Baumgartner *et al.*
Communication and Change (2026) 2:1
https://doi.org/10.1007/s44382-026-00021-8

Communication and Change

RESEARCH Open Access

Agenda-setting studies in public policy: Origins, development, and new possibilities for coding in the age of AI

Frank Baumgartner[1*], Shaun Bevan[2] and Miklós Sebők[3]

*Correspondence:
frankb@unc.edu

[1] Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill, United States
[2] University of Edinburgh, Edinburgh, United Kingdom
[3] poltextLAB, ELTE Centre for Social Sciences, Budapest, Hungary

Abstract

We assess the impact of the 1972 work of McCombs and Shaw in communications as well as in political science and describe the literature that derived from it, with a focus on methodological innovations. We explain that transformations in the field made possible by increasingly powerful computer technology, the analysis of text as data, automated classification systems, and the development of artificial intelligence (AI) technologies. We describe possible future directions of the field given these developments.

Keywords: Agenda-setting, Public policy, Mass communications, Automatic classification, Text-as-data, Artificial intelligence

Introduction

Maxwell McCombs and Donald Shaw were methodological innovators as well as theoretical ones. Further, and coincidentally, they were institutional colleagues of Frank Baumgartner (at the University of North Carolina) and Bryan D. Jones (at the University of Texas).[1] In this article, we focus on the links between the work of these sets of collaborators; the cross-pollinations among studies of agenda-setting in the fields of communications and political science or public policy; the methodological changes that have affected the fields since the publication of McCombs & Shaw's, 1972 article; and possible future directions in the age of Artificial Intelligence (AI). Wherever the field of agenda-setting may go, it has undergone massive shifts in the fifty-plus years since McCombs and Shaw set the agenda for the study of agenda-setting.

Given our own expertise, compared to that of others in this Special Issue, we focus on the development of agenda-setting studies in the wake of McCombs and Shaw with a focus on public policy, not mass communication. (McCombs alone and with collaborators provided updates on developments in the field; see McCombs, 1993, 2004, McCombs & Weaver, 1985, and McCombs & Zhu, 1995.) We focus on methodological

[1] Baumgartner and Shaw served together on the dissertation committee of Milad Minooie (who received his PhD from UNC in 2018); Jones and McCombs, both originally from Alabama, ended their careers at the University of Texas.

Springer

trends, in particular the trend toward projects of larger empirical scope over time. First, we discuss the connections and common origins of the communications and political science versions of agenda-setting research. Next, we describe the history and growth of the US-based Policy Agendas Project and its eventual expansion into a network of dozens of political systems internationally through the growth of the Comparative Agendas Project. This process required the development of a consistent, robust, comprehensive, and flexible classification system for policy topics. The methodological challenge in this was substantial, and our experience with it helped develop machine-learning programs that were critical in developing text-as-data techniques and methods that are key to AI-related tools and methods. We describe our collective experiences developing these methods and the research questions they have allowed us to address, building on the work of earlier scholars, including McCombs and Shaw, but extending their insights based on much larger empirical points of reference. Finally, we discuss the possible futures of these areas of research in the future, with a focus on the possibilities of AI.

Agenda-setting studies in public policy and mass communication

The political science and mass communication literatures in agenda-setting developed more in common than separately. The first footnote in McCombs & Shaw's, 1972 article is to Berelson et al.'s 1954 book, *Voting*. Both fields were concerned with how voters got their information, the rise of television, and the nationalization of the mass-media environment in the post-war period. Few methodological differences distinguished scholars in political science and communications, as both were interested in where voters got their information.

Some differences emerged among specialists, of course, as authors such as Schattschneider (1960) and Bachrach and Baratz (1962) developed ideas of how the political agenda is set, how conflicts are constructed, and how some issues are kept off the political agenda; these are not listed in the footnotes of McCombs and Shaw (1972), indicating the beginning of a division among specialists in the two fields. Roger Cobb and Charles Elder's (1972) book, *Participation in American Politics: The Dynamics of Agenda-Building*, was also focused on the construction of the political agendas through elite-level competition. It was published in the same year as McCombs and Shaw's foundational article, but addressed the agenda-setting question from a different perspective. As Wolfe, Jones et al. (2013) discussed, the two fields later developed in quite different ways as the two disciplines addressed overlapping questions with different methods, concerns, and increasingly separate audiences. Indeed, the main divide may have been not between political science and mass communications scholars, but between those interested in how voters gather information (a common point of interest for behavioral political scientists as well as for mass communication scholars) and a new group of political science and public policy scholars who wanted to know where ideas came from, how the political agenda was set, and why some issues were more likely to be on the governmental agenda than others.

Jack L. Walker Jr. deserves some of the credit for moving the field of agenda-setting forward in political science in his controversial 1966 article "A Critique of the Elitist Theory of Democracy" in which he took aim at the most dominant perspective on US politics at the time, pluralism. He pointed out that that theory's focus on competition among elite actors seemed content with rank-and-file citizens playing little role,

other than choosing occasionally among competing elites. He pushed the literature forward with his 1977 article "Setting the Agenda in the US Senate," explicitly focusing on how the political agenda of the US Senate is set, and which Senators were more active in setting it. By that point, the communications and public policy literatures were becoming quite distinct.

Social movement scholars such as John McCarthy and Mayer Zald (1978) developed a parallel literature in the field of collective action and social movements putting an emphasis on the abilities of social movement organizations such as the women's movement, civil rights groups, and environmental activists to set the national agenda and to bringing new issues to the fore.

John Kingdon (1984) further reinforced the multiple sources of the public agenda with his influential study of how policymakers in Washington know what to focus on by looking at what their contemporaries are doing. Frank Baumgartner and Bryan Jones (1993) built on this work but used a different methodological approach, allowing them to study the development of the public agenda over many decades, rather than taking a snapshot of it at a single point in time.

One important development in the history and likely future of studies of agenda-setting is the scope of the empirical work. We have come a long way since the hand-coding, use of scissors and glue, literally to cut our articles from printed sources to compile content analysis files, and transcription of data to 80-character IBM punch cards. McCombs and Shaw (1972) interviewed 100 undecided voters across five voting precincts in Chapel Hill, NC between September 18 and October 6, 1968, asking them ""What are you most concerned about these days? That is, regardless of what politicians say, what are the two or three main things which you think the government should concentrate on doing something about?" (McCombs & Shaw, 1972, 178). Their research team also conducted a content analysis of local and national newspapers, magazines, and the evening television news. The scope of the work was typical for studies of the time (for example see the pathbreaking work of James Prothrow and Charles Grigg (1960), who interviewed 244 registered voters in Ann Arbor, Michigan and Tallahassee, Florida; national level survey research was just getting underway (see Campbell et al. 1960) and these authors ruled it out "simply on the ground of costs" (Prothro and Grigg 1960, 284).

John Kingdon (1984) spent several years in Washington, or travelling frequently to Washington, to conduct 247 interviews with policymakers over the period of 1976 to 1979 (Kingdon, 1984, 220). Frank Baumgartner and Bryan Jones (1993) developed methods of content analysis of media coverage, congressional hearing abstracts, and other printed documents in order to trace the policy histories and agenda status of nine public policy issues over several decades. This work began a transition from interviews to content analysis of publicly available written sources. McCombs and Shaw had done both, as had Kingdon. Today, interviews with political leaders are relatively rare; many on-line platforms allow surveys (and survey experiments) with nationally representative samples; cross-national research is more common; and it is common to base studies on millions of digitized government or media documents (see for example Baumgartner, Breunig, and Grossman, eds. 2019). We will discuss below what the future might bring.

Development of the comparative agendas project

Immediately on completion of *Agendas and Instability in American Politics* (1993), Frank Baumgartner and Bryan Jones set out to create a research infrastructure that did not previously exist. They applied for grants from the National Science Foundation, eventually (not immediately!) leading to some success and to the creation of the databases that now form the basis of the Policy Agendas Project; for example, a record of the date, policy focus, and committee associated with every congressional hearing since World War Two (see Jones & Baumgartner, 2012 for a description of the development of this theoretical and methodological tradition). Their 2005 *Politics of Attention* was the first book to make use of such large data infrastructure for the study of agenda setting in public policy.

Shortly after the publication of this book, scholars in Europe began to replicate the databases and to question the limitations of a theory that appeared particularly if not peculiarly American; the Comparative Agendas Project was born (see Baumgartner, Breunig, and Grossman, eds. 2019 for information on these developments). A key issue in expanding what had been a US-only database to one that could be generalized was to understand the full range of policy activities of modern governments. The first expansion, to Denmark through the dedicated work of Christoffer Green-Pedersen, led to a number of discoveries: Denmark engages in a lot more oversight of the fishing industry than does the US government; it has a royal family and an official church, topics that were not in the American classification system; on the other hand its government spends no time managing millions of acres of public lands in the arid west, and when it is involved in housing policy, it does so directly rather than through such mechanisms as secondary markets for mortgages through such agencies as Freddie Mac and Fannie Mae. So there were some differences when the agendas project moved from one country to the next. On the other hand, the vast majority of the topics of attention of the US government had a direct counterpart in the Danish one. Each country spends time paying attention to the health of the economy and the inflation rate; each supports agriculture in various ways; each is concerned with gender equality; each has programs to ensure water and air quality; each has to build and maintain roads and railroads; each has a hospital network; each has a defense and foreign policy establishment. Today, the Comparative Agendas Project brings together scholars from more than two dozen countries all using a common classification system. The development of this common classification system was fundamental to our ability to compare across countries and over time. Its history also sheds light on the methodological innovations that have affected the study of agenda setting and on where we may be going in the age of AI.

The methodological challenge of a common classification system

As the CAP expanded to many more projects, several other differences between cases appeared and at first those differences were addressed by each project individually without comparison in mind. This needed to be addressed. The following section briefly explains the process of reconciling differences in coding by focusing on the challenges of creating a common classification system, the CAP Master Codebook.

It concludes by exploring how the decisions necessary to meet those challenges have both helped and complicated the use of computer assisted tools as well as AI in the coding process.

The beginning

Grouping policy areas like agriculture and transportation is based on context and knowledge. The same is true for AI and for humans with differences in available information and context leading to considerable variation in how policy areas are conceived. This represents the main challenge for comparative research and classification as no two contexts nor knowledge-bases are the same.

The Policy Agendas Project started with the coding of US congressional hearings which broadly represent all the work of the US national government leading to the creation of a robust coding system (see Jones & Baumgartner, 2005). [2] Despite starting in the US, a focus on comparison between datasets and over time was at the core of creating the coding system. As new datasets were added, changes were made to the coding system, but they always followed a simple rule, that existing codes may never be combined, but that new codes can be created to match new policy areas or to further distinguish existing codes.

CAP projects aimed to uphold a view towards comparison, but they needed help. The first piece of help was inherent to the design of coding system which is limited by design to focus on "policy not targets". A target for a policy may be a financial crisis, climate change, or international threats. While many policies aimed at one of these targets may be in specific policy areas, like the environment for climate change, many other policies will be aimed at the target. For instance, policies focused on the target of climate change may include insulation schemes (housing), more efficient roadways or vehicles (transportation), new rules for the use of fertilizers (agriculture) and so on. In this way, CAP coding is always focused on the policy alone, not the reason or justification for the policy. While perhaps the most common criticism of CAP data with many scholars more interested in targets (e.g. Dowding et al., 2016), this limited approach kept the coding focused and much less open to interpretation. Coders did not have to attempt to determine the reason for the policy, or what problem the policy was designed to solve; they needed only to classify it as housing, transportation, or agriculture.

Coding through compromise

Even with a limited focus for coding and following the rules for adding new codes, some variation among national project codebooks inevitably developed. Germany had to address reunification, Spain the transition to democracy, the United Kingdom its monarchy, and so on. It was not always logical or practical to address these often unique policy areas through existing codes when working within a project. As a result, by July 2014 there were more than 450 subtopics across 15 projects, a doubling of the 225 subtopics in the original US codebook. Projects had their rules for codes, but how those

[2] For a more complete retelling of the history as well as additional details like the skipped master topic numbers see Bevan, 2019.

codes worked across projects was unclear and a common comparative coding system was needed to keep the promise of the CAP, making it comparative.

A compromise was therefore needed, and, in the end, this would take the form of the Master Codebook that prioritized comparability over level of detail. While the majority of subtopics introduced by projects offered slight variations on existing codes, how all this fit together required a shift back to basics and in a sense the sorting or coding of existing codes into common conceptual bins. To accomplish this, projects created an English language codebook, held face to face meetings with the director of the Master Codebook, Shaun Bevan, and completed a common coding exercise aimed at the most variable policy areas. This mixture of assessments of existing coding, codebooks, and what could best be described as interview data with project leads and researchers helped create a clear picture of what subtopics fit into a larger group, or least-common-denominator policy. Several drafts of the Master Codebook were created until a full version was presented and discussed with representatives from nearly every project. Following some minor revisions as well as clarifications, the Master Codebook with 21 major topics and a slimmer 213 subtopics was adopted in 2014.

The devil in the details

While intense, it was generally a straightforward procedure to create the Master Codebook; still, the work had to clear several hurdles. Many common and in context easy to understand issues like culture and fishing led to complex debates as contextually the importance and the scale of policy in these and other areas varied immensely based on project, data type, and time period.

While there are many examples, immigration is probably both the most important and the timeliest (given that we are drafting this article this paper in the fall of 2025). With most of the data in the CAP focusing on the twentieth and twenty-first centuries, immigration has almost always been a policy area focused on by national governments. That said, it is easy to forget that the often-high levels of public concern over immigration as well as the intense politics of the issue have not always been the case. When the US project started, immigration was broadly seen as a labor issue, with concerns over illegal workers as well as the need for seasonal agricultural workers. As the CAP got started years later, two other understandings emerged, one seeing immigration as an issue of civil rights, and one perceiving immigration as a policy area unto itself in part due to the role and importance of the issue in the formation of the European Union (Guiraudon, 2000). The Master Codebook treats immigration as its own policy area, and comparatively it certainly is just that. While often closely associated with other core policy areas based on context such labor in the US and civil rights in countries like the UK both during the twentieth century, the context of immigration has clearly shifted in those cases in the twenty-first century.

Ultimately sorting through these differences in context and understandings led to a Master Codebook that did not look exactly like the codebook from any one project. Instead, it offered a comparative understanding of policy, and this aspect still sometimes eludes users of the data and even project members. It is easy to forget that it is valid for a code to have very few or even no usage in a particular context, such as various defense subtopics in Switzerland, and a high level of usage in another context, such as

those same subtopics in the US. Fishery management is a large part of the Danish policy agenda. Appointing bishops in the official national church or approving the budget for the royal family are the object of some activity in certain countries but are absent in others. Of course, subtopics that are rarely used in any national context are of little use analytically, are hard for human coders to learn, and generally even harder for automated classifiers to master. But just because they are rarely used in one place or context does not mean that they will be equally rare in another place.

Separating domestic policy

Mentions of policy in other countries without the involvement of the country being researched most often received a regional subtopic code. These codes were used to separate domestic from foreign matters. In some cases, projects instead used dummy variables to indicate a focus on another country or region and otherwise coded the policy content. This also allowed for the same separation but additionally coded the foreign focused item for policy. Unfortunately, projects had different understandings of the world, breaking up regions in different ways and with different rules on recognizing the legitimacy of nations. Therefore, when compared across all projects only the separation of foreign and domestic was reliably coded. Additionally, separating the data in this way broke a core rule of the codebook, namely coding on policy. This has led to mistakes by human coders but has proven far more difficult for computer tools which have generally been designed to only focus on policy and not how to recognize this exception to the rule.

Coding and the introduction of computer assistance

Most of the data discussed to build the Master Codebook was coded by hand, by expertly trained coders following the rules of project codebooks and with leaders holding discussions on how to address unique or troublesome cases. Machine learning, specifically supervised learning was introduced to the CAP in the late noughts (e.g. Hillard et al., 2008) and built upon through several tools targeted at and developed by the CAP community.

Machine learning refers to computational methods that enable computers to learn patterns from data without being explicitly programmed with rules for every scenario. Supervised learning trains algorithms on labeled examples—for instance, teaching a model to recognize 'Health' policy by showing it thousands of pre-coded health-related documents—so it can then classify new, unlabeled documents into the same predefined categories. In contrast, unsupervised learning discovers patterns in unlabeled data without predetermined categories, identifying clusters or topics based solely on statistical similarities in the text. For CAP research, supervised learning became essential because the project's theoretically-grounded codebook requires documents to be classified into specific, predefined policy categories (such as 'Macroeconomics' or 'Defense') rather than allowing algorithms to generate their own emergent groupings that might not align with meaningful policy distinctions.

Efforts with supervised learning eventually led to RTextTools (Jurka et al., 2012) a user-friendly means for using supervised learning to assist in coding, which led to an expansion of datasets. Using a mixture of different supervised learning algorithms, many

different approaches were taken and tried with these tools. Some of these included the use of existing coded datasets, partial coding of new datasets, or coding of a portion of yearly data from new datasets to train models. Projects had different levels of success in using these tools and most often used a hybrid approach such as accepting codes where multiple algorithms agreed then reviewing and hand checking any disagreements. The two main challenges mentioned above, codes which were used infrequently, and the difficulty in identifying domestic policy often created the greatest number of problems. The tools were very useful, but also clearly limited. They helped with coding a high volume of data and helped highlight problematic policy topics for closer inspection by project teams. Clearly there was room for improvement.

Developing automatic classifiers

The transition from computer-assisted human coding to fully automatic classifiers represented a critical juncture in the evolution of empirical agenda-setting research. While RTextTools and similar supervised learning approaches provided valuable assistance to human coders, the exponential growth in digital government documents as well as media and social media content demanded more sophisticated solutions. In this section, we trace the development of automatic classifiers within the Comparative Agendas Project, examining both the technical innovations and the methodological challenges that emerged as we moved toward increasingly autonomous classification systems.

From dictionaries to generative AI

The quest for ever-improving machine coding solutions aims to achieve what is still widely considered the "gold standard" of multiclass classification schemes (such as the CAP codebook): at least double-blind human coding with high inter-coder reliability (ICR). The notion of a gold standard in content analysis, established through studies like Krippendorff's (1980) foundational work on content reliability and Neuendorf's (2002) comprehensive coding framework, refers to human-coded datasets with high ICR that serve as the ground truth for training and evaluating automated systems. In the CAP context, this gold standard was traditionally established from multiple coders independently classifying the same documents and resolving disagreements through discussion (as was suggested, inter alia, by Mikhaylov et al. (2012) for political text analysis). Establishing a single coherent benchmark is made difficult by the variety of human ICR measures by different country projects (such as agreement percentages, Cohen's kappa, Fleiss' kappa). But in the CAP handbook by Baumgartner et al. (2019, pp. 79, 102, 170) country projects reported a range of 79–97% of agreements on major topics and 76–85% on subtopics (for the Croatian project) while the French project utilized an internal threshold of 85% for major topics (2019, p. 94). These are rates consistent with Landis and Koch's (1977) threshold for "substantial agreement" in categorical data.

Each subsequent wave of machine coding approaches aimed at presenting a solution for both the validity (whether classifications capture genuine policy content) and reliability (consistency across repeated classifications) of machine labels. Accuracy—the proportion of correct classifications—and the F1 score, which balances precision and recall through their harmonic mean, became the dominant metrics for evaluating classifier performance.

Table 1 Four waves of automated CAP classification

Dimension	Wave 1: Rule-Based (Mid-2000s)	Wave 2: Traditional ML (Early 2010 s)	Wave 3: Deep Learning (Late 2010 s)	Wave 4: Generative AI (2020s)
Core Technology	Dictionary methods, keyword lists	SVM, Naïve Bayes, Maxent	RNNs, BERT, XLM-RoBERTa (transformers)	GPT-4, Claude, Llama (generative AI)
Text Representation	Manual keyword matching	Word frequencies (bag-of-words)	Contextual embeddings	Generative token predictions
Training Requirements	Manual rule creation	5,000–10,000 + labeled examples	500–5,000 labeled examples for fine-tuning	Minimal (zero-shot) to hundreds (few-shot)
Key Innovation	Human-codified domain knowledge	Automated pattern learning from frequencies	Bidirectional context understanding	In-context learning from instructions
Preprocessing Needs	Manual keyword curation	Extensive (stemming, stopwords, feature engineering)	Minimal (handled by model)	None (raw text)
F1 Performance (English with extreme variations depending on data size, domain etc.)	0.30–0.60 s	0.60–0.80 s	0.75–0.90 +	0.44–0.82 (variable by approach)
Multilingual Capacity	Language-specific (manual translation)	Limited (separate models per language)	Strong (100 + languages, one model)	Strong (100 + languages, one model)
Context Understanding	None (keyword presence only)	None (word counts only)	Strong (sequential dependencies)	Very strong (semantic understanding)
Interpretability	Very high (visible keyword rules)	High (visible features/weights)	Moderate (attention weights)	Low (black box) but explainable outputs
Primary Limitation	No context, manual labor intensive	Context-blind, sensitive to preprocessing	Requires substantial training data	Inconsistent without fine-tuning or prompt engineering
Reproducibility	Very high	High	High	Variable (API changes, proprietary models)
Typical Accuracy Trade-off	Precision vs. recall imbalance	Class imbalance issues	Balanced across topics	Variable quality across topics
Example Tools/Methods	Manual keyword dictionaries	RTextTools, Maxent	Babel Machine (XLM-RoBERTa)	ChatGPT, Claude, fine-tuned open-source LLMs

Initially, text-as-data research faced two fundamental machine learning approaches with distinct logics for capturing policy topics in large textual databases (Grimmer & Stewart, 2013, p. 268). Unsupervised learning algorithms identify mathematical patterns in document collections without predetermined categories—for instance, analyzing congressional hearings to detect recurring word clusters that might reveal shifting policy priorities or emerging issue areas. These methods impose statistical structure on text but cannot guarantee that discovered patterns align with theoretically meaningful policy categories.

In contrast, supervised learning relies on human-coded training data to classify documents into predefined categories, such as teaching an algorithm or large language model (LLM) to distinguish between CAP's 'Macroeconomics' (topic 1) and 'Health' (topic 3) codes by learning from thousands of previously classified bills, hearings, or newspaper articles. CAP's commitment to a theoretically-grounded codebook made supervised approaches essential, as unsupervised methods might identify

statistically coherent but conceptually meaningless clusters—grouping documents by latent variables such as rhetorical style or temporal period rather than policy content. This fundamental choice shaped every subsequent technical decision in developing CAP's automatic classification systems (even though some experiments tried to leverage both logics within a single research design, see e.g. Béchara et al., 2021).

The quest to automate CAP coding has progressed through four distinct technological waves, each offering different but cutting-edge approaches to understanding policy text in its respective period (see Table 1). (We should also note that this rough and simplified periodization is based on the take-up of technology in the wider CAP community as opposed to the original invention of these solutions, as, for instance, the history of research on neural networks goes back decades.) In CAP research, the first two waves often overlapped (2008–2015): The first consisted of dictionaries and other rules-based methods and the second involved traditional machine learning used algorithms like Support Vector Machines (SVM) that analyzed word frequencies; third-wave neural networks (2016–2022) employed contextual embeddings and the transformer architecture from models like BERT to understand words in relation to each other; fourth-wave generative AI (from 2023 on) uses LLMs that can both classify and explain their "reasoning" (which is, unlike human reasoning, still based on predicting the next most statistically probable token).

Each wave represents not just technical progress but a fundamental reconceptualization of how machines can understand the complex, contextual nature of policy language. Along with human coding, which is still at least partly deployed in the majority of CAP research (as witnessed by Baumgartner et al., 2019), each of these approaches continues to co-exist in published research.

The first wave began with rule-based dictionary methods that manually cataloged keywords for each policy area, but quickly evolved into a second wave that incorporated traditional machine learning algorithms like SVM and Naïve Bayes, which, while more sophisticated, still treated documents as collections of word frequencies and required extensive preprocessing and feature engineering. These second-wave supervised machine learning algorithms like SVM required extensive training datasets (often thousands of examples) to learn statistical patterns through feature engineering. Third-wave transformer models like BERT needed fine-tuning on labeled data to adapt pre-trained contextual linguistic information (called embeddings). Fourth-wave generative AI employs 'teaching' through prompt engineering and in-context learning, where models classify based on codebook instructions and minimal examples without task-specific retraining—a fundamentally different paradigm that trades data-intensive training for the iterative refinement of natural language instructions.

The first wave, emerging in the late 2000 s (starting with works such as Hillard et al., 2008; Quinn et al., 2006; Purpura & Hillard, 2006), began with rule-based approaches using dictionary methods similar to sentiment analysis codebooks consisting of positive, negative, and neutral labels (Albaugh et al., 2013). Researchers manually compiled lists of keywords and phrases associated with each policy area—for instance, cataloging terms like "missile," "troops," and "Pentagon" to identify defense policy. Over time, these dictionary-based approaches gave way to traditional machine

learning algorithms like SVM and Naïve Bayes, which treated documents as word frequency collections (Breeman et al., 2009; Karan et al., 2016).

RTextTools, and its predecessor, Maxent by Jurka (2012), were the first systematic attempts to provide automated CAP labels to text inputs, and they relied on these algorithms. However, both dictionary and algorithmic methods required extensive pre-processing (such as the removal of 'stopwords' unrelated to policy) and feature engineering—manually crafting rules about which words or phrases signaled particular policy areas (Burscher et al., 2015). Furthermore, Denny and Spirling (2018, p. 4) demonstrated that results could be "extremely sensitive" to seemingly innocuous preprocessing decisions such as stemming, stopword removal, or lowercase conversion. Later first-wave innovations attempted semi-supervised learning, combining small sets of labeled data with larger unlabeled corpora and through lexicon-building, achieving in some cases improved accuracy but still struggling to keep up with groups of well-trained researchers (see e.g., Kreutz & Daelemans, 2021, p. 62). These first-wave solutions struggled with CAP's imbalanced datasets, where topics like healthcare appear far more frequently than specialized areas like transportation. While rule-based dictionaries offered interpretability and algorithm-based research achieved reasonable accuracy for CAP's major topic-based classification tasks (Loftis & Mortensen, 2020), their performance on CAP's 200 + minor topics rarely yielded systematically valid and generalizable results, exposing the fundamental limitation of treating language as mere word counts without understanding context or meaning.

A core issue was a trade-off between precision and recall: the correct labelling of a single observation, on the one hand, and the retrieval of all observations of a given class (education or defense), on the other (Sebők & Kacsuk, 2021). Overall, as a review of the results by Sebők et al., (2022, p. 3626) shows, algorithm-based research in the first wave typically reached accuracy/F1 scores of up to the high 0.80 s but with very low values (in the low 0.30 s) for some applications and very uneven cross-class performance and mainly applied to English language data from the U.S. It took another technological innovation, the emergence of deep learning approaches, to allow for the linguistic and geographical, as well as text domain-based expansion of machine coding for the CAP codebook.

In this second wave, recurrent neural networks (RNNs) and later transformer architectures offered revolutionary possibilities for CAP classification by capturing sequential dependencies and contextual relationships crucial for understanding policy documents. The introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019) fundamentally transformed text classification by allowing models to understand words bidirectionally—considering both preceding and following context simultaneously. Unlike first-wave approaches that required extensive feature engineering to account for n-grams, BERT's pre-training on massive text corpora enabled it to develop a deep contextual understanding that could be fine-tuned for specific tasks like CAP classification. Building on BERT's model, and its optimized successors, Conneau et al. (2020) introduced the multilingual XLM-RoBERTa.

It was this latter model, trained on 100 languages with multiple terabytes of internet data, that enabled cross-linguistic policy classification without requiring separate models for each language. This breakthrough was particularly significant for the Babel

Machine project (Sebők et al., 2025), an international collaboration of CAP scholars led by poltextLAB in Hungary. The Babel Machine leveraged XLM-RoBERTa-large's 550 million parameters to build a suite of 100 + models for various domains (from legislative speech to social media), languages, and countries. These domains of policy agenda-setting are different in multiple respects relevant for modeling including terminology (more formal for laws than for tweets) and text length (different for newspaper articles and manifestos). Still, language (small such as Hungarian vs. big, such as English from a language technology perspective), country and period are just as important factors having an effect on the accuracy of more general models. This is the reason why the Babel Machine employs 100 + different models as opposed to a single universal solution.

In what can still be considered to be a state-of-the-art replicable machine coding performance at the time of writing, the models of the Babel Machine achieved F1 scores exceeding 0.90 for major topic classification of English-language congressional speeches, with weighted macro F1 scores surpassing 0.75 for 24 out of 41 language-domain pairs tested. The cross-linguistic capabilities of these models revealed fascinating insights into the representation of policy concepts. The Babel Machine demonstrated remarkable transfer learning as it produced high accuracy on original language training sets machine translated— primarily from English and Hungarian—to other languages, suggesting the model learned abstract policy concepts rather than language-specific patterns. Mate et al. (2023) demonstrated that combining machine translation with multilingual models could extend CAP classification to languages with minimal training data. Besides making its models public with a registration, the Babel Machine also contributed to the proliferation of CAP research through its fully automated web service (babel.poltextlab.com), which processes uploaded datasets and returns classified results via email within minutes for datasets of 10,000 observations or larger, democratizing access to advanced classification capabilities for researchers without computational resources or technical expertise to deploy such models independently.

The comparative advantages of transformer-based models over first-wave algorithmic approaches catalyzed an expansion of machine coding research using the CAP codebook across diverse contexts and languages beyond XLM-RoBERTa and the Babel Machine. Laurer et al. (2024) demonstrated that BERT-NLI models trained on merely 500 texts matched the performance of traditional algorithms requiring approximately 5,000 training examples, representing a ten-fold reduction in annotation requirements on eight diverse political science tasks. By leveraging both "language knowledge" from pretraining and "task knowledge" from natural language inference training, their approach enabled zero-shot classification capabilities, allowing models to predict classes without any training examples through verbalized hypotheses that mirrored human codebook instructions (see Bucher and Martini 2024).

Frantzeskakis and Seeberg (2023, p. 630) applied a BERT model to create the first comprehensive legislative database for 13 African countries. They used hand-coded bills from Nigeria to train the neural network and validated its cross-country validity using hand-coded laws from Ghana and bills from Zimbabwe. Using a 90%−10% train-test split, they achieved an F1 score of 88.4% for Nigerian data and 84.2% and 80%, for Ghana and Zimbabwe, respectively. The promising results of these pieces of research were corroborated by other studies using similar codebooks, such as that of the Comparative

Manifestos Project (Volkens et al., 2020). Licht (2023), for instance, demonstrated that multilingual sentence embeddings could achieve good cross-lingual performance, especially when their resources to collect labeled data are limited.

The third wave of automation, emerging in the early 2020s and hallmarked by the introduction of ChatGPT by OpenAI, featured generative AI models offering revolutionary capabilities through in-context or zero-shot learning—the ability to classify documents based on instructions without requiring extensive, task-specific training data. However, the initial promise of effortless, high-quality coding quickly met with the reality of performance limitations. On the one hand, studies showed that "off-the-shelf" models struggled with complex, deductive coding tasks. For instance, Gunes and Florczak (2025) found that using models like GPT-4 with minimal human intervention resulted in only moderate accuracy, with weighted F1 scores as low as 0.44. Even the superior use-case, which combined GPT 4 and Gemini 1.5 Pro achieved 0.82 weighted F1 score on the 83% of the data in which the two models agreed. When benchmarked against the Babel Machine, however, they found the latter to have an accuracy advantage of 13–16 percentage points depending on setup. This timid evaluation of generative AI models was echoed by Halterman and Keith (2025), who concluded that open-weight LLMs have significant limitations in following complex political science codebooks in a zero-shot context, sometimes yielding F1 scores as low as 0.21 for a Manifesto Project experiment (but 0.65 and 0.57 for some other political science tasks).

On the other hand, this performance gap is not a definitive failure but rather highlights a new set of methodological trade-offs and opportunities. The study by Gunes and Florczak demonstrated that a pragmatic human-in-the-loop framework, where humans resolve disagreements between two high-performing models, could elevate the weighted F1 score to 0.82 on most of the data. An alternative path to high performance comes from fine-tuning; Carammia et al. (2024) showed that smaller, fine-tuned open-source models can achieve accuracy comparable to massive proprietary systems, all while ensuring the scientific cornerstones of reproducibility, transparency, and data privacy. The quality of the human-AI interaction itself is also a critical factor. Hila and Hauser (2025) found that sophisticated prompt engineering that decomposes a complex task into a series of logical steps can dramatically improve validity, achieving levels of "substantial agreement" between the AI and human coders with an accuracy rate of 0.78.

All this reveals the fundamental pros and cons of the generative era. The advantages are clear: LLMs can match the performance of typical human coders and enable researchers to analyze political text at an unprecedented scale, as demonstrated by Rytting et al. (2023) and Salloum et al. (2025) for the CAP coding of congressional hearing summaries and social media posts on Bluesky, respectively. The primary disadvantage is that this potential is rarely unlocked "out of the box." Achieving high accuracy requires a strategic investment, not in massive training datasets of the past, but in clever human-AI workflows, meticulous prompt engineering, or the fine-tuning of more transparent, but usually more resource-intensive open-source models than traditional LLMs.

The central trade-off is thus between the unparalleled flexibility of fourth-wave generalist generative AI models (like GPT-4 and Claude) that can perform tasks through prompting alone, and the specialized, reproducible accuracy that can be achieved through the more computationally feasible methodological solution of third-wave

transformer models (like BERT and XLM-RoBERTa) that require fine-tuning on labeled data but offer greater transparency and control.

The limits of AI (and Humans): why perfect classification remains elusive

Recent advances in automated CAP classification have achieved remarkable F1 scores for at least the major topic classification of English-language political text. This performance approaches human-level validity, raising the question: what separates us from perfect classification? The answer lies not in a single technical barrier but in fundamental tensions between theoretical approaches, the constraints of practical projects, and the inherent ambiguity of political language. Understanding these limits requires examining at least six interconnected challenges that affect the generalizability of machine policy coding across countries, languages, historical periods, political systems, and text domains (from print to social media and legislative speech).

These six critical areas are: the ambivalent status of human gold standards, the varying informativeness of units of observation, the constraints of some of the CAP's project coding rules, the scarcity of high-quality training data, the constrained nature of pre-set codebooks, and the potential of emerging model architectures. Each area reveals how seemingly technical limitations often reflect deeper theoretical or methodological tensions in how we conceptualize and measure policy content. The Babel Machine's extensive deployment across diverse contexts provides empirical evidence for these challenges while also suggesting pathways toward solutions.

First, the notion of a "gold standard" assumes human coding provides ground truth for training and evaluation. Yet, in practical projects, human intercoder reliability rarely exceeds 90% for major topics without costly reconciliation procedures by expert teams. Paradoxically, machine classifications sometimes appear more consistent than human coding when evaluated against theoretical definitions—machines don't suffer from fatigue, distraction, or evolving interpretations over multi-year projects. It is easier to hard-code rules establishing reliability–such as "policy, not targets"–to computer code than to humans. This creates an epistemological puzzle: when human coders disagree with each other at rates of 10–25%, and machines produce classifications that align with codebook definitions but diverge from specific human judgments, which classification is "correct"? The Babel Machine's outputs frequently reveal cases where automated classification appears more empirically justified than human coding, particularly for text segments where human coders may have mistakenly relied on perceived or limited contextual knowledge (due to, for instance, limitations on understanding historical policy terminology) not present in the text itself. Conversely, the processes that could lead to at least a new human–machine hybrid gold standard (let alone a fully machine-based one) are far from elaborated in the literature.

Second, this issue points to the importance of context in deciding CAP labels, where context is often a function of the length of the text available for the human and machine coders. This unit of observation problem proves particularly acute when classifying document titles versus full text, as is customary in a large swath of CAP research. Newspaper article or Congressional bill titles often provide insufficient information for valid and reliable classification. New York Times article titles such as "A Grim Milestone for the US" or "In Los Angeles, trying to atone for water sins" can lead down many different,

legitimate paths depending on the follow-up text. Similarly, the Congressional bill titled "Infrastructure Investment and Jobs Act" (H.R. 3684 of the 117th Congress) could concern macroeconomics, labor, technology, energy or even transportation, strictly based on its title (in fact, it started its Congressional career as a transport-oriented policy package).

Babel Machine experiments show that classification accuracy improves significantly when a lead paragraph or full article text is available compared to titles alone. Yet most historical CAP data relies on titles or brief descriptions due to the impossibility of manually coding full text at scale. This ties in to two additional problems which put a fundamental ceiling on achievable accuracy with humans or machines: the multitopic nature of some longer political texts coupled with the brevity of others, and the limited availability of high-quality human training data tailor-made for all potential usages of the CAP codebook across time, space, input text domains, and languages.

This leads to our third challenge: in many cases, texts contain a mixture of topics, such as an article discussing climate change that simultaneously addresses environmental regulations, economic impacts on fossil fuel industries, and technological innovations in renewable energy, all within a space of a few paragraphs. CAP's foundational rule requiring single codes per observation may conflict with the multi-dimensional nature of some policy documents, especially for longer texts such as bills and laws. The aforementioned "Infrastructure Investment and Jobs Act" spans transportation, environment, technology, labor, and macroeconomic policy areas, and forcing a single classification for such omnibus legislation violates both theoretical validity and practical utility (as a dataset of a few thousand bills may turn into one of millions of observations on the page or paragraph level). The challenge intensifies when time-tested and high-quality legislative summaries are unavailable from official services.

The Babel Machine addresses this multitopic challenge through three complementary solutions. First, it offers flexibility in selecting the unit of observation, recognizing that what works for State of the Union addresses—which can be meaningfully segmented into quasi-sentences—may be inappropriate for omnibus bills that resist clean division. Users can choose between analyzing full documents, paragraphs, or sentences depending on their research needs and the nature of their texts. Second, the system introduced a "non-policy content" category (code 999) alongside the traditional 21 CAP major topics, acknowledging that not all text segments contain classifiable policy content—a prevalent issue in media coverage that includes editorial commentary or human-interest framing. All Babel Machine models were retrained to recognize and appropriately code these non-policy segments rather than forcing them into ill-fitting policy categories. Third, the system provides top-3 classifications with associated confidence scores through softmax probability distributions. This probabilistic approach better captures policy complexity: a climate bill might be classified as 60% environmental policy, 30% macroeconomic policy, and 10% technology policy. Over millions of observations, these probability distributions provide more accurate population-level estimates of the original theoretical interest of policy attention than forced single classifications, even if they challenge traditional CAP coding conventions that assume mutual exclusivity among policy categories.

Fourth, training data scarcity is an inherent limitation in achieving high machine coding performance across all languages and domains. While English congressional data

includes hundreds of thousands of coded examples, many languages have fewer than 10,000 training documents, insufficient for robust deep learning. The Babel Machine experiments with synthetic data generation—using LLMs to create training examples that mimic real policy documents—showing promise for low-resource languages. However, synthetic data can amplify biases and may not capture the full complexity of authentic policy discourse, particularly for culturally specific policy framings that don't translate directly across political systems.

Fifth, the CAP codebook represents a remarkable achievement in creating a unified framework for comparative policy analysis across dozens of political systems. Yet, its very strengths create inherent limitations that constrain automated classification performance. Boydstun's (2013) media codebook addressed one of these challenges by adding codes (such as Sports and Recreation) prevalent in a new domain of application, newspapers, but misclassified when the original CAP codebook was mechanically applied. A similar problem is related to international relations coding (topic 19), which can be disaggregated into two separate tasks for better results: named entity recognition to distinguish geographical relevance and policy topic coding, which should have the same results regardless of where a central bank decision happened. As a solution, the Babel Machine offers additional named entity recognition beyond CAP codes and options to use media codebooks when appropriate.

Finally, as they did throughout the three waves of machine coding for CAP, emerging model architectures offer potential breakthroughs on an ongoing basis. The Babel Machine's recent experiments combine outputs from multiple generative AI providers—GPT-5, Claude 4, Llama 4—using weighted voting schemes that leverage each model's strengths. Alternatively, within model Mixture-of-experts (MoE) approaches dynamically route different types of documents to specialized sub-models: one expert for legislative text, another for media coverage, a third for executive communications. Future implementations might combine ensemble voting with true MoE architectures. These hybrid approaches suggest that near-perfect classification may emerge not from any single algorithmic breakthrough but from intelligently combining an ensemble of multiple strategies, each optimized for specific aspects of the classification challenge.

Conclusion

At the advent of transformer models, Wilkerson and Casas (2017, p. 543) argued "that researchers do not need to be computer programmers or statistical methodologists to use text as data methods in their research". The way computerized text analysis transformed political science research allowed them to "have the ability to explore massive amounts of politically relevant text using increasingly sophisticated tools" without turning into methodologists. Nearly a decade later, the field has largely lived up to this prophecy: Generative and fine-tuned models can now classify political text with near-human levels of accuracy at previously unimaginable scales.

Yet this achievement has revealed a new irony, one that Wilkerson and Casas aptly presaged: While substantive researchers did not have to turn into methodologists to leverage AI-supported tools, "They do need to be attentive to the same concerns about validity and reliability that apply to all methods". The true bottleneck, it turns out, was never purely computational. Instead of purely celebrating the new opportunities of AI

models to implement more valid, reliable, and cost-effective projects faster, researchers now grapple with deeper questions: What constitutes valid classification, how should we handle the inherent ambiguity of political language, and whether forcing singular interpretations onto multifaceted texts advances our theoretical goals. In this sense, the bottleneck has migrated from processing capacity to epistemological clarity. Time may reveal that the real challenge was never simply teaching machines to read like humans, but instead accepting that both humans and machines are imperfect readers of an inherently complex political world. Until then, the task is not a perfect classification, but a transparent, reproducible, and theoretically grounded approximation method that acknowledges its limitations while enabling the large-scale textual research Wilkerson and Casas once envisioned.

Fifty years since McCombs and Shaw sent us all down this path with their remarkable insights, much has changed and the scope of research that is possible is certainly unprecedented. But the need for solid theory and clear methods has not changed.

Acknowledgements
We thank the special issue journal editors for their invitation to submit.

Authors' contributions
FB, SB, and MS shared writing of the entire manuscript with FB taking the lead on organization and the first section on the literature; SB leading on the section about the development of the CAP codebook; and MS on the development of automated text classification systems. All authors reviewed and approved of the full manuscript.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
This article does not contain any study with human subjects.

Competing interests
The authors declare no competing interests.

Received: 4 September 2025 Revised: 10 December 2025 Accepted: 4 February 2026
Published online: 26 February 2026

References

Albaugh, Q., Sevenans, J., Soroka, S., & Loewen, P. J. (2013). The automated coding of policy agendas: A dictionary-based approach. *6th Annual Comparative Agendas Project (CAP) Conference,* 1–22.
Bachrach, P., & Baratz, M. (1962). The two faces of power. *American Political Science Review, 56*, 947–952.
Baumgartner, F. R., & Jones, B. D. (1993). *Agendas and Instability in American Politics*. University of Chicago Press.
Baumgartner, F. R., Breunig, C., & Grossman, E. (Eds.). (2019). *Comparative policy agendas: Theory, tools, data*. Oxford University Press. https://library.oapen.org/handle/20.500.12657/52225
Béchara, H., Herzog, A., Jankin, S., & John, P. (2021). Transfer learning for topic labeling: Analysis of the UK House of Commons speeches 1935–2014. *Research & Politics, 8*(2), 1–10. https://doi.org/10.1177/20531680211022206
Berelson, B., Lazarsfeld, P., & McPhee, W. (1954). *Voting*. University of Chicago Press.
Bevan, S. (2019). Gone fishing: The creation of the comparative agendas project master codebook. In *Frank R. Baumgartner, Christian Breunig, and Emiliano Grossman (eds.) Comparative policy agendas*. Oxford: Oxford University Press.
Boydstun, A. E. (2013). *Making the news: Politics, the media, and agenda setting*. University of Chicago Press.
Breeman, G. E., Then, H., Kleinnijenhuis, J., van Atteveldt, W., & Timmermans, A. (2009). Strategies for improving semi-automated topic classification of media and parliamentary documents. *67th Annual Meeting of the Midwest Political Science Association*. 1–14. https://library.wur.nl/WebQuery/wurpubs/385462
Bucher, M.J. and Martini, M. 2024. *Fine-tuned "Small" LLMs (Still) outperform zero-shot generative AI in text classification*. arXiv preprint arXiv:2406.08660.
Burscher, B., Vliegenthart, R., & De Vreese, C. H. (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. https://doi.org/10.1177/0002716215569441
Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. (1960). The American Voter. New York: John Wiley and Sons.

Carammia, M., Iacus, S. M., & Porro, G. (2024). *Rethinking scale: The efficacy of fine-tuned open-source LLMs in large-scale reproducible social science research (No. arXiv:2411.00890)*. arXiv. https://doi.org/10.48550/arXiv.2411.00890
Cobb, R. W., & Elder, C. D. (1972). *Participation in American politics: The dynamics of agenda-building*. The Johns Hopkins University Press.
Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised cross-lingual representation learning at scale (version 2)*. arXiv. https://doi.org/10.48550/ARXIV.1911.02116
Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26*(2), 168–189.
Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423
Dowding, K., Hindmoor, A., & Martin, A. (2016). The comparative policy agendas project: Theory, measurement and findings. *Journal of Public Policy, 36*(1), 3–25.
Frantzeskakis, N., & Seeberg, H. B. (2023). The legislative agenda in 13 African countries: A comprehensive database. *Legislative Studies Quarterly, 48*(3), 623–655. https://doi.org/10.1111/lsq.12404
Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028
Guiraudon, V. (2000). European integration and migration policy: Vertical policy making as venue shopping. *Journal of Common Market Studies, 38*(2), 251–271.
Gunes, E., & Florczak, C. K. (2025). Replacing or enhancing the human coder? Multiclass classification of policy documents with large language models. *Journal of Computational Social Science, 8*(2), 31. https://doi.org/10.1007/s42001-025-00362-2
Halterman, A., & Keith, K. A. (2025). *Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts (version 2)*. arXiv. https://doi.org/10.48550/ARXIV.2407.10747
Hila, A., & Hauser, E. (2025). *Assessing the reliability of large language models for deductive qualitative coding: A comparative study of ChatGPT interventions (version 1)*. arXiv. https://doi.org/10.48550/ARXIV.2507.14384
Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics, 4*(4), 31–46. https://doi.org/10.1080/19331680801975367
Jones, B. D., & Baumgartner, F. R. (2005). *The politics of attention: How government prioritizes problems*. University of Chicago Press.
Jones, B. D., & Baumgartner, F. R. (2012). From there to here: Punctuated equilibrium to the general punctuation thesis to a theory of government information processing. *Policy Studies Journal, 40*(1), 1–19.
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.
Jurka, Timothy P., Loren Collingwood, Amber E. Boydstun, Emiliano Grossman and Wouter van Atteveldt. 2012. *RTextTools: Automatic text classification via supervised learning*. R package version 1.3.9. http://CRAN.R-project.org/package=RTextTools
Karan, M., Šnajder, J., Širinić, D., & Glavaš, G. (2016). Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *N. Reiter, B. Alex, & K. A. Zervanou (Eds.), Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (pp. 12–21)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2102
Kingdon, J. W. (1984). *Agendas, alternatives, and public policies*. Little, Brown.
Kreutz, T., & Daelemans, W. (2021). A semi-supervised approach to classifying political agenda issues. *Proceedings of the 1st Workshop on Computational Linguistics for Political Text Analysis (CPSS-2021), 59–64*. https://hdl.handle.net/10067/1852840151162165141
Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Laurer, M., Atteveldt, W., Casas, A., & Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis, 32*(1), 84–100. https://doi.org/10.1017/pan.2023.20
Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis, 31*(3), 366–379. https://doi.org/10.1017/pan.2022.29
Loftis, M. W., & Mortensen, P. B. (2020). Collaborating with the machines: A hybrid method for classifying policy documents. *Policy Studies Journal, 48*(1), 184–206. https://doi.org/10.1111/psj.12245
Mate, A., Sebők, M., Wordliczek, L., Stolicki, D., & Feldmann, Á. (2023). Machine translation as an underrated ingredient? Solving classification tasks with large language models for comparative research. *Computational Communication Research, 5*(2), 1–34. https://doi.org/10.5117/CCR2023.2.6.MATE
McCarthy, J. D., & Zald, M. N. (1978). Resource mobilization and social movements: A partial theory. *American Journal of Sociology, 82*, 1212–1241.
McCombs, Maxwell E., and David H. Weaver. 1985. Towards a merger of gratifications and agenda setting research. In K. E. Rosengren, L. A. Wenner, & P. P. Palgreen (Eds.), *Handbook of political communication*. Newbury Park, CA: Sage.
McCombs, M. E. (1993). The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas. *Journal of Communication, 43*(2), 58–67.
McCombs, M. E. (2004). *Setting the agenda: The mass media and public opinion*. Blackwell Publishing Inc.
McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly, 36*(2), 176–197.
McCombs, M. E., & Zhu, J. H. (1995). Capacity, diversity and volatility of the public agendas: Trends from 1954 to 1994. *Public Opinion Quarterly, 59*(4), 495–525.
Mikhaylov, S., Laver, M., & Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis, 20*(1), 78–91. https://doi.org/10.1093/pan/mpr047

Neuendorf, K. A. (2002). *The content analysis guidebook* (2.). SAGE Publications.
Prothrow, J. W., & Grigg, C. M. (1960). Fundamental principles of democracy: Bases of agreement and disagreement. *Journal of Politics, 22*(2), 276–294.
Purpura, S., & Hillard, D. (2006). Automated classification of congressional legislation. *Proceedings of the 2006 International Conference on Digital Government Research*. 219–225.
Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. 2006. An automated method of topic-coding legislative speech over time with application to the 105th-108th U.S. senate. In *Midwest political science association meeting*.
Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). Towards coding social science datasets with language models. arXiv:2306.02177. https://arxiv.org/abs/2306.02177
Salloum, A., Quelle, D., Iannucci, L., Bovet, A., & Kivelä, M. (2025). Politics and polarization on Bluesky (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.03443
Schattschneider, E. E. (1960). *The semi-sovereign people*. Holt, Rinehart and Winston.
Sebők, M., & Kacsuk, Z. (2021). The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach. *Political Analysis, 29*(2), 236–249. https://doi.org/10.1017/pan.2020.27
Sebők, M., Kacsuk, Z., & Máté, Á. (2022). The (real) need for a human touch: Testing a human–machine hybrid topic classification workflow on a New York Times corpus. *Quality & Quantity, 56*(5), 3621–3643. https://doi.org/10.1007/s11135-021-01287-4
Sebők, M., Máté, Á., Ring, O., Kovács, V., & Lehoczki, R. (2025). Leveraging open large language models for multilingual policy topic classification: The Babel machine approach. *Social Science Computer Review, 43*(2), 295–317. https://doi.org/10.1177/08944393241259434
Volkens, A., Burst, T., Krause, W., Lehmann, P., Matthieß, T., Merz, N., Regel, S., Weßels, B., Zehnter, L., & Wissenschaftszentrum Berlin Für Sozialforschung (WZB). (2020). Manifesto project dataset (version 2020b) . *Manifesto project*. https://doi.org/10.25522/MANIFESTO.MPDS.2020B
Walker, Jack L., Jr. (1966). A critique of the elitist theory of democracy. *American Political Science Review 60, 2*:285–295, 391–392.
Walker, J. L., Jr. (1977). Setting the agenda in the U.S. Senate: A theory of problem selection. *British Journal of Political Science, 7*, 423–445.
Wilkerson, J. D., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science, 20*(1), 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542
Wolfe, M., Jones, B. D., & Baumgartner, F. R. (2013). A failure to communicate: Agenda setting in media and policy studies. *Political Communication, 30*(2), 175–192.

Publisher's Note