Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts

Vanja Mladen Karan[*] **Jan Šnajder**[*] **Daniela Širinić**[†] **Goran Glavaš**[*]
[*]Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
{mladen.karan,jan.snajder,goran.glavas}@fer.hr

[†]Faculty of Political Science, University of Zagreb, Croatia
dsirinic@fpzg.hr

Abstract

Policy agenda research is concerned with measuring the policymaker activities. Topic classification has proven a valuable tool for policy agenda research. However, manual topic coding is extremely costly and time-consuming. Supervised topic classification offers a cost-effective and reliable alternative, yet it introduces new challenges, the most significant of which are the training set coding, classifier design, and accuracy-efficiency trade-off. In this work, we address these challenges in the context of the recently launched Croatian Policy Agendas project. We describe a new policy agenda dataset, explore the many system design choices, and report on the insights gained. Our best-performing model reaches 77% and 68% of F_1-score for major topics and subtopics, respectively.

1 Introduction

Understanding politics means understanding what political actors are saying and writing (Grimmer and Stewart, 2013), i.e., understanding the *content* of the messages. Accordingly, *content analysis* plays an important role in political science (Holsti, 1969; Weber, 1990; Krippendorff, 2012). Probably the most prominent form of content analysis is *topic classification*. In topic classification, the individual documents are assigned to a limited set of categories. Once documents have been assigned categories, they can be searched more efficiently than when using traditional keyword-based methods. Moreover, categories are a prerequisite for the analysis of patterns and changes in political content across time. As noted by, among others, Hillard et al. (2007), reliable topic classification can save significant research time.

One strand of research in which topic classification has proven beneficial is the analysis of policy agendas (Kingdon and Thurber, 1984): the set of issues arising in the decision-making process. The main idea is that the frequency with which the issues occur in political texts can be used as a measure of policy attention. This strand of research has been particularly influenced by the Policy Agendas Project (PAP), initiated by Bryan Jones and Frank Baumgartner in 1993, with the intention to track changes in policy activity within particular areas of policy-making over longer periods of time (John, 2006).[1] The main issue PAP addressed is that of reliably measuring the policymaker activities across time. To this end, PAP developed an exhaustive and consistent codebook comprised of 19 major topic and 225 subtopic codes, by which all policymaker activities were categorized. Building on this idea, the Comparative Agendas Project (CAP) (Bevan, 2014) extended the PAP codebook, originally developed for the United States.[2] While PAP was focused on ensuring longitudinal measurement reliability, CAP extended this methodological framework to also study policy changes comparatively, across time and space (countries). The CAP codebook consists of 21 major topics and more than 200 subtopics, used for coding of political texts for over 18 countries. Consequently, CAP-coded data have been used as the primary source for a number of policy agenda studies (e.g., Baumgartner et al. (2006)), and have been a foundation for one of the largest and most productive research networks in political science.

The perennial problem of topic classification – and content analysis in general – is the sheer volume of political texts. Manual coding is extremely time-consuming and costly, and thus does not scale

[1] http://www.policyagendas.org

[2] http://www.comparativeagendas.info

to large text collections. Consequently, as pointed out by Grimmer and Stewart (2013), analyzing large text collections is impossible for all but the most well-funded projects. Moreover, manual coding can be unreliable and inconsistent. For this reason, social scientists are increasingly relying on automated topic classification (ATC) (Purpura and Hillard, 2006; Quinn et al., 2006; Hillard et al., 2008; Quinn et al., 2010). ATC has two compelling advantages over human coding (Benoit, 2011): reliability and efficiency.

From a computational perspective, ATC is an instance of a more general text categorization task (Sebastiani, 2002), which falls within the purview of natural language processing and machine learning. The task is typically framed as a supervised machine learning problem, either multi-class (a single topic per document) or multi-label (multiple topics per document). Note that policy agenda research typically adopts the single-topic approach.

While arguably more efficient than human coding, ATC does come with its problems. First and foremost, ATC does not get around the problem of validity: ATC generally cannot detect nuances in the text as well as a human can, thereby limiting the validity of content analysis results. Secondly, there are a number of practical challenges involved in setting up a high-performance ATC system. Building an ATC system requires a high-quality manually coded dataset with a sufficiently large coverage. Furthermore, there are a lot of design choices involved, which greatly affect the system's performance. In the end, one does typically not want to compromise the quality otherwise obtainable by human coding, which means that a trade off has to be found between accuracy and human coding effort. This can be done by estimating the confidence of classifier decisions for each individual document, and then forwarding to a human coder the (hopefully small) subset of documents for which the decision confidence is low. For this to work, however, we need reliable estimates of classifier confidence, which turns out to be far from trivial.

In this work, we address the above challenges in the context of automatic topic classification of Croatian political texts. We first present a new dataset, built within the Croatian Policy Agendas Project, and a first such dataset for Croatian. The dataset has been manually coded according to the CAP codebook, with additional measures taken to ensure reliability. An additional challenge lies in the fact that the dataset consists only of titles, which further exacerbates the data sparsity problem. We use this dataset to train and evaluate a number of text classification models, also experimenting with two problem-specific extensions. Finally, we consider various confidence estimation strategies. The main research questions we answer are as follows: (1) Can we use the hierarchical structure of our topic scheme to improve classification performance?; (2) Can we make use of idiosyncratic coding rules?; and (3) What confidence estimation strategy gives best accuracy-efficiency trade-off? We hope that the lessons learned from these experiments will be useful to others working on the same or similar task for other languages.

The rest of the paper is structured as follows. In the next section, we briefly review the related work on ATC. In Section 3, we describe the Croatian Policy Agendas Project and the corresponding dataset. Section 4 focuses on the classification models. In Section 5, we present the experimental results. Section 6 concludes the paper and outlines future work.

2 Related Work

The use of supervised topic classification for policy agenda research has been introduced by Purpura and Hillard (2006). The authors presented a system that classifies the Congressional Bills according to the PAP codebook. Their system is a two-level support vector machine (SVM) with word features weighted by pointwise mutual information. The authors conclude that the system performs "about as well as humans would be expected to perform."

In subsequent work, Hillard et al. (2008) experiment with a number of classifiers (Naïve Bayes, SVM, BoosTexter, and MaxEnt), achieving high prediction accuracies across the different algorithms, with SVM emerging as the winner (88.7% and 81.0% accuracy on major topics and subtopics, respectively). Furthermore, they experiment with voting ensembles and investigate the accuracy-efficiency trade-off. While their experiments indicate that the improvement by ensemble voting is negligible, they also indicate that combining classifier decisions provides a key indication of classification confidence, which in turn can be used to lower the cost of improving accuracy. In particular, they demonstrate that inspecting and manually coding 20% of bills (about 1300 documents) where all three classifiers disagree boosts accuracy from

78% to 87%. Similarly, Collingwood and Wilkerson (2012) show that accepting decisions where at least three classifiers agree results in 86% average agreement at about 85% coverage.

The key idea behind the accuracy-efficiency trade-off is to reject the automatic classification of documents on which the classifiers exhibit low confidence. The alternative way to mitigate the cost of human coding is to incorporate the classifier in the coding process up front, in a so-called active learning setup. In active learning, the classifier confidence is used as a signal to guide the human coder which documents to code next, yielding larger accuracy improvements with lower coding effort. Hillard et al. (2007) show that, when compared to random sampling, active learning leads to a statistically significant 3% accuracy increase on the Congress Bills dataset.

Albeit our work focuses on supervised topic classification, for completeness we note that there exists a valuable body of work on the use of unsupervised topic classification from political texts. This strand of research mostly revolves around the use of topic models (Blei, 2012), e.g., (Quinn et al., 2006; Quinn et al., 2010; Grimmer, 2010). Other lines of research consider the estimation of category proportions instead of assigning single topics to documents (Hopkins and King, 2010), as well as the use of dictionaries for single- and multi-topic classification (Albaugh et al., 2013).

3 The Croatian Policy Agendas Project

The Croatian Policy Agendas project was launched with the aim of better understanding the changes in policy activity and policy priorities in a new democracy. The project is part of a large body of political agenda research that started with the Policy Agendas and Congressional Bills projects in the United States (E Adler and Wilkerson, 2006; John, 2006), and which has recently evolved into the Comparative Agendas Project (CAP) – a growing network of national projects in 17 countries. All national projects focused on manual topic coding of various policy documents such as legislation, political speeches, judicial decisions, media content, or public opinion. Regardless of the type of documents and observations, all materials were coded according to the CAP master codebook with 21 top-level (major) topic codes (shown in Table 1) and over 200 subtopic codes. The standardized coding system enables (1) the capturing of the policy focus of each observation, regardless of its source (Bevan, 2014), and (2) comparison of policy agendas across countries and regions.

Code	Major topic
1	Domestic Macroeconomic Issues
2	Civil Rights, Minority Issues, and Civil Liberties
3	Health
4	Agriculture
5	Labor and Employment
6	Education
7	Environment
8	Energy
9	Immigration and Refugee Issues
10	Transportation
12	Law, Crime, and Family Issues
13	Social Welfare
14	Community Development and Housing Issues
15	Banking, Finance, and Domestic Commerce
16	Defense
17	Space, Science, Technology, and Communications
18	Foreign Trade
19	International Affairs and Foreign Aid
20	Government Operations
21	Public Lands, Water Management, and Territorial Issues
23	Cultural Policy Issues

Table 1: Top-level policy topics (major topics)

3.1 Data Collection

The data gathering for the Croatian Policy Agendas project began in June 2015 and has so far resulted in a collection consisting of titles[3] of (1) all documents published by the National Gazette from January 1990 to December 2015 (all legal acts of the Parliament, the Government, and the President), (2) all agendas of the Croatian Parliament and Croatian Government, and (3) parliamentary questions. All document titles were merged into a single dataset, totaling over 100,000 title units. A subset of these were chosen for manual topic coding. It is worth pointing out that a large portion of documents from our collection are restricted access documents (e.g., minutes of the Government cabinet meeting), hence working with titles is the only option in such cases. In contrast, for publicly accessible documents, the content analysis could also be extended to full texts; we leave this option for future work.

[3] Whenever possible, CAP datasets include a link to original documents and complementary text that was used for classification. In some countries, full access to digitized documents was possible. In most cases, however, including Croatia, only document titles were available.

Measure	CS #1	CS #2	CS #3	CS #4
Percent agreement	81.5	81.2	80.6	85.4
Fleiss' κ	0.61	0.61	0.60	0.70
Krippendorff's α	0.61	0.62	0.60	0.70

Table 2: Calibration inter-annotator agreement

Measure	Phase #1	Phase #2	Phase #3
Percent agreement	51.2	79.7	83.0
Cohen's κ	0.51	0.79	–
Fleiss' κ	–	–	0.87
Number of coders	2	2	3

Table 3: Inter-annotator agreement

3.2 Coding Procedure

We devised the coding procedure so to ensure high reliability of the data. To this end, we split the coding procedure into several sessions, with checkpoints between them. The coding was carried out by thirteen students of political science and legal studies. After the initial training session, whose purpose was to introduce the students to the task and explain the coding guidelines, all thirteen students coded four small calibration sets, each consisting of 50 titles (a total of 200 titles). The calibration step allowed us to (1) identify which topics require a more detailed explanation and provision of examples from the codebook and (2) measure the inter-annotator agreement (IAA). We show the IAA on the four calibration sets (CS) in Table 2.

After the calibration session, we prepared a sample of document titles for further coding. To ensure that there is a sufficient variation across subtopics, we used stratified random sampling to select 7300 titles, accounting also for the source of the document (National Gazette, parliamentary sessions agenda, government weekly meetings agenda, or parliamentary questions). This introduces a variance across the topics and document types, which differ greatly in vocabulary and form of the titles.

The main coding session was carried out in four phases. First, each document title was coded independently by two out of thirteen students, where students were asked to take notes and tag the examples they consider problematic. In the second phase, we split the thirteen students into four groups and considered only the titles where coders disagreed in the first coding phases, as well as titles tagged as problematic by at least one of the coders (even if they agreed on the code). Each title on which the coders disagreed or which was tagged as problematic in the second phase was independently coded by two out of four groups. In the third coding phase, three political sciences experts independently coded all titles where codings by two student groups differed. Finally, the disagreements remaining after the third coding phase were discussed and resolved by consensus by the three experts. Table 3 shows the IAA measures for each of the coding phases. We make the manually coded dataset freely available.[4]

Table 4 gives some examples from the dataset. Particularly interesting are the titles that belong to the 00 subtopic (General): these are either (1) too general to be categorized in any of the more specific subtopics or (2) pertaining to two or more different subtopics. Also interesting is the 99 subtopic (Other), assigned to titles on a well-defined subtopic not covered by the CAP codebook.

4 Topic Classification Models

Following Purpura and Hillard (2006) as well as Hillard et al. (2008), we frame the topic classification task as a supervised multi-class classification problem. Solving this problem involves a number of design choices: choosing from among different machine learning algorithms, multi-class classification schemes, and methods to handle hierarchy. While our study is far from exhaustive, we do explore a reasonable number of options.

4.1 Text Preprocessing

We apply the typical text categorization preprocessing pipeline: we tokenize all documents, lemmatize the words using an automatically acquired morphological lexicon built by Šnajder et al. (2008), and remove all stopwords (non-content words). We chose to lemmatize because Croatian is an inflectionally rich language, and prior research (Malenica et al., 2008) has shown that lemmatization improves classifier performance. We do not apply any further preprocessing such as parsing, as syntactic features are very sparse and would require much more data to yield any benefits.

4.2 Algorithms and Schemes

There are three approaches to multi-class classification. One option is to use a classifier that can naturally handle multiple classes, such as the Naïve Bayes. The other two options rely on decomposing

[4] `http://takelab.fer.hr/data/apa`

Title (Croatian)	Title (English)	Code	Major topic / Subtopic
Odluka o imenovanju ministra financija	Appointment decision for the finance minister position	1500	Finance / General
Odluka o suglasnosti za povećanje cijena električne energije	Decision of approval for the increase in electricity prices	802	Energy / Electrical Energy
Pravilnik o socijalnom zbrinjavanju useljenika i povratnika	Regulation of social care for immigrants and returnees	1399	Social Welfare / Other
Zakon o postupanju s nezakonito izgradenim zgradama	Law on the treatment of illegally constructed buildings	1401	Community Development / Housing
Pravilnik o praćenju emisija onečišćujućih tvari u zrak iz nepokretnih izvora	Regulation of tracking air pollutants emissions from immobile sources	705	Environment topic / Air Pollution

Table 4: Example titles and their codes from the Croatian Policy Agendas Project data set

a multi-class problem into a series of binary classification problems. The one-vs-one (OVO) scheme works by training one binary classifier for each pair of classes. The prediction for an instance is obtained by voting of the individual binary classifiers. In contrast, in the one-vs-rest (OVR) scheme, we train for each class one binary classifier separating that class from all the other classes. An instance is classified into the class for which the corresponding classifier confidence is the highest.[5] The OVO and OVR schemes apply a divide-and-conquer strategy as they break up one difficult multi-class problem into many smaller and simpler binary problems. However, the downside of these schemes is that they introduce a large number of classifiers, consequently making the training resource-intensive.

In this work we consider a number of different algorithms and schemes, as follows.

LR-OVO. For this model, we use a binary logistic regression classifier implemented in the LIBLINEAR package (Fan et al., 2008), coupled with the OVO scheme.[6] To avoid overfitting, we optimize the hyperparameter C on a held-out validation set. Moreover, we perform implicit feature selection using L1-regularization, enforcing feature sparsity. The logistic regression classifier predicts class probability, which can be used directly as a measure of classification confidence. To accommodate the multi-class setup, we compute the confidence for class c as the average of confidences of all pairwise classifiers that include c.

LR-OVR. This model is the same as LR-OVO, but employs the OVR multi-class scheme. The confidence for class c is simply the confidence of the binary classifier corresponding to that class.

GNB. A Naïve Bayes (NB) model with numerical feature vectors, where the class likelihoods are modeled using Gaussian distributions. We make the usual simplifying assumption of a diagonal and shared covariance matrix. We note that for text classification a multinomial NB is more often used than a Gaussian NB. The motivation for using a Gaussian version is that we wanted all our classifiers to work with identical (numeric) feature vectors.

XGB. We experiment with the extreme gradient boosting algorithm (Chen and Guestrin, 2016). It is a decision tree-based algorithm, which aims to obtain "strong" classifiers by combining a large number of "weaker" ones. To avoid overfitting, we optimize the *eta* and *numrounds* hyperparameters on a held-out validation set.

4.3 Hierarchical Classification

The CAP codebook is a two-level taxonomy, featuring 21 major topics and more than 200 subtopics. Although we are ultimately interested in classifying documents into subtopics, we can leverage the hierarchical structure to decompose the multi-class problem into two separate classification problems, one for each hierarchy level. The assumption is that the separate problems are easier to solve than the original joint problem.

In line with common practice, we use the top-down level-based approach, in which one flat classifier is trained for each level of the hierarchy. We train a classifier to discriminate between major topics and, for each major topic, one classifier to discriminate between its subtopics. At prediction time, the straightforward approach would be to apply the

[5] We note that there are many variants of the OVO and OVR schemes; the interested reader is referred to (Galar et al., 2011) for an overview.

[6] We also experimented with the SVM algorithm from the same library and found the logistic regression to perform slightly better on our dataset. For the sake of brevity, we omit the SVM results.

top-level classifier first to obtain the major topic, and then apply the corresponding second-level classifier to obtain the subtopic. The obvious downside is that the error propagates: if the model makes a mistake at the major topic level, it cannot be undone. To mitigate this, we linearly combine the confidences from both levels:

$$f(t, s_t) = conf_1(t) + \alpha \cdot conf_2(s_t) \tag{1}$$

where t is a major topic, s_t is its subtopic, and $conf_n$ is the confidence of the classifier at level n. Using the joint confidence derived by f softens the strict two-level split and may alleviate error propagation issues. Furthermore, it allows us to weigh decisions from different levels differently. The intuition behind this is that we expect decisions on the first level to be more confident as (1) there is more training data and (2) the differences between major topics are more prominent than the differences among subtopics within one major topic.

In our models, we calculate f for all possible major topic/subtopic pairs and classify the document into the subtopic that maximizes f. We optimize α on a held-out validation set. We denote the hierarchical versions of our models as LR-OVO-H, LR-OVR-H, LR-GNB-H, and LR-XGB-H. To account for the possibility that a non-hierarchical approach works better on our dataset, we also build a flat LR-OVR model trained directly on all 208 subtopics, denoted LR-OVR-F.

4.4 Features

We use the same set of features for our models:

- Lemmas – we weigh each lemma l using the tf-idf weighting scheme:

  $$\mathit{tfidf}(l) = \mathit{freq}(l) \cdot \frac{|D|}{|\{d \mid l \in d\}|} \tag{2}$$

  where *freq* is the frequency of l in the document, while D is the set of all documents;
- Bigrams – binary features for 300 most frequent bigrams in the data set;
- Word2vec – we use distributed word representations proposed by Mikolov et al. (2013), derived by applying the *word2vec* tool on the hrWaC web-corpus (Ljubešić and Erjavec, 2011). Following Mitchell and Lapata (2010), we compute the composed semantic representation of a document as the sum the vectors of its content words. The resulting vector of length 300 is fed as input to our models.

While we do not perform explicit feature selection, it is performed implicitly by the L1-regularization in LR-based models, and also in the XGB model, which embeds feature selection.

4.5 Postprocessing Rules

The second extension we consider is the application of postprocessing rules. These are meant to enforce two specific coding principles, also prescribed in the coding guidelines:

1. If two or more subtopics are equally represented in a document, or the document content is rather general, then it should be assigned the General (00) subtopic;
2. If a document does not fit well into any of the existing subtopics, but the document content is not general, then it should be assigned the Other (99) subtopic.

We map these to two postprocesing rules:

1. If, for a given document, the ratio of confidences for the top two subtopics is above a threshold θ_1, the document is labeled with the General (00) subtopic;
2. If the highest confidence subtopic for a given document is below θ_2, then the document is labeled with the Other (99) subtopic.

Each rule is parametrized by a threshold that is tuned on a held-out validation set.

4.6 Confidence Estimation

Validity is of central concern to any content analysis study. To preserve validity, researchers will often be willing to trade off coding efficiency for topic classification accuracy. As demonstrated by Hillard et al. (2008), as well as Collingwood and Wilkerson (2012), significant improvements in accuracy can be obtained by leveraging the insights about classification confidence.

In machine learning parlance, the accuracy-efficiency trade-off is known as *classification with reject option* (Herbei and Wegkamp, 2006). In many practical applications, it is better if the classifier refrains from making a prediction unless it is sufficiently confident. Intuitively, the accuracy and rejection are related; according to Chow (1970), the error rate decreases monotonically with increasing the rejection rate. The key, then, is devising the optimal optimal rejection rule.

In our experiments, we wish to control the number of documents, N, rejected by the classifier. These documents will be forwarded to a human coder, and hence directly determine the coding costs. We implemented four rejection strategies.

Single threshold. This simple strategy relies on classifier confidence estimates. Documents are ranked by confidences and the bottom-ranked N documents are rejected.

Ensemble disagreement. Classifier ensembles (Dietterich, 2000) provide a natural way of estimating confidences by means of agreement levels. The main idea is to reject the instances on which a certain number of classifiers disagree. While this strategy has been shown efficient by Collingwood and Wilkerson (2012), it does not control for the number of rejections. We therefore use a slightly different strategy, also considered by Hillard et al. (2008): using a 3-classifier ensemble, we sample the desired number of documents from the set of document on which at least one classifier disagrees. In the experimental section, to account for the randomness of the sampling, we run the procedure 100 times and report the average performance.

Ensemble threshold. Inspired by Fumera and Roli (2004), we compute the total confidence of a 3-classifier ensemble as a product of the individual classifiers' confidences.

Optimized thresholds. This is a more elaborate rejection strategy that leverages the hierarchical structure as well as confidences between subtopics. A document is rejected if either:

1. Its major topic confidence is less than a threshold p_1. The intuition here is that, if a prediction has low confidence on the major topic level, then it is most likely erroneous;
2. Both its subtopic confidence is less than p_2 and the difference to the second-highest confidence subtopic is less than p_3. The intuition is that, in addition to the classifier confidence, what signals classification error are the situations in which the confidences for the two most confident classes are too close.

We optimize thresholds p_1, p_2, and p_3 on a held-out validation set to maximize accuracy score, while fixing the maximum number of documents the model is allowed to reject.

5 Experimental Evaluation

In this section, we report on the results for the different classification models and rejection strategies on the Croatian Policy Agendas Project dataset.

5.1 Setup

To obtain more reliable performance estimates, we use 5-fold cross-validation, and report the mean and standard deviation of each evaluation measure across the five folds. We report micro- and macro-averaged F1-scores (denoted F_1^μ and F_1^M, respectively), Cohen's kappa coefficient (Cohen, 1960), and the AC1 coefficient (Gwet, 2002). All model hyperparameters are tuned using grid search on a held-out validation set.

5.2 Classification Accuracy

Classification performance for all our models is given in Table 5. The LR-based models outperform the other two considered models on both hierarchy levels. We observe that, on the major topic level, the OVR-based models considerably outperform OVO-based models. However, on the subtopic level, both approaches perform comparably. Another observation is that, on the subtopic level, the best models are those that use hierarchy.

In addition to the individual models, we also experiment with an ensemble comprised of LR-OVR-H, LR-OVO-H, and XGB-H classifiers. The ensemble employs the majority voting strategy, while in case of ties it falls back to the prediction of the best-performing individual classifier (LR-OVR-H). The ensemble performs comparably to, or numerically outperforms, the LR-OVR-H model. The best micro F1-score is 0.77 and 0.68 for the major topic and subtopics, respectively.

In Table 6 we present results of the best-performing individual model (LR-OVR-H) for the major topics. We observe that those major topics on which the classifier performs the worst are also those with the least number of training instances. Table 7 shows the performance of LR-OVR-H on the 10 best-performing subtopics. As for the worst-performing subtopics, these have a score of 0 due to data sparsity (less than 15 training instances). These include, e.g., *Juvenile Crime (1206)* and *Rural Housing (1404)*, each with only 7 instances.

5.3 Thresholds

Our models use a number of thresholds for hierarchical classification and postprocessing rules, op-

	Subtopics				Major topics			
Model	F_1^μ	F_1^M	κ	AC1	F_1^μ	F_1^M	κ	AC1
GNB-H	0.41 ± .01	0.31 ± .01	0.40 ± .01	0.41 ± .01	0.57 ± .01	0.50 ± .01	0.53 ± .01	0.57 ± .01
LR-OVO-H	0.61 ± .01	0.50 ± .01	0.61 ± .01	0.61 ± .01	0.75 ± .01	0.69 ± .02	0.72 ± .01	0.75 ± .01
XGB-H	0.58 ± .02	0.46 ± .03	0.57 ± .02	0.58 ± .02	0.71 ± .01	0.69 ± .03	0.68 ± .01	0.71 ± .01
LR-OVR-H	0.65 ± .01	0.55 ± .02	0.65 ± .01	0.65 ± .01	**0.77** ± .01	**0.75** ± .01	**0.75** ± .01	**0.77** ± .01
LR-OVR-F	0.65 ± .01	0.54 ± .01	0.65 ± .01	0.65 ± .01	0.74 ± .01	0.71 ± .02	0.72 ± .01	0.74 ± .01
Ensemble	**0.68** ± .01	**0.56** ± .01	**0.67** ± .01	**0.68** ± .01	**0.77** ± .01	**0.75** ± .02	**0.75** ± .01	**0.77** ± .01

Table 5: Classifiers' performances and standard deviations on major topics (22) and subtopics (208)

Topic	# docs	F_1	κ
Macroeconomics (1)	410	0.72 ± .05	0.71 ± .05
Civil Rights . . . (2)	224	0.76 ± .05	0.75 ± .05
Health (3)	295	0.82 ± .01	0.82 ± .01
Agriculture (4)	397	0.77 ± .03	0.75 ± .03
Labor . . . (5)	202	0.76 ± .04	0.75 ± .04
Education (6)	222	0.84 ± .04	0.83 ± .04
Environment (7)	199	0.73 ± .04	0.72 ± .04
Energy (8)	225	0.87 ± .03	0.86 ± .03
Immigration . . . (9)	29	0.51 ± .15	0.51 ± .15
Transportation (10)	356	0.80 ± .02	0.79 ± .02
Law, Crime . . . (12)	711	0.82 ± .02	0.80 ± .02
Social Welfare (13)	191	0.68 ± .06	0.67 ± .06
Community (14)	245	0.76 ± .03	0.75 ± .03
Banking (15)	566	0.75 ± .01	0.73 ± .02
Defense (16)	437	0.75 ± .04	0.74 ± .04
Space, Science (17)	184	0.75 ± .02	0.74 ± .02
Foreign Trade (18)	206	0.73 ± .03	0.73 ± .03
International (19)	623	0.77 ± .02	0.75 ± .02
Government op. (20)	1253	0.74 ± .01	0.68 ± .01
Public lands (21)	298	0.84 ± .03	0.84 ± .04
Cultural Policy . . . (23)	91	0.67 ± .09	0.67 ± .09
Other (99)	14	0.59 ± .10	0.59 ± .10

Table 6: Results by topic on the major topic level

Topic	# docs	F_1	κ
Drugs . . . (342)	21	0.96 ± .05	0.96 ± .05
Gender . . . (202)	22	0.95 ± .10	0.95 ± .10
Court . . . (1204)	344	0.92 ± .02	0.92 ± .02
Alternative . . . (806)	22	0.91 ± .11	0.91 ± .11
Price control . . . (110)	26	0.89 ± .05	0.89 ± .05
Trade . . . (1802)	19	0.87 ± .19	0.87 ± .19
Census . . . (2013)	25	0.86 ± .17	0.86 ± .16
Monetary . . . (104)	30	0.86 ± .03	0.86 ± .03
Drinking water . . . (701)	20	0.85 ± .11	0.85 ± .11
Water . . . (2104)	171	0.85 ± .04	0.84 ± .04

Table 7: Results for 10 best-predicted subtopics

timized on held-out datasets. Some insights into model behavior and nature of the task can be obtained by inspecting the optimal threshold values.

The optimal values for the postprocessing rules' thresholds are such that the rules are effectively never activated. This is likely because the cases where the rules could improve the accuracy are much less frequent than those where they could harm, so overall it is better never to activate them.

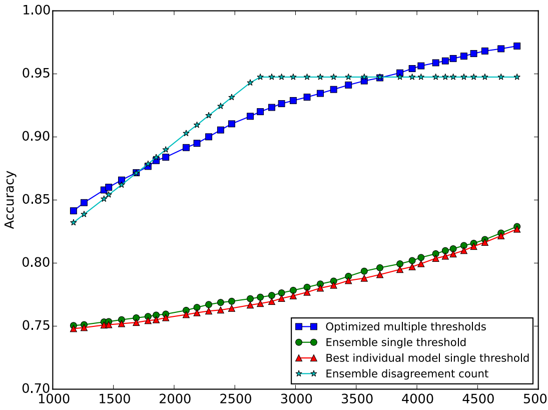

Figure 1: Acceptance-rejection curves for the different rejection strategies

The optimal value for the hierarchy threshold is very low (α=0.01). This suggests that, when calculating the joint confidence, much more weight is given to the major topic decision. This result is in line with the expectation that major topic classifiers are more reliable than subtopic classifiers.

5.4 Rejection Strategy

We evaluate the different rejection strategies to see which one offers the best accuracy-effort trade-off. To quantitatively assess this trade-off, we adopt the Accuracy-Rejection curves (ARC) proposed by Nadeem et al. (2010). The ARC shows the accuracy of a classifier as a function of its rejection rate (number of documents forwarded to human coders). A good rejection strategy will reach high accuracy levels even for low rejection rates.

The plots for various strategies described in Section 4.6 are given in Figure 1. The strategy of optimizing several thresholds to yield maximum accuracy significantly outperforms the two single-threshold strategies. Moreover, it performs comparably to the ensemble disagreement-based approach, even though it requires only a single classifier. The ensemble disagreement approach levels out at about 2600 documents because that is the number of documents that satisfy its agreement condition; even if the maximal allowed number of documents to reject is higher, it can never reject more than 2600. Before that point, however, it provides the optimal rejection strategy. These results suggest that it might be beneficial to combine the ensemble disagreement and optimized thresholds strategies. The results also show that, if relying on the ensemble disagreement strategy, manually checking about 30% of the data set (2300/7300) would yield a substantial improvement in accuracy from 77% to 90%.

6 Conclusion

We addressed the task of supervised topic classification of Croatian political texts, undertaken as part of the recently launched Croatian Policy Agendas Project. We built a new dataset consisting of 7300 titles, manually coded according to the Comparative Agendas Project codebook. On this dataset, we experimented with a number of machine learning models, and investigated to what extent the models can benefit from including hierarchy information or postprocessing rules. We learned that, on this dataset, a hierarchical approach indeed performs better. Rules however, did not bring any improvement to our models. We also experimented with different rejection strategies, aiming to optimize the accuracy-efficiency trade-off. We find that an ensemble disagreement-based method and our proposed method that optimizes multiple thresholds perform comparably well.

A possible venue of future work is the combination of different rejection strategies. Another promising possibility is the use of the most recent state-of-the-art models for text classification such as convolutional neural networks (CNN) or recurrent neural networks (RNN). Finally, it would be interesting to see whether the performance could be improved further by supplying full document texts and additional meta-data.

Acknowledgments. The Croatian Policy Agendas Project is funded by the European Social Fund; more details can be found on `www.cepis.hr`. This work was also in part supported by the Unity Through Knowledge Fund of the Croatian Science Foundation, under the grant 19/15: "EVEnt Retrieval Based on semantically Enriched Structures for Interactive user Tasks (EVERBEST)". We thank the reviewers for their helpful suggestions.

References

Quinn Albaugh, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. The automated coding of policy agendas: A dictionary-based approach. In *6th Annual Comparative Agendas Conference, Atnwerp, Beligum*.

Frank R Baumgartner, Christoffer Green-Pedersen, and Bryan D Jones. 2006. Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7):959–974.

William L Benoit. 2011. Content analysis in political communication. *The sourcebook for political communication research: methods, measures, and analytical techniques*, pages 268–279.

Shaun Bevan. 2014. Gone fishing: The creation of the comparative agendas project master codebook. Technical report, Mannheim: Mannheimer Zentrum für Europäische Sozialforschung.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*.

Chao K Chow. 1970. On optimum recognition error and reject tradeoff. *Information Theory, IEEE Transactions on*, 16(1):41–46.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Loren Collingwood and John Wilkerson. 2012. Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, 9(3):298–318.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer.

Scott E Adler and John Wilkerson. 2006. Congressional bills project - technical report. Technical report, University of Washington.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Giorgio Fumera and Fabio Roli. 2004. Analysis of error-reject trade-off in linearly combined multiple classifiers. *Pattern Recognition*, 37(6):1245–1265.

Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761–1776.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Kilem Gwet. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, 1(6):1–6.

Radu Herbei and Marten H Wegkamp. 2006. Classification with reject option. *Canadian Journal of Statistics*, 34(4):709–721.

Dustin Hillard, Stephen Purpura, and John Wilkerson. 2007. An active learning framework for classifying political text. In *Annual Meeting of the Midwest Political Science Association, Chicago*.

Dustin Hillard, Stephen Purpura, and John Wilkerson. 2008. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46.

Ole R Holsti. 1969. *Content analysis for the social sciences and humanities*. Addison-Wesley.

Daniel J Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Peter John. 2006. The policy agendas project: a review. *Journal of European Public Policy*, 13(7):975–986.

John W Kingdon and James A Thurber. 1984. *Agendas, alternatives, and public policies*, volume 45. Little, Brown Boston.

Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology*. Sage.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer.

Mislav Malenica, Tomislav Šmuc, Jan Šnajder, and Bojana Dalbelo Bašić. 2008. Language morphology offset: Text classification on a Croatian–English parallel corpus. *Information processing & management*, 44(1):325–339.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2010. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *MLSB*, pages 65–81.

Stephen Purpura and Dustin Hillard. 2006. Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on Digital government research*, pages 219–225. Digital Government Society of North America.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2006. An automated method of topic-coding legislative speech over time with application to the 105th-108th us senate. In *Midwest Political Science Association Meeting*, pages 1–61.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Jan Šnajder, Bojana Dalbelo Bašić, and Marko Tadić. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management*, 44(5):1720–1731.

Robert Philip Weber. 1990. *Basic content analysis*. Number 49. Sage.