Learning curves for automating content analysis: How much human annotation is needed?

Emi Ishita
Research and Development Division,
Kyushu University Library,
Kyushu University
Higashi-ku, Fukuoka, 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

Yoichi Tomiura
Faculty of Information Science and
Electrical Engineering,
Kyushu University
Fukuoka, Japan
tom@inf.kyushu-u.ac.jp

Yasuhiro Takayama
National Institute of Technology,
Tokuyama College,
3538 Gakuendai, Shunan,
Ymagaguchi 745-8585 Japan
takayama@tokuyama.ac.jp

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

***Abstract*—In this paper, we explore the potential for reducing human effort when coding text segments for use in content analysis. The key idea is to do some coding by hand, to use the results of that initial effort as training data, and then to code the remainder of the content automatically. The test collection includes 102 written prepared statements about Net neutrality from public hearings held by the U.S Congress and the U.S. Federal Communications Commission (FCC). Six categories used in this analysis: *wealth, social order, justice, freedom, innovation* and *honor*. A support vector machine (SVM) classifier and a Naïve Bayes (NB) classifier were trained on manually annotated sentences from between one and 51 documents and tested on a held out of set of 51 documents. The results show that the inflection point for a standard measure of classifier accuracy (F_1) occurs early, reaching at least 85% of the best achievable result by the SVM classifier with only 30 training documents, and at least 88% of the best achievable result by NB classifier with only 30 training documents. With the exception of *honor,* the results show that the scale of machine classification would reasonably be scaled up to larger collections of similar documents without additional human annotation effort.**

Keywords—automatic content analysis; machine learning; human values; learning curve; computational social science

I. INTRODUCTION

Content analysis is a widely used method among social scientists. The typical social science research process consists of the following steps; (1) theorizing, including identifying research questions and collecting a corpus, (2) creating a typology of the phenomena to be studied and coding guidelines for training additional coders, (3) a pilot study to refine both the typology and the coding guidelines, (4) coding the entire corpus, and (5) quantitative analysis using appropriate statistical techniques. Human effort is required for all steps, although it may in many cases be augmented by software, such as the use of qualitative data analysis software for steps (3) and (4) and statistical software packages for step (5). The process is often iterative in the early stages, with the coding frame evolving as new phenomena are encountered. The process typically ultimately converges, so after some point the human effort is principally devoted to examining content and assigning codes from an existing coding frame. It is this later phase in step (4), the assignment of existing codes to existing content, following patterns that have already been established and for which numerous examples exist from early coding, that may in some cases be amenable to automation. The potential benefits from automating this second stage are substantial because coding larger collections provides greater potential for finer-grained analysis, and automatic coding is affordably scalable to very large document collections. In our previous work [1], we have shown that three classifiers (k Nearest Neighbor, Naïve Bayes, and Support Vector Machine) trained using about 8,000 manually annotated sentences (from 101 documents) can be used to automatically code held out sentences (from other documents) with reasonable accuracy for five human values. The SVM classifier yielded the best results among the three, with an F_1 score of 0.7068.

In this paper, we explore the potential for reducing human effort when coding text segments for use in content analysis. The key idea is to do some coding by hand, to use the results of that initial effort as training data, and then to code the remainder of the content automatically. In the previous work, we have explored how well a classifier would work when trained on nearly the entire test collection, but we don't yet know if that much training data is actually needed. In this paper, the question that we ask is that how much training data would have actually been needed to obtain results similar to those produced by a classifier with the maximum available amount of training data. Alternatively, we might ask how many documents must be manually coded to achieve at least 90% of the best F_1 value that we could achieve (with 101 documents). Our results show that our most accurate classifier, the SVM, can often achieve that threshold with as few as 30 training documents.

II. RELATED WORK

In our research, we focus on human values [2], particularly in terms of detecting human values invoked in opinionated text [3], [4]. Specifically, our research involves studying the relationship between human values and attitudes toward the controversies such as the Fukushima nuclear accidents or the Park51 project in downtown Manhattan [3]. We have shown that parts of this process can be effectively automated using machine learning [4].

Content analysis is one of the approaches used to study human values [5]. We examined the role of human values in shaping the Net neutrality debate through a content analysis of testimonies from U. S. Senate, House, and FCC hearings on Net neutrality [6], [7]. One of the authors coded sentences in 102 prepared statements for six human values [7]. A manual coding process is expensive and time consuming, so applying such a process at Web scale would be infeasible. The goal of this paper is to explore the potential of automated methods to address this challenge.

There is a modest but growing literature on automatic content analysis. Yan, McCracken, and Crowston [8] built a software tool to assist social scientists performing content analysis. Their semi-automatic system leveraged natural language processing and machine learning techniques (specifically a SVM classifier) for initial automatic coding. They used a gold standard corpus that includes 408 email messages, in which sentences may by assigned more than one code. There are a total of 39 codes in their coding scheme. The average recall they achieved over all 39 codes is 0.702. In contrast, the average overall precision is 0.078.

Evans et al. [9] reported results of an experiment designed to test the strengths and weakness of alternative approaches for classifying the positions and interpreting the content of advocacy briefs submitted to the U.S. Supreme Court. They found that the Wordscores introduced by Laver et al. [10] and various models in a Naïve Bayes classifier performed well. They evaluated these classifiers based on precision, recall and accuracy as measures.

Wallace et al. [11] proposed automatically annotating transcripts of patient-provider interactions with topic codes via machine learning. They used a Conditional Random Field (CRF) to model utterance topic probabilities. They evaluate their approach using kappa, accuracy, precision, recall, and F-measure. As these studies illustrate, for studies involving automatic content analysis it is common to evaluate classifiers using precision, recall and the F-measure (the harmonic mean of precision and recall). In this paper, we focus principally on the F_1 measure.

III. APPROACH

In this section we describe the test collection and introduce our experiment design and evaluation measures.

A. *Test Collection*

The collection consists of 102 written prepared statements about Net neutrality from public hearings held by the U.S. Congress and the U.S. Federal Communications Commission (FCC) [7]. The categories used in the content analysis were selected from the Meta-Inventory of Human Values (MIHV) [12]. After four rounds of developmental annotation, we selected six MIHV categories; *wealth, social order, justice, freedom, innovation,* and *honor*, all of which were fairly common in our collection (with a prevalence above 4%). One of the authors of this paper, a social scientist, then annotated each of the sentences in the 102 documents with zero or more MIHV categories. In 102 documents, total 9,890 sentences were manually annotated. Table I shows examples of annotations for some sentences.

We use a classifier to assign labels representing human values to each sentence. For our experiments, we identified each sentence using TreeTagger [13]. Sentences including more than 40 words or no non-stopwords were removed. After removing words in the SMART stopword list [14] from those sentences, those sentences were stemmed by the Porter stemmer [15]. No other feature selection techniques were applied, both because full-vocabulary SVM's have been shown to be effective for text classification and because the relatively short sentence lengths already raise some concerns about sparsity. After pre-processing, 8,660 sentences remain in the test collection. Table II shows the how many of the 8,660 sentences were manually annotated with each of the six values. There are 1,545 sentences that were not annotated for any value, and all but two of the 102 documents have at least one sentence with no value.

A total of 20 documents were annotated by a second annotator, yielding the kappa values shown in Table II [16].

TABLE I. EXAMPLE SENTENCES.

Categories	**Sentence**
freedom, social 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.
innovation, freedom	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.
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.
social 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.

TABLE II. ANNOTATOR AGREEMENT (KAPPA) AND PREVALENCE FOR EACH MIHV CATEGORY.

Categories	***Kappa***	***#sentences***	***#docs***
wealth	0.621	3,156	102
social order	0.688	2,503	102
justice	0.423	2,267	99
freedom	0.628	2,155	101
innovation	0.714	1,018	94
honor	0.437	317	80

B. *Experimental Design*

In our earlier work, we have sought to characterize the best possible classification effectiveness for a variety of widely used classifier designs [1]. To do this, we built classifiers for each value, each trained with binary category annotations for all sentences in some 101-document subset of the 102 documents. We then tested those classifiers on the sentences from the one remaining document. This process was repeated 102 times, once with each of the 102 documents as the held out test document. This approach, 102-fold cross-validation, produces one classification result for each document. We tried Support Vector Machine (SVM), Naïve Bayes (NB) and k-Nearest Neighbor classifiers from University of Waikato's Weka toolkit [17], obtaining the best results (F_1=0.7068, averaged over all six categories) from a linear kernel SVM classifier and the second best results (F_1=0.6333) from the NB classifier. We therefore report SVM and NB cross-validation results as upper baselines for the learning curve experiments in this paper.

The purpose of this paper is to determine whether that much training data is actually needed. Specifically, we examine how many documents are required to obtain classifier performance similar to that which could be obtained using 102-fold cross-validation. The classifiers are trained using different numbers of training documents (one to 51 documents), and the trained classifiers are then used to assign value labels to each sentence in the remaining 51 documents. For each classifier (i.e., for each number of training documents) we compute precision, recall and F_1 to make a learning curve for F_1.

We ran a linear kernel SVM and NB in Weka for some number *i* of training documents, as follows:

a) The collection was divided into two disjoint groups of 51 documents (Group A and B).

b) i documents from Group A were randomly selected and used to train a classifier.

c) That trained classifier then assigned values to each sentence in the 51 documents of Group B.

d) Groups A and B were swapped (Group A becoming testing) and steps b) and c) were repeated.

e) The 102 resulting classification decisions were used to compute F_1.

In step a), interval sampling was applied to divide the two groups. The documents were arranged as order of annotatitons. Group A always includes documents numbered 1, 3, 5, ... 101 and Goup B includes documents numbered 2, 4, 6, ..., 102. Steps b) through e) were repeated 10 times, and for tabular display the ten F_1 values are averaged. Steps b) through e) were repeated for i = {1, 2, 3, …., 51} training documents.

C. *Evaluation Meaure*

We compute precision as number of correctly assigned categories divided by the number of assigned categories, recall as the number of correctly assigned categories divided by the number of human-annotated categories, and F_1 as the balanced harmonic mean of precision and recall. Table III shows the contingency table matrix for binary classification.

TABLE III. CONTIGENCY TABLE FOR EVALUATION MESURE.

		Annotator	
		Positive	**Negative**
Classifier	**Positive**	a	b
	Negative	b	d

These evaluation measures are computed as follows:

$$Precision = \frac{a}{a+b}$$

$$Recall = \frac{a}{a+c}$$

$$F_1 = \frac{1}{(\frac{1}{Precision} + \frac{1}{Recall})/2}$$

IV. RESULTS

Figure 1 shows F_1 scores for the SVM classifier with different number of training documents for each of the six values. For example, the F_1 scores for *wealth* produced by an SVM with one randomly selected training document vary between 0.011 and 0.463, depending on which training document was randomly selected. As expected, F_1 scores with 10 randomly selected training documents are more stable, carrying between 0.529 and 0.606.

Figure 2 and 3 show the resulting learning curve for each of the six categories from the SVM and NB classifiers, respectively. Tables III and IV show the same information for specific numbers of training documents (including the baseline result for 101 training documents), by SVM and NB classifiers. As we have seen in our prior work, *honor* is a difficult category for automated classifiers [1], perhaps because of the relative sparsity of positive training examples. We therefore focus principally on the other five categories for the remainder of this paper.

As we have previously reported, the SVM classifier archives better F_1 values than the NB classifier for each of the six categories with 102-fold cross-validation. Comparing corresponding values from Tables III and IV, we can now see

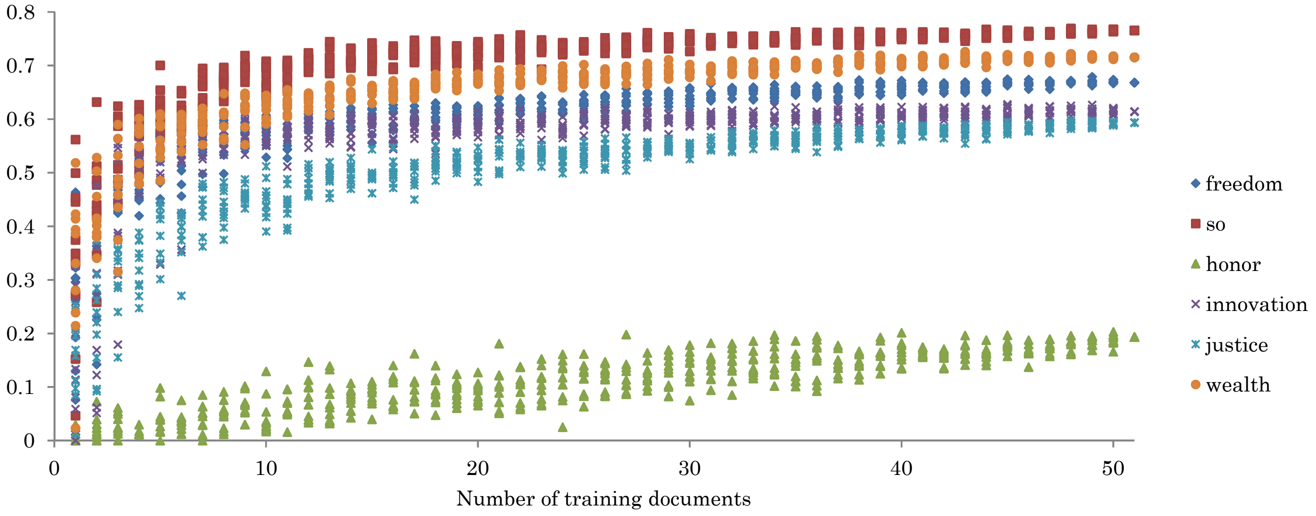

Fig. 1. F_1 for SVM classifier with different numbers of training documents (best viewed in color)

that the same pattern of dominance by the SVM is evident for each smaller number of training documents at which we computed averages. Interestingly, the same is not true of *honor*. As Figures 2 and 3 illustrate, the relative order of the F_1 scores between the six categories become stable early, providing clear evidence for the relative difficulty of the classification tasks. The *social order* category consistently yields the best results, followed by *wealth*, then *freedom* and *innovation*, then *justice*, and finally *honor*. This ordering does not follow the relative frequency of value categories in the collection. It is interesting to note that the F_1 for the NB classifier rises markedly faster than the F_1 for the SVM classifier, however, for both *justice* and *honor*. Although not similar in relative frequency, these two categories do have markedly lower kappa values. We interpret this as perhaps indicating that the NB classifier is able to more easily accommodate inconsistent training annotations when only a limited amount of training data is available. This observation could be important when applying these techniques in severely cost-constrained or time-constrained settings.

From Table IV we can see that SVM classifier achieve a mean F_1 (over 10 runs) that is above 90% of the high (102-fold cross-validation) baseline with 50 training documents for all five of the categories other than *honor*; it achieves that 90% threshold with 30 documents for four of those five categories; and with 20 documents for three of those four categories. From Table V, the NB classifier achieves a mean F_1 that is above the corresponding high (102-fold cross-validation NB) baseline with 20 documents for all five of the categories other than *honor*. Nonetheless, the NB classifier's high baseline is markedly lower than the high baseline of the SVM classifier, and the SVM classifier is thus the better choice overall.

TABLE IV. MEAN F_1 FOR DIFFERENT NUMBERS OF TRAINING DOCUMENTS, SVM (% IS OF 101-DOC RESULT).

	Number of training documents for SVM				
	10 docs	***20 docs***	***30 docs***	***50 docs***	***101 docs***
wealth	0.620 (84%)	0.673 (91%)	0.689 (93%)	0.716 (96%)	0.742
social order	0.680 (87%)	0.727 (93%)	0.746 (95%)	0.766 (98%)	0.784
justice	0.449 (70%)	0.518 (80%)	0.546 (85%)	0.592 (92%)	0.645
freedom	0.576 (82%)	0.611 (87%)	0.640 (91%)	0.670 (96%)	0.700
innovation	0.572 (89%)	0.589 (91%)	0.600 (93%)	0.615 (95%)	0.645
honor	0.045 (17%)	0.093 (35%)	0.135 (50%)	0.189 (71%)	0.267

TABLE V. MEAN F_1 FOR DIFFERENT NUMBERS OF TRAINING DOCUMENTS, NB (% IS OF 101-DOC RESULT).

	Number of training documents for NB				
	10 docs	***20 docs***	***30 docs***	***50 docs***	***101 docs***
wealth	0.630 (93%)	0.653 (97%)	0.659 (98%)	0.666 (98%)	0.674
social order	0.716 (93%)	0.743 (96%)	0.750 (97%)	0.762 (97%)	0.770
justice	0.485 (90%)	0.512 (95%)	0.519 (96%)	0.523 (96%)	0.541
freedom	0.565 (94%)	0.587 (97%)	0.595 (99%)	0.600 (99%)	0.603
innovation	0.517 (89%)	0.531 (91%)	0.545 (94%)	0.555 (94%)	0.581
honor	0.118 (53%)	0.171 (77%)	0.195 (88%)	0.206 (88%)	0.222

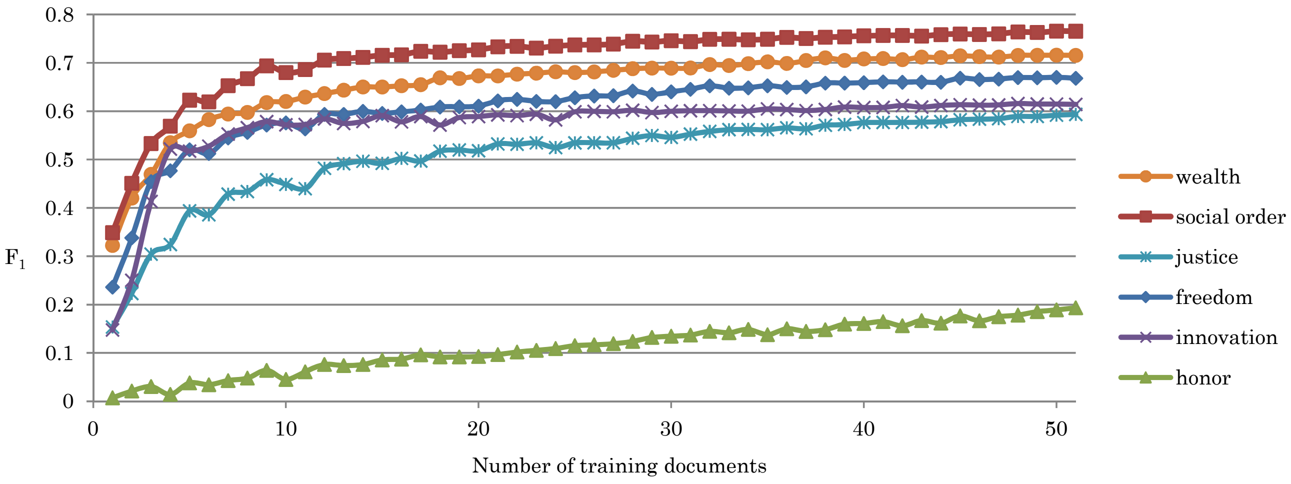

Fig.2. Learning curves for SVM classifier

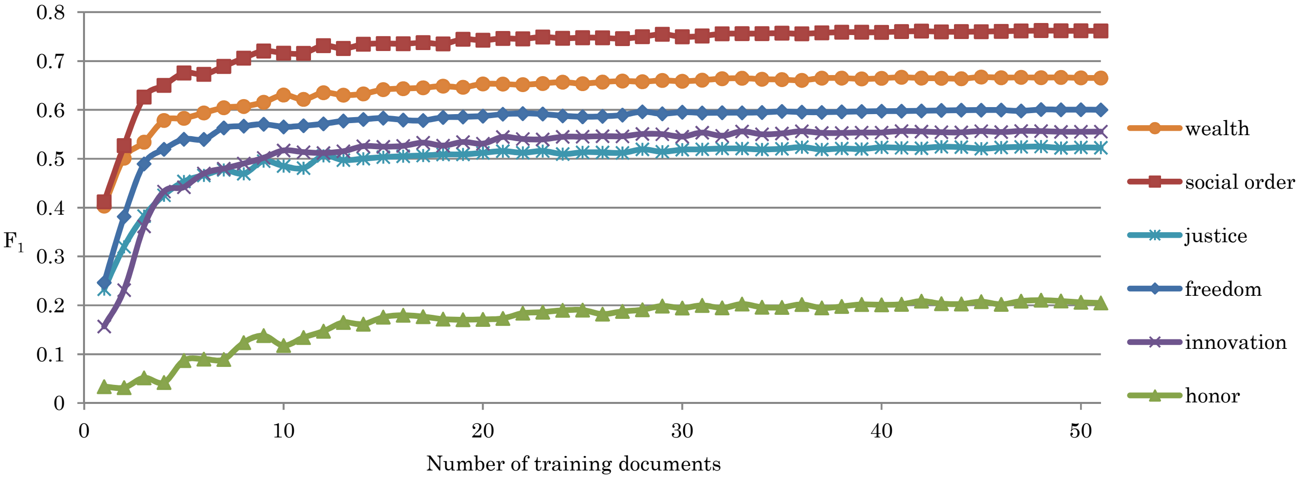

Fig. 3. Learning curves for NB classifier

V. CONCLUSION AND NEXT STEPS

We have shown that with our human values test collection, reducing the workload of human annotators from 101 documents to 30 or so would have yielded only modest reductions in classification accuracy, at least as measured by F_1. Because reducing human effort is our ultimate goal, we see this as a promising result.

Even more importantly, with the exception of *honor*, our learning curves are fairly flat between 50 and 101 training documents; extrapolating from that suggests that we could now reasonably scale up our research to larger collections of similar documents without additional human annotation effort.

We are now working to further improve over these baseline classifiers. In our initial work on this challenge, we have found that some improvement is possible from the use of word n-grams in addition to the single words that we have used in these experiments, but obtaining good results with that approach requires careful attention to balancing sparsity and informativeness [16].

As we noted in Section II, the precision, recall, and F_1 measures that we have used have been widely reported (e.g., [8], [9], [11]). These measures focus on aggregating counts of errors on individual instances, but other approaches to evaluation are also possible. In particular, task-specific measures such as those proposed by Hopkins and King offer potential for generating complementary insights [18]. As Hopkins and King explain "Although computer scientists have methods for automated content analysis, most are optimized to classify individual documents, whereas social scientists instead want generalizations about the population of documents, such as the proportion in a given category". They then demonstrated a method that gave approximately unbiased estimates of category proportions, even when the optimal classifier performs poorly by instance-oriented measures.

This suggests a future direction for our work. Building on this insight, we can clearly use the F-measure to determine which classifiers are clearly less good than others. Once we have some classifiers that are fairly good at classifying instances, we can then begin to look at which of those yield the most reliable estimates of category proportions [6], [7]. From there, the next logical step will be to actually use our classifier results in social science experiments, comparing the conclusions that can be drawn to those that are already available based on human coding. In this way, we can affordably balance the strengths of several approaches to evaluation.

ACKNOWLEDGMENT

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.

REFERECES

[1] Y. Takayama, Y. Tomiura, E. Ishita, Z. Wang, D. W. Oard, K. R. Fleischmann, A.-S. Cheng, "Improving Automatic Sentence-Level Annotation of Human Values Using Augmented Feature Vectors," Proceedings of Conference of the Pacific Association for Computational Linguistics (PACLING 2013), in CD-ROM, 2013

[2] K. R. Fleischmann. “Information and Human Values,” Synthesis Lectures on Informaiton Concepts, Retrieval, and Services. Morgan & Claypool Publishers, pp.100, 2013.

[3] T. C. Templeton, and K. R. Fleischmann, “The Relationship between Human Values and Attitudes toward the Park51 and Nuclear Power Contraversies,” Proceeding of the American Society for Information Science and Technology (ASIST2011), vol. 48, no.1 pp.1-10, 2011, DOI: 10.1002/meet.2011.14504801172

[4] T. C. Templeton, K. R. Fleischmann, and J. Boyd-Graber, “Simulating audiences: Automating Analysis of Values, Attitudes, and Sentiment,” Proceedings of the Third IEEE International Conference on Social Computing, 2011.

[5] K. R. Fleischmann, D. W. Oard, A.-S. Cheng, J. Boyd-Graber, T. C. Templeton, E. Ishita, J. A. Koepflear, and W. A. Wallace. “Content Analysis for Values Elicitation,” Proceedings of the ACM SIGHI 2012 Conference on Human Factros in Computing Systems, Workshop on Methods for Accounting for Values in Human-Centered Computing, 2012.

[6] 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 Techonology, vol. 63, no. 7, pp.1360-1373, 2012

[7] A.-S. Cheng. "Values in the Net Neutrality debate: applying content analysis to testimonies from public hearings," University of Maryland Theses and Dissertations / Information Studies Theses and Dissertations, 2012, http://hdl.handle.net/1903/12701

[8] J. L. S. Yan, N. McCracken, and K. Crowston, "Semi-Automatic Content Analysis of Qualitative Data," iConference 2014 Proceedings, pp.1128-1132, 2014, DOI:10.9776/143999

[9] M. Evans, W.V. McIntosh, J. Lin, and C. L. Cates, "Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research," 1st Annual Conference on Empirical Legal Studies Paper, 2006, http://dx.doi.org/10.2139/ssrn.914126,

[10] M. Laver, K. Benoit and J. Garry, "Extracting Policy Positions from Political Texts Using Words as Data," The American Political Science Review, vol. 97, no. 2, pp. 311-331, 2003

[11] B. C. Wallace, M. B. Laws, K. Small, I. B. Wilson, and T. A. Trikalinos, "Automatically Annotating Topics in Transcripts of Patient-provider Interactions via Machine Learning.," Medical Descison Making, vol. 34, no.4, pp.503-512, 2014, doi: 10.1177/0272989X13514777

[12] A.-S. Cheng and K. R. Fleischmann, “Developing a Meta-Inventory of Human Values,” Proceedings of the American Society for Information Science and Technology (ASIST2010), vol. 47, no. 1, pp.1-10, 2010.

[13] H. Schmid, "Probabilic part-of-speech tagging using decision trees," Proceeding of International Conference on New Methods in Language Processing, 1994.

[14] G. M. Salton, and M. J. McGill, "The SMART and SIRE Experimental Retrieval Systems," Readings in Information Retrieval (1997), Moragan Kaufmann, pp.381-399, 1980.

[15] M. F. Poter, "An algorithm for suffix stripping," Readings in Information Retrieval (1997), Morgan Kaufmann, pp.313-316, 1980.

[16] Y. Takayama, Y. Tomiura, E. Ishita, D. W. Oard, K. R. Fleischmann, and A.-S. Cheng, "A Word-Scale Probabilistic Latent Variable Model for Detecting Human Values," Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (ACM CIKM 2014), pp.1489-1498, 2014.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, " The WEKA Data Mining Software: An Update," SIGKDD, vol. 12, no. 1, pp. 10-18, 2009.

[18] D. G. Hopkins, and G. King, " A method of automated nonparametric content analysis for social science," American Journal of Political Science, vol. 54, no. 1, pp.229-247, 2010, DOI: 10.1111/j.1540-5907.2009.00428.x