# Investigating Multi-Label Classification for Human Values

**Emi Ishita**
Faculty of Media & Information Resources
Surugadai University, Japan
emi@surugadai.ac.jp

**Douglas W. Oard**
Coll. of Info. Studies/UMIACS
University of Maryland, USA
oard@umd.edu

**Kenneth R. Fleischmann**
**An-Shou Cheng**
**Thomas Clay Templeton**
College of Information Studies
University of Maryland, USA
{kfleisch,anscheng,clayt}@umd.edu

**ABSTRACT**
This paper describes the development of a scalable process for people and machines working together to identify sections of text that reflect specific human values. A total of 2,005 sentences from 28 prepared testimonies presented before hearings on Net neutrality were manually annotated for one or more of ten human values using an annotation frame based on experience annotating similar content using the Schwartz Values Inventory. Moderately good agreement suggests that meaningful distinctions can often be drawn by human annotators. Several k-Nearest-Neighbor classifiers were compared in this preliminary study, yielding results that appear promising and that clearly point to productive directions for future work.

**Keywords**
Computational social science, automatic text classification.

**INTRODUCTION**
When we want to know what people are thinking, we can ask them (which is the basis for survey research), or we can look at what they say and write (broadly, "content analysis"). The explosive growth of Internet-accessible user-generated content has the potential to shift the balance between those complementary approaches towards progressively greater use of content analysis (Cheng et al., 2008). In particular, content analysis permits longitudinal analysis that is unmatched by other research techniques (Morris, 1994). For example, Cheng et al. (2010) identified relationships between sentiment toward Net neutrality and human values (beliefs pertaining to desirable end states or modes of conduct that transcend specific situations), and detected changes in those values between presentations to different audiences at different times. Extending that sort of analysis to far larger settings (e.g., discussion of the adoption of specific technologies in the trade press over multiple time scales) requires that content analysis be extended to larger text collections than manual annotation alone could possibly accommodate. Fleischmann et al. (2009) therefore proposed developing tools that allow human annotators and automated classification techniques to work together to identify the expression of human values in large collections. Humans and machines bring complementary strengths to this task: machines are less adept at classifying novel phenomena, but human annotation exhibits substantial variation (e.g., due to differing conceptualizations of the task over time, or due to fatigue). Our ultimate goal is therefore to develop a human-machine system that can consistently perform accurate annotation at a scale far greater than would be possible through human effort alone.

Our narrower goal in this paper is to begin to investigate the quantity vs. quality tradeoff that results from using a limited amount of human annotation effort to enable automatic annotation of additional content. Although our ultimate goal is to annotate texts from a broad range of sources, including news, blogs, and transcribed speech, we focus here on classification of content from a single source based on learning from annotations made by a single human annotator. We have, therefore, created a "test collection" in which sentences have been annotated with human values. In the next section, we briefly review related work on use of automatic classification in social science research. We then introduce our test collection (which we are happy to share), describe the classifiers that we have tried, and present evaluation results. We conclude with thoughts on next steps.

**RELATED WORK**
Although the use of machine learning for text classification has been well studied by computational linguistics and information retrieval researchers, there has to date not been a great deal of uptake of these techniques in the social sciences. Notably, Scott & Smith (2005) used Leximancer to automate the annotation of newspaper articles based on hand annotation of some short segments. As in computational linguistics, variants of item-level accuracy are the most often reported intrinsic measures of classifier accuracy, but Hopkins & King (2007) developed measures based on measured bias in aggregate results for an opinion analysis system that they applied to create aggregate data on opinions expressed in blogs about presidential candidates. Some computational linguistics research has focused on annotating sentiment as a form of subjectivity (e.g., Wiebe, Wilson, & Cardie, 2005). Pang and Lee (2008) extended the range of subjectivity in language to



*ASIST 2010,* October 22–27, 2010, Pittsburgh, PA, USA.

values, or "what a person or group of people consider important in life" (Friedman et al., 2006).

## TEST COLLECTION

We have built a test collection from 28 written statements prepared in advance by witnesses invited to testify before "Net neutrality" hearings held by the U.S. Senate Committee on Commerce, Science, and Transportation on February 7, 2006 (U.S. Senate, 2006), and by the Federal Communications Commission (FCC) on April 17, 2008 (FCC, 2008). Initially, we tried using the Schwartz (1992) Value Inventory (SVI), which forms the basis for a widely used survey instrument, as an annotation frame (Cheng et al., 2010). The SVI's 56 categories proved too fine-grained for our analysis, and using the higher-level categories provided by the SVI's three-level ontology did not substantially improve inter-annotator agreement. We therefore developed a specialized annotation frame, one informed by the SVI, which contains 10 categories: Effectiveness, Human Welfare, Importance, Independence, Innovation, Law and Order, Nature, Personal Welfare, Power, and Wealth. To simplify automatic classification and calculation of inter-annotator agreement, we chose to annotate sentences rather than arbitrary passages. A sentence might reflect one or more values, or it might reflect no values. Human annotators view sentences in context, while (for now) our automated classifiers do not.

The principal annotator (the fourth author of this paper) manually annotated all 2,294 sentences in the 28 documents. Table 1 shows some examples. This yielded a total of 3,403 category annotations over 1,783 sentences that were annotated with at least one category; the 511 sentences with no assigned categories were removed. The number of annotations per sentence in the resulting corpus ranges from 1 to 7, with a median of 2 and a mean of 1.91. The average sentence length is 24.5 words. We selected 4 documents for annotation by four additional annotators (including the third and fifth authors of this paper). There are a number of cases in which one annotator annotated no values for a sentence while the other annotator annotated one or more values. There are, of course, also cases in which both annotators annotated different values. These differences in annotation may be due to differences in interpretation, or to simple errors. As Artstein & Poesio (2008) recommend, we use multiple- $\pi$ (computed on binary agreement for each category and then macro-averaged over categories) for characterizing the chance-corrected agreement between multiple annotators:

$$\pi = \frac{A_0 - A_e}{1 - A_e}$$

Let $n_{ik}$ be the number of times item $i$ is classified in category $k$. Each category $k$ contributes $\binom{n_{ik}}{2}$ pairs of agreeing judgments for item $i$; the amount of agreement $agr_i$ for item $i$ is therefore the sum of $\binom{n_{ik}}{2}$ over categories $k \in K$, divided by $\binom{c}{2}$, the total number of judgment pairs per item. The overall observed agreement is the mean of $agr_i$ for all items $i \in I$

$$A_0 = \frac{1}{i}\sum_{i \in I} \text{agr}_i = \frac{1}{ic(c-1)}\sum_{i \in I} \sum_{k \in K} n_{ik}\,(n_{ik} - 1)$$

$$A_e^{\pi} = \sum_{k \in K} \left(\hat{P}(k)\right)^2 = \sum_{k \in K} \left(\frac{1}{ic} n_k\right)^2 = \frac{1}{(ic)^2} \sum_{k \in K} {n_k}^2$$

where $i$ is the number of items (i.e., the number of sentences), $c$ is the number of annotators, and $n_k$ is the total assignments for category $k$ (i.e., yes or no). For our five annotators and the full set of 2,294 sentences, multiple- $\pi$ is 0.306, which corresponds to "fair" agreement according to Landis & Koch (1977), and which is well below what is normally considered satisfactory for training and evaluating automated systems in computational linguistics. As presently formulated, this is a difficult task for humans. We are, therefore, now developing more carefully specified annotation guidelines. For the remainder of this paper we use the principal annotator's annotations as ground truth.

| Labels | Sentence |
|---|---|
| Effectiveness, Independence | Its open nature has enabled those with unique interests and needs to meet and form virtual communities like no tool before it. |
| Human Welfare, Independence, Power, Wealth | It has also empowered consumers as citizens and as entrepreneurs. |
| Effectiveness, Innovation | Consumers are increasingly creative in the way that they use these new technologies - nowhere more so than here in Silicon Valley. |

**Table 1. Examples of values annotation.**

## MULTI-LABEL CLASSIFICATION

Multi-label classification raises two principal challenges: (1) how should multiple labels be used during training?, and (2) how many labels should a classifier assign? Tsoumakas and Katakis (2007) suggest five methods for selecting training instances when multiple labels are present. We tried two: (Train 1) replicating each sentence that has more than one label and assigning a unique label to each replication, and (Train 2) selecting only the most selective label (in our case, the label with the lowest frequency in the training set) for each training sentence. This distinction affects only training; in both cases, the machine's task is to assign the right set of labels to each sentence in the evaluation set. We had tried two other methods from Tsoumakas and Katakis (2007) previously with disappointing results (on the SVI category set): creating an aggregate for each label combination in the training set, or limiting the training set to sentences with a single label.

We used *k*-Nearest-Neighbor (*k*NN) classifiers (with *k*=1, 3, 5, 10, 15, …, 40) from the University of Waikato's Weka toolkit. Preliminary experiments showed that stemming to

be helpful so we used the Snowball implementation of the Porter stemmer. Terms occurring four or fewer times in the entire collection were removed. We tried two ways of weighting the *k* examples: equal-weighted (voting) and inverse-distance weighted (*w*=1/distance). In the tables below, we refer to these as "vote" and "iw." Weka's *k*NN classifiers rank candidate labels. We used two methods for deciding how many labels to select: Oracle and Threshold. For the (unfair) Oracle condition, we assigned the same number of labels as in the evaluation data, thus producing an approximate upper bound on the accuracy of our classifier. If sentence *s* has *i* labels in the ground truth, we simply select Weka's *i* most probable labels. In case of ties, all labels with the same classifier-assigned score are chosen.

For the Threshold condition, we learned a threshold on the score assigned by the *k*NN classifier. To set this threshold, we divided the test collection into three sets; 80% for training the *k*NN classifier (the "training set"), 10% for learning the threshold (the "devtest" set), and the remaining 10% for evaluation (the "evaluation" set). We set the threshold using the following steps; 1) learn a classifier using the training set, 2) automatically assign label probabilities to each devtest sentence, 3) select the threshold to optimize *F* on devtest, 4) automatically assign label scores to each sentence in the evaluation set, and then 5) select all labels with a score higher than the threshold. We repeat both (Oracle and Threshold) with 10 disjoint evaluation sets, reporting 10-fold cross-validation results.

### Results

We are interested in both false negatives and false positives, so we elected to compare the performance of each method by first computing macro-averaged precision (number of correctly assigned categories over number of assigned categories) and recall (number of correctly assigned categories over number of human-annotated categories) and then computing the balanced *F* measure (the harmonic mean of precision and recall). Table 2 shows the macro-averaged F for automatic classification by Train 1 and Train 2 on the Oracle condition; Train 1 does slightly better than Train 2. "Threshold" is the macro-averaged *F* for classification using a score threshold to select the number of categories to be assigned. For these experiments, we swept the threshold between 0.01 and 0.25 in increments of 0.01 separately for each fold and selected the threshold that yielded the best *F* on the devtest set for that fold. The best macro-averaged *F* is 0.48 (for *k*=25, vote). This result is quite close to the best comparable result for the Oracle condition (0.45 for *k*=25, vote). We therefore conclude that our simple threshold selection method is reasonably effective. The results are relatively insensitive to *k*, and both weighting schemes yield similar results.

| | Oracle | | Oracle | | Threshold | |
|---|---|---|---|---|---|---|
| | Train 1 | | Train 2 | | Train 1 | |
| ***k*** | **vote** | **iw** | **vote** | **iw** | **vote** | **iw** |
| 15 | 0.46 | 0.46 | 0.43 | 0.43 | 0.43 | 0.44 |
| 20 | 0.48 | 0.48 | 0.45 | 0.44 | 0.44 | 0.44 |
| 25 | **0.48** | 0.47 | **0.46** | 0.45 | **0.45** | 0.45 |
| 30 | 0.48 | 0.47 | 0.45 | 0.45 | 0.45 | 0.45 |
| 35 | 0.47 | 0.47 | 0.45 | 0.45 | 0.45 | 0.45 |

**Table 2. Classification accuracy using *k*NN (*F*).**

### ANOTHER LOOK AT HUMAN-SYSTEM AGREEMENT

Although *F* is a widely reported measure for classifier accuracy, it is rather opaque as an absolute measure; the better use of *F* is as a basis for comparing alternative classification techniques. Our ultimate goal suggests a natural comparison: what *F* would another human performing the same task achieve? The results in Table 2 suffer from two artificialities that are useful during development, but would make such a comparison less informative: (1) they focus only on sentences to which at least one label was assigned in the ground truth, and (2) the cross-validation was performed without regard to which document a sentence came from (since our system presently takes no advantage of any context beyond the sentence).

In order to establish comparable conditions, we selected the same four prepared statements that were annotated by the four additional human annotators; this yields 227 sentences in the evaluation set, including sentences with no assigned values. As before, we use the principal annotator's annotations as ground truth, and we compute *F* for our system and for each other assessor (again, computing *F* on binary agreement for each category and then macro-averaging over categories). For the classification system, the remaining 24 testimonies constitute the training data. For training, we use only sentences with at least one annotated value, and we use the same threshold learned in the cross-validation experiment. For the results in Table 2 we had always selected at least one value, but here we would allow the threshold to exclude all categories.

Table 3 shows the macro-averaged *F* for each annotator (numbered 2 … 5) and our best *k*NN threshold classifier from the earlier experiments (*k*=25, vote). Clearly there is considerable room for improvement, with the best human annotators achieving *F* values more than twice as high as our present system. Our preliminary analysis points to problems resulting from the use of a single threshold. As Table 4 shows, our automated system is overly sensitive, as the system assigned at least one label to every sentence.

| Second Annotator | | | | 25-NN Vote |
|---|---|---|---|---|
| 2 | 3 | 4 | 5 | |
| 0.64 | 0.66 | 0.70 | 0.39 | 0.31 |

**Table 3. Comparing human and system annotation (*F*)**

| | Ground Truth | 2 | 3 | 4 | 5 | 25-NN Vote |
|---|---|---|---|---|---|---|
| **Effectiveness** | **24** | 31 | 13 | 27 | 0 | 222 |
| **Human Welfare** | **51** | 36 | 40 | 60 | 17 | 111 |
| **Importance** | **67** | 48 | 54 | 54 | 6 | 196 |
| **Independence** | **49** | 46 | 22 | 63 | 36 | 124 |
| **Innovation** | **16** | 39 | 17 | 31 | 13 | 38 |
| **Law and Order** | **24** | 42 | 15 | 40 | 19 | 82 |
| **Power** | **49** | 37 | 28 | 42 | 7 | 135 |
| **Wealth** | **23** | 65 | 12 | 25 | 1 | 212 |
| None | **58** | 49 | 102 | 59 | 138 | 0 |

**Table 4. Number of sentences per category.**

## NEXT STEPS

There are several ways in which we might improve our classifiers. For example, we can train a first-stage classifier to determine when no value should be assigned. We also need to explore a broader range of classifiers, including a cohort of support vector machines and/or maximum entropy classifiers (e.g., one per category). We are not yet taking advantage of patterns of category co-occurrence in the training data, but we are exploring the use of language modeling techniques for that purpose. We can also almost certainly improve our feature set by exploiting broader context, by feature selection, or by some form of smoothing or dimensionality reduction. Moreover, we might also explore the use of non-lexical features. We have to date focused on intrinsic evaluation measures for individual decisions. Although these measures are useful during development, extrinsic evaluation on actual content analysis tasks will be important for our intended applications in social science research. Moreover, for future work with intrinsic measures we will be particularly interested in characterizing classifier bias, not just in measuring improvements in mean values. While much remains to be done, our results suggest that that scalable approaches that are sufficiently accurate for some social science applications may be achievable.

## ACKNOWLEDGMENTS

Thanks to the PopIT team, especially annotators Karen Viruez-Munoz and Mardy Shualy. This work was supported in part by NSF IIS-0725459, and Japan Grant-in-Aid for scientific research (C) 22500220.

## REFERENCES

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555-596.

Cheng, A.-S., Fleischmann, K.R., Wang, P., & Oard, D.W. (2008). "Advancing Social Science Research by Applying Computational Linguistics." *ASIS&T*, Columbus, OH.

Cheng, A.-S., Fleischmann, K.R., Wang, P., Ishita, E., & Oard, D.W. (2010). Values of Stakeholders in the Net Neutrality Debate: Applying Content Analysis to Telecommunications Policy, *HICSS*, Kauai, HI.

FCC. (2008). *Broadband Network Management Practices Public Hearing.* Palo Alto, CA, April 17.

Fleischmann, K.R., Oard, D.W., Cheng, A.-S., Wang, P., & Ishita, E. (2009). Automatic Classification of Human Values: Applying Computational Thinking to Information Ethics, *ASIS&T*, Vancouver, BC, Canada.

Friedman, B., Kahn, P.H., Jr., & Borning, A. (2006). Value Sensitive Design and Information Systems. In P. Zhang & D. Galletta (Eds.), *Human-Computer Interaction in Management Information Systems: Foundations* (pp. 348-372). New York: M. E. Sharpe.

Hopkins, D., & King, G., (2007). *Extracting Systematic Social Science Meaning from Text*, available at http://gking.harvard.edu/files/words.pdf.

Landis, J.R., & Koch. G.G., (1977). The Measurement of Observer Agreement on Categorical Data, *Biometrics*, 33, 159-174.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages and limitations, *Journal of Management*, 20(4), 903-931.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(1-2): 1-135.

Schwartz, S.H. (1992). Universals in the Content and Structure of Values, in M. Zanna, ed., *Advances in Experimental Social Psychology*, Acad. Press, 25, 1-66.

Scott, N., & Smith, A.E. (2005). Use of Automated Content Analysis Techniques for Event Image Assessment, *Tourism Recreation Research*, 30(2), 87-91.

Tsoumakas, G. & Katakis, I. (2007). Multi Label Classification: An Overview, *International Journal of Data Warehousing and Mining*, 3(3), 1-13.

U.S. Senate. (2006). *Senate Committee on Commerce, Science and Transportation Hearing on Network Neutrality*. Feb 7.

Wilson, J., Wilson, T., & Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39(2-3): 165-210.