The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach

Miklós Sebők[1] and Zoltán Kacsuk[1,2]

[1] Centre for Social Sciences, Hungarian Academy of Sciences, Budapest, Hungary.
Email: sebok.miklos@tk.mta.hu
[2] Hochschule der Medien, Stuttgart, Germany

Abstract

In this article, we present a machine learning-based solution for matching the performance of the gold standard of double-blind human coding when it comes to content analysis in comparative politics. We combine a quantitative text analysis approach with supervised learning and limited human resources in order to classify the front-page articles of a leading Hungarian daily newspaper based on their full text. Our goal was to assign items in our dataset to one of 21 policy topics based on the codebook of the Comparative Agendas Project. The classification of the imbalanced classes of topics was handled by a hybrid binary snowball workflow. This relies on limited human resources as well as supervised learning; it simplifies the multiclass problem to one of binary choice; and it is based on a snowball approach as we augment the training set with machine-classified observations after each successful round and also between corpora. Our results show that our approach provided better precision results (of over 80% for most topic codes) than what is customary for human coders and most computer-assisted coding projects. Nevertheless, this high precision came at the expense of a relatively low, below 60%, share of labeled articles.

Keywords: machine learning, statistical analysis of texts, Comparative Agendas Project, multiclass classification, automated content analysis

1. Introduction

In the 21st century , machine learning (ML) has become one of the cutting-edge subfields of quantitative political science. According to Grimmer (2015, 82), using ML to "make causal inferences is one of the fastest growing and most open fields in political methodology." Besides the prediction of roll call votes (Bonica 2018) and international conflicts (Colaresi and Mahmood 2017), ML is also widely used for discovery, such as detecting electoral fraud (Levin, Pomares, and Alvarez 2016). In light of these developments, it is safe to say that ML has become a standard tool in the toolkit of political analysis. Nevertheless, combined with other fast-developing areas of research, such as text mining, it also offers new solutions to the methodological problems of the creation of Big Data datasets which serve as the basis for a swathe of contemporary quantitative political analysis.

Despite these methodological advancements, some of the most important international collaborative projects in comparative politics still rely on human effort in creating Big Data databases. This is true of one of the premier such enterprises, the Comparative Agendas Project (CAP—Baumgartner, Breunig, and Grossman 2019). The CAP project assigns 21 "major" policy topics from education to defense to observations from a number of different data sources such as newspaper articles (Boydstun 2013) or laws.[1] These efforts have mostly relied on double-blind human coding although a few experimental papers supplanted human coding with a dictionary-based method

[1] Although in this article we use media data, for which additional codes (such as "Weather" or "Sports") are available in some datasets, for the sake of comparability in our analysis of media data we use the original, core set of codes.

Political Analysis (2021)
vol. 29: 236–249
DOI: 10.1017/pan.2020.27

Published
11 November 2020

Corresponding author
Miklós Sebők

Edited by
Jeff Gill

(Albaugh *et al*. 2013), a mixture of dictionary-based and ML approaches (Albaugh *et al*. 2014), or "pure" ML (Burscher, Vliegenthart, and De Vreese 2015; Karan *et al*. 2016).

What is common in these studies is that precision and/or F1 scores,[2] these major metrics of ML efficiency, surpassing a value of 80% (which is sometimes considered to be the benchmark of validity), remained elusive. Furthermore, non-English language applications of these methods remained few and far between (except for the abovementioned Dutch and Croatian cases), and they mostly rely on a single corpus to demonstrate the ostensible effects (with the notable exception of Burscher *et al*. 2015). As a result, the automated or semiautomated classification of data is still the exception rather than the rule in the CAP and many similar content analysis endeavors.

In this article, we present an ML-based solution for matching the performance of the gold standard of double-blind human coding when it comes to the multiclass classification of imbalanced classes of policy topics in newspaper articles. Such imbalanced database structures are not only a prevalent feature of CAP, they are also common in comparative politics in general. Therefore, our solution may have added value to projects beyond the use case presented in this article. Our primary performance metric is precision as our aim is to arrive at a valid classification of articles. By churning out true positives, an ML-augmented project may significantly reduce the amount of items that can only be handled by trained annotators, and, therefore, the success of such a project immediately benefits large-scale coding undertakings. Hence, the proposed process is inspired by the need to keep human coding costs as low as possible, while extracting the largest possible gain per invested human coding hour.

We call our approach a "hybrid binary snowball" (HBS) workflow based on the three defining characteristics of the proposed solution. Since our aim is to get the biggest return on manpower we use a hybrid or semiautomatic process (for a similar approach, see Loftis and Mortensen 2020). Second, our approach simplifies multiclass classification by creating a setup in which each code is assigned based on a pairwise ("binary") comparison with all other codes. Third, we apply a snowball method to augment the training set with machine-classified observations after each successful round of classification and also to create a training set for classifying another in-domain corpus.

We tested the HBS workflow for classifying items in textual databases with unbalanced classes on Magyar Nemzet (MN), a Hungarian right-wing daily by using a training set generated from the left-leaning Népszabadság (abbreviated as NS).[3] The input data are the full text of front-page articles as they appeared on the front-page of the two newspapers. We used our NS corpus to train a model for classifying articles in our "virgin" MN corpus. As is clear from this description, no human coding was applied to this second dataset—only ex post validation entailed manual work on behalf of our research team.

On the one hand, the output result of over 83% precision is comparable to the intercoder reliability values of human coding-based projects. On the other hand, the total percentage of texts classified of around 58.2% of all articles (a proxy for recall, which cannot be calculated for our sample-based approach) requires further refinements of our research design. Based on these results, the HBS process offers a viable, scalable, and potentially domain-independent solution to multiclass classification problems in comparative politics and beyond.

In what follows, we first review the relevant literature. Next, we provide an outline of our proposed workflow for tackling the task of multiclass classification with unbalanced classes for our corpora. This is followed by a presentation of the results from our case study of intercorpus snowballing. Next, we discuss our results in terms of their robustness in light of simulation results

[2] Precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of the total amount of relevant instances that were actually retrieved. The F1 score is the harmonic mean of the precision and recall.

[3] Replication data and code is provided on the Political Analysis Dataverse, see Sebők and Kacsuk (2020).

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

and possible improvements over the default version of HBS. We conclude by an assessment of our contributions to the literature and avenues for further research.

2. Literature Review

Despite recent methodological advancements in the field of ML and text mining, the most important international collaborative projects in comparative politics still rely on human effort in creating Big Data databases. Notwithstanding the criticism (Mikhaylov, Laver, and Benoit 2008) aimed at some projects, such as the Comparative Manifesto Project/MARPOR (Volkens, Bara, and Budge 2009), the gold standard of annotation and coding for large-scale endeavors covering multiple countries and languages remains double-blind human coding. This is true of one of the premier such enterprises in comparative politics, the CAP as well.

As an international network of scholars studying the dynamics of public policy agendas (Baumgartner, Green-Pedersen, and Jones 2013), CAP relies on large-scale databases of textual data and a codebook which caters to country-specific needs but, at the same time, maintains the comparability of project-level data (Bevan 2019). The CAP codebook covers 21 "major" policy topics from education to defense along with a total of over 200 subtopics for each major topic. The first and most developed country projects (such as the U.S. Policy Agendas Project or the UK team) predominantly relied on double-blind human coding, and most local teams follow this methodology to this day. This is certainly feasible for smaller datasets (such as those related to laws, the number of which usually remains in the range of a few thousand per government cycle). Furthermore, the need for consistency of coding over time, between agendas and across countries and languages, puts a premium on human judgment.[4]

This is not to say that no precursors are available for contemporary attempts at using computer-assisted methods for the multiclass classification task of CAP. Efforts at the "computer-assisted" (Hillard, Purpura, and Wilkerson 2008; Collingwood and Wilkerson 2012; Lucas *et al*. 2015), "automated" (Quinn *et al*. 2006; Young and Soroka 2012; Flaounas *et al*. 2013), or "semiautomated" (Breeman *et al*. 2009; Jurka 2012) topic/thematic/sentiment classification (or content analysis, coding) of documents in comparative politics have produced results that speak to the relative usefulness of these methods.

In agenda research proper, multiple papers presented computer-assisted coding results either based on a dictionary-based method (Albaugh *et al*. 2013), a mixture of dictionary-based and ML approaches (Albaugh *et al*. 2013), or "pure" ML (Burscher *et al.*2015; Karan *et al*. 2016). What is common in these studies is that precision and/or F1 scores surpassing a rate of 80% (which is considered to be acceptable for many human coding-based projects) remained elusive. Furthermore, non-English language applications of these methods remained few and far between (except for the abovementioned Dutch and Croatian cases), and they mostly rely on a single corpus to demonstrate the ostensible effects (with the notable exception of Burscher *et al.*2015 and Loftis and Mortensen 2020).

As a result, the semiautomated classification of data is still the exception rather than the rule in CAP and many similar content analysis projects and even if it is used, it plays a support role besides human coders. A cautious approach is certainly warranted: "while human-based content analysis is accused of being unreliable, computer-based content analysis is castigated for missing out in semantic validity" (Volkens *et al*. 2009, 236). Besides such general reservations, the relative sparsity of computer-assisted coding may be partly due to the relatively low precision and/or recall results of previous studies, the inefficiencies of purely dictionary-based or mechanical methods of computer-assisted coding that characterized earlier research, or simply a lack of evangelization of ML in comparative content analysis research.

[4] We thank the anonymous reviewer for this comment.

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

3. HBS: A Human-Machine Hybrid Workflow for Multiclass Classification

The aim of our project is to match the performance of the gold standard of double-blind human coding when it comes to the multiclass classification of imbalanced classes of policy topics in newspaper articles. Such imbalanced class distributions are common in comparative politics, and, therefore, our solution may have added value to projects beyond the use case presented in this article.

Our primary performance metric is precision as our aim is to arrive at a valid classification of articles. The proposed process is inspired by the need to keep human coding costs as low as possible, while extracting the largest possible gain per invested human coding hour. This guiding principle informs the structure of our process. Table 1 presents an overview of the structural components, or modules, of the HBS workflow as well as some technical features of the analysis.

The main elements of the proposed workflow for solving unbalanced multiclass classification problems are the hybrid, the binary, and the snowball aspects (the latter is utilized in two different ways: as intracorpus and as intercorpus snowballing). First, we offer a hybrid solution that draws

Table 1. Elements of the HBS workflow solution.

	Human contributions	Machine tasks	Technical features of machine tasks
Hybrid	Initial human coding	Text preprocessing	Stopwords removed Stemming Minimum token length: two characters Minimum document frequency for tokens: 5 Weighting: tf-idf
	Ex post validation	Classification with supervised machine learning	Classifier: support vector machine (SVM)
			Default model parameters changed: max iterations: 10 and regularization parameter: 0.1
Binary	Simplified human validation task (not multiclass, only correct/incorrect)	Decomposition of multiclass classification problem into a set of binary ones	Ensemble classifier: bagging-type
			Treating class imbalance: Undersampling of negatives (without replacement)
Snowball: Intracorpus	Human validation of training set expansion: Unvalidated results also added to training set based on validated samples' precision	Ensemble voting solution for in-process training set expansion	Intracorpus snowballing: moving nonvalidated classified elements to the training set
Snowball: Intercorpus	Researcher managed process of using validated results from one HBS setup-corpus as training set for another	Possibility for integrated script in future work	

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

on human coding and ML in a strategic way, by exploiting their relative strengths. For same domain classification (where language and data sources are given), double-blind human coding is only needed for a small-scale training set and subsequent sample-based validation. With these limited efforts—as it is shown in this paper—a research team can undertake the task of content analysis of multiple newspapers of the same language regardless of the size and time frame of the underlying dataset. Although all supervised learning approaches that depend on initial human coding for the creation of a training set could be seen as being already hybrid in nature, our workflow differs in that we actively prioritize the allocation of human working hours to validation over initial coding (in line with the work of Loftis and Mortensen 2020).

Second, our process simplifies multiclass classification by creating a setup in which each code is contemplated for an observation in a pairwise comparison with all other codes. That is, in the first step, our algorithm assigns either “macroeconomics” (the first policy topic code in the codebook) or “other” to each observation in the sample. This comparison, then, is repeated for each subsequent topic code from human rights to culture. The decomposition of the multiclass classification problem into a set of binary ones therefore is based on the usage of pairwise comparisons. The “binary” element also contributes by simplifying the choice for human validation from a multiclass assignment to one that entails the assignment of the “correct” or “incorrect” label. This setup provides an easier task for human annotators from a cognitive perspective (see also Loftis and Mortensen 2020).

Third, we apply a snowball method to augment the training set with machine-classified observations after successful rounds of classification. This we call intracorpus snowballing, and in the following, we detail a setup where the newly labeled elements are added to the training set automatically. There is a further aspect of the way we take advantage of snowballing, namely using an already classified corpus to work on another corpus without the need for an initial round of human coding to create a starting training set for the new corpus. This latter aspect we refer to as intercorpus snowballing.

Both aspects of snowballing enable a speed up in the performance and savings in invested human work hours for our workflow. In this way, we are leveraging a relatively small human-coded training set for the computer-assisted creation of a full-scale training set and, then, apply this latter training set on a virgin corpus, thereby paving our way to an ever growing corpus/corpora of classified articles.

Thus, the solution we propose for complex, multicorpora projects with limited human resources is a sequential combination of HBS workflows. The choices for specific HBS setups (allocating human resources to coding or validation, etc.) and parameters are utilitarian, in the sense that they depend on the results of experimentation and the budget and other specifics of the project in which HBS is applied. By linking the workflows for specific HBS setup-corpus pairs into an overarching intercorpus learning framework (within, at least, the same domain, in terms of language or period or substantive classes), one can leverage validated results from one workflow as a training set for another.

The primary contribution of HBS based on these elements is to provide a workflow solution to real-life scientific projects with limited budgets and manpower. The key added value of the HBS process is that it uses concepts and methods in a particular way, as elements of the same workflow. This workflow itself, and not any of its constituent parts, offers a solution to the imbalanced class classification problem that researchers in the CAP community and beyond face. It also has to be emphasized that all three elements can be considered to be more workflow-based than statistical in nature. HBS as a workflow can actually accommodate all sorts of specific ML or preprocessing (see feature set) solutions and decisions, which can be tested in parallel with an eye toward better performance.

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

4. The Empirical Application of HBS to the Use Case of MN

We tested this HBS method of classification by using two Hungarian newspaper corpora: NS[5] and MN. The input data in both cases were the full text of articles as they appeared on the front page of the two newspapers. As stated above, our research design is geared toward maximizing ML precision for a virgin corpus (MN) based on a training set created out of an independent corpus (NS). Intercorpus snowballing refers to using the NS data as an input for the coding process of the virgin MN corpus. This is a feasible choice given that we remained in the domain of Hungarian newspaper front pages for the same period.

Our experiments with various feature sets (tf, tf-idf) pointed toward the relative insignificance of the chosen feature set vis-á-vis other structural choices and parameters such as the utilization of ensemble learning, bagging, and treating class imbalance (see details in the next sections and also in technical column of Table 1).

In initial tests, support vector machine (SVM) classifiers outperformed other commonly used algorithms (such as Naive Bayes, Random Forest, etc.), thus it was decided that our efforts were to be aimed at enhancing the results of an SVM-based workflow.[6] Finally, we decided to employ a technique related to (but not the same as) active learning, the snowball method, to gradually increase our training set and thereby achieve better results.

The coding of the MN corpus (34,670 articles) was based on a complex process involving both intercorpus and intracorpus snowballing (see Figure 1). We used the coded NS corpus as our training set, with the whole MN corpus acting as the virgin test set to be classified. Since running a full round of coding for all code categories we were working with would take 3–4 days on the desktop computer, we used a Spark cluster for which the same process took roughly 30 minutes long (see Pintye, Kail, and Kacsuk 2019).

During our exploratory analysis, one of the problems that stood out was related to consistently high precision but very low recall values, especially for smaller topics. This is a common problem for text classification tasks on highly imbalanced multiclass data (Kumar and Gopal 2010) especially when using SVM models, which are sensitive to the distribution of positives and negatives in the training set. The two most common ways of handling this problem are either increasing the number of positives through oversampling (including multiples of the same elements) or decreasing the number of negatives by undersampling (Lango and Stefanowski 2018).

We chose to go with the latter solution, which is also in line with the results of Kumar and Gopal (2010), who also worked on textual data. The simple rule we implemented checked whether the ratio of positives to negatives in the training set was less than 10%, if yes, then the algorithm would take a sample (without replacement) from the negatives[7] and use only those for the actual training set.

This approach of using proportional training set sizes had a dramatic positive impact on our recall values; however, at the same time, our precision started to plummet. We decided to tackle this problem by introducing a committee voting method with multiple samples being run for each topic before a conclusion is reached in relation to the final classification decision. This led to a bagging-type[8] ensemble setup, which is again a very common counterpart to the correction of the ratio of positives to negatives by sampling; especially effective when combined with undersampling (Kumar and Gopal 2010; Lango and Stefanowski 2018). We added one final component to

[5] For a detailed description of this corpus see the research note by Mészáros (2018) and other chapters on media and results in the book edited by Boda and Sebők (2018).

[6] Although the decision to work with only SVMs has proved to be fruitful for the MN corpus, future work should revisit the difference between the outputs for various classification algorithms within the framework of HBS.

[7] If the number of positives was less than 1% of the original training set size, then the sample size was equal to 10% of the original training set size, otherwise it was equal to the number of positives multiplied by 10.

[8] The reason we refer to it as a bagging-type ensemble as opposed to just bagging, is because bagging employs sampling with replacement, whereas our algorithm uses sampling without replacement. We tried both versions of the ensemble, and sampling without replacement proved to be slightly better, even if only by a very small margin.

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

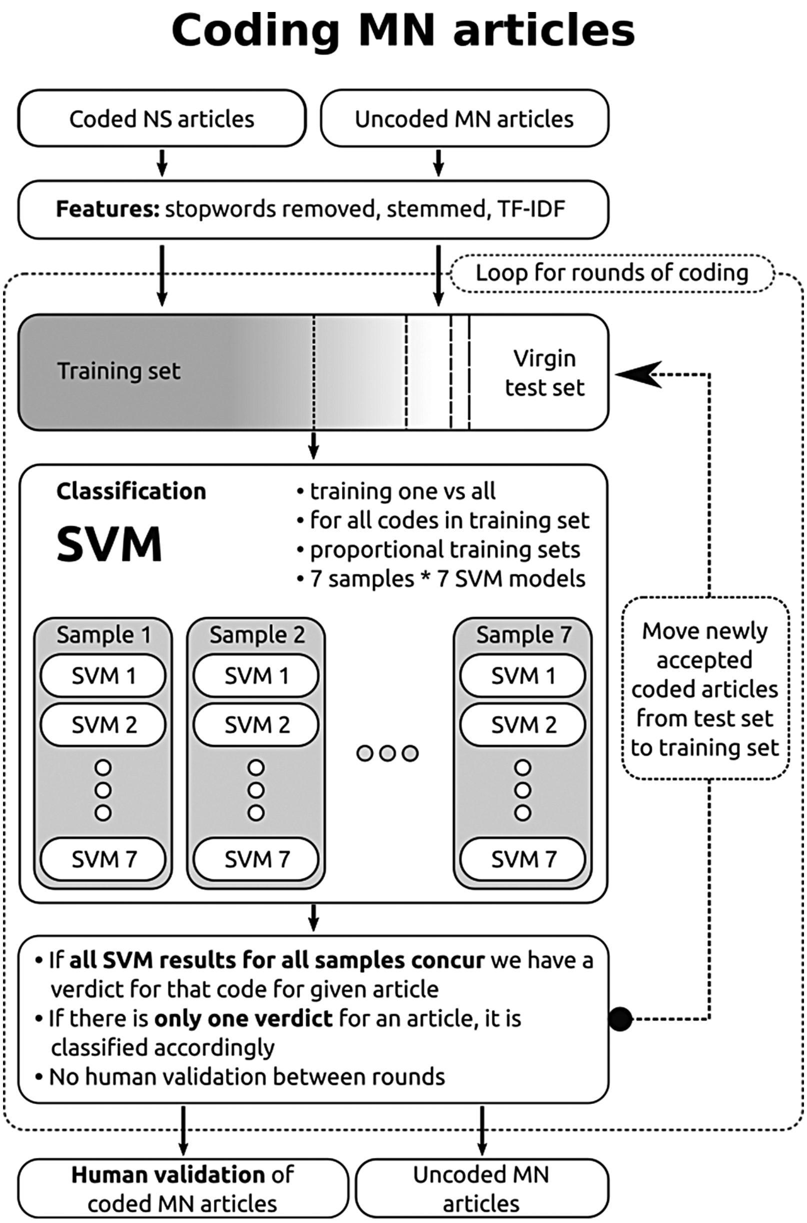
Coding MN articles
Coded NS articles
Uncoded MN articles
Features: stopwords removed, stemmed, TF-IDF
Loop for rounds of coding
Training set
Virgin test set
Classification
SVM
• training one vs all
• for all codes in training set
• proportional training sets
• 7 samples * 7 SVM models
Sample 1
SVM 1
SVM 2
SVM 7
Sample 2
SVM 1
SVM 2
SVM 7
Sample 7
SVM 1
SVM 2
SVM 7
Move newly accepted coded articles from test set to training set
• If all SVM results for all samples concur we have a verdict for that code for given article
• If there is only one verdict for an article, it is classified accordingly
• No human validation between rounds
Human validation of coded MN articles
Uncoded MN articles

Figure 1. HBS in action: A coding process for a virgin corpus with in-process ensemble voting.

this setup to account for another element of randomness in the process, the way the SVM model would reach its equilibrium. We decided to also repeat the SVM training and classification for each sample multiple times.

Based on our experiments, we settled on seven samples with seven iterations of SVM training and classification (see Figure 2). For each code category, if all 49 SVM results concurred in classifying an article as belonging to that category, the votes of the SVMs would become a verdict. And for each article that had only one verdict at the end of a coding round, the verdict would become the code category assigned to the article.

Newly classified articles were automatically added to the training set to be used for the next round, and by the end of round 3, the number of coded MN articles was 20,194. At this point, a random sample was drawn from each code category based on the number of classified articles. For code categories with less than a hundred articles, all were selected. For code categories with a higher number of articles, a sample size was selected that would allow for a −5%/+5% confidence

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

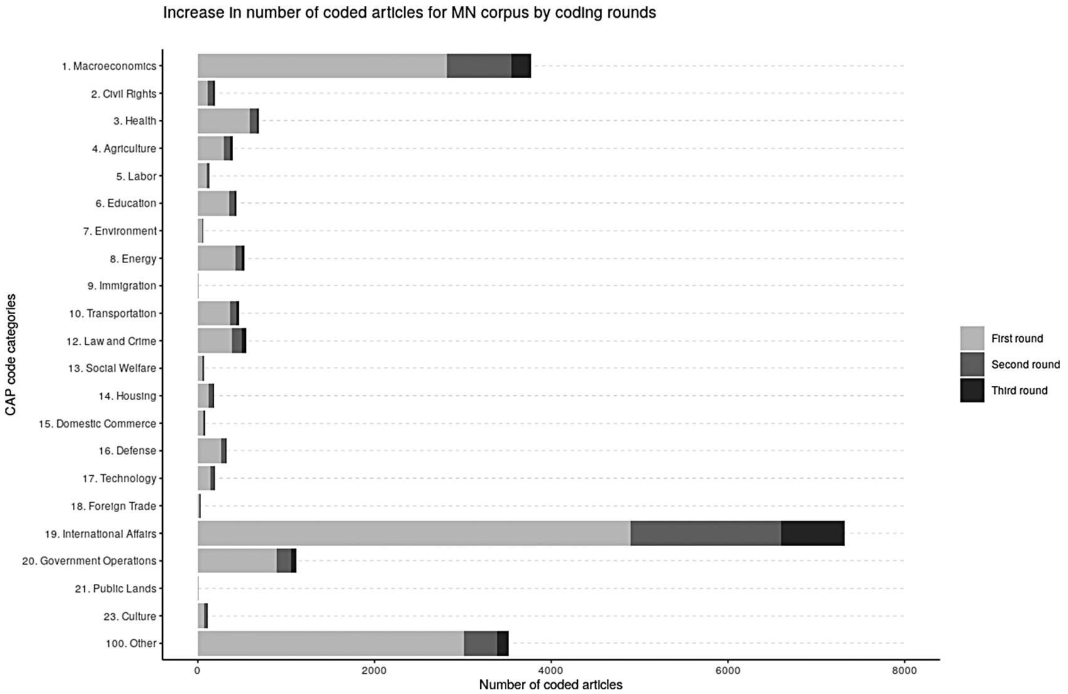

Figure 2. The increase of the coded articles for MN by coding rounds.

interval with 95% confidence (we return to this point with Figure 4 below). These samples were then validated by human experts.

5. Results for the Hungarian Media Corpus Use Case of the HBS Process

As we described in the previous section, we tested our HBS process of ML-based classification on a use case of Hungarian media corpora. In this section, we present the preliminary results from a test run of HBS in Apache Spark on a cluster of workers.

Our aim was to classify front-page newspaper articles of MN without resorting to any type of initial human coding. Furthermore, we set our benchmark for precision at the levels usually associated with coding processes based on double-blind human coding (80% precision for intercoder reliability of labels). Figure 2 shows the increase in the number of coded articles for each major topic code for the three rounds of our test run.

What is clear from Figure 2 is that, for most codes, each round added new coded articles. The additions of the second and third rounds show a declining marginal rate. Further tests are required from a validation perspective to tease out the optimum number of rounds before capping the process. The optimum number would equal the last round which still adds new articles to the coded set.

Since the aim of our process was to arrive at a valid classification of articles, the key metric to consider is precision. Figure 3 provides an overview of our results by topic codes. This figure shows results that are significantly higher than those reported by similar previous studies. For most major topics, the relevant rate of precision is over 80%, which is in line with results from "gold-standard" human coding processes. The confidence interval of our precision estimations surpasses 75% for the majority of topic codes which speak to the overall validity of this case of the HBS process. A closer look at the negative outliers also offers promising code-specific solutions for increasing precision (see Section 7).

The size of individual major topics in terms of their estimated share of the total shows that CAP coding is also an important aspect of such coding processes. Figure 4 indicates a loose connection between code size and precision, which indicates that the HBS is not biased toward small or

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

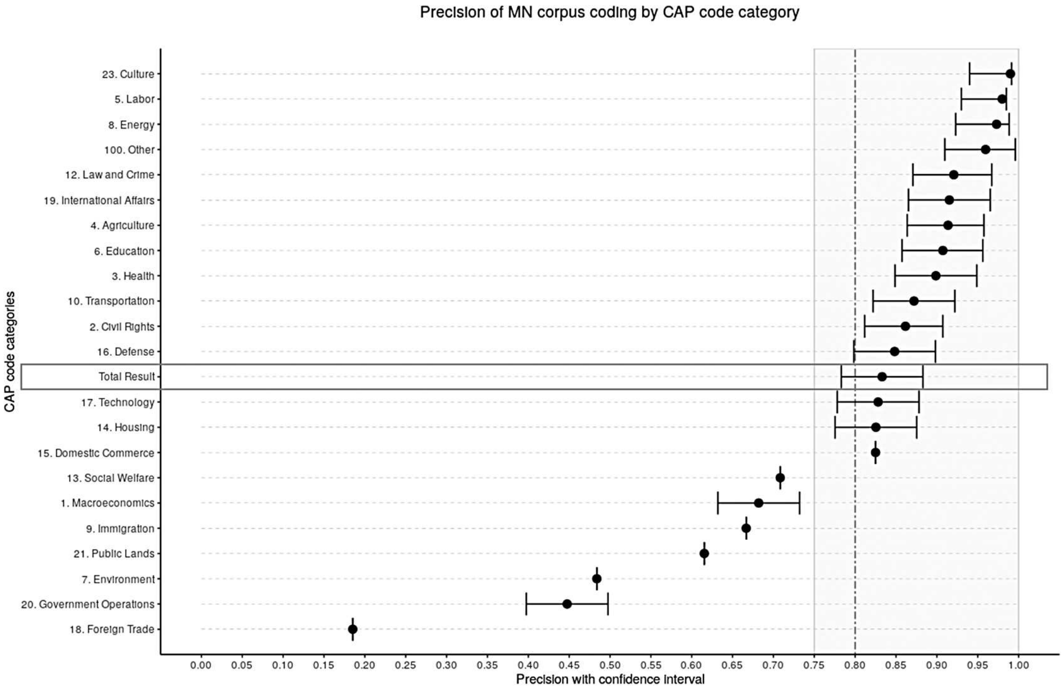

Figure 3. Precision of MN corpus coding by CAP major topic.

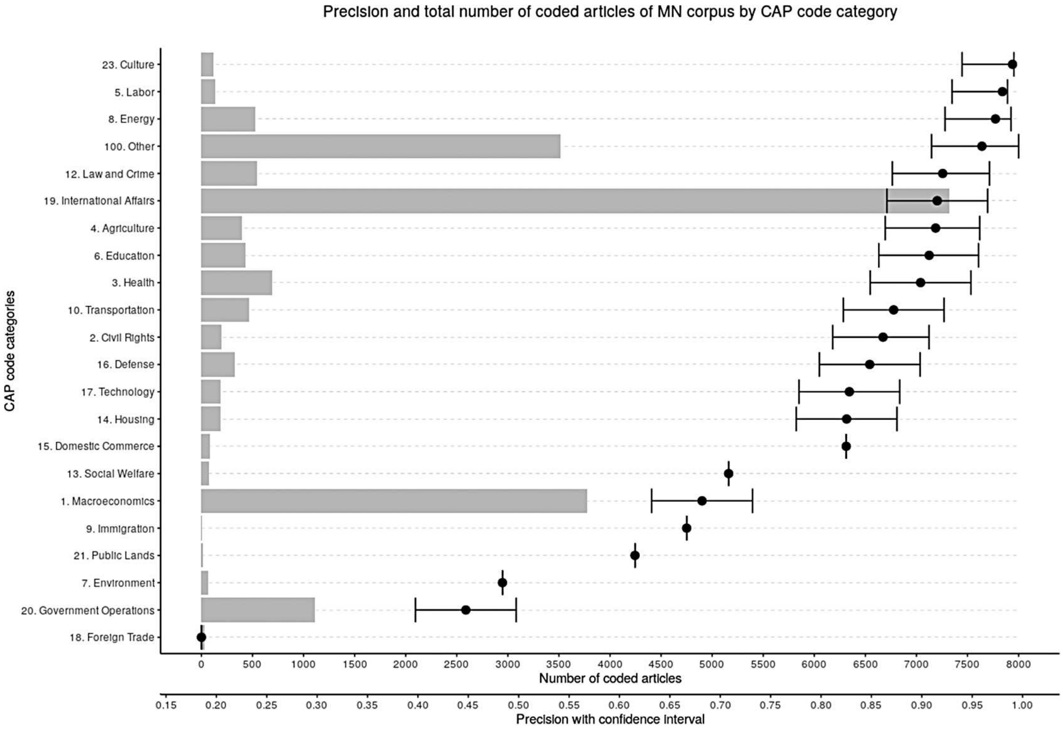

Figure 4. Precision of MN corpus coding by CAP major topic.

big code categories. Finally, we calculated estimates for code-specific total percentage of texts classified (our proxy for recall) scores based on the distributional characteristics of the coded NS set (see Figure S1 in Supplementary Materials). While we can make no statements with absolute certainty regarding the relevance of NS topic-recalls for MN topics, the scores related to more populous categories (such as International affairs or Macroeconomics) may serve the fine tuning of our case-specific HBS process.

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

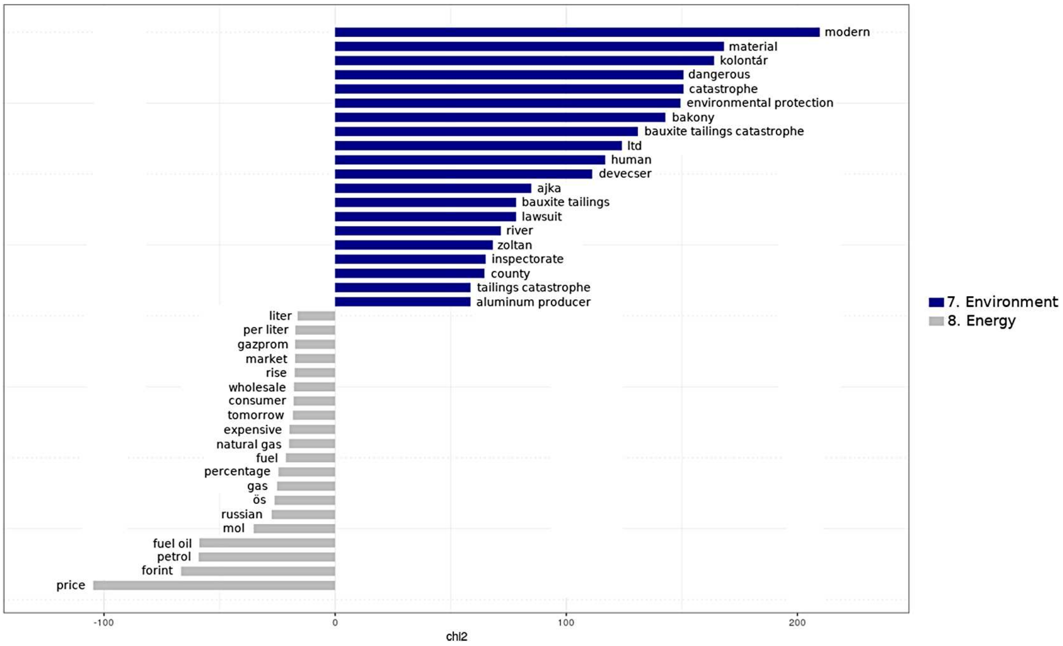

Figure 5. Representative words of the boundary topics of environment and energy.

6. Discussion

According to our results for many topic codes of a virgin corpus, the HBS process yielded gold-standard-rate precision values with very limited human involvement. In this, our analysis partially fulfilled the goal set out as the central research problem for this article. However, for many code categories, such as the environment, macroeconomics, or government operations, precisions rate lagged behind what is acceptable in most projects. One of the reasons for subpar precision may be related to the issue of *boundary topics*.[9]

The delimitation is more difficult between specific code-pairs than for others. This is both due to the inherent overlap between certain policy areas (such as taxation and social policy, where the latter is often implemented through the former through tax credits or other forms of reductions) and also their vocabulary. This grey zone between topics also produces problems specific to the CAP research agenda as articles which cover multiple topics can only be assigned—as per the rules of coding—to a single class. One such example would be a piece which covers both international affairs and international trade with respect to a bilateral country relation. Finally, in some cases, the CAP codebook itself does not provide clear instructions as to which topic a rare issue would belong to (such instructions are reserved for more prevalent themes).

Based on our analysis, topic-pairs which would seemingly be linked by a thematic boundary do not always pose a problem for classification. This may be related to the fact that, more often than not, certain topics are covered from a specific angle in newspapers. Take the example of the two classes of environment and energy which are clearly related in news reports (think oil spills). Figure 5 presents the representative words (in translation from Hungarian) for both classes by using the "keyness" score of the quanteda package in R. This is a metric for features that occur differentially across different categories.

What is clear from this figure is that the coverage of environment in MN in the given period is strongly focused on a single event: the tailings catastrophe in Devecser, near the town of Ajka. Indirectly, this is also indicative of the relatively low share of environmental topics on the front page of MN given that the overwhelming majority of words associated with this topic are related

[9] We thank the anonymous reviewer for the comments regarding boundary issues.

to this event. This also holds for the other topics in question. Energy is mostly associated with a single issue as well: the price of gasoline. In sum, this seemingly related topic-pair is in fact fairly distinct in the articles that got categorized through the HBS process.

This is not to say that boundary topics do not cause a problem for HBS classification, only that this impact is limited. Let us consider two segments from two different articles (the first from 2003, the second from 2008).

(1) The Constitutional Court will consider the "hospital law." The Court will start hearing the case of the hospital law, which was initiated by multiple parties and professional bodies.

(2) Decision on hospitals. This morning, the Constitutional Court will hand down a decision on the hospital law. This piece of legislation concerns the transformation of the health care system and multiple organizations petitioned the CC for a hearing on the regulation.

These two segments (and the entire text of the articles in question) were related to the same issue: the hearing of the hospital law case by the Constitutional Court. Nevertheless, the first segment was categorized by the HBS process to 3-Health care, while the latter to 20-Government operations. This is where even trained human coders may produce unreliable results as the themes of health care and the functioning of the judicial branch of government are both present in the same article. For each similar case, we would need an in-depth analysis of the word counts associated with multiple categories. This would allow for inserting dictionary-based rules into the HBS process with a view toward enhancing reliability and, therefore, better precision.

Another option for clearing up ambiguities is to create more coherent topics in the codebook. In fact, CAP codebooks utilize more refined minor topics within the general major topic. A case in point is the sprawling category of macroeconomics which is basically a collection of multiple, distinct subtopics such as taxation or government debt. Therefore, our initial expectation was that it would return lower than average results (which, in fact, it did). In principle, and depending on project features, the separation of minor topics from certain major topics as standalone categories may produce superior results, but this hunch would have to be tested in future work.

Besides handling the complexities of boundary topics, another way to improve the precision scores of HBS within a given research design is by testing alternative setups and comparing their performance. First, our results show that in future iterations of the HBS workflow, it would make sense to use a more sophisticated active learning approach.[10] Second, our data confirms the superior performance of SVMs vis-á-vis the Naive Bayes algorithm, especially if we consider recall beyond our key metric of precision. Third, our simulations highlight the effectiveness of using multiple iterations of sample-algorithm pairs. Here, we found that optimum performance for precision comes at the expense of recall. In future work, simulations based on different feature sets (tf vs. tf-idf vs. word embeddings) as well as grid searches of parameters could contribute to the fine-tuning of HBS to project-specific needs.

As a final note, it is important to stress that any results of such experiments will be related to the effectiveness of individual aspects of the complex HBS solution (see Table 1). Furthermore, parameter choice and other methodological decisions should be tailored to the needs of the given project, and simulations should also be run for the adequate corpus and setup for the research design in question. HBS is a framework for solving imbalanced multiclass classification problems, but its specific form should always be dependent on the context in which it is applied.

7. Conclusion

This article presented a new workflow solution for content analysis in comparative politics. We combined a quantitative text-mining approach with supervised learning methods and limited

[10] We thank the anonymous reviewer for extensive comments on how to utilize active learning in the HBS setting.

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

human coding in order to classify the front-page articles of a Hungarian daily newspaper. Our precision results for this virgin corpus surpassed those of similar available studies and are competitive with the "gold standard" of double-blind human coding. These exceptionally high-precision scores came at the expense of an estimated recall (or the number of coded articles as a share of the total number) of below 60% for the virgin dataset.

In light of these positive results, our process offers a fourfold contribution to the literature. First, we assign a key but limited role for human coding in our workflow. This hybrid, or semi-automated, approach brought added value vis-á-vis fully automated approaches. Second, the snowball method of increasing training set size allows for leveraging a relatively small initial training set for the computer-assisted creation of a full-scale coded corpus. This latter can then be applied to a within-domain virgin corpus, in a process called intercorpus snowballing.

Third, our ML process design is built on binary coding (that is, one vs. all) as opposed to the multiclass setup which in our tests, all other things being equal, significantly improved our precision results. Finally, the proposed technical environment for implementing the HBS process offers a solution for projects with limited budgets but which have an access to commercial or academic cloud infrastructure.

As we stated in the introduction, the aim of our project was to match the performance of the gold standard of double-blind human coding when it comes to the multiclass classification of policy topics of newspaper articles. Our performance metric of choice was precision as the automated and valid coding of even a half of the items in a Big Data size corpus readily creates direct and tangible benefits for large-scale projects. While the application of HBS to a Hungarian language newspaper corpus offers promising results, a number of options could still be exploited beyond optimizing setups and parameters.

One such option is adding a "finishing" step to the coding process that addresses project-specific needs. The hybrid approach of human and ML-based coding can be extended to dictionary-based methods as well. Using regular expressions could fill a gap in the process by correcting the systematic errors of ML. In our case, the relatively inefficient coding for the major topic of environment is partly caused by associating the word "design" with the topic by our algorithms.

Future work could also focus on moving beyond the low-hanging fruits and applying a disjointed, topic-specific approach of HBS. In this respect, neural networks may offer superior solutions vis-á-vis standard ML algorithms. These could be utilized, for instance, in identifying errors in the underlying human-coded training sets which lower the quality of intracorpus and intercorpus snowballing and, therefore, have a significant effect on end results. A potential solution for low-precision categories would be to adjust the weights of words which are discriminative for each topic (see Figure 5).[11] Another potential solution would be to relax the constraint related to the assignment of a single code to each item.

Finally, and in reference to a similar project by Loftis and Mortensen (2020), we believe that HBS has the potential to solve multiclass classification problems with unbalanced classes beyond the domain of our current corpora and the research task of the CAP. It is our expectation that the strategic use of human coding for creating limited training sets and for performing ex post validation will provide a competitive edge for our hybrid process vis-á-vis pure ML-based approaches, even if these latter utilize ensemble coding relying on different ML algorithms (which may not be a crucial element of success). The tactics of snowballing and the simplification of the multiclass problem to a binary one is not specific to the CAP challenge either. These features of HBS position this approach well for future tests on alternative corpora or even cross-domain classification tasks.

[11] We thank the anonymous reviewer for useful comments regarding this issue.

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press

Acknowledgments

We thank Evelin Mészáros, Zsanett Pokornyi, and Csaba Molnár for their excellent research assistance and the Hungarian Comparative Agendas Project (cap.tk.mta.hu) for providing the underlying textual databases. We are grateful to Amber Boydstun and Ákos Máté for their help. We acknowledge the role of the MTA Cloud (https://cloud.mta.hu/) infrastructure and the Occopus cluster deployment tool (http://occopus.lpds.sztaki.hu/) and express our gratitude to the team—and especially Enikő Nagy and István Pintye—of the Laboratory of Parallel and Distributed Systems at ELKH SZTAKI. We are thankful for the participants of APSA, EPSA, the Comparative Agendas Annual Conference, the Zurich Text as Data Workshop and the POLTEXT conference for their stimulating feedback. We thank the three anonymous reviewers for their insightful comments.

Funding

This work was supported by the National Research, Development and Innovation Office (NKFIH) grant (ÁJP-K-109303, FK-123907); the Incubator and Flagship grants of ELKH TK, and the Bolyai Research Grant for Miklós Sebők.

Data Availability Statement

The replication materials for this paper can be found at Sebők and Kacsuk (2020).

References

Albaugh, Q., J. Sevenans, S. Soroka, and P. J. Loewen. 2013. "The Automated Coding of Policy Agendas: A Dictionary-Based Approach." In *6th Annual Comparative Agendas Project (CAP) Conference*, *Antwerp, Belgium*.

Albaugh, Q., S. Soroka, J. Joly, P. Loewen, J. Sevenans, and S. Walgrave. 2014. "Comparing and Combining Machine Learning and Dictionary-Based Approaches to Topic Coding." In *7th Annual Comparative Agendas Project (CAP) Conference*, *Konstanz, Germany*.

Baumgartner, F. R., C. Breunig, and E. Grossman, eds. 2019. *Comparative Policy Agendas: Theory, Tools, Data*. Oxford: Oxford University Press.

Baumgartner, F. R., C. Green-Pedersen, and B. D. Jones. 2013. *Comparative Studies of Policy Agendas*. Oxon: Routledge.

Bevan, S. 2019. "Gone Fishing: The Creation of the Comparative Agendas Project Master Codebook." In *Comparative Policy Agendas: Theory, Tools, Data*, edited by F. R. Baumgartner, C. Breunig, and E. Grossman, vol. 17. Oxford: Oxford University Press.

Boda, Z., and M. Sebők, eds. 2018. *A magyar közpolitikai napirend: Elméleti alapok, empirikus eredmények (The Hungarian Policy Agenda: Theoretical Foundations and Empirical Results)*. Budapest: MTA TK PTI.

Bonica, A. 2018. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning." *American Journal of Political Science* 62(4):830–848.

Boydstun, A. E. 2013. *Making the News: Politics, the Media, and Agenda Setting*. Chicago: University of Chicago Press.

Breeman, G. E., H. Then, J. Kleinnijenhuis, W. van Atteveldt, and A. Timmermans. 2009. "Strategies for Improving Semi-Automated Topic Classification of Media and Parliamentary Documents." Paper prepared for the 2nd Annual Comparative Policy Agendas (CAP) Conference, The Hague, The Netherlands.

Burscher, B., R. Vliegenthart, and C. H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize Across Contexts?" *The ANNALS of the American Academy of Political and Social Science* 659(1):122–131.

Colaresi, M., and Z. Mahmood. 2017. "Do the Robot: Lessons from Machine Learning to Improve Conflict Forecasting." *Journal of Peace Research* 54(2):193–214.

Collingwood, L., and J. Wilkerson. 2012. "Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods." *Journal of Information Technology & Politics* 9(3):298–318.

Flaounas, I., et al . 2013. "Research Methods in the Age of Digital Journalism: Massive-Scale Automated Analysis of News-Content—Topics, Style and Gender." *Digital Journalism* 1(1):102–116.

Grimmer, J. 2015. "We are all Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48(1):80–83.

Hillard, D., S. Purpura, and J. Wilkerson. 2008. "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research." *Journal of Information Technology & Politics* 4(4):31–46.

Jurka, T. P. 2012. "Maxent: An R Package for Low-Memory Multinomial Logistic Regression with Support for Semi-Automated Text classification." *The R Journal* 4(1):56–59.

Karan, M., J. Šnajder, D. Sirinic, and G. Glavaš. 2016. “Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts.” In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Kumar, M. A., and M. Gopal. 2010. “A Comparison Study on Multiple Binary-Class SVM Methods for Unilabel Text Categorization.” *Pattern Recognition Letters* 31(11):1437–1444.

Lango, M., and J. Stefanowski. 2018. “Multi-Class and Feature Selection Extensions of Roughly Balanced Bagging for Imbalanced Data.” *Journal of Intelligent Information Systems* 50(1):97–127.

Levin, I., J. Pomares, and R. M. Alvarez. 2016. “Using Machine Learning Algorithms to Detect Election Fraud.” In *Computational Social Science: Discovery and Prediction (Analytical Methods for Social Research)*, edited by R. Alvarez. Cambridge: Cambridge University Press.

Loftis, M. W., and P. B. Mortensen. 2020. “Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents.” *Policy Studies Journal* 48(1):184–206.

Lucas, C., R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley. 2015. “Computer-Assisted Text Analysis for Comparative Politics.” *Political Analysis* 23(2):254–277.

Mészáros, E. 2018. “A média közpolitikai napirendjének felügyelt gépi tanulással történő elemzése (The supervised machine learning analysis of the policy agenda in the media).” In *A magyar közpolitikai napirend: Elméleti alapok, empirikus eredmények (The Hungarian Policy Agendas: Theoretical Foundations and Empirical Results)*, edited by Z. Boda and M. Sebők, 31–52. Budapest: MTA TK PTI.

Mikhaylov, S., M. Laver, and K. Benoit. 2008. “Coder Reliability and Misclassification in Comparative Manifesto Project Codings.” In *66th MPSA Annual National Conference*.

Pintye, I., E. Kail, and P. Kacsuk. 2019. “Big Data and Machine Learning Framework for Clouds and its Usage for Text Classification.” In *IWSG’2019*, *Ljubljana*.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2006. An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th–108th US Senate." *Midwest Political Science Association Meeting*.

Sebők, M., and Z. Kacsuk. 2020. “Replication Data for: The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach.” https://doi.org/10.7910/DVN/CFHOCU, Harvard Dataverse, V1.

Volkens, A., J. Bara, and I. Budge. 2009. “Data Quality in Content Analysis. The Case of the Comparative Manifestos Project.” *Historical Social Research/Historische Sozialforschung* 1:234–251.

Young, L., and S. Soroka. 2012. “Affective News: The Automated Coding of Sentiment in Political Texts.” *Political Communication* 29(2):205–231.

https://doi.org/10.1017/pan.2020.27 Published online by Cambridge University Press