A Study in Rule-Specific Issue Categorization for e-Rulemaking

Claire Cardie
Faculty of Computing and Information Science
Cornell University
Ithaca, NY USA
cardie@cs.cornell.edu

Cynthia Farina
Law School
Cornell University
Ithaca, NY USA
cynthia-farina@lawschool.cornell.edu

Adil Aijaz
Department of Computer Science
Cornell University
Ithaca, NY USA
aa362@cornell.edu

Matt Rawding
Information Science Program
Cornell University
Ithaca, NY USA
mdr36@cornell.edu

Stephen Purpura
Information Science Program
Cornell University
Ithaca, NY USA
sp559@cs.cornell.edu

ABSTRACT

We address the e-rulemaking problem of categorizing public comments according to the issues that they address. In contrast to previous text categorization research in e-rulemaking [5, 6], and in an attempt to more closely duplicate the comment analysis process in federal agencies, we employ a set of rule-specific categories, each of which corresponds to a significant issue raised in the comments. We describe the creation of a corpus to support this text categorization task and report interannotator agreement results for a group of six annotators. We outline those features of the task and of the e-rulemaking context that engender both a non-traditional text categorization corpus and a correspondingly difficult machine learning problem. Finally, we investigate the application of standard and hierarchical text categorization techniques to the e-rulemaking data sets and find that automatic categorization methods show promise as a means of reducing the manual labor required to analyze large comment sets: the automatic annotation methods approach the performance of human annotators for both flat and hierarchical issue categorization.

1. BACKGROUND AND INTRODUCTION

Every year federal agencies publish in the Federal Register [on-line version: http://www.gpoaccess.gov/fr/index.html] several thousand documents on which they seek public comment. Most of these are proposed rules in areas including: environmental protection; agriculture standards; drug, workplace, and consumer safety; import and export controls; air, highway, and water-based transportation safety; communications; and various federal grant and aid programs. Federal statutes, especially the Administrative Procedure Act [5 U.S. C. §§551 et seq.], generally require such regulations to go through this "notice and comment" process before they can become final and binding on the public. In addition, agencies may request comments on a category of documents known as "guidance." These documents, which often closely resemble proposed rules in form, have different names[1] but share the characteristic that they are advice, or warning, about how the agency will exercise its power or interpret the law, rather than binding rules themselves. Sometimes, the agency is legally required to seek public comment before it finalizes guidance; other times, it simply chooses to do so. Finally, there is a third, miscellaneous category of documents in the Federal Register on which comments are solicited. These include such things as draft statements of environmental or other impacts of proposed agency action. Again, the agency may be required by one of its statutes to allow the public to comment, or it may be doing so as a matter of policy.

Until the 1990s, all comments came to the agency in hard copy — through hand delivery, conventional, or express mail. As electronic transmission and then the Internet became more generally available, agencies began to receive comments first by fax and then by e-mail. Most recently, agencies have provided web portals for comment submission. Although a few agency-specific sites remain, most have been superseded by a central portal, www.regulations.gov, which now provides access to all agencies' proposed rules and guidance, as well as to some of the documents in the third, miscellaneous category. Comments can be submitted to all agencies through this portal. (Commentors can continue to use the older submission methods as well.) Transfer of the notice-and-comment process to the web — and, more broadly, the use of information technology to support any step in the rulemaking process — is known as *electronic rulemaking (e-rulemaking)*.

At the close of the public comment period (typically, 30–60 days) the comments must be reviewed to determine what issues they contain. The comment process is not a vote; its purpose is not to tally the commentors' preferences for or against the proposal. By law, agencies must act based on factors, and to further objectives, specified by their autho-

[1]These include such widely used names as "statements of policy" and "interpretive rule," as well as more agency-specific names as "circular" and "bulletin."

rizing statutes. Along with the proposed rule, guidance, or other document on which it seeks comment, the agency is supposed to reveal its underlying data, as well as its legal and policy rationale. Ideally, the comments will address the substance of the proposal, and discuss how well the agency has met the statutory factors and objectives. Ultimately, the agency's responsibility (enforceable by the courts in many cases) is to issue a statement accompanying any final action it takes; this statement demonstrates its attention to the comments by responding to significant criticisms they contain, and explaining why it rejected alternatives they suggest [10].

Reviewing the comments to determine what relevant issues they raise can present substantial challenges for agencies. Sometimes they are working under a deadline for final decision, set by their statute or a court order. Even with no formal time limits, the process is often intense and laborious. As we observed working with rulewriters and analysts in two units within the Department of Transportation,[2] analysts read the comments and manually mark, code, summarize or partially re-type portions. These "annotations" identify the relevant issues raised by commentors, and organize the various references to each in a fashion that facilitates analysis by the entire group working on the rule (or other proposal). This process ultimately leads to preparation of the accompanying final statement. As the number, or number-plus-complexity, of comments increases, the process of finding, extracting, and organizing material raising relevant issues becomes proportionately more challenging. Indeed, agencies that have the resources to do so frequently hire outside contractors to read and summarize large comment sets.

The current paper. This paper reports results in a project to determine the degree to which automatic issue categorization can facilitate reviewing public comments: given a comment set, the automated system should determine for each sentence in each comment, which of a group of pre-defined issues it raises, if any. We build on the work of Kwon & Hovy [5] and Kwon et al.[6], which applies machine learning-based text categorization techniques (see Sebastiani [8] for an overview) to automate the comment sorting process. In particular, Kwon et al. [5, 6] first develop a set of eight general topic codes — ECONOMIC, ENVIRONMENT, GOVERNMENT RESPONSIBILITY, HEALTH, LEGAL, POLICY, POLLUTION, and TECHNOLOGY and train a machine learning algorithm (they use a support vector machines (SVMs) [11]) to classify individual sentences according to the topics they address. Using a set of 160 comments divided appropriately into training, development, and test sets, they report F-measure scores of 0.30 to 0.83 depending on the topic, with an average F-measure score of 0.67. SVMs significantly outperforms three baselines that assign to each sentence (a) all topics, (b) the most common topic, and (c) any topic with a morphological variant of its name in the sentence. In addition, the system performance approaches that of human annotator agreement (0.72 F-measure).

Rather than use a small, closed set of general topic codes, however, we investigate the possibility of categorizing sentences according to the usually much larger, and possibly hierarchical, set of *rule-specific issues* employed by rulewriters as they sort and analyze the comments. In this manner, we aim to replicate more closely what agency personnel now do manually. The longer range goal is to employ automatic issue categorization to speed up the (required) manual review of public comments by grouping similar comment snippets so that rulewriters can read and respond to them as a whole. Another application would facilitate reply comment periods by allowing agencies to rapidly provide the public with first-round comments sorted by issue, to aid and channel responsive submissions.

In the sections below, we begin by presenting the first in a series of sentence-level text categorization corpora to be developed in this project by the Cornell e-Rulemaking Initiative (CeRI)[3]. We describe the creation and annotation of the corpus, focusing on characteristics of the notice-and-comment domain that engender a nontraditional text categorization corpus and a correspondingly difficult machine learning task. Interannotator results are presented for a group of six annotators.

We next investigate the application of both standard and hierarchical text categorization techniques to the e-rulemaking data sets. We find that automatic text categorization methods show promise as a means of reducing the manual labor required to analyze public comment sets: the sentence-level issue annotation techniques approach the performance of human annotators for both flat and hierarchical issue categorization and outperform a baseline that selects the most common category for each sentence. The categorization scheme includes 17 issues, some of which can be further divided to create a set of 39 fine-grained issues. Using an overlap measure of agreement, human annotators achieve interannotator agreement scores of 64.7% and 46.4% for the 17 and 39 issues, respectively. Measured across three issue categorization data sets, the best-performing automatic categorization technique is competitive with the interannotator agreement results, reaching levels of 59-66% and 42-56% accuracy for the 17 and 39 issues, respectively.

2. RELATED WORK

In recent years, researchers have begun to investigate a range of methods from natural language processing, information retrieval, and machine learning for a number of e-rulemaking sub-tasks. Yang & Callan [12, 13], for example, extend duplicate detection methods from information retrieval to handle "e-postcard campaigns" — e-mail campaigns organized by special interest groups that supply constituents with electronic form letters for submission during the comment period. When comments were submitted on paper, modifying the form letters was difficult — the letter would need to be re-typed to add or remove text. As a result, most form letters were exact duplicates of one another. These are fairly easy to identify and need be analyzed for content only once. In electronic form, however, form letters are very easy to change, and it is exactly these changed snippets that agency rulewriters want to locate to determine if the modification introduces substantive information not present in the orig-

[2]The Federal Transit Authority and the Office of Civil Rights.

[3]URL: ceri.law.cornell.edu.

inal. The Yang & Callan [12, 13] work develops automatic methods to identify these *near-duplicate* submissions and to delineate the modified portions from the original letter.

Kwon et al. [5, 6] investigate the use of natural language processing methods to identify the main claims of a comment and then categorize them according to whether they support the proposed rule, oppose the proposed rule, or are proposing a new idea.

Most relevant for the current paper is the work of Kwon et al. [5, 6] on topic categorization of public comments. As discussed above, our work differs from theirs in that we categorize sentences in public comments according to a large set of rule-specific issues rather than a small set of general topics. We also investigate hierarchical categorization techniques in addition to standard flat text categorization methods.

Although we do not aim to make any advances in the area of text categorization in this paper, we clearly rely on previous work in this area and describe it in Section 5.

3. CORPUS CREATION

In this work, we treat issue categorization as a problem in text categorization and apply inductive learning techniques from the field of machine learning. This is the standard framework employed in the area of automated text categorization [8]. In particular, we employ *supervised* learning algorithms that require an initial "training phase" in which the learning algorithm is provided with many examples of the task to be learned. In our case then, we require a corpus of public comments that has been manually annotated at the sentence-level according to the rule-specific issue(s) that it addresses, if any. The details of the corpus creation process are described next.

Working with analysts from the Federal Transit Authority (FTA) in the Department of Transportation, we identified two interlinked sets of comments, both involving a group of guidance "circulars" the agency proposed to issue. Such circulars are a type of document on which the FTA frequently seeks public comments. Here, the proposed advice involved grants under three federal statutes that fund local transportation services for the elderly, disabled persons, and low income persons commuting to work.[4] FTA had been seeking public input at several stages of developing this guidance.

We used comments from the final two comment periods: March 15–May 22, 2006[5] and September 6–November 6, 2006[6]. Based on the judgment of the agency official primarily responsible that comments from both periods raised the same issues, we treat them as a single set. A total of 290 comments were submitted (211 + 79). Many of the comments were not submitted electronically. When scanned by the agency, several became image-based PDFs that could not be converted to machine-readable form. Also, some commentors filed comments with identical text; we retained only a single version of such duplicate comments. As a result of these adjustments, we were left with 267 comments. These comprise the CeRI FTA Grant Circulars Corpus.

Next, we constructed a list of 38 issues likely to be raised in the comments. This list was derived by consulting both the actual issue summaries prepared by the FTA analyst when she reviewed the comments, and the Federal Register notice seeking comments, which explained the proposal in detail and highlighted various aspects. The issues are organized into a shallow categorization hierarchy in which the 38 issues are leaf nodes. Seventeen form the first level; five of these expand into two or more sub-issues at level two. The issue hierarchy, expressed in the abbreviated form used within the annotation tool[7], is shown in Figure 1. NONE is a special category (shown as the 39th "issue"). It is automatically assigned to sentences deemed by the annotator to address none of the rule-specific issues. The expanded form of the issue set, with brief explanation, appears in the Appendix.

The annotation team comprised six law students in their final year of study. They were deliberately selected because of their general academic performance and, particularly, their work with the legal member of the research team (Farina) in a course on the federal regulatory process. However, none of the annotation team, nor anyone else involved in the project, had expertise in the substantive areas or regulatory programs involved in the guidance. After an initial three-week training period in which all students annotated the same comments and then discussed their selections as a group, they began annotation. Sporadic follow-up discussion occurred throughout the annotation period about the meaning and/or scope of specific issues, with clarifying information then being circulated to the entire group. The students annotated comments according to the 39 fine-grained issues.

The annotation tool allows for annotation at the word-, phrase-, sentence-, or paragraph- level. After an initial period of individual annotator discretion, it was determined that annotation would occur at the sentence level. As a result, all issue annotations are automatically projected to sentences. In addition, the fine-grained issue annotations can be converted to their corresponding top-level issue as needed for any of our analyses. Finally, any sentences the student annotator left unmarked are automatically assigned the label NONE. Annotators were free to assign more than one issue to a single span of text. Multiple annotations, however, were rare (4% of sentences in the corpus).

In all, there are 11,094 sentences in the corpus. On average, there are 41.55 sentences per comment. The shortest comment has one sentence; the largest has 1420 sentences.

3.1 Interannotator Agreement Results

146 of the 267 comments were used for the interannotator agreement study, with an average of 2.66 annotators per comment. Because there can be multiple issues per sentence and the annotators covered different numbers and subsets of the documents, we currently measure interannotator agreement using a basic agreement (AGR) measure (rather than

[4]Docket No. FTA-2006-24037: Elderly Individuals and Individuals With Disabilities, Job Access and Reverse Commute, and New Freedom Programs: Coordinated Public Planning Guidance for FY 2007 and Proposed Circulars.
[5]FTA-2006-24037-002.
[6]FTA-2007-24037-0222.
[7]Mitre's Callisto.

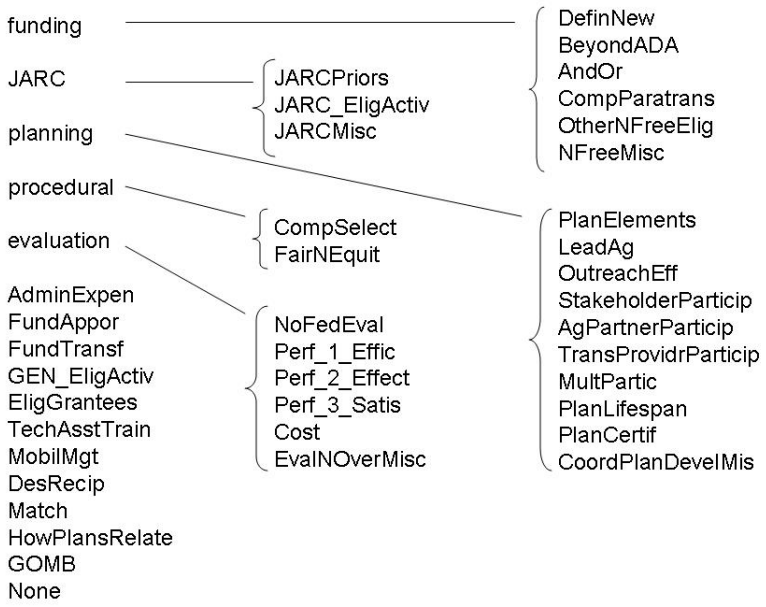

Figure 1: Rule-specific Issue Hierarchy. There are 17 top-level categories and 39 leaf categories.

Fleiss' kappa)[8]: for all pairs of annotators across all comments that were annotated by both annotators, we calculate the percentage of sentences for which the annotators assign overlapping issue labels. In most cases, this amounts to checking for an exact issue match (since 96% of the sentences are assigned a single issue). Table 1 shows the AGR score across all pairs of annotators for the full set of 39 issues, the top-level of the issue hierarchy (17 issues), the five hierarchical issues, and the five hierarchical issues plus NONE. Along with the AGR scores, we show the coverage of each issue set across all sentences of the corpus.

When calculated across the full set of 39 issues (38 issues plus NONE), interannotator agreement scores are quite low (see row 1 of the table), indicating either that more training is required for the annotators or that there is inherent difficulty in interpreting the meaning and applicability of each issue. The latter possibility is addressed in the next section. Annotation of just the 17 top-level issues (row 2) ameliorates the problem to some degree — agreement increases to 64.7% across all sentences in the corpus. Even higher levels of agreement (69.3%) can be obtained if annotation is limited to just the five hierarchical issues at the top-level (row 3) although this issue subset covers only 35.7% of the sentences in the corpus. Annotating these five issues as well as NONE's, however, allows for agreement scores approaching 70% and sentence coverage of 86.7%.[9]

[8]In current work, we have moved to the more reliable Cohen's and Fleiss' kappa for measuring interannotator agreement [4].

[9]Coverage is measured on the "aggregate gold standard" described in the next section.

4. DISTINCTIVE FEATURES OF THE PUBLIC COMMENT CATEGORIZATION TASK

Formulation of the comment categorization problem as a text categorization task raises a number of non-standard and/or difficult issues for text categorization algorithms. We enumerate these below.

Sentence-level Categorization. Although most text categorization tasks make decisions on entire documents, issues in the e-rulemaking domain are expressed, and annotated, at the sentence level or below. This is problematic because categorization of short texts is known to be quite a bit harder than categorization of longer texts [2, 7, 14].

Multiple Issues per Sentence. Typically, the lengthier comments submitted to the agency are written by lawyers or other persons well-experienced in the legal and/or substantive regulatory domain. They tend to contain long, complex sentences. These stylistically dense sentences may also be packed with meaning, and so may be annotated with multiple issues. Handling such sentences might call for (1) phrase-, rather than sentence-, level annotation (by both the human annotators and the text categorization algorithms); (2) expansion of the issue set to include new labels that cover multiple issues; or (3) changes in the text categorization algorithm. Yet any of these would likely cause a corresponding drop in performance.

Our policy for handling multi-issue sentences is laid out in Section 6.

Fairly Large, Hierarchical Issue Set. Proposed rules, guidance and other documents that generate a sufficient amount

	Agreement (%)	Coverage (%)
39 issues	46.4	100
17 top-level issues	64.7	100
5 hierarchical issues	69.3	35.7
5 hierarchical issues plus NONE	68.4	86.7

Table 1: Interannotator agreement scores when annotating w.r.t. different subsets of issues. The table also shows the percentage of sentences in the CeRI FTA Grant Circulars Corpus that each issue set covers.

of public comment to warrant the help of automatic issue categorization almost invariably raise a large number of issues. The 38-issue list we used for this corpus appears to be within the range we expect in future corpora. Hence, the e-rulemaking domain will typically present a large multi-class text categorization problem, which is generally more difficult than a binary classification problem. For one thing, because of the substantial skew in frequency with which issues are discussed (see below), insufficient numbers of training examples are likely to occur for some issues. In addition, at least some portion of the issues is likely to be hierarchically related. As discussed in the next section, the hierarchical nature of categories can both help and complicate the process of training accurate text classifiers (see, e.g. Dumais & Chen [1]).

The NONE Category. The NONE category is likely to be difficult for the machine learning algorithms in part because the associated comment sections can cover a wide variety of topics. Commentors often raise a variety of points for or against the proposal or the entire process about which they feel strongly but which the agency does not consider germane. Some of these non-germane topics will appear frequently and predictably; but many will be random and unpredictable.

As explained below, we will treat the NONE category specially in our hierarchical text categorization scheme.

Multiple Gold Standards. There are at least three types of gold standards one could generate from public comment issue categorization corpora like the FTA Grant Circulars Corpus. The first is an "aggregate" gold standard comprised of comments whose annotations have been reconciled by a pair of annotators. The second type would more closely approximate what we understand, from our agency partners, to be real-world agency practice. When more than one analyst reviews a comment set to find, extract, and organize the issue references for subsequent analysis and preparation of the accompanying final statement, these analysts typically divide the issues among themselves: each reads all the comments, taking responsibility for collecting material as to his or her allotted issues. As a result, there typically is not more than one "annotator" per issue in the real-world. The gold standard under this annotation scheme would then be the union of the issue-specific annotations of each analyst.

We have adopted yet a third strategy for creating a gold standard for the purposes of this paper. In particular, we are interested in investigating the ability of the text categorization algorithms to learn to duplicate the annotations produced by an arbitrary agency analyst. As a result, we treat the annotations of each annotator as a separate gold standard, producing six separate corpora.

Skewed Distribution Across Issues. The FTA Grant Circulars data exhibits substantial skew in terms of the distribution of sentences that address each issue, further complicating the learning task. Table 2 shows the distribution of issue annotations across sentences in the aggregate gold standard described just above. No rule-specific issue (NONE) was selected for fully 51% of the 11696 sentences in the gold standard. No other first-level or second-level category approaches this level of coverage. Discounting NONE, distribution of the remaining 16 top-level issues is still problematic, with coverage ranging from 0.2% (32 sentences) for GEN_ELIGACTIV to 10.2% (1193 sentences) for PLANNING. Our agency partners indicate that this is standard for most rulemakings.

Our only attempt in the current work to deal with the skewed category distribution is to treat NONE as a special category in the hierarchical categorization algorithm (see Section 5).

Domain Knowledge Slippage. Proposed rules, guidance and other documents on which agencies seek public comment often deal with issues that cannot be adequately understood without fairly sophisticated legal, scientific and/or technical knowledge. We believe the extraordinary demands for domain expertise posed by these kinds of text may introduce a real, but difficult to estimate, degree of confusion among non-expert annotators when an aggregate gold standard is used. Even after their initial period of training and group annotation, the upper-level law students annotating the FTA Grant Circulars Corpus struggled to establish nuances of meaning, as well as the precise scope, of many of the 38 issues. Further exacerbating these direct consequences of the lack of domain knowledge, many of the commentors were, like the agency, well-acquainted with the statutes, programs and policies involved. This shared knowledge enabled them to shortcut formal references and explanations that would have helped non-experts make categorization decisions.[10] Thus it is likely to be very difficult to obtain training sets with high levels of agreement across large issue sets for these kinds of texts using student or other non-expert annotators. We currently do not try to identify or correct for domain knowledge slippage.

Dynamically Changing Issue Set. According to our agency collaborators, their analysts can determine virtually all of

[10]Such "repeat players" are a feature of virtually every rulemaking and typically write the longest, most issue-laden and — according to agency rulewriters — "useful" comments.

Issue	Coverage (%)
funding	8.7
DefinNew	1.1
BeyondADA	1.8
AndOr	1.5
CompParatrans	0.6
OtherNFreeElig	3.2
NFreeMisc	0.5
JARC	4.6
JARCPriors	0.5
JARC_EligActiv	3.9
JARCMisc	0.2
planning	10.2
PlanElements	2.1
LeadAg	0.6
OutreachEff	1.3
StakeholderParticip	0.5
AgPartnerParticip	0.7
TransProvidrParticip	2.0
MultPartic	0.7
PlanLifespan	0.7
PlanCertif	0.3
CoordPlanDevelMisc	1.3
procedural	6.2
CompSelect	6.0
FairNEquit	0.2
evaluation	6.0
NoFedEval	0.3
Perf_1_Effic	1.1
Perf_2_Effect	0.8
Perf_3_Satis	0.8
Cost	0.2
EvalNOverMisc	2.8
AdminExpen	1.2
FundAppor	0.9
FundTransf	0.6
GEN_EligActiv	0.2
EligGrantees	0.3
TechAsstTrain	1.1
MobilMgt	1.5
DesRecip	2.8
Match	0.5
HowPlansRelate	2.8
GOMB	0.9
NONE	51.0

Table 2: Issue Distribution. Table shows the percentage of sentences (in the aggregate gold standard) that are labeled with each rule-specific issue.

the substantive issues that will arise in the comments even before the comments begin to arrive. Oftentimes, the proposed rule itself lays out the set of issues that the agency would like feedback on. Unexpected issues, however, sometimes arise, and existing issues might need to be further subdivided during the annotation process. We have ignored these complications in our current study.

Variation in Comment Quality, Scope and Form. Since comments are posted by entities ranging from law firms and trade or professional associations — both of which tend to have expertise in the area of the proposed rule — to relatively non-expert members of the public, the comments themselves vary in their clarity and their use of legal and technical terminology.

Knowledge Transfer Across Rulemakings. For text categorization techniques to be a feasible solution for rule-specific issue categorization, the amount of manually annotated training data (i.e. comments annotated by the rulewriters and analysts themselves) should be kept to a minimum. For this reason, text categorization methods that allow for inductive transfer across related rulemakings will need to be employed and developed [9] so that new rulemakings can benefit from previous rulemakings. We have also left this issue for future work.

5. THE TEXT CATEGORIZATION METHODS

In spite of the difficulties raised in the previous section, we have made progress in applying text categorization techniques to the CeRI FTA Grant Circulars Corpus.

We have investigated both *flat* and *hierarchical* text categorization methods. Flat (standard) text categorization techniques make no assumptions about relationships among categories and require that training texts are labeled according to a pre-defined, non-hierarchical list of categories. The vast majority of research in text categorization falls under this paradigm [8] and a state-of-the-art machine learning technique to use in this situation is a support vector machine (SVM) [11]. SVMs find a hyper-plane that separates the positive and negative training examples with a maximum margin in the vector space. For our flat text categorization algorithm, we use the multi-class version of SVM-light [3], which allows us to train a single categorization model that distinguishes among the 39 fine-grained rule-specific issues.

In contrast, *hierarchical* text categorization methods try to exploit a hierarchical categorization scheme when attempting to classify texts. Again, our goal is not to develop new hierarchical categorization algorithms, but instead to apply one such state-of-the-art technique to determine if our real-world categorization task will succumb to automatic methods. As a result, we loosely follow the approach proposed in Dumais & Chen [1], which trains separate sets of SVMs for each level of the categorization hierarchy. Unlike Dumais & Chen, however, we employ multi-class SVM classifiers whenever possible rather than train a collection of binary SVM classifiers. In addition, we do not experiment with combining the scores of first- and second-level classifiers. Finally, we found that treating NONE as a special case improves performance, resulting in what is essentially a three-level cate-

gorization scheme:

NONE**-classifier:** At the top of the hierarchy is a binary classifier that distinguishes sentences that address NONE of the issues from those that address *some* issue.

level-1 classifier: At the next level, is a multi-class classifier that distinguishes among the remaining 16 top-level issues for each sentence (i.e. excluding NONE).

level-2 classifiers: At the lowest level, we train one binary or multi-class classifier to distinguish among the leaf classes for each of the five hierarchical classes of level-1 (FUNDING, JARC, PLANNING, PROCEDURAL, and EVALUATION).

In the hierarchical setting, test sentences are processed by first applying the binary NONE-classifier. If the sentence is deemed non-NONE, then the level-1 and possibly a level-2 classifier is applied depending on the issue specificity required. When categorizing according to the 17 top-level issues, therefore, only the NONE-classifier and level-1 classifier are applied; when categorizing according to the 39 fine-grained issues, all three levels of classifier are applied.

6. EXPERIMENT METHODOLOGY

Following the real-world e-rulemaking setting that we are trying to emulate, we create six gold standards, one for each law student annotator. In contrast to many real-world comment analysis scenarios, however, each annotator was instructed to annotate the comment set w.r.t. all 39 of the fine-grained issues (rather than concentrate his or her annotation on a subset of the issues). This results in data sets notably smaller in size than would be the case if we combined the annotations of all of the annotators: although every annotator is responsible for covering the entire issue set, each annotator was assigned a relatively small set of comments for annotation.

Because the training data for each annotator is in short supply, we apply Porter stemming and stopword elimination on the term-based feature vectors. Minimally, this will aid in generalization at the lexical level.

During training and testing, we treat sentences with multiple issues as separate instances, one for each assigned issue. As a result, we will get at most one of the alternative instances correct in the test data. Note that this method of handling multiple-issue sentences differs from the Kwon et al. [5, 6] work, in which the learning algorithms are developed with multiple-issue sentences in mind. Their evaluation measures, in turn, differ from ours — they employ F-measure where we are able to use accuracy.

We investigated SVMs, naive Bayes, and conditional random fields (CRFs) under a variety of parameter settings and using 5-fold cross-validation. We use word-based feature vectors for the sentence-based training and test instances and report here only the results for the top-performing model — SVMs — under its best parameter settings determined using the training data (standard tfidf term weighting, and an RBF kernel). The c (complexity) and g (RBF kernel) parameters are also determined using the training data. We report results for half (3) of the annotators; results for the remaining three annotators are very similar.

7. RESULTS

Results are shown in Table 3. Note that the three annotators have labeled overlapping, but different, subsets of the FTA Grant Circulars comment set. The results across annotators are, therefore, not directly comparable.

Possibly the most notable result of Table 3 is that flat categorization techniques outperform hierarchical categorization when categorizing at both the coarse- and fine-grained issue levels. (Differences are statistically significant at the 0.05 level.) This has also been the case in previous research on hierarchical vs. flat text categorization: it has been more difficult than expected to produce hierarchical categorization methods that outperform their flat text categorization counterparts. (See Dumais & Chen [1] for a discussion.)

Nevertheless, the flat and hierarchical text categorization algorithms significantly outperform randomly assigned predictions when classifying sentences according to coarse-grained issues (achieves 5.9% accuracy) and according to fine-grained issues (achieves 2.6% accuracy). Both approaches also significantly outperform a classifier that always selects the most frequent issue that appears in the training set. (The performance of the most frequent issue baseline varies for each of the three annotation sets — from 24% accuracy for annotator3 to 35% for annotator1 and annotator2.)

Possibly more important is the fact that performance approaches our current interannotator agreement results — 46.4% for coarse-grained issue categorization and 64.7% for fine-grained issue categorization. This provides a promising indication that improvements in individual and inter-rater reliability in the training data will produce similar gains for automated text categorization techniques.

Figure 2 summarizes results for the top-performing fine-grained issues based on the Annotator 1 Corpus. Categorization results for NONE were obtained from the NONE classifier; results for the other issues were obtained from the hierarchical categorization system, which performs better than the flat categorization system for almost all categories except NONE. For each of the issues, we see that categorization accuracy approaches or exceeds the interannotator agreement score for the issue and that the accuracy can be relatively high even for issues with low coverage in the corpus. This bodes well for future work where agreement scores are expected to be higher.

8. CONCLUSIONS

We have presented the first results to date on rule-specific issue categorization in e-rulemaking. We provide detailed information on the creation of a public comment data set that has been manually annotated according to rule-specific issues at the sentence level. This is the first in a series of similar corpora to be developed by our e-rulemaking initiative.

We also presented results on automatic issue categorization using standard and hierarchical text categorization tech-

	Annotator 1 Corpus	Annotator 2 Corpus	Annotator 3 Corpus
flat categorization 39 fine-grained issues	0.45	0.56	0.42
hierarchical 39 fine-grained issues	0.43	0.53	0.38
flat categorization 17 coarse-grained issues	0.59	0.66	0.63
hierarchical 17 coarse-grained issues	0.56	0.60	0.60

Table 3: Flat and hierarchical categorization results. Results are 5-fold cross-validation accuracies.

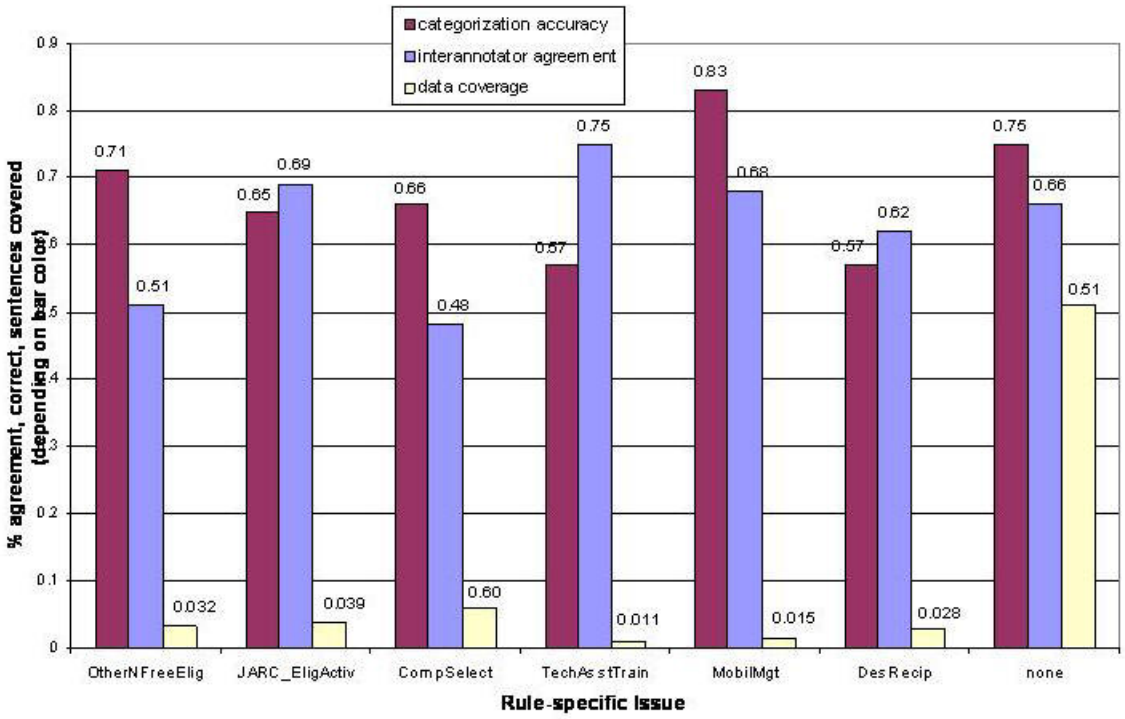

Figure 2: Per-Issue Performance for the Best-Performing Fine-Grained Issues for the Annotator 1 Corpus. Categorization results for NONE were obtained from the NONE classifier; results for the other issues were obtained from the hierarchical categorization system. In addition to categorization accuracy, we show interannotator agreement scores and data coverage for the issue.

niques. As in existing research, we find that flat categorization outperforms our attempts at hierarchical text categorization. Nevertheless, both approaches offer promise for the e-rulemaking domain in that they approach the levels of human interannotator agreement for the current data set.

9. ACKNOWLEDGMENTS

This work was supported by NSF Grant IIS-0535099.

10. REFERENCES

[1] S. Dumais and H. Chen. Hierarchical classification of web content. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263, New York, NY, USA, 2000. ACM.

[2] V. Hatzivassiloglou, J. Klavans, and E. Eskin. Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, pages 203–212, University of Maryland, College Park, MD, 1999. Association for Computational Linguistics.

[3] T. Joachims. *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.

[4] K. Krippendorff. *Content analysis: An introduction to its methodology (2nd Ed.)*. Sage Publications, 2004.

[5] N. Kwon and E. Hovy. Information acquisition using multiple classifications. In *Proceedings of the Fourth International Conference on Knowledge Capture (K-CAP 2007)*, 2007.

[6] N. Kwon, E. Hovy, and S. Shulman. Multidimensional text analysis for erulemaking. In *Proceedings of the 7th Annual International Conference on Digital Government Research*, 2006.

[7] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386, New York, NY, USA, 2006. ACM.

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

[9] D. Silver, G. Bakir, K. Bennett, R. Caruana, M. Pontil, S. Russell, and P. Tadepalli. *Inductive Transfer : 10 Years Later*. NIPS 2005 Workshop, 2005.

[10] P. Strauss, T. Rakoff, and C. Farina. *Administrative Law*. 10th edition, 2003.

[11] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[12] H. Yang and J. Callan. Near-duplicate detection for erulemaking. In *Proceedings of the Fifth National Conference on Digital Government Research*, 2005.

[13] H. Yang and J. Callan. Near-duplicate detection by instance-level constrained clustering. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[14] S. Zelikovitz and H. Hirsh. Improving short text classification using unlabeled background knowledge. In P. Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 1183–1190, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US.

APPENDIX

The proposed guidance circulars relate to grants under three federal statutory programs:

- New Freedom – grants for "new" public transport services and alternatives "beyond those required by" the ADA (Americans with Disabilities Act) that help with transportation for the disabled
- Elderly Individuals & Individuals with Disabilities Job Access & Reverse Commute (JARC)
- Grants for transportation of welfare recipients and other low income persons to and from jobs

A subsequent statute, the Safe, Accountable, Flexible, Efficient Transportation Equity Act: a Legacy for Users (SAFETEA-LU), links these three together through requirements that funded projects be "derived from a locally developed, coordinated public transit-human services transportation plan" developed through a specified process of stakeholder participation.

The 38 issue tags (plus NONE) used to annotate the FTA Grant Circulars corpus are described briefly in the table below.

FUNDING (under New Freedom program)		
	DefinNew	what qualifies as “new” services under “New Freedom”
	BeyondADA	when services are “beyond those required by” the ADA
	AndOr?	are these cumulative or alternative requirements?
	CompParatrans	eligibility of complementary paratransit services for funding
	Other NFreeElig	fundable activities not covered in prior categories
	NFreeMisc	issues not covered by any prior category
JARC (program issues)		
	JARCPriors	eligibility status of prior JARC-funded projects
	JARC-EligActiv	other eligible projects
	JARCMisc	other issues under this program
PLANNING (development of coordinated plan)		
	PlanElements	elements of plan
	Lead Ag	lead Agency
	OutreachEff	required public outreach efforts by grantees
	StakeholdrPartic	required stakeholder participation in plan development
	AgPartnerPartic	agency partner participation in plan development
	TransProvidrPartic	transportation provider participation in plan development
	MultiPartic	multiple participants in plan
	PlanLifespan	lifecycle and duration
	PlanCertif	certifying that funded projects come from a plan
	CoordPlanDevelMiS	issues not covered by any prior category
PROCEDURAL (aspects)		
	CompSelect	competitive selection process
	Fair&Equit	fair and equitable distribution of grant funds among organizations
EVALUATION (and oversight strategies)		
	NoFedEval	federal govt should not be setting evaluation measures
	Perf1Effic	performance measure #1: efficiency (more rides provided)
	Perf2Effect	performance measure #2: effectiveness (more communities served)
	Perf3Satisf	perform measure #3: satisfaction
	Cost	relevance of cost as evaluation measure
	EvalNOverMisc	evaluation & oversight issues not covered by any prior category
AdminExpen		what administrative expenses can be charged to the grants
FundAppor		how grant funds can be apportioned
FundTransf		permissible transfers of grand funds
Gen Elig Act		eligible activities under the grant (used only if no statute-specific tag applies)
EligGrantees		who is eligible to receive grants
TechAsst&Train		federally provided technical assistance & training for states & transit agencies
MobilMgt		mobility management and capital funding
DesRecip		selection of designated recipient of grant monies
Match		matching funds requirements
How Plans Relate		relationship of coordinated plan to statewide and metropolitan transportation planning
GOMB (”Get off my back”)		federal government should defer to state and local decision-making
NONE		tag automatically supplied to any sentence annotators left unmarked