ways.

5. LATENT VALUE MODEL

In this section, we propose a new method for detecting human values by using a statistical language model we call our *Latent Value Model* (*LVM*, for short), that estimates the posterior probability of sentence $\mathbf{w}$ having values v using Gibbs sampling in a Markov Chain Monte Carlo framework [14]. In order to investigate relationships between words for detecting values as discussed in section 4, we take the effect of the preceding word into account in our *LVM*.

5.1 Preparation and Notation

A sentence $\mathbf{w}$ is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, ..., w_N)$, where w_n is the n-th word in the sequence. The sentence $\mathbf{w}$ has *sentence-level values* v, where $v \in \{0,1\}^6$ = {000000, 000001, ..., 111111}. Each bit in the sequence pattern represents one of the six values. The pattern 000000 means the corresponding sentence does not have any values.

We introduce latent variables $\mathbf{x}$ into the model to represent the value(s) associated with each word in the sentence. If the word w_n has a value x, the sentence $\mathbf{w}$ also has the value x. On the other hand, if no word w_n in a sentence $\mathbf{w}$ has value x, then the sentence $\mathbf{w}$ does not have value x. In addition, we assume that each word in a sentence has at most two values. The sequence of the values corresponding to the sentence $\mathbf{w}$ is denoted by $\mathbf{x} = (x_1, x_2, ..., x_N)$, where x_n is the *word-level value(s)* of the word w_n.

We restrict x_n to be an element in χ, where χ = {000000, 000001, 000010, 000011, 000100, ..., 110000}. The cardinality of χ is 22. We denote 000000 as μ_0, 000001 as μ_1, ..., and 110000 as μ_{21}, for convenience of notation. Restricting the number of values with which a word can be associated limits sparsity. Whether an at-most-two model is a good choice is an empirical question. In preliminary experiments, single-value models perform poorly and three values models show no further improvement.

The sentence-level values v are the result of logical bitwise OR operation $\oplus$ for all $x_n (1 \leq n \leq N)$. The sequence of word-level value(s), therefore, is restricted to the following $\chi^N(v)$, when the sentence-level values v are given.

$$\chi^N(v) = \{(x_1, x_2, \ldots, x_N) \in \chi^N | (x_1 \oplus x_2 \oplus \ldots \oplus x_N) = v\}.$$

For example, sentence-level values for the sentence *"Congress enact safeguards to preserve American consumers' longstanding freedom of Internet content choice."* calculated as *social order* (from the word "safeguards") and *freedom* (from the word "freedom" and "choice").

We also introduce another type of latent variables $\mathbf{y} = (y_1, y_2, ..., y_N)$ into the model. The *context indicator* y_n expresses whether the previous word w_{n-1} influences the value of w_n or not. When the values associated with word w_n are subject to the influence of the previous word w_{n-1}, y_n takes the numerical value 1. This design choice is motivated by syntactic structures such as noun+noun and adjective+noun, or semantic disambiguation associated with verb+noun and verb+adjective. Otherwise, y_n takes 0 (The values associated with w_n are determined by only w_n itself).

5.2 Model and Estimation of Values

5.2.1 Model

For the word sequence (w_{n-1}, w_n), the context indicator y_n follows a Bernoulli distribution $Bern(\theta_1^{(w_{n-1}, w_n)})$ and its parameter $\theta_1^{(w_{n-1}, w_n)}$ follows a Beta distribution with the parameters (α_0, α_1):

$$\frac{1}{Beta(\alpha_0, \alpha_1)} (1 - \theta_1)^{\alpha_0 - 1} \theta_1^{\alpha_1 - 1}.$$

$Beta(\alpha_0, \alpha_1)$ is a Beta-function. When y_n takes the value 0, the values associated with the word w_n follow a multinomial distribution $Multi(\phi_0^{w_n}, \phi_1^{w_n}, ..., \phi_{21}^{w_n})$ and its parameters $\phi_{0:21}^{w_n}$ follow a Dirichlet distribution with the pa-

foreach $n = 1, 2, ..., N$ **do**
  (i) **draw** *context indicator* y_n:
    $y_n|\theta_1^{(w_{n-1},w_n)} \sim Bern(\theta_1^{(w_{n-1},w_n)})$
  (ii) **draw** the *word-level value(s)* x_n:
    if $y_n = 1$ **then**
      $x_n|w_{n-1}, w_n, \phi \sim Multi(\phi_{0:21}^{(w_{n-1},w_n)})$,
    else $x_n|w_n, \phi \sim Multi(\phi_{0:21}^{(w_n)})$.
sentence-level values become:
  $v = (x_1 \oplus x_2 \oplus \ldots \oplus x_N)$.

Figure 1: Generative process of *LVM*.

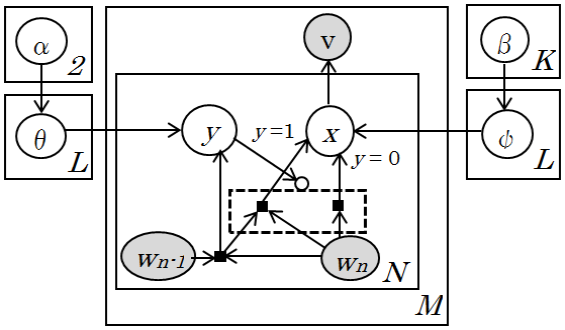


Figure 2: Graphical representation of *LVM*.

rameters $(\beta_0, \beta_1, ..., \beta_{21})$. Hereafter, we use $\phi_{0:21}^{w_n}$ and $\beta_{0:21}$ for notational convention. When the context indicator y_n takes 1, the values associated with the word w_n follow a multinomial distribution $Multi(\phi_0^{(w_{n-1},w_n)}, \phi_1^{(w_{n-1},w_n)}, ..., \phi_{21}^{(w_{n-1},w_n)})$ and its parameters $\phi_{0:21}^{(w_{n-1},w_n)}$ follow a Dirichlet distribution with the parameters $\beta_{0:21}$. The generative process for the sequence of sentence-level value patterns $\mathbf{x} = (x_1, x_2, ..., x_N)$, the sequence of context indicators $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ and the sentence-level human value(s) v for a sentence $\mathbf{w} = (w_1, w_2, ..., w_N)$ in our proposed *LVM* is as follows in Figure 1. In Figure 1, priors $\theta_u^{(w_{n-1},w_n)}$ are called *contextual affinities*, $\phi_{0:21}^{w_n}$ and $\phi_{0:21}^{(w_{n-1},w_n)}$ are *word-level associations*.

Our *LVM* is represented as a graphical model using conditioning with *gates* [26] in Figure 2. The outer plate represents sentences, while the inner plate represents generation of word-level values from a pair of the words or a single word by its context. The dotted box inside the inner plate shows the determination of previous word's influence depending on the context indicator y. The sentence-level value v is an aggregation of word-level values $\mathbf{x}$ for the corresponding sentence $\mathbf{w}$. In Figure 2, L is the number of distinct (w_{n-1}, w_n), W is the number of vocabulary and K is fixed at 22.

That is, our proposed model can be represented by the following equation (1).

$$P(\mathbf{x}, \mathbf{y}|\mathbf{w}, \theta, \phi) = \prod_{n=1}^{N} P(y_n|w_{n-1}, w_n, \theta) \times P(x_n|y_n, w_{n-1}, w_n, \phi), \quad (1)$$

where w_0 is the special symbol (\$) expressing the sentence head, and y_1 is always 0. The probabilities $P(y_n|w_{n-1}, w_n, \theta)$ and $P(x_n|y_n, w_{n-1}, w_n, \phi)$ are defined as follows:

$$P(y|a, b, \theta) = \begin{cases} \theta_0^{(a,b)}; & y = 0 \\ \theta_1^{(a,b)}; & y = 1 \end{cases},$$

$$P(x = \mu_j|y, a, b, \phi) = \begin{cases} \phi_j^{(b)}; & y = 0 \\ \phi_j^{(a,b)}; & y = 1 \end{cases}. \quad (2)$$

For simplifying notation, the symbol a represents the word w_{n-1}, and the symbol b represents w_n, the previous word of w_n in the equation (2), and the same style notation shall apply hereafter. The constant μ_j in equation (2) is the j-th possible word-level value(s) pattern as described in section 5.1.

We assume the following properties about the relation between words and their values:

(1) Most words do not have any values,

(2) For most two-word sequences, the values associated with the second word are probabilistically determined by that second word alone, without influence from the previous word.

We adopt a Bayesian approach to embed these properties in our model. The prior distribution of $(\theta_0^{(a,b)}, \theta_1^{(a,b)})$ is 2-dimensional Dirichlet distribution (beta distribution) $Dir(\alpha_0, \alpha_1)$. To reflect the property (2) above, we set the meta-parameters α_0 and α_1 as follows: $0 < \alpha_0, \alpha_1 < 1$ and $\alpha_0 > \alpha_1$. The prior distributions of $(\phi_0^{(b)}, \phi_1^{(b)}, ..., \phi_{21}^{(b)})$ and $(\phi_0^{(a,b)}, \phi_1^{(a,b)}, ..., \phi_{21}^{(a,b)})$ are 22-dimensional Dirichlet distributions $Dir(\beta_0, \beta_1, ..., \beta_{21})$. To reflect the property (1) above, we set the meta-parameters as follows:
$0 < \beta_0, \beta_1, ..., \beta_{21} < 1$ *and* $\beta_0 > \beta_1 + \beta_2 + ... + \beta_{21}$
Furthermore, to keep the number of meta-parameters small, we set the following restrictions:

$$\alpha_0 = \alpha, \alpha_1 = \gamma\alpha, \beta_0 = \alpha, \beta_i = \gamma\alpha/21 \quad (i = 1, 2, ..., 21, 0 < \alpha < 1 \text{ and } 0 < \gamma < 1). \quad (3)$$

Thus, the free meta-parameters are only α and γ.

When the word-level values $\mathbf{x}$ is determined, the sentence-level values v is uniquely determined as $v = x_1 \oplus x_2 \oplus ... \oplus x_N$. Therefore, the probability of $(\mathbf{x}, \mathbf{y}, v)$ given $\mathbf{w}$ is then:

$$P(\mathbf{x}, \mathbf{y}, v|\mathbf{w}, \theta, \phi) = \begin{cases} P(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \theta, \phi); & \mathbf{x} \in \chi^N(v) \\ 0; & otherwise \end{cases}.$$

The probability of $(\mathbf{x}, \mathbf{y})$ given $(\mathbf{w}, v)$ is therefore:

$$P(\mathbf{x}, \mathbf{y}|\mathbf{w}, v, \theta, \phi) \propto \begin{cases} P(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \theta, \phi); & \mathbf{x} \in \chi^N(v) \\ 0; & otherwise \end{cases}. \quad (4)$$

5.2.2 Estimation of Values

Let $(\mathbf{W}, V)$ be a collection of sentences and their values. $\mathbf{W} = (\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(M)})$, where $\mathbf{w}^{(m)}$ is the m-th sentence, and $V = (v_1, v_2, \ldots v_M)$, where v_m is the value(s) of the m-th sentence. The n-th word of $\mathbf{w}^{(m)}$ is denoted $w_n^{(m)}$, and the length of m-th sentence is denoted N_m. The collection $(\mathbf{x}^{(1)}\mathbf{x}^{(2)}, ..., \mathbf{x}^{(M)})$ is denoted $\mathbf{X}$, and the collection $(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}..., \mathbf{y}^{(M)})$ is denoted $\mathbf{Y}$ in a like manner. We

can get the probability of $(\mathbf{X}, \mathbf{Y})$ given $(\mathbf{W}, V)$ from (4) as follows:

$$P(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}, V, \theta, \phi) = \prod_{m=1}^{M} P(\mathbf{x}^{(m)}, \mathbf{y}^{(m)} \mid \mathbf{w}^{(m)}, v_m, \theta, \phi)$$

$$\propto \begin{cases} \prod_{a,b} \prod_{u \in \{0,1\}} \{\theta_u^{(a,b)}\}^{C_Y((a,b),u)} \\ \times \prod_b \prod_t \{\phi_t^{(b)}\}^{C_X(b,t,0)} \\ \times \prod_{a,b} \prod_t \{\phi_t^{(a,b)}\}^{C_X((a,b),t,1)} \\ \quad ; \ \mathbf{x}^{(m)} \in \chi^{N_m}(v_m) \ \ for\ all\ m \\ 0 \ ; \ otherwise \end{cases},$$

where $C_Y((a,b),u)$ is the number of times u has been assigned to a two-word sequence (a,b) as the value of context indicator y, $C_X(b,t,0)$ is the number of times value μ_t has been assigned to word b without the influence of the previous word, and $C_X((a,b),t,1)$ is the number of times value μ_t has been assigned to the word b with the influence of the previous word a.

When $\mathbf{x}^{(m)} \in \chi^{N_m}(v_m)$ for all $m = 1, 2, ..., and\ M$, we get the following formula by calculating the marginal probability:

$$\begin{aligned} &P(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}, V, \alpha, \gamma) \\ &= \int P(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}, V, \theta, \phi) \pi(\theta \mid \alpha, \gamma) \pi(\phi \mid \alpha, \gamma) d\theta d\phi \\ &\propto \prod_{a,b} \frac{\Gamma(\sum_{u \in \{0,1\}} \alpha_u)}{\prod_{u \in \{0,1\}} \Gamma(\alpha_u)} \frac{\prod_{u \in \{0,1\}} \Gamma(C_Y((a,b),u) + \alpha_u)}{\Gamma(\sum_{u \in \{0,1\}} \{C_Y((a,b),u)) + \alpha_u\})} \\ &\quad \times \prod_b \frac{\Gamma(\sum_t \beta_t)}{\prod_t \Gamma(\beta_t)} \frac{\prod_t \Gamma(C_X(b,t,0) + \beta_t)}{\Gamma(\sum_t \{C_X(b,t,0) + \beta_t\})} \\ &\quad \times \prod_{a,b} \frac{\Gamma(\sum_t \beta_t)}{\prod_t \Gamma(\beta_t)} \frac{\prod_t \Gamma(C_X((a,b),t,1) + \beta_t)}{\Gamma(\sum_t \{C_X((a,b),t,1)) + \beta_t\})}, \end{aligned} \quad (5)$$

where $\pi(\theta \mid \alpha, \gamma)$ and $\pi(\phi \mid \beta, \gamma)$ are the prior distribution of θ and the prior distribution of ϕ, respectively, and $\Gamma(\cdot)$ is the gamma function.

We can estimate the value(s) for a sentence $\mathbf{w}$ that has N words:

$$\begin{aligned} values(\mathbf{w}) &= \operatorname{argmax}_v P(v \mid \mathbf{w}, \alpha, \gamma) \\ &= \operatorname{argmax}_v \sum_{\mathbf{y} \in \{0,1\}^N} \sum_{\mathbf{x} \in \chi^N(v)} P(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \alpha, \gamma). \end{aligned}$$

Also we can estimate that $\mathbf{w}$ has the j-th value (denoted by $(v)_j$) of the six, when

$$\sum_{v:(v)_j=1} \sum_{\mathbf{y} \in \{0,1\}^N} \sum_{\mathbf{x} \in \chi^N(v)} P(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \alpha, \gamma) \geq \tfrac{1}{2}. \quad (6)$$

That is, whether a sentence has the j-th value is determined for each j separately. This means that the comparison among SVM, sLDA and *LVM* is fair enough because both judgements do not take into account combination of values.

5.2.3 Posterior Probabilities by Gibbs Sampling

We need the probabilities $P(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \alpha, \gamma)$ for every $\mathbf{x}$ and $\mathbf{y}$ to estimate the values of a sentence $\mathbf{w}$. These are the predictive posterior probabilities after giving the training data $(\mathbf{W}, V)$, $\hat{P}(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \mathbf{W}, V, \alpha, \gamma)$, to be exact. We obtain the following predictive posterior probabilities. We give the derivation of equation (7), the explanation of $\hat{P}(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \mathbf{W}, V, \mathbf{X}, \mathbf{Y}, \alpha, \gamma)$ and its calculation in Appendix B.

$$\begin{aligned} &\hat{P}(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \mathbf{W}, V, \alpha, \gamma) \\ &= \sum_{\mathbf{X}} \sum_{\mathbf{Y}} P(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}, V, \alpha, \gamma) \cdot \hat{P}(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \mathbf{W}, V, \mathbf{X}, \mathbf{Y}, \alpha, \gamma). \end{aligned} \quad (7)$$

By the law of large numbers, equation (7) can be approximated as follows:

$$\frac{1}{T} \sum_{t=1}^{T} \hat{P}(\mathbf{x}, \mathbf{y} \mid \mathbf{w}, \mathbf{W}, V, \mathbf{X}(t_o + t), \mathbf{Y}(t_o + t), \alpha, \gamma). \quad (8)$$

In equation (8), $(\mathbf{X}(s), \mathbf{Y}(s))$ is the s-th sample of $(\mathbf{X}, \mathbf{Y})$ that is drawn according to the posterior probability $P(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}, V, \alpha, \gamma)$, given by a Gibbs sampler. We get the conditional probability used in Gibbs sampling from equation (5) as follows:

$$When\ \ (x_1^{(m)} \oplus ... \oplus x_{n-1}^{(m)} \oplus \mu_j \oplus x_{n+1}^{(m)} \oplus ... \oplus x_{N_m}^{(m)}) \neq v_m :$$
$$P(x_n^{(m)} = \mu_j, y_n^{(m)} = u \mid \mathbf{X}_{-(m,n)}, \mathbf{Y}_{-(m,n)}, \mathbf{W}, V, \alpha, \gamma) = 0.$$

$$When\ \ (x_1^{(m)} \oplus ... \oplus x_{n-1}^{(m)} \oplus \mu_j \oplus x_{n+1}^{(m)} \oplus ... \oplus x_{N_m}^{(m)}) = v_m :$$

$$\begin{aligned} &P(x_n^{(m)} = \mu_j, y_n^{(m)} = 0 \mid \mathbf{X}_{-(m,n)}, \mathbf{Y}_{-(m,n)}, \mathbf{W}, V, \alpha, \gamma) \\ &\quad = \frac{C_Y^{-(m,n)}((w_{n-1}^{(m)}, w_n^{(m)}), 0) + \alpha_0}{\sum_{u \in \{0,1\}} \{C_Y^{-(m,n)}((w_{n-1}^{(m)}, w_n^{(m)}), u) + \alpha_u\}} \\ &\qquad \times \frac{C_X^{-(m,n)}(w_n^{(m)}, j, 0) + \beta_j}{\sum_t \{C_X^{-(m,n)}(w_n^{(m)}, t, 0) + \beta_t\}}, \\ &P(x_n^{(m)} = \mu_j, y_n^{(m)} = 1 \mid \mathbf{X}_{-(m,n)}, \mathbf{Y}_{-(m,n)}, \mathbf{W}, V, \alpha, \gamma) \\ &\quad = \frac{C_Y^{-(m,n)}((w_{n-1}^{(m)}, w_n^{(m)}), 1) + \alpha_1}{\sum_{u \in \{0,1\}} \{C_Y^{-(m,n)}((w_{n-1}^{(m)}, w_n^{(m)}), u) + \alpha_u\}} \\ &\qquad \times \frac{C_X^{-(m,n)}((w_{n-1}^{(m)}, w_n^{(m)}), j, 1) + \beta_j}{\sum_t \{C_X^{-(m,n)}((w_{n-1}^{(m)}, w_n^{(m)}), t, 1) + \beta_t\}}, \end{aligned}$$

where $\mathbf{X}_{-(m,n)}$ is $\mathbf{X}$ from which $x_n^{(m)}$ is removed, and $C_X^{-(m,n)}(\cdot)$ is a count that does not include the current assignment of $x_n^{(m)}$. The same holds for $\mathbf{Y}_{-(m,n)}$ and $C_Y^{-(m,n)}(\cdot)$ with $y_n^{(m)}$.

6. EXPERIMENTS

In this section, we describe our experiment design, report classifier effectiveness, and compare our automated results to those of a human annotator.

6.1 Experiment Design

We use 102-fold document-scale cross-validation (except in Table 3, where in preliminary experiments we had not grouped sentences by document). 102-fold cross-validation seeks to model the case in which some set of 101 documents have been annotated as training data and we are interested in the degree to which the machine can automatically code all future documents. To select the meta-parameters for each fold, we use 100 documents for development training and one held-out document for development testing. We perform a parameter sweep by training on all sentences in the development training set and then testing on all sentences in the one development testing document to select the meta-parameters α and γ that yield the best F_1, sweeping both parameters across 0.05, 0.1, 0.2, 0.5 and 0.9. The

101-document training set is trained using the best α and γ, and the resulting model is used to classify the sentences in the test set.

For Gibbs sampling we used 50,000 trials. Thirty percent of those trials were treated as the burn-in period. We used 1-for-3 samples of them as ($\mathbf{X}(t_0+1)$, $\mathbf{Y}(t_0+1)$), ($\mathbf{X}(t_0+2)$, $\mathbf{Y}(t_0+2)$), and so on to calculate equation (8). These parameters were empirically determined in preliminary experiments on development data. We apply the same process to determine the frequency threshold η for bigram features (use if frequency $\geq \eta$, a meta-parameter for SVM[1]) and to determine meta-parameters for sLDA.[2]

In order to examine influence of the previous word, we compare our *LVM* with *LVM*($y_n = 0$) which is our model without any influence from the previous word (i.e., with the context indicator y_n in the equation (1) always zero. The meta-parameters were same as for *LVM*.) We also compare our models with two types of SVM as fair baselines, SVM(w) and SVM(w, b). SVM(w) uses only word features, and SVM(w, b) uses word and bigram features. We use 2nd-degree polynomial kernel for SVM(w) and linear kernel for SVM(w, b), that kernels are determined respectively in experiments.

sLDA [3] is a general supervised method but it inherited the property of LDA [4, 14] which is a generative model for "documents" so that multiple topics are responsible for the words occurring in a single document. When we apply sLDA our test corpus, we assume that one sentence is regarded as a document. This setting might lose reliability of sLDA's behavior, because the expected number of words which have values in a sentence is a few. However, sLDA is a representative supervised probabilistic model, so we investigate how it works in the actual experiment.

6.2 Results

Table 3 shows results for SVM(w), SVM(w, b), sLDA, *LVM* (y_n=0) and *LVM*. For the comparison between SVM(w, b) and *LVM*, the difference of the error rate in the average F_1 between them, $z = 5.98 > Z_{0.975}$ (= 1.96) suggests that the equality can be rejected at significance level 0.05 by a z-test [9] [19]. *LVM* is much better than sLDA. We can see that even *LVM* (y_n=0) outperforms SVM(b) and SVM(w, b) (significantly for SVM(b), but not significantly for SVM(w, b)).

Table 4 shows classifier effectiveness by 102-fold document cross-validation. As can be seen, *LVM* apparently outperforms SVM(w, b). This is also true for sLDA, even when the number of topics is set to 22, which is the closest approximation to our model. Note that we omit *honor* from these micro-averaged results in Table 3 and 4 because no classifier did well for that category due to a scarcity of annotations for that value in our corpus, as illustrated in Table 2.

In sLDA, the response is regressed on the topic proportions, while the SVM calculates the weights for the response directly from words. We believe the reason why sLDA works so badly is as follows: (1) it is a model for "document" but not for "sentence" as we mention in the section 6.1; (2) linear regression of the latent variables for words to explain the response is not as well suited to our very sparse data as our estimation of the sentence-level values by a bitwise OR of the word-level values is.

[1]http://chasen.org/ taku/software/TinySVM/
[2]http://www.cs.cmu.edu/ chongw/slda/

Value	Precision		Recall		F_1	
	SVM	*LVM*	SVM	*LVM*	SVM	*LVM*
wealth	0.735	**0.816**	**0.871**	0.681	**0.797**	0.743
s-order	**0.775**	0.748	0.759	**0.820**	0.767	**0.782**
justice	0.664	**0.739**	0.464	**0.544**	0.546	**0.627**
freedom	0.681	**0.780**	**0.768**	0.704	0.722	**0.740**
innov	**0.764**	0.736	**0.720**	0.640	**0.741**	0.685
honor	0.395	**0.571**	**0.553**	0.094	**0.461**	0.162
average	0.712	**0.772**	**0.732**	0.668	**0.722**	0.716

Table 6: Human "classifier" and *LVM* effectiveness (same 20 test docs., micro-averaged).

In Table 5 shows per-category effectiveness measures for the SVM and for our *LVM*, respectively. For each comparison across the two classifiers, the bolded value is the higher of the two results. This is always true for F_1, even in the case of the category with the fewest training examples, *honor*. As Table 4 shows, SVM(w) and SVM(w, b) achieve nearly identical F_1 with 102-fold document cross-validation (the same condition reported in Table 5, which models the actual annotation process), with SVM(w) yielding $F_1 = 0.7166$ and SVM(w,b) yielding 0.7154. We therefore chose SVM(w) with the numerically higher score as the illustrative baseline for Table 5.

The value *honor* is omitted from the averages in Tables 3 and 4 because we focus our analysis of those tables on relative comparisons between usable classifiers. As Table 5 shows, the recall for *honor* is too low (0.26, meaning about 3 of every 4 cases are missed) for practical application. Table 5 also shows that our *LVM* achieves markedly better precision and recall (and thus better F_1) on *honor* than does SVM(w), so including *honor* in the micro-averages would not have changed the direction of the improvement that Tables 3 and 4 currently show.

To better understand the behavior of *LVM* on this collection, we have looked into the estimated word-level values as the first step of qualitative analysis. The social scientists collaborating on this research identified cue words used to invoke particular values during the annotation process. For example, *"American consumers will lose basic Internet freedoms, the engine of innovation will be hobbled, and our global competitiveness will be compromised"* which is annotated with *freedom*, *innovation*, and *wealth* as sentence-level values. The values names serve as good cue words, and *LVM* assigned the appropriate values for the words "freedom" and "innovation". As for *wealth*, *LVM* estimated that "competitiveness" has the word-level value *wealth* with influence from the previous word "global". We assumed that each word in a sentence has at-most-two values, and *LVM* aggregates the word-level values above then correctly estimated all three sentence-level values for the sentence. We plan to conduct more detailed qualitative analysis in our future work.

6.3 Comparison with Human Annotation

Because human values are unobservable private states rather than observable facts [36], we see the annotator's task as rendering an opinion about which values a statement reflects, and the system's task as replicating that result. As our inter-annotator agreement in Table 2 indicates, well trained and qualified people will sometimes make different judgments about the same sentence. To see how our *LVM* compares

Method	Precision	Recall	F_1
SVM(w)	0.7924	0.6802	0.7320
SVM(w, b)	0.7784	0.6988	0.7365
sLDA	0.7016	0.4821	0.5715
LVM($y_n = 0$)	0.7916	0.6931	0.7391
LVM	**0.8000**	**0.7132**	**0.7542**

Table 3: Classifier effectiveness (micro-averaged, w/o *honor*, 3 ×10-fold sentence cross-validation).

Method	Precision	Recall	F_1
SVM(w)	0.7784	0.6638	0.7166
SVM(w, b)	0.7535	0.6809	0.7154
sLDA	0.6875	0.4591	0.5506
LVM($y_n = 0$)	**0.7930**	0.6869	0.7361
LVM	0.7885	**0.6909**	**0.7365**

Table 4: Classifier effectiveness (micro-averaged, w/o *honor*, 102-fold document cross-validation).

The meta-parameters for sLDA: α= 0.05, 0.1, 0.2, 0.5 or 0.9 (fixed at initial α), the number of topics K=16, 22, 32, 64, 96 or 128. The meta-parameters for SVM(w, b): Bigram frequency threshold η = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ∞ (= w/o bigrams).

	Precision		Recall		F_1	
Value	SVM(w)	*LVM*	SVM(w)	*LVM*	SVM(w)	*LVM*
wealth	0.7859	**0.7934**	0.6977	**0.7392**	0.7392	**0.7654**
social order	**0.8235**	0.7803	0.7587	**0.8174**	0.7898	**0.7984**
justice	0.7275	**0.7800**	**0.5558**	0.5492	0.6302	**0.6446**
freedom	0.7461	**0.7927**	0.6654	**0.6742**	0.7035	**0.7287**
innovation	**0.8139**	0.8023	0.5629	**0.5817**	0.6655	**0.6744**
honor	0.4324	**0.6051**	0.2019	**0.2593**	0.2753	**0.3631**
average	0.7730	**0.7849**	0.6510	**0.6737**	0.7068	**0.7251**

Table 5: Per-category effectiveness (102-document cross-validation, micro-averaged).

with human annotator on a per-category basis, we ran experiments with the 20 documents (2,430 sentences) annotated by a second annotator as described in Section 3.

For this experiment, we trained *LVM* on the remaining 82 documents with meta-parameters: $\alpha = 0.2, \gamma = 0.9$ (most frequently selected meta-parameters during document cross-validation). For comparability, we treat the first annotator's annotations of those 20 documents as correct, and we compute effectiveness as if the second annotator were a classifier. The results are shown in Tables 6.

Although human performance is not necessarily an upper bound on performance (because the classifier has more access to evidence about how one annotator makes decisions than another human would), we see it as a useful reference because the utility of our classifier depends on its relative costs and benefits when compared to the alternative for coding at large scales, which would be to hire many annotators. Our results show that automation can achieve results similar to human annotation, but at a lower cost (in terms of human effort).

The difference of the error rate in the average F_1 between human and *LVM*, $z = 0.465 \leq Z_{0.975}$ suggests that the equality cannot be rejected in significance level 0.05. This means that *LVM* effectiveness is statistically indistinguishable from the human classifier. As can be seen, *LVM* does about as well as our human second annotator on average, and it does substantially better in both precision and recall (and thus in F_1) than the second annotator on *justice*. Notably, *honor* is markedly less problematic for the human second annotator than for *LVM*.

7. CONCLUSION

We have proposed a word-level probabilistic latent variable model for detecting the sentence-level human values reflected in prepared statements on a contentious political issue. The model treats the words in a sentence as having been chosen based on specific human values, and the values reflected by each sentence thus can be estimated by aggregating the values associated with each word. We have achieved the highest reported sentence classification effectiveness $F_1 = 0.737$ in 102-document cross-validation, which is a 3% relative improvement over SVM(w) that does not take account of sequential dependencies between words, as our model does. *LVM* also improved over SVM(w, b), which uses bigram features.

Our model can determine the human value(s) x_n for the word w_n in light of the influence of the previous word w_{n-1}. It is natural to next consider that word w_n's value(s) x_n might also be influenced by the both previous word w_{n-1} and following word w_{n+1}. This more complex model may suffer from sparsity, however. We might also explore using longer-distance syntactic dependencies found by a dependency parser, but since dependency parsing is imperfect, proximity features will likely continue to offer some benefit.

Acknowledgments

This work was supported in part by NSF IIS-0725459, Japan grant-in-aid for scientific research (B) 25280118, and DARPA contract HR0011-12-C-0015. Thanks go to Scott Block for serving as the second annotator for this project.

8. REFERENCES

[1] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[2] D. N. Bengston, W. T. J., and D. P. Fan. Shifting forest value orientations in the United States, 1980-2001: A computer content analysis. *Environmental Values*, 13(3):373–392, 2004.

[3] D. M. Blei and J. D. Mcauliffe. Supervised topic models. *Neural Information Processing Systems*, 17(4):537–544, 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[5] A.-S. Cheng. *Values in the Net neutrality debate: Applying content analysis to testimonies from public hearings*. Unpublished dissertation. University of Maryland, College Park, MD, 2012, 2012.

[6] A.-S. Cheng and K. R. Fleischmann. Developing a meta-inventory of human values. In *Proceedings of the American Society for Information Science and Technology (ASIST2010)*, volume 47(1), pages 1–10, 2010.

[7] A.-S. Cheng, K. R. Fleischmann, P. Wang, E. Ishita, and D. W. Oard. 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(7):1360–1373, 2012.

[8] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[9] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[10] K. R. Fleischmann. *Information and Human Values*. Morgan & Claypool, Reading, Massachusetts, 2014.

[11] K. R. Fleischmann, D. Oard, A.-S. Cheng, J. Boyd-Graber, T. C. Templeton, E. Ishita, J. A. Koepfler, and W. A. Wallace. Content analysis for values elicitation. In *Proceeding of the ACM SIGCHI 2012 Conference on Human Factors in Computing Systems, Workshop on Methods for Accounting for Values in Human-Centered Computing*, 2012.

[12] K. R. Fleischmann, D. W. Oard, A.-S. Cheng, P. Wang, and E. Ishita. Automatic classification of human values: Applying computational thinking to information ethics. In *Proceedings of the 72nd Annual Meeting of the American Society for Information Science and Technology*, 2009.

[13] B. Friedman, J. Kahn, P. H., and A. Borning. *Value Sensitive Design and Information Systems, In Human-computer interaction in management information systems: Foundations, 348-372*. M. E. Sharpe, New York, 2006.

[14] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5228–5235, 2004.

[15] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. *Advances in Neural Information Processing Systems*, 17, 2005.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[17] H.-F. Hsieh and S. Shannon. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288, 2005.

[18] E. Ishita, D. W. Oard, K. R. Fleischmann, A.-S. Cheng, and T. C. Templeton. Content analysis for values elicitation. In *Proceedings of the American Society for Information Science and Technology (ASIST2010)*, volume 47(1), pages 1–4, 2010.

[19] N. Japkowicz and M. Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[20] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *LS-8 Report 23, University of Dortmund*, 1998.

[21] T. Joachims. *Learning to classify text using support vector machines*. Springer Science+Business Media, New York, 2002.

[22] L. R. Kahle, B. Poulos, and A. Sukhdial. Changes in social values in the united states during the past decade. *Journal of Advertising Research*, 28:35–41, 1988.

[23] C. Kluckhohn. *Values and Value-Orientations in the Theory of Action: An exploration in definition and classification, In Parsons, T. and Shils, E. eds., Toward a general theory of action*. Harvard University Press, 1951.

[24] J. R. Landis and G. G. Koch. A one-way components of variance model for categorical data. *Biometrics*, 33:671–679, 1977.

[25] B. Liu. *Opinion Mining and Sentiment Analysis. In Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Data-Centric Systems and Applications, 459-526*. Springer-Verlag Berlin Heidelberg, 2011.

[26] T. Minka and J. Winn. Gates: A graphical notation for mixture models. *MSR-TR-2008-185, Microsoft Research*, 2008.

[27] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[28] M. F. Porter. An algorithm for suffix stripping. *In Readings in Information Retrieval (1997), Morgan Kaufmann*, pages 313–316, 1980.

[29] M. Rokeach. *Introduction to Information Retrieval. Chapter 13-15*. Cambridge University Press, 2008.

[30] G. M. Salton and M. J. McGill. The smart and sire experimental retrieval systems. *In Readings in Information Retrieval (1997), Morgan Kaufmann*, pages 381–399, 1980.

[31] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[32] S. H. Schwartz. Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50(4):19–45, 1994.

[33] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[34] T. C. Templeton, K. R. Fleischmann, and J. Boyd-Graber. Simulating audiences: Automating analysis of values, attitudes, and sentiment. In *Proceedings of the Third IEEE International Conference on Social Computing*, 2011.

[35] B. Verplanken and R. W. Holland. Motivated decision making: effects of activation and self-centrality of values on choices and behavior. *Journal of personality and social psychology*, 82(3):434–447, 2002.

[36] J. M. Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287, 1994.

APPENDIX

A. ANNOTATION SCHEME FOR VALUES

Table 7 shows the way we defined each annotated human value [5].

Value	Definition
freedom	The condition of being free of restraints and encouraging competition; allowing individuals to have their own beliefs and to make their own choices; freedom from interference or influence of another or others; the quality of being autonomous and independent.
honor	Understanding of who you are and how you are perceived by others; a feeling of pride in oneself or one's organization, group, or nation and belief in one's own worth; accomplishment that is honored, esteemed, respected or well regarded by yourself or others.
innovation	The capacity to create or discover new things and new ideas that contribute to the advancement of knowledge and/or technology.
justice	The state of being treated equally and fairly, especially having the same rights, status, and opportunities; the process of settling a matter properly and fairly for all parties according to their capabilities and needs, especially protecting the weak and correcting any injustice; need for equal or fair distribution of resources, information, benefits, burdens, and power among the members of a society.
social order	Using the power of the government, military and/or legal system to protect the stability of society and/or to protect people from possible harms mentally or physically; acting in accordance with laws, regulations, and social norms.
wealth	An explicitly stated concern with or interest in pursuing economic goals such as money, material possessions, resources, and profit; focusing on the market value of a change, decision, or action; allocating resources appropriately and/or efficiently.

Table 7: Definition and Annotation Scheme of Values.

B. PREDICTIVE POSTERIORS

The predictive posterior probabilities are calculated by integrating out θ and ϕ as follows:

$$\begin{aligned}&\hat{P}(\mathbf{x},\mathbf{y}\mid\mathbf{w},\mathbf{W},V,\alpha,\gamma)\\&=\int P(\mathbf{x},\mathbf{y}\mid\mathbf{w},\theta,\phi)\cdot\pi(\theta,\phi|\mathbf{W},V,\alpha,\gamma)d\theta d\phi\\&=\sum_{\mathbf{X}}\sum_{\mathbf{Y}}\tfrac{P(\mathbf{X},\mathbf{Y},V|\mathbf{W},\alpha,\gamma)}{P(V|\mathbf{W},\alpha,\gamma)}\\&\quad\int P(\mathbf{x},\mathbf{y}\mid\mathbf{w},\theta,\phi)\tfrac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi|\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V|\mathbf{W},\alpha,\gamma)}d\theta d\phi,\end{aligned}$$

where $\pi(\theta,\phi\mid\mathbf{W},V,\alpha,\gamma)$ is the posterior probability after giving the training data $(\mathbf{W},V)$.

In the last formula above,

$$\int P(\mathbf{x},\mathbf{y}\mid\mathbf{w},\theta,\phi)\frac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi\mid\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V\mid\mathbf{W},\alpha,\gamma)}d\theta d\phi$$

is the predictive posterior probability after observation ($\mathbf{W}$, V, $\mathbf{X}$, $\mathbf{Y}$), $\hat{P}(\mathbf{x},\mathbf{y}\mid\mathbf{w},\mathbf{W},V,\mathbf{X},\mathbf{Y},\alpha,\gamma)$.

For instance, when $\mathbf{w}=(a,\,b)$, $\mathbf{x}=(\mu_j,\,\mu_k)$ $(a\neq b)$, and $\mathbf{y}=(0,1)$, $\hat{P}(\mathbf{x},\mathbf{y}\mid\mathbf{w},\mathbf{W},V,\mathbf{X},\mathbf{Y},\alpha,\gamma)$ can be calculated by integrating out θ and ϕ as follows:

$$\begin{aligned}&\int\theta_0^{(\$,a)}\theta_1^{(a,b)}\phi_j^{(a)}\phi_k^{(a,b)}\frac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi\mid\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V\mid\mathbf{W},\alpha,\gamma)}d\theta d\phi\\&=1\times\frac{C_Y((a,b),1)+\alpha_1}{\sum_{u\in\{0,1\}}\{C_Y((a,b),u)+\alpha_u\}}\\&\times\frac{C_X(a,j,0)+\beta_j}{\sum_t\{C_X(a,t,0)+\beta_t\}}\times\frac{C_X((a,b),k,1)+\beta_k}{\sum_t\{C_X((a,b),t,1)+\beta_t\}}.\end{aligned}$$

Also, when $\mathbf{w}=(a,\,a)$, $\mathbf{x}=(\mu_j,\,\mu_k)$ $(\mu_j\neq\mu_k)$, and $\mathbf{y}=(0,0)$, $\hat{P}(\mathbf{x},\mathbf{y}\mid\mathbf{w},\mathbf{W},V,\mathbf{X},\mathbf{Y},\alpha,\gamma)$ can be calculated:

$$\begin{aligned}&\int\theta_0^{(\$,a)}\theta_0^{(a,a)}\phi_j^{(a)}\phi_k^{(a)}\frac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi\mid\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V\mid\mathbf{W},\alpha,\gamma)}d\theta d\phi\\&=1\times\frac{C_Y((a,a),0)+\alpha_0}{\sum_{u\in\{0,1\}}\{C_Y((a,a,u)+\alpha_u\}}\\&\times\frac{C_X(a,j,0)+\beta_j}{\sum_t\{C_X(a,t,0)+\beta_t\}}\times\frac{C_X((a,k,0)+\beta_k}{\sum_t\{C_X(a,t,1)+\beta_t\}}.\end{aligned}$$

However, in the case of $\mathbf{w}=(a,\,a)$, $\mathbf{x}=(\mu_j,\,\mu_j)$, and $\mathbf{y}=(0,0)$, $\hat{P}(\mathbf{x},\mathbf{y}\mid\mathbf{w},\mathbf{W},V,\mathbf{X},\mathbf{Y},\alpha,\gamma)$ becomes:

$$\begin{aligned}&\int\theta_0^{(\$,a)}\theta_0^{(a,a)}\phi_j^{(a)}\phi_j^{(a)}\frac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi\mid\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V\mid\mathbf{W},\alpha,\gamma)}d\theta d\phi\\&=1\times\frac{C_Y((a,a),0)+\alpha_0}{\sum_{u\in\{0,1\}}\{C_Y((a,a),u)+\alpha_u\}}\\&\times\frac{C_X(a,j,0)+\beta_j+1}{\sum_t\{C_X(a,t,0)+\beta_t\}+1}\times\frac{C_X((a,j,0)+\beta_j}{\sum_t\{C_X(a,t,1)+\beta_t\}},\end{aligned}$$

because of the property of the Γ function: $\Gamma(z+2)=(z+1)z\Gamma(z)$. When there are more than two occurrences for one unique word, we have to take into account a large number of combinations for the theoretically-derived calculation. Then we approximate above calculation as:

$$\begin{aligned}&\int\{\phi_j^{(a)}\}^2\frac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi\mid\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V\mid\mathbf{W},\alpha,\gamma)}d\theta d\phi\cong\\&\left\{\int\phi_j^{(a)}\frac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi\mid\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V\mid\mathbf{W},\alpha,\gamma)}d\theta d\phi\right\}^2\end{aligned}$$

By this approxmation, the predictive probability in the last case above becomes as follows:

$$\begin{aligned}&\int\theta_0^{(\$,a)}\theta_0^{(a,a)}\phi_j^{(a)}\phi_j^{(a)}\frac{P(\mathbf{X},\mathbf{Y},V,\theta,\phi\mid\mathbf{W},\alpha,\gamma)}{P(\mathbf{X},\mathbf{Y},V\mid\mathbf{W},\alpha,\gamma)}d\theta d\phi\\&=1\times\frac{C_Y((a,a),0)+\alpha_0}{\sum_{u\in\{0,1\}}\{C_Y((a,a),u)+\alpha_u\}}\\&\times\frac{C_X(a,j,0)+\beta_j}{\sum_t\{C_X(a,t,0)+\beta_t\}}\times\frac{C_X(a,j,0)+\beta_j}{\sum_t\{C_X(a,t,1)+\beta_t\}}.\end{aligned}$$

We found the difference between the theoretically-derived and the approximate calculation was not statistically significant in preliminary experiments. We therefore used the approximate calculation in our actual implement for efficiency reasons.