Cost-Effective Learning for Classifying Human Values

Emi Ishita,[1] Satoshi Fukuda,[1] Toru Oga,[1] Yoichi Tomiura,[1]
Douglas W. Oard,[2] and Kenneth R. Fleischmann,[3]
[1] Kyushu University, Fukuoka 819-0395, Japan
[2] University of Maryland, College Park, MD 20742, USA
[3] University of Texas at Austin, Austin, TX 78705, USA
ishita.emi.982@m.kyushu-u.ac.jp

Abstract. Prior work has found that classifier accuracy can be improved early in the process by having each annotator label different documents, but that later in the process it becomes better to rely on a more expensive multiple-annotation process in which annotators subsequently meet to adjudicate their differences. This paper reports on a study with a large number of classification tasks, finding that the relative advantage of adjudicated annotations varies not just with training data quantity, but also with annotator agreement, class imbalance, and perceived task difficulty.

Keywords: Text Classification, Content Analysis, Human Values, Annotation Cost.

1 Introduction

Modern approaches to automated text classification (i.e., assigning documents to predefined categories) typically rely on supervised machine learning. Many machine learning classifiers have been developed, including Support Vector Machine (SVM), Naïve Bayes, and Decision Tree. A recent innovation has been the development of classifiers either employing deep learning directly, or employing features learned in that way (e.g., using fastText). The training data from which classifiers are learned is typically created using human annotation. Building training data with sufficient scale and data quality can be time-consuming, and thus expensive. Moreover, scale and data quality are often in tension, since single annotation can achieve greater scale, while multiple annotation can achieve higher data quality. Prior work has shown that there are cases in which single-annotation at scale can produce a better classifier than multiple annotation [1, 2]. Our focus in this paper is to explore this question in the context of building classifiers for human values.

We proposed a three-stage process for labeling sentences in newspaper editorials that address a specific topic with the human values that those sentences express or reflect [1]. That process included: (1) identifying documents that address the topic being studied, (2) identifying "value sentences" that express or reflect one or more human values, and (3) assigning human value categories to those value sentences. Experimental results for the first task, on/off topic document identification, showed that classifier accuracy can be improved early in the training process by having each

annotator label different documents, but that later in the process it becomes better to rely on a more expensive multiple-annotation process, and in particular one in which annotators subsequently meet to adjudicate their differences.

In this paper, we use the same collection as in [1], focusing now on the third task, assigning human values categories to each value sentence as the text classification task. Because this task is done at sentence scale, we can construct learning curves over larger sets of items. Moreover, we can do this for several classifiers, one for each of six human values. Using exploratory data analysis, we find that the best approach – single annotation, multiple annotation, or a sequential combination of the two – depends on a number of factors.

In this paper we introduce our extended test collection in Section 2, we present our experiments in Section 3, and we conclude in Section 4 with some remarks on next steps.

2 Extending the Test Collection

We chose to study human values in Japanese newspaper editorials that address the nuclear power debate in Japan [1]. The Great East Japan Earthquake on March 11, 2011 damaged to the Fukushima Daiichi nuclear power plant, resulting in one of the most consequential nuclear emergencies of our time [1]. After the disaster, various discussions have occurred regarding, for example, incident response in nuclear power plants, government and corporate reactions, how residents coped with the disaster, reactivation or decommissioning of nuclear power plants, and nuclear power plant inspections. The collection includes 750 editorials from the Mainichi Shimbun CD-ROM [3] from 2011-2016, each of which include 原発(an abbreviation for nuclear power plant) or 原子力(nuclear power). For on/off topic identification, 448 of the 750 editorials were randomly selected, and 239 on-topic editorials were ultimately manually identified (based on adjudicated annotations from two annotators). We randomly selected 120 editorials from this corpus.

2.1 Selecting the Human Values to Study

Human values can be defined as "guiding principles of what people consider important in life" [4]. Human values are an object of study in a wide range of fields, from social psychology [5] to human-computer interaction [6], and play an important role in the information field [7], including in prior studies of the nuclear power debate [8].

We started by defining a set of eight human values based on four broad factors we expect people would value in a crisis situation. The first question involves *responsibility*: whether people focus on results or on emotions, feelings and integrity. The second involves *order*: whether people focus on social order or individual choices. The third involves *interest*: whether people focus on safety or on wealth. The fourth involves *welfare*: whether people focus on the benefit to society or self-enhancement. Table 1 defines eight human values that anchor those four contrasts.

Table 1. Definitions of the human value categories.

Human Value	Definition
Consequence	Values on judgement or evaluation based on results including future prospects (e.g. outcomes, objectives, targets) or macro/long-term perspectives.
Intention	Values on emotion or feelings including impression, attitude, empathy, prudence, and sincerity; The quality of being honest and integrity; adherence to moral principles.
Social Order	Values on social structure, including rules, norms, common sense and expectations as well as social responsibility; Institutional, legal, and political decisions involving governments and states.
Freedom	Value of individual freedom and choices; the state of being unconstrained; freedom from interference or influence by others;
Safety	Values of safety and security; the state of being free from danger, injury, threat or fear; measures to prevent accidents and hazards.
Wealth	Values on pursuing any economic goals, such as money, material possessions, resources, and profit including business activities.
Human Welfare	Values on fulfilling benefits common to human beings and related to society as a whole; Clear benefits to the public.
Personal Welfare	Values on personal needs, growth, and self-actualization.

2.2 Coding Process

After a training session using a held out set of 43 editorials, two annotators (the first and third authors of this paper, both of whom are native speakers of Japanese) independently annotated each sentence in 20 editorials as value sentences or fact sentences, and then assigned human value categories to the value sentences that were identified. Each value sentence could be labeled with one or more human values. After each set of 20 editorials was annotated, the two coders discussed their differences and created adjudicated annotations by consensus, subsequently updating the written annotation guidelines before starting on the next set of 20 editorials. They repeated this process six times. Table 2 shows English translations of some example sentences with the manually assigned human value categories (the sentences that were actually annotated were in Japanese). Table 3 shows Cohen's Kappa scores as measures of inter-annotator agreement for each human value category in each round. These Kappa scores generally increase in later rounds, although Consequence is a notable exception.

Table 2. Example sentences with associated human values (English translations).

Human Value	Example Sentence
Consequence, Intention, Safety, Human Welfare	Fight against radiation problem without a prospective solution, worries and anxieties of return to home town, and despair to hometown loss.
Social Order, Safety, Wealth	The resolution paper points out that in order to run the nuclear power plant, safety measures should be given priority over cost.
Intention, Social Order, Safety	First of all, the government should to explain fully to remove the resident's anxiety of radiation.

Table 3. Cohen's Kappa for each human value in each round.

	Round	1	2	3	4	5	6	Means
	#docs	20	20	20	20	20	20	
	#sentences	584	532	541	550	565	540	
Human Value	Consequence	0.14	0.16	0.04	0.13	0.33	0.18	0.16
	Intention	0.35	0.39	0.25	0.54	0.43	0.56	0.42
	Social Order	0.32	0.42	0.49	0.41	0.54	0.43	0.43
	Freedom	-	0.00	0.00	1.00	0.00	0.00	0.20
	Safety	0.55	0.52	0.42	0.72	0.60	0.60	0.57
	Wealth	0.64	0.60	0.61	0.46	0.63	0.71	0.61
	Human Welfare	0.12	0.52	0.32	0.43	0.50	0.50	0.40
	Personal Welfare	0.04	0.15	0.37	0.00	0.00	0.00	0.09
	N/A	0.07	0.08	-0.01	0.14	0.18	-0.01	0.07

3 Constructing Learning Curves

The annotated sentences were used to train and evaluate SVM classifiers for automated annotation of human values.

3.1 Experimental Design and Setting

Japanese words are not separated by spaces, so JUMAN version 7.01 [9] was used to tokenize each sentence. All of the resulting words were used as features for the classifier, after removing period and comma characters. Sentence contained an average of 22 words in Rounds 2 to 5. We implemented linear kernel SVM classifiers using TinySVM [10].

Figure 1 illustrates the experiment design that we used to create learning curves. Annotated sentences in Round 2 to 5 were used as training data, with annotated sentences from Round 6 used as evaluation data; sentences from Round 1 were not used in order to minimize annotator learning effects. Documents in each round set were

randomly ordered for training, but sentences within a document were used in the order in which they occurred.

We plot leaning curves by placing the number of annotations on the horizontal axis and the F_1 for that number of annotations on the vertical axis. Because adjudication requires two independent annotations, we count each adjudicated annotation as two annotations when plotting learning curves (Of course, the actual time to obtain adjudicated annotations include discussion time, but here we as account only for the two annotations). We plot two kinds of learning curves: Adjudicated and Hybrid. For Adjudicated, we use the adjudicated annotations for training. For Hybrid, we alternate between Annotator A's or Annotator B's annotations for training. In every case, we use Adjudicated annotations for evaluation.

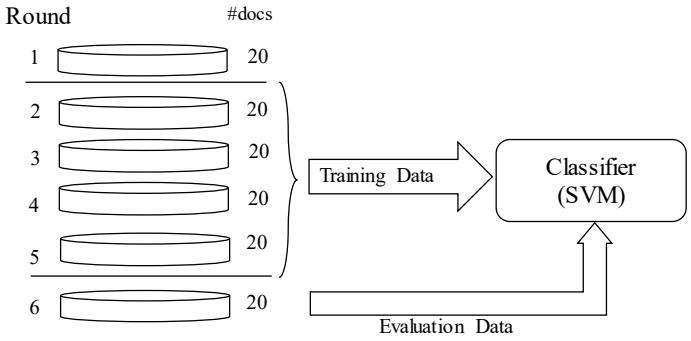

Fig. 1. Experiment design for creating learning curves.

Table 4 shows the distribution of annotations in the training and evaluation data. For example, in the 2,188 sentences in the adjudicated training data that were labeled for Consequence, 959 positive examples (44% of the total) have that label. Personal Welfare and Freedom have fewer than 50 positive examples in the adjudicated training data, which is too sparse for the construction of informative learning curves. We therefore focus on six categories (See Table 4) for our experiments.

Table 4. Distribution of positive examples for each human value category.

Consequence	Adjudicated	Annotator A	Annotator B
Train (2188)	959 (44%)	758 (35%)	667 (31%)
Eval (540)	270 (50%)	239 (44%)	159 (30%)
Social Order			
Train (2188)	1570 (72%)	1473 (67%)	1445 (66%)
Eval (540)	431 (80%)	426 (79%)	367 (68%)
Wealth			
Train (2188)	289 (13%)	251 (12%)	263 (12%)
Eval (540)	118 (22%)	105 (19%)	107 (20%)
Intention			
Train (2188)	205 (9%)	152 (7%)	167 (8%)
Eval (540)	66 (12%)	43 (8%)	60 (11%)
Safety			
Train (2188)	719 (33%)	663 (30%)	586 (27%)
Eval (540)	249 (46%)	263 (49%)	199 (37%)
Human Welfare			
Train (2188)	224 (10%)	196 (9%)	196 (9%)
Eval (540)	73 (14%)	52 (10%)	61 (11%)

3.2 Results

Figure 2 shows six pairs of learning curves, each of which shows how the mean F_1 (over 100 random shuffles) varies with the number of annotations. Three broad patterns are evident. For Consequence and Social Order, adjudicated training is consistently the better choice. The opposite is true for Safety and Human Welfare, with hybrid training consistently being the better choice. Perhaps the most interesting cases are Safety and Human Welfare for which a crossover is evident, with hybrid training being better initially, but eventually adjudicated training becomes the better choice. This third pattern was the one that Ishita et al. had seen for the on/off topic identification task [1].

Based on these results, it is clear that the relative advantage of adjudicated annotations varies with more than just training data quantity. Table 5 shows some other factors that might affect classifier performance. Here value categories are sorted in decreasing order of Net Adjudicated Advantage (the mean difference in F_1 between adjudicated and hybrid training data). Positive examples is the fraction of positive examples in adjudicated training data. Annotator Agreement is characterized two ways: (1) as averaged Kappa over Rounds 2 to 5 (higher is better), and (2) as the increase in positive examples after adjudication (e.g., Annotator A annotated 35% of sentences as positive for Consequence, Annotator B annotated 31% as positive, and after adjudication 44% of sentences were positive, an average absolute increase of 11%). Task Difficulty is the self-reported difficulty by the two annotators. For exam-

ple, Annotator B stated that "Consequence and Social Order were comparatively hard because these categories cover broader concepts and an annotator has to interpret the context and meaning of whole sentence. On the other hand, Wealth or Intention are relatively easy because annotators can assign those categories when they find specific words related to these categories."

As Table 5 shows, Consequence and Social Order exhibit the largest number of positive examples, the highest adjudication increase, and the greatest task difficulty; both benefit from multiple-annotation adjudicated training. Intention and Wealth, by contrast, have relatively few positive examples, a correspondingly low adjudication increase, and the lowest task difficulty; they consistently benefit from single-annotation Hybrid training. Safety and Human Welfare also have relatively few positive examples and a correspondingly small adjudication increase, and a more modest level of task difficulty; they exhibit crossover, with Hybrid initially the better choice. Self-reported task difficulty is difficult to quantify objectively, so the jury is still out on how we might predict whether a crossover will occur. But a lower prevalence of positive examples does seem correlated with some benefit to starting with hybrid annotation.

Table 5. Exploratory data analysis for correlates with learning curve type.

	Net Adjudication Advantage		Annotator Agreement (kappa)		Adjudication Increase		Positive Examples		Task Difficulty	
	Pos.	Neg.	High	Low	High	Low	Many	Few	Hard	Easy
Consequence	+0.09			0.10	11%		44%		-	
Social Order	+0.02		0.46		5%		74%		-	
Safety		-0.01	0.57		4%			33%	~	~
Intention		-0.03	0.46			2%		9%		+
Human Welfare		-0.05	0.45			1%		10%	~	~
Wealth		-0.09	0.60			2%		13%		+

4 Conclusion

There are many ways in which one might try to minimize the number of annotations needed to learn a good classifier. Examples include active learning [12], estimation of annotation quality [2], or relying on single rather than multiple annotation, the focus of this paper. Through experiments with classifiers with six human values, we have observed that this simple single-annotation approach seems well suited to categories with relatively few positive training examples. In future work we plan to experiment with a broader range of techniques for improving the cost-effectiveness of human annotation. We also plan to provide our annotation results of human values for Japanese newspaper editorials for use by other researchers.

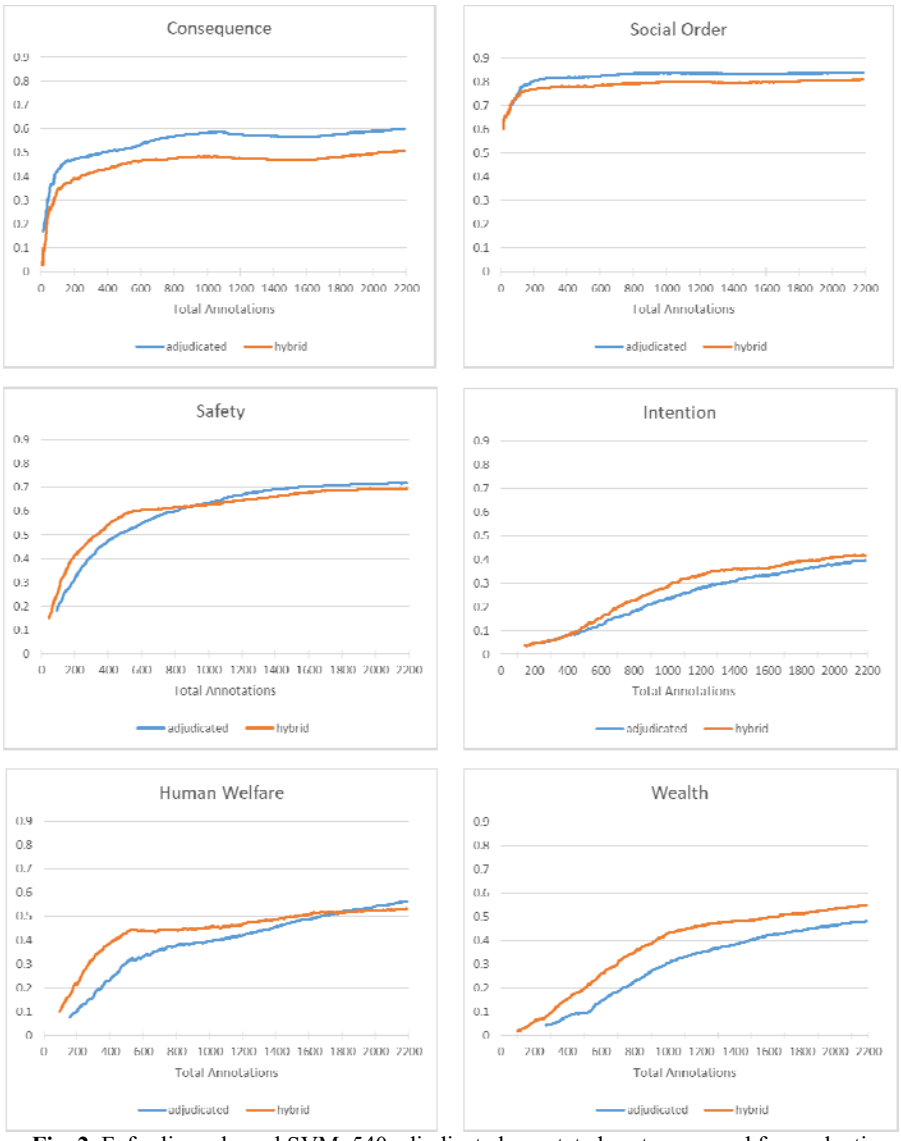

Fig. 2. F_1 for linear kernel SVM, 540 adjudicated annotated sentences used for evaluation, average of 100 random shuffles within each round.

Acknowledgements. This work has been supported in part by JSPS KAKENHI Grant Number JP18H03495.

References

1. Ishita, E., Fukuda, S., Oga, T., Oard, D.W., Fleischmann, K.R., Tomiura, Y., Cheng, A-S.: Toward three-stage automation of annotation for human values. Proceedings of 14th iConference 2019, 188-199 (2019).
2. Khetan, A., Lipton, Z.C., Anandkumar, A., Learning from noisy singly-labeled data. In: Proceedings of ICLR 2018. 15p. (2018), https://arxiv.org/abs/1712.04577, last accessed 2019/9/23.
3. CD-Mainichi Shimbun Data Collection 2011 version; 2012 version; 2013 version; 2014 version; 2015 version; and 2016 version.
4. Cheng, A.-S., Fleischmann, K.R., Wang, P., Ishita, E., Oard, D.W.: The role of innovation and wealth in the net neutrality debate: A content analysis of human values in congressional and FCC hearings. Journal of the American society for information science and technology, 63, 1360-1373 (2012).
5. Schwartz, S.H.: Value orientations: Measurement, antecedents and consequences across nations. In: R. Jowell, C. Roberts, R. Fitzgerald, and G. Eva (eds.), Measuring attitudes cross-nationally: Lessons from the European social survey, pp.169-203. London: Sage. (2007). DOI: 10.4135/9781849209458. n9.
6. Friedman, B., Kahn, P. H. Jr., and Borning, A.: Value sensitive design and information systems. In: P. Zhang and D. Galletta (eds.) Human-computer interaction and management information systems: Foundations, pp.348-372. Armonk, NY: M.E. Sharpe. (2006). DOI: 10.1002/9780470281819.ch4.
7. Fleischmann, K.R.: Information and Human Values. San Rafael, CA: Morgan & Claypool. (2014).
8. Templeton, T.C. and Fleischmann, K.R.: The relationship between human values and attitudes toward the Park51 and nuclear power controversies. In: Proceedings of the 74th annual meeting of the American society for information science and technology, New Orleans, LA. (2011). DOI: 10.1002/ meet.2011.14504801172.
9. JUMAN (a user-extensible morphological analyze for Japanese), http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN, last accessed 2019/9/23.
10. TinySVM: Support Vector Machines, http://chasen.org/~taku/software/TinySVM/, last accessed 2019/9/23.
11. Grimmer, J., Stewart, B.M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21, 267-297 (2013).
12. Goudjil, M., Koudil, M., Bedda, M., Ghoggali, N.: A novel active learning method using SVM for text classification. International Journal of Automation and Computing, 15(3), 290-298 (2018).