Journal of Social Policy (2026), **55**, 385–404
doi:10.1017/S0047279424000382

ARTICLE

Goodbye human annotators? Content analysis of social policy debates using ChatGPT

Erwin Gielens[1], Jakub Sowula[2] and Philip Leifeld[3]

[1]Department of Sociology, Tilburg University, Tilburg, the Netherlands;, [2]University of Tübingen, Germany and Bern University of Teacher Education, Bern, Switzerland and [3]Department of Social Statistics, University of Manchester, Manchester, UK
Corresponding author: Erwin Gielens; Email: e.e.c.gielens@tilburguniversity.edu

(Received 8 May 2024; revised 23 October 2024; accepted 10 November 2024; first published online 03 January 2025)

Abstract
Content analysis is a valuable tool for analysing policy discourse, but annotation by humans is costly and time consuming. ChatGPT is a potentially valuable tool to partially automate content analysis for policy debates, largely replacing human annotators. We evaluate ChatGPT's ability to classify documents using pre-defined argument descriptions, comparing its performance with human annotators for two policy debates: the Universal Basic Income debate on Dutch Twitter (2014–2016) and the pension reforms debate in German newspapers (1993–2001). We use the API (GPT-4 Turbo) and user interface version (GPT-4) and evaluate multiple performance metrics (accuracy, precision and recall). ChatGPT is highly reliable and accurate in classifying pre-defined arguments across datasets. However, precision and recall are much lower, and vary strongly between arguments. These results hold for both datasets, despite differences in language and media type. Moreover, the cut-off method proposed in this paper may aid researchers in navigating the trade-off between detection and noise. Overall, we do not (yet) recommend a blind application of ChatGPT to classify arguments in policy debates. Those interested in adopting this tool should manually validate bot classifications before using them in further analyses. At least for now, human annotators are here to stay.

Keywords: content analysis; text analysis; policy debates; policy discourse

> *"The cat is out of the bag. Universal Basic Income is 760 euro each month. Unemployment benefits, sick pay and other benefits will be abolished. To be used for self-realisation. Crazy. #Tegenlicht"*

This is an example of how a Dutch socialist politician argued against Universal Basic Income. As many scholars nowadays recognise, such arguments and ideas within

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is included and the original work is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

policy discourses are crucial to understanding policy reform processes (e.g. Beland, 2019; Prior, Hughes and Peckham, 2012; Schmidt, 2008). Political coalitions are constructed and communicated through arguments – or positions – expressed in the media. Similarly, policymakers and other stakeholders may rally around 'good ideas' that draw attention in media debates (e.g. Willems and Beyers, 2023). Moreover, media coverage pressures political actors to (not) implement policy reforms (e.g. Jensen and Wenzelburger, 2021). Whether seen as the glue that binds coalitions or as the driver of policy change, policy discourse has become a major concern for those interested in policy processes.

The value of policy discourse analysis is demonstrated by its broad applicability, for example, in analyses on environmental policy (Gutierrez Garzon et al., 2022), welfare policies (Blum and Kuhlmann, 2019; Theiss, 2023), urban mobility (Towns and Henstra, 2018) or educational policy (Symeonidis, Francesconi and Agostini, 2021). Moreover, those using the discourse network analysis framework also aim to identify coalitions on the basis of arguments employed in policy debates (e.g. Eder, 2023; Fergie et al., 2019; Gupta et al., 2022; Leifeld, 2013; Markard, Rinscheid and Widdel, 2021). Policy discourse analysis routinely relies on content analysis. In the context of policy discourse, content analysis is used to annotate several elements, such as actors, arguments or policy solutions, and actors' (dis)agreement with arguments. Here, we focus on the task of identifying policy solutions or arguments (i.e. 'thematic analysis'; see Nguyen-Trung, 2024) as the most important and difficult-to-code element. The difficulty with identifying arguments lies not only in the first stage, wherein a coding scheme is defined through the interpretation and comparison of texts, but also in the time and effort spent in applying the crystallised coding scheme to large amounts of documents. This specific subtask of content analysis is sometimes referred to as a 'directed' or 'supervised' stage of content analysis (e.g. Hsieh and Shannon, 2005; Petchler and Gonzáles-Bailon, 2015). While the automated classification of arguments would vastly increase the speed and scale of analysing policy discourse, attempts to automate this process have thus far not produced very accurate results (e.g. Ceron et al., 2024; Haunss et al., 2020; Lapesa et al., 2020).

A solution to this problem might lie in the much-discussed language model ChatGPT.[1] Much like bag-of-words classification (Kowsari et al., 2019; see also Grimmer and Stuart, 2013) and bidirectional encoder representations from transformers (BERT) models, ChatGPT can be instructed to classify texts in terms of containing a particular topic. The use of ChatGPT as a tool for automated content analysis might have enormous potential as a tool to automatically analyse the contents of text documents. First, the time used for manual content analysis can be dramatically sped up by using large language models (LLMs) such as ChatGPT. Second, as with other automated approaches, ChatGPT can handle large amounts of data, allowing researchers to extend their scope across larger periods of time or multiple contexts. Manual annotation is limited to a sample of the corpus to validate the results. Third, unlike typical natural language processing techniques, using ChatGPT's user interface does not require special programming skills or statistical knowledge. Any researcher with one or more well-defined topics could extend their content analysis in a corpus of virtually unlimited size. As such it opens a world of opportunities for social scientists working in the qualitative or survey traditions.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

In this paper, we thus seek to validate the classification abilities of ChatGPT in the context of policy debates in Germany (pension reforms, see Leifeld, 2016) and the Netherlands (Universal Basic Income, see Gielens, Roosma and Achterberg, 2022). Comparing these cases represents a strong empirical test as it grants insight into the generality of ChatGPT's abilities, namely in handling policy debates in different languages and of different lengths, contexts and complexity. We first assess the reliability of the model by repeating classifications over several iterations. We then compare the classifications provided by ChatGPT with the human annotations of these datasets. In the following sections, we first present the existing literature on content analysis relying on LLMs such as ChatGPT. Next, we describe the datasets and our methodological approach. We then present the results before moving to a discussion of the implications and limitations of our analysis.

ChatGPT for text analysis

Political scientists have been fascinated with the potential for automated text analysis for some time (e.g. Grimmer and Stewart, 2013; see also Slapin and Proksch, 2008). More recently, the field has turned to large language models to efficiently identify, for example, ideological placement (Rheault and Cochrane, 2020), political emotions (Widmann and Wich, 2023) and political manifestos (Laurer et al., 2024; Licht, 2023). Miller, Linder and Mebane (2020) explored an active labelling strategy where manual classification is aided by a text algorithm to select relevant documents (see also Alshami *et al.,* 2023).

Researchers have now turned to exploring the text-analytic abilities of ChatGPT in applications of text analysis. For example, Prakash et al. (2023) have used the model to identify topics in a collection of memes by clustering texts or images on the basis of similarities in images and words. In their application, ChatGPT 'outperformed well-established topic models across three distinct datasets' (p.8). A rapidly expanding number of studies have also used ChatGPT for content analysis, much like in the current application. Huang, Kwak and An (2023) have used the model to classify tweets containing hate speech, finding that 'ChatGPT correctly identified 80% of the implicit hateful tweets in our experimental setting' (p.4). Wang et al. (2021) find that the chatbot classifies news topics with an accuracy ranging between 77.5 per cent and 87.5 per cent depending on the labelling strategy. Gilardi, Alizadeh and Kubli (2023) analysed topics and frames in tweets and news articles, showing that GPT3 classifications matched trained annotators in around 60 per cent of cases. Moreover, they find that the chatbot substantially outperforms crowd-workers recruited on Mturk (see also Horn, 2019). Morgan (2023) and Turobov et al. (2024) investigate how GPT performs relative to manual annotation in thematic analysis and topic classification of focus groups and United Nations policy documents, respectively, but they do not provide measures of classification performance. In a wide array of classification tasks, Ziems et al. (2023) show that GPT4 performance differs strongly between types of utterances. The model does well at identifying stance and ideology but performs poorly in classifying, for example, misinformation and implicit hate speech. These applications demonstrate the potential and limitations of ChatGPT in clustering and classifying policy debates.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

With this contribution, we add to this emerging field of study in two ways. First, we extend the validation of this method to the application of policy debates. While some attention has indeed gone to our sources of interest – newspaper articles and tweets – prior applications have involved specific datasets on hate speech (Huang, Kwak and An, 2023), content moderation (Gilardi, Alizadeh and Kubli, 2023) and general news topics (Wang et al., 2021) that do not necessarily translate to the classification of arguments in policy debates. In addition to being valuable for students of policy discourse, validating this classification task contributes to demonstrating the general applicability of ChatGPT in classification problems.

Second, we add to prior research by introducing more fine-grained evaluation methods. Existing studies have relied heavily on accuracy and inter-rater reliability metrics (Gilardi, Alizadeh and Kubli, 2023; Huang, Kwak and An, 2023; Wang et al., 2021; cf. Ziems et al., 2023). When datasets are unbalanced – that is, there are more non-occurrences (0) than occurrences (1) of arguments – accuracy estimates are biased towards classifying non-occurrences (e.g. Juba and Le, 2019). We therefore also account for precision and recall, indicating the rate of false positives and false negatives in bot-labelled argument occurrences. A more detailed explanation of these metrics is included in the methods section.

Lastly, we generate relevant insights by performing our empirical test on different models (GPT-3.5 Turbo and GPT-4 Turbo) and through different types of access (user interface [UI] and the application programming interface [API]). This allows for speculation about future improvements in the content analysis abilities of ChatGPT and gives practical hints for researchers interested in relying on ChatGPT in future research. We expect significant improvements with the newer version but similar performance between different access types (API versus UI). Although we do not expect major differences in performance between API and UI, the comparison is valuable for potential end-users for reasons of accessibility (API is paid and requires programming skills) and time (API is considerably faster).

Datasets

We used two distinct datasets, each representing a unique case study with its own sets of characteristics. The first dataset is a collection of tweets on a Universal Basic Income (UBI) in the Netherlands posted between 2014 and 2016. The second dataset is composed of newspaper articles on the German pension reforms published between 1993 and 2001. We specifically relied on datasets analysed in previous peer-reviewed studies (see Gielens, Roosma and Achterberg, 2022; Leifeld, 2016) to ensure the quality of the human coding, which is used as a comparative benchmark for ChatGPT's coding performance. For each of these datasets, we assess performance for the ten most frequently adopted arguments.

The arguments and descriptions were initially created by human coders in the publications. These studies inductively developed and refined coding schemes and were reviewed by second researchers to test their reliability. The tweets dataset

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

adopted a formal inter-coder reliability procedure, yielding an average Cohen's Kappa of $\kappa = 0.430$ across arguments. This can be considered 'fair to good' (Fleiss, 1981, p. 218), especially considering the high number of arguments and lack of context present in tweets. The newspaper dataset assessed reliability by having a second researcher evaluate all labels, with discussion and revision in cases of disagreement.

We used the description of arguments provided in the codebooks of these studies to test how successfully the manual coding can be replicated using ChatGPT. We used argument descriptions in the original language, which we took from the source data of the content analysis. However, because the argument descriptions were sometimes very technical, we shortened and simplified the description where possible. These descriptions will be used as input for the classification task. The descriptions of the arguments in their original language, as they are used in the classification task, are available in Appendix A. The English translation of these argument descriptions is presented in Tables 1 and 2.

Dataset 1: Dutch Twitter data on universal basic income

Dataset 1 consists of 5,128 Dutch Twitter users discussing the benefits and disadvantages of the Universal Basic Income (UBI) policy proposal. Gielens, Roosma and Achterberg (2022) manually analysed the content of these tweets on 3 days of peak attention between 2014 and 2016. They identified fifty-six arguments in favour or against UBI within these tweets. A detailed description of the dataset and coding process can be found in Gielens, Roosma and Achterberg (2022). Twitter data are an interesting case study for our analysis due to the messy nature of tweets. Tweets are short messages with a platform-specific writing style, difficult to understand in isolation. Machines often find it hard to clearly identify topics in this type of data (Duarte, Llanso and Loop, 2017), so a good performance of ChatGPT would be encouraging. The codebook for this dataset contains arguments related to a specific policy. The ten most frequently occurring topics are included in our analysis (see Table 1).

Dataset 2: newspaper articles on the German pension policy reform

Dataset 2 includes statements collected from German newspapers in the time frame of January 1993 to May 2001, preceding the German Riester pension reform. The dataset includes 7,249 statements about sixty-eight concepts from 1,879 articles of political actors in this period, which were also identified by human coders. A detailed description of the dataset can be found in Leifeld (2016). The second dataset serves as a stark contrast to the first dataset, as newspaper articles are written by experts trained to write clear texts that are not restricted by very short character limitations, and are thus able to present a more coherent picture than possible in a tweet. Accordingly, the structured and expertly crafted nature of the statements provides a reliable comparable benchmark to the 'messy' tweets in dataset 1, making it an interesting point of departure for our analysis. The codebook for this dataset contains proposed policy solutions related to the financing gap in the pension system, rather than arguments for or against one

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Table 1. Top ten arguments from the UBI tweets (descriptions translated from Dutch) and the number of sampled tweets (*N*) containing these arguments

	N	Description
Experiment	419	We should experiment with UBI. We should investigate the effects of UBI. This topic is about starting a new pilot, not about experiments that took place in the past.
Deregulation	244	Implementing UBI leads to a simple and efficient social security in two ways. (1) UBI replaces many supplements and benefits, leading to a smaller government. (2) UBI simplifies social security by abolishing regulations.
Freeriding	244	People will quit their jobs or work less when they receive a UBI. People will become lazy or move to another country. UBI encourages laziness and is mainly appealing to scroungers: these people do not deserve a benefit. UBI leads to more welfare fraud.
Affordable	218	UBI is affordable. Calculations show how UBI can be financed. UBI saves on costs of civil servants and administration. UBI is affordable because the costs of crime and healthcare will go down.
Political support	186	Politicians are in favour or against UBI. Politicians and political parties are afraid to discuss UBI. UBI belongs on the political agenda. Some municipalities are in favour of UBI. There are also interest groups that lobby for UBI.
Social benefits	184	UBI is different from social assistance benefits and retirement benefits. UBI and social assistance both guarantee a minimum income. However, UBI does not require a reciprocal act (*tegenprestatie*). Also, everybody receives a UBI, not just the unemployed.
Free money	183	UBI is free money for everybody. The expression 'free money' is taken from a book written by Rutger Bregman. It is also the title of a documentary made by *Tegenlicht*.
Unrealistic	178	UBI is unattainable. UBI is a utopian idea. UBI is a fairytale. UBI is unrealistic and idealist.
Security	156	UBI reduces poverty. The benefit is sufficient to cover essential living costs. UBI guarantees an income level above the poverty line. Poor people will benefit from UBI. UBI is a basic human right.
Wellbeing	130	UBI reduces stress and improves your health. UBI makes people more healthy and more happy.

specific policy. The two datasets hence cover a broad spectrum of codes found in policy discourse analysis. The ten most frequently occurring topics are included in our analysis (see Table 2).

Sampling

To save time and money, we selected a stratified random sample of tweets and newspaper articles. We used a stratified random sample rather than a simple random sample to ensure that the categories we used for coding were well-represented in the analytical sample. For each of the ten selected arguments, we sampled 50 per cent of the documents containing that argument, removing duplicates. Documents in each dataset are randomly shuffled to minimise bias.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Table 2. Top ten arguments from the Pension reform newspaper articles (translated descriptions from German) and the number of sampled newspaper articles (*N*) containing these arguments

	N	Description
Private capital	235	The state provides incentives for additional private retirement savings. By partially privatising the pension system, the pressure exerted by demographic population aging and occupational trends can be alleviated because the contribution level and the pension level are not as closely tied at the population level anymore.
Pension cuts	228	Lowering the level of pensions will sustain the pension system. Defined contributions and a lower pension level may, depending on the amount by which pensions are cut, imply that a compensation for pensioners by other means becomes necessary.
Contributions	179	Increasing contribution levels will sustain the pension level. A moderate increase of the contribution rate would be an adequate response to population aging.
Pay as you go	130	This category includes all statements where actors stated that they either favoured or rejected the contribution-based PAYG system. The concept represents the status quo. Rejecting the PAYG system is not necessarily the same as calling for a private capital cover system because there could be other forms of system change, such as at rate pensions.
Early retirement	115	Too many people are retiring early. To sustain the pension system we should prevent people from retiring early. Measures should be taken to reduce unemployment amongst those nearing retirement. Part-time work contracts (*Altersteilzeit*) for those nearing retirement should be discouraged.
Subsidies	115	External sources of money can help to retain pension levels without increasing contributions. This money could be collected via a newly proposed ecological tax, increasing the value added tax, or simply by taking the money from the national tax budget.
Demographic factor	104	Pension problems are due to an ageing population. Therefore, pension levels should be adjusted for the number of elderly in society.
Widower cuts	79	Lowering or even abolishing pensions for widowers or the invalid releases financial pressure on the pension system.
Low earners	73	Low-income workers do not contribute to the pension system through payroll tax. These people should be included in the PAYG system to increase the number of contributors.
Economic development	69	Pension levels should be adjusted for changes in economic growth. Pension levels should decrease in times of crisis and increase in times of growth.

We sampled 50 per cent of tweets containing each argument included in our analysis. For example, the deregulation argument was identified in 244 tweets, so the sample contains 122 tweets mentioning the deregulation argument. Our sample for dataset 1 contains 1,282 tweets. We sampled 25 per cent of newspaper articles containing each argument included in our analysis. This fraction is lower because newspaper articles are long and more expensive and time consuming to process. Our sample for dataset 2 contains 537 newspaper articles.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Methods

Our methodological approach consists of three steps. First, we designed the instruction (prompt) for ChatGPT for the API version (GPT-4 Turbo) and the user interface (GPT-4). Rather than writing a different prompt for each debate, we developed prompts that can be used for any policy debate. Next, we assess the performance of the model in terms of reliability and validity. Since GPT-4 turbo (API) is more capable and less error-prone than GPT-4 (OpenAI, 2024), we used GPT-4 Turbo as the reference model for our analyses and relied on the standard version of GPT-4 (user interface) as a complementary model.

Prompt engineering

ChatGPT's performance largely depends on the prompt and context provided. We formulated an extensive prompt for directed qualitative content analysis, which was developed by relying on a trial-and-error process. A rapidly expanding literature on prompt engineering has emerged to test which instructions are most effective across a variety of tasks. Broadly speaking, there are four core elements of prompt engineering: providing context, asking a question, setting model parameters, and providing output constraints. Specifying the context of a question is important to guide performance. Ekin (2023) notes that specifying a knowledge domain (e.g. Universal Basic Income) and a role for the chatbot (e.g. a social policy expert) improves performance. Clavié et al. (2023) find that adding a (any) name to the role of the bot also improves performance. Asking a question is best done in single sentences, breaking up complicated instructions into multiple prompts (Wu and Hu, 2023). In such multi-turn dialogue, Clavié et al. (2023) further point out that better results are obtained when asking whether the chatbot understood the instruction and providing positive feedback in between prompts. Moreover, they find that prompts that compare options A and B elicit deeper reasoning. Regarding parameters, Wu and Hu (2023) find that setting a lower temperature increases the focus and reduces the randomness of replies (see 'model parameters' section below). Similar advice for prompt engineering is provided by OpenAI, who advocate for similar strategies and tactics such as writing clear instructions, splitting complex tasks into simpler sub-tasks ('prompt chaining') or providing reference texts and examples.[2]

> Persona: You are a professional researcher named Jakub. You are an expert on qualitative content analysis. You are always focussed and rigorous.
>
> Task Description: Analyse [language] [document_type] for arguments related to [policy_name]. [policy_description]. The analysis will identify whether [document_type] contain arguments for or against [policy_name].

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

On the basis of these suggestions and several trial-and-error adjustments, we used the prompt below. First, we set the following system instructions:

> For each [document_type], provide a classification for each argument in an HTML table.
> Do not include the text of the [document_type] in the table. Only report the classification values.
> The HTML table has 5 rows, one per [document_type].
> The HTML table has 10 columns, one per argument.
> The elements of the table are '0' and '1'. Indicate '1' if the [document_type] discusses aspects of the specified argument and '0' is the [document_type] does not discuss the specific argument.
> Here is an example of the required output format:
> [example_output]

Then we entered the following messages in a prompt chain:

> Determine whether a [document_type] discusses each of the following ten arguments:
> [[arguments]]
> [document_type] contain an argument if the author opposes the argument and also when the author argues in favour of the argument.
> [policy_name] need not be mentioned explicitly in the [document_type] to relate to the argument. A [document_type] can discuss more than one argument.
> You will now be provided with 5 [document_type] separated by a new line.
> [[documents]]

Finally, we spent a substantial part of the prompt specifying the output format:

By entering values for the [variables] this prompt can be altered to any policy type and any number of arguments. The fully written prompt versions used for the two datasets can be found in Appendix B. The full script and other documentation (e.g. full prompt for user interface) will be available via the Open Science Framework.

Reliability

ChatGPT is a generative model – also known as probabilistic or non-deterministic – which means that answers can differ if asked in different chats or when asking the same question twice in the same chat. Therefore, before comparing the model classification with human annotations, we tested the reliability of the model by simply re-iterating the procedure $k = 5$ times and correlating the resulting vectors of zeros and ones. To evaluate the reliability of these replications, we calculate the phi correlation, a binary measure of association, for each unique combination of replications. The phi correlation is symmetrical, so we have $[k^*(k-1)]/2 = 10$

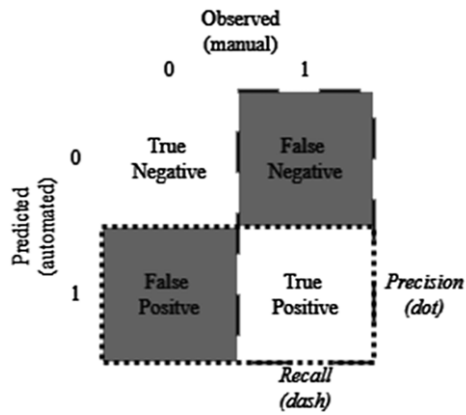
https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

unique chat combinations. The reliability of the chatbot can be inferred from the mean and variance of these correlations.

Validity

We evaluate the validity of the automated classification by comparing the assigned labels with human annotations. We rely on three popular and intuitive metrics to assess the performance compared to human annotation (e.g. Powers, 2020). The calculation of these metrics is visualised in Figure 1.

a) Accuracy equals the overall percentage of agreement between the chatbot and human coders. It is calculated by dividing the number of true positives and true negatives by the number of documents. It is important to emphasise that, in this context, accuracy means agreement with the human coder, who is also not flawless in interpreting true intents.
b) Precision reflects the amount of 'noise' in the documents classified as containing an argument. It is computed as the percentage of true positives amongst all chatbot positives. This is the same as the inverse false positive rate.
c) Recall shows how often the chatbot detects an argument in documents that contain this argument according to the human annotators. Recall is conceptually analogous to statistical power.



Note: Precision = TP / (TP+FP); Recall = TP / (TP+FN); Accuracy = (TN +TP) / (TN +TP + FN + FP)

Figure 1. Conceptual diagram of a confusion matrix.

Model parameters

Results are computed using the gpt-4-turbo-preview model. We set the model temperature to 20 per cent as suggested by, for example, Gilardi, Alizadeh and Kubli (2023, p.3). The temperature influences the 'randomness' of word predictions, so

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

that low temperatures are more likely to select words with high probability of occurring next in the sentence (e.g. Davis et al., 2024). In practice, low temperatures lead to more deterministic outcomes, sometimes referred to as more focussed and fact-based. For our purpose, setting a low temperature aids in getting a 'clean' response from the model, that is, a set of classifications without any additional text. Documents are supplied in 'batches' of five, meaning that we repeat the prompt with five new documents until all documents have been processed.[3]

Comparing approaches

We run the analysis with both the user interface (UI) and the application programming interface (API). The UI is an HTML-based website that can be run in a browser, and much like the Android application, the way in which people would normally interact with ChatGPT. The API provides back-end access to the model, typically used by application developers.[4] Via programming languages such as Python and R, we can send prompts directly to the servers of ChatGPT. The main advantage is the ability to automate requests, rather than inserting each batch manually. Additionally, the API has exclusive access to model parameters such as version, role and temperature settings. Thus far, no studies compared the performance of UI and API. This is unfortunate since the UI is relatively cost-effective (we used a Plus subscription) and much easier to use for those with little programming experience. The main disadvantage of the UI is the effort spent manually submitting input and extracting output as well as the message limit (at the time of the analysis: forty messages/3 hours). The API, on the other hand, allows for prompt-chaining (repeating a set of messages rather than sending one big message) and a flexible batch size (we supplied five documents per run). However, API access is technical to set up and can get expensive – especially for the later models – depending on the size of the corpus. At the time, one run of each dataset described above cost approximately 10 euros, at a rate of $0.01/$0.03 per 1 k input/output tokens. Below we discuss the difference in terms of performance of these access points.

Secondly, we implemented a 'cut-off' approach to try and reduce uncertainty in classifications. Wang et al. (2021) used a similar 'few-shot' approach where classifications (in this case logit estimates) are repeated, finding that GPT3 outperforms single repetitions in terms of accuracy. We repeated the estimation five times, obtaining five argument classifications for each document. A document is then classified as containing an argument when it is classified as such in at least 3/5 repetitions. This approach mimics the method typically employed in logistic regression, where an outcome is predicted to be present when the predicted probability is above 50 per cent.

Results

Figure 2 shows the reliability of ChatGPT in classifying arguments in our two datasets. Overall, the classifications are quite reliable. Without distinguishing between arguments, classifications are correlated with $\varphi = 0.71$ for newspaper articles and $\varphi = 0.84$ for tweets. Results are thus somewhat more reliable for (shorter) tweets than newspaper articles.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

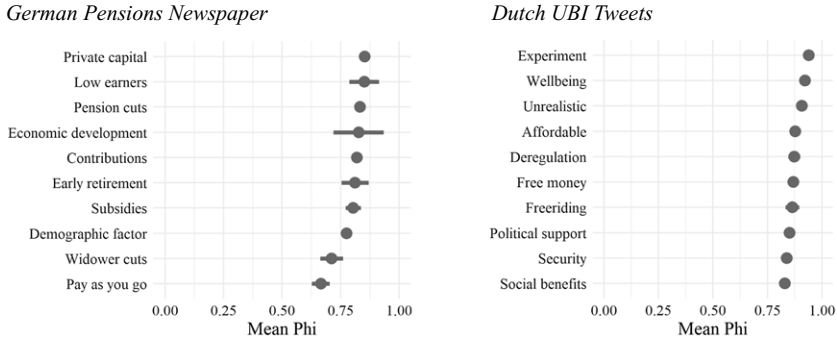


Figure 2. Average phi correlation between repeated ChatGPT classifications.

The lower reliability for newspaper articles is due to greater variation within and between arguments for newspaper articles. Results are consistently less reliable for the argument contributions, with an average correlation between runs of $\varphi = 0.67$. Within arguments we see that correlations differ within the argument economic development, with the lowest correspondence at $\varphi = 0.66$ and the highest at $\varphi = 0.95$. Especially for newspaper articles, then, we find that the chatbot may return somewhat different classifications between runs.

We now turn to the validity of the classifications in terms of accuracy and recall. The total average performance of the chatbot compared to human annotators is presented in Figure 2. For the newspaper dataset, 78.7 per cent of all bot classifications match the human annotations. For the tweets dataset, accuracy is even higher: 91.7 per cent of all classifications are in agreement between humans and the chatbot. For interpretability, we report the average accuracy between replications, which is a good indication given that variation between replications is less than 1 per cent. These accuracy scores are good, especially given the complexity of the task (Figure 3).

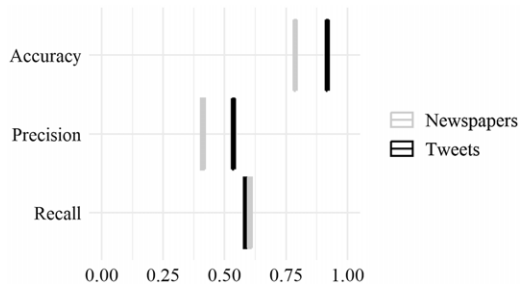


Figure 3. Total performance metrics.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Total precision estimates, however, are underwhelming. For newspaper articles we find a precision rate of 41.3 per cent, meaning that 58.7 per cent of articles are false positives: they contain an argument according to the chatbot but not the human annotators. Classification of tweets is slightly less noisy: precision is 58.6 per cent. If we again take human annotations as true positives, then 41.4 per cent of chatbot classifications are false positives or noise. If this is indeed the case, this shows that argument classifications from ChatGPT must be cautiously used.

Total recall values are not impressive but acceptable. For the newspaper dataset, we find that 60.1 per cent of all occurrences found by humans are also identified by the chatbot. Inversely, this means that 39.9 per cent of bot-identified arguments are false negatives, found by humans but not by the chatbot. Similarly, for the tweets dataset, we find that 58.6 per cent of arguments found by humans are also detected by the chatbot, with a corresponding false negative rate of 41.4 per cent. These total recall values illustrate that, while accuracy may be high, ChatGPT often misses occurrences of arguments that are found by humans. This may be due to the complexity of the task, and the limited context provided in the argument description. Still, a detection rate of 60 per cent would be acceptable in some cases, especially when these are the most obvious occurrences.

Moreover, a part of the recall problem is due to considerable variation between arguments (see Figure 4). For both datasets, we find that some arguments are much more easily detected than others. For the arguments private capital and freeriding, for example, 80.3 per cent and 80.8 per cent, respectively, of human annotations are identified by the chatbot. In contrast, for economic development and security, only 24 per cent and 39.3 per cent of human labels are identified, respectively. We suspect this variation may be due to complex statements that refer to the argument, and perhaps due to the difficulty of devising a good argument description.

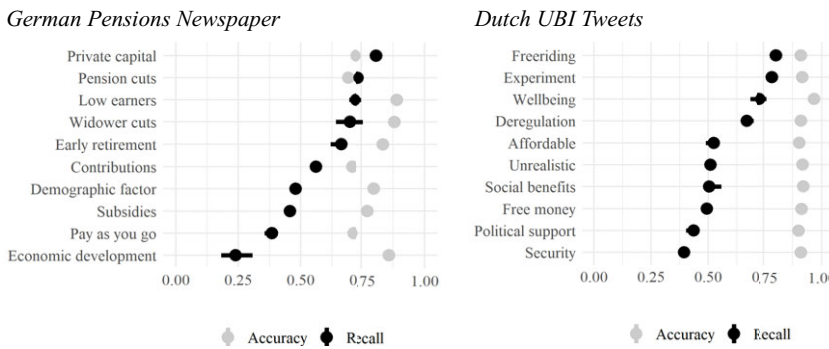


Figure 4. Average accuracy and recall scores based on human-annotated documents ($k = 5$).

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

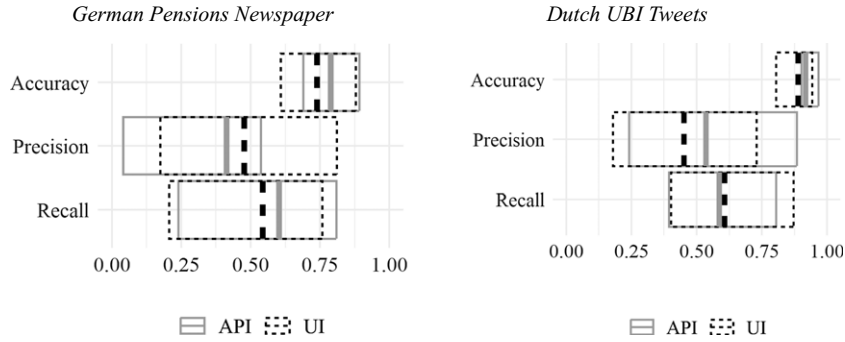


Figure 5. Comparison of performance between UI and API approaches.
Note: The range of the crossbars correspond to the best- and worst-performing argument per metric.

Comparing approaches

At first sight, the user interface (UI) performs slightly worse than the average API run, despite identical argument descriptions and near-identical prompts. Bear in mind, however, that the comparison is somewhat unstable because we did not perform multiple runs for the user interface. Multiple runs for the UI approach take time and effort but would reduce the degree of random variation in results. However, given that variation between runs is limited (as shown in Figure 1), we still believe this comparison is valid when it concerns large differences in performance (say, more than 5 percentage points).

Figure 5 shows the total accuracy, precision and recall per dataset. The upper and lower crossbars indicate the range in performance between arguments. Total accuracy is comparable between the API and UI methods, for both datasets. Precision provides a mixed signal for performance between methods: the API yields more precise estimates on the tweets data (a difference of 8.5 percentage points), but less precise estimates on the newspaper data (a difference of 6.4 percentage points). Recall is somewhat better in the average API run compared with the UI run, but only for the newspaper dataset. Overall, we conclude that the API performs roughly similar to the UI.

Finally, we evaluate what we have called the cut-off method, where an argument is deemed present or absent when 3/5 repetitions agree. Interestingly, the method introduces a trade-off between precision and recall compared with the average API run. Accuracy is practically equivalent between the two methods, for both newspapers and tweets. In both datasets, especially in the newspaper dataset, the cutoff approach substantially increases precision at the cost of reducing recall. In other words, using a cut-off value decreases the level of noise in the classification (i.e. fewer false positives) but loses detection power (i.e. more false negatives). In both datasets, the gain in precision almost exactly matches the loss in recall. The fact that this happens in both datasets suggests that this is not a matter of coincidence. With this method, researchers have the option to reduce noise in the estimates when it is deemed more important than detecting all arguments present (Figure 6).

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

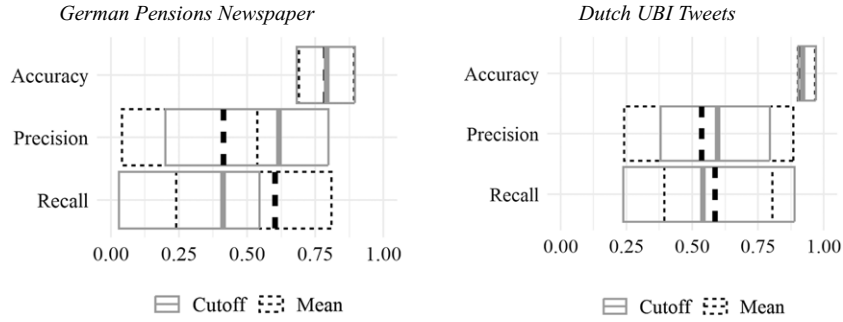


Figure 6. Performance of cut-off method compared with average API.
Note: The range of the crossbars correspond with the best- and worst-performing argument per metric.

Discussion

This study evaluated the performance of ChatGPT (GPT-4 Turbo and GPT-4) as a tool to perform directed content analysis for policy debates on two very different data sources: Dutch tweets on Universal Basic Income and German newspaper articles on the Riester pension reform. Our results show relatively high levels of accuracy and reliability for the tweets dataset and – to a slightly lesser extent – for the newspaper articles. However, there are three main concerns when using ChatGPT to automate content analysis.

First, while overall results are positive, we show that bot-labelled argument occurrences contain a fair amount of noise (precision) and fail to detect a good number of human-labelled occurrences (recall). The primary reason for the discrepancy between accuracy versus precision and recall is that the occurrence of arguments, much like the occurrence of hate speech, is naturally 'imbalanced'. There are bound to be more documents that do not contain one specific argument than documents that do contain that argument, a situation that is only exacerbated when the number of arguments increases. Even the most adopted arguments in the UBI debate, for example, only occur in around 1–2 per cent of all tweets under investigation. Since accuracy measures the correct classification of both occurrences and non-occurrences, without any discrimination between the two, the value is highly determined by ChatGPT's ability to correctly classify non-occurrences or 'true negatives'. Precision and recall are better suited to evaluate the model's ability to identify arguments' occurrences, because they essentially disregard the correct identification of non-occurrences (see methods section for details). To avoid an overly optimistic evaluation of the method (Gilardi, Alizadeh and Kubli, 2023; Huang, Kwak and An, 2023; Wang et al., 2021; cf. Ziems et al., 2023), we therefore argue that it is vital to include statistics on precision and recall when validating such classification tasks.

Second, since performance varies strongly between arguments, we suggest that the method is better suited for arguments with a clear description and well-defined associated vocabulary in the documents it intends to classify. Interestingly, most performance indicators turned out to be better for the Twitter data compared with the newspaper articles. This is insofar surprising as the 'messiness' of tweets and the well-crafted nature of newspaper articles could lead one to expect the opposite.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Given that both original texts were not English, we conclude that rather than the nature of the text itself, context length might be a bigger problem for ChatGPT. Better results might thus be obtained by dividing newspaper articles and other policy documents into more 'digestible' chunks for ChatGPT. Alternatively, however, the variance in performance between arguments may also be caused by human mistakes in the coding process.

Third, when relying on language models, manual labelling of a subset of the debate will still be required to (a) identify arguments and their descriptions to establish a codebook and (b) validate the classifications generated by the chatbot. When more general topics suffice, a researcher may consider using topic models to find the most important arguments in the discussion. When arguments' definitions are available and validation is of little concern, language model classification could be directly applied. For most applications, however, we recommend manually annotating a subset of the data. This subset can be used to build a grounded coding scheme and to evaluate the performance of the chatbot using the same techniques as elaborated in this article. For those interested in applying and developing this method, we have published the scripts and tweets data on the Open Science Framework for public use.[5]

So, does ChatGPT herald the end of human annotators? While it is tempting to make such grand statements, the fair answer is no. First, humans are (at least for now with the current models) better at doing content analysis than the most prominent large language model. Nonetheless, there is reason to believe that the models will continue to improve. Like Ziems et al. (2023, p. 21), we observed a substantial improvement in performance between models GPT4-Turbo (API) and GPT4 (UI). Moreover, improvements were even more substantial between GPT4-Turbo (API) and GPT3.5-Turbo (API), especially in terms of precision (see Appendix C). In turn, this suggests that future models may equal or even surpass the current golden standard of human annotators. However, the realisation of such improvements will depend on the model's future ability to prevent collapse when trained on recursively trained data (Shumailov et al., 2024).

Like any other, this study also has its limitations, which point to avenues for future research. Our approach is limited to the automated application of a pre-defined coding scheme to identify arguments. The method is not supposed to generate codebooks or identify arguments without some prior definition. This also means that the challenge of automatically identifying actor positions towards arguments – that is, whether they agree or disagree with an argument – remains to be addressed in future research. We did not compare performance with methods such as topic modelling, specialised LLMs or competitors of ChatGPT. Haunss et al. (2020) and Lapesa et al. (2020) employed natural language processing (NLP) methods combining transformers and recurrent neural networks to predict arguments for policy discourse analysis. Ceron et al. (2024) found comparable performance to these methods to the results reported here. While prior studies seem to suggest that GPT4 performance for political texts is good, training and evaluating a dedicated language model for arguments in policy debates remains a valuable course of action.

We hope this study will serve as a starting point, providing the baseline results against which future strategies for improvement can be assessed. Several strategies

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

to further improve the performance of ChatGPT classifications are imaginable. First, variance of performance between arguments may indicate not just different levels of intrinsic difficulty of arguments, but potentially also heterogeneity in the set of descriptions. While we refrained from extensively tweaking prompts and argument descriptions to avoid 'overfitting' descriptions on the data, more detailed attention to prompt design and argument descriptions may further enhance performance in the future. One approach is 'template refinement', that is, reducing the number of arguments by grouping them into clusters (King et al., 2018; Nguyen-Trung, 2024). We also look to prompt engineering – that is, the specific formulation of instructions – a rapidly developing field seeking to optimise LLM performance through prompting techniques (e.g. Clavié et al., 2023; for an application to classification tasks, see Thomas et al., 2023). Fine-tuning models may also provide further improvements to classifications. Wang et al. (2021) found that fine-tuned LLMs outperform the basic GPT3 model. Ziems et al. (2023) found that fine-tuned classifiers outperform GPT4 on some types of text (e.g. misinformation) but not political texts on ideology and stance. Considering the drastic improvements between GPT3 and GPT4, however, the main performance gains are likely to result from model updates.

Second, humans and LLMs are both imperfect in establishing the true intent of authors, and thus in establishing the ground truth. One way to reduce noise in classifications is to examine the misclassified documents, correct any human mistakes in the coding and rerun the classification procedure (e.g. Nguyen-Trung, 2024). This procedure provides insight into which arguments and texts are hard to classify, and simultaneously bolsters the model performance. The extent to which these improvements are effective depends on the number of flaws in human coding. The improvements in performance gained by applying this technique remains to be investigated.

Third, while the comparison of two datasets shows the general ability of ChatGPT to classify policy debates, the differences in performance between datasets may be attributed to a range of factors. The two datasets differ in the type of argument, the political context, the length of the text, the language and the sampled fraction of documents, as well as potentially the technical complexity of the policy issue. A comparison with more cases is needed, however, to isolate the exact sources of the variation in results. Future research should therefore investigate which of these differences explain the differential performance.

As a final remark, while we see future potential, caution is warranted. Data privacy concerns as well as possible climate impacts (e.g. water usage; see Li et al., 2024) and potential political bias (McGee, 2023; Rozado, 2023) must not be neglected when using ChatGPT. Overall, however, we hope that future improvements in abilities will consolidate large language models as an important tool for social scientists.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0047279424000382

Competing interests. The author(s) declare none.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Notes

1 https://chat.openai.com/ (accessed 28 August 2024).
2 https://platform.openai.com/docs/guides/prompt-engineering.
3 Batches of five documents proved to be more reliable than batches with more documents when accessing ChatGPT via the user interface (GPT-4), often leading to generative errors. Generative errors should occur less frequently via the API access, as GPT-4 Turbo is more resistant to such errors (OpenAI, 2024b). For reasons of comparability, batches of five documents were also used in the API version, although we assume that a higher batch number is possible there.
4 https://platform.openai.com/docs/api-reference/introduction.
5 http://doi.org/10.17605/OSF.IO/MR5D2.

Bibliography

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E., & Zayed, T. (2023). Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. *Systems*, **11**(7), 351.

Béland, D. (2019). *How ideas and institutions shape the politics of public policy*. Cambridge University Press.

Blum, S., & Kuhlmann, J. (2019). Stories of how to give or take–towards a typology of social policy reform narratives. *Policy and Society*, **38**(3), 339–355.

Ceron, T., Barić, A., Blessing, A., Haunss, S., Kuhn, J., Lapesa, G., Padó, S., Papay, S., & Zauchner, P. F. (2024). Automatic analysis of political debates and manifestos: successes and challenges. In R. Goebel, W. Wahlster & Z. Zhou (Eds.), *Conference on advances in robust argumentation machines* (pp. 71–88). Springer Nature Switzerland.

Clavié, B., Ciceu, A., Naylor, F., Soulié, G., & Brightwell, T. (2023). Large language models in the workplace: a case study on prompt engineering for job type classification. In E. Metais, F. Meziane, V. Sugumaran, W. Manning & S. Reiff-Margianiec (Eds.), *International conference on applications of natural language to information systems* (pp. 3–17). Springer Nature Switzerland.

Davis, J., Van Bulck, L., Durieux, B. N., & Lindvall, C. (2024). The temperature feature of ChatGPT: modifying creativity for clinical research. *JMIR Human Factors*, **11**(1), e53559.

Duarte, N., Llanso, E., & Loup, A. (2017). Mixed messages? The limits of automated social media content analysis. Center for Democracy and Technology. https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf

Eder, F. (2023). Discourse network analysis. In P. Mello & F. Ostermann (Eds.), *Routledge handbook of foreign policy analysis methods* (pp. 516–535). Taylor & Francis.

Ekin, S. (2023). Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Authorea Preprints*.

Fergie, G., Leifeld, P., Hawkins, B., & Hilton, S. (2019). Mapping discourse coalitions in the minimum unit pricing for alcohol debate: a discourse network analysis of UK newspaper coverage. *Addiction*, **114** (4), 741–753.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. (2nd ed.). John Wiley.

Gielens, E., Roosma, F., & Achterberg, P. (2022). More than a free lunch: a content analysis of the controversies surrounding universal basic income on Dutch Twitter. *Social Policy and Society*, 1–21.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, **120**(30), e2305016120.

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.

Gupta, K., Ripberger, J., Fox, A., Jenkins-Smith, H., & Silva, C. (2022). Discourse network analysis of nuclear narratives. In M. Jones, M. McBeth & E. Shanahan (Eds.), *Narratives and the Policy Process: Applications of the Narrative Policy Framework* (pp. 13–39). Montana State University Library.

Gutierrez Garzon, A. R., Bettinger, P., Abrams, J., Siry, J. P., & Mei, B. (2022). Forest sustainability in state forest management plans: a content analysis. *Journal of Sustainable Forestry*, **41**(1), 92–113.

Haunss, S., Kuhn, J., Padó, S., Blessing, A., Blokker, N., Dayanik, E., & Lapesa, G. (2020). Integrating manual and automatic annotation for the creation of discourse network data sets. *Politics and Governance*, **8**(2), 326–339.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Horn, A. (2019). Can the online crowd match real expert judgments? How task complexity and coder location affect the validity of crowd-coded data. *European Journal of Political Research*, **58**(1), 236–247.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, **15**(9), 1277–1288.

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Jensen, C., & Wenzelburger, G. (2021). Welfare state reforms and mass media attention: Evidence from three European democracies. *European Journal of Political Research*, **60**(4), 914–933.

Juba, B., & Le, H. S. (2019). Precision-Recall versus Accuracy and the Role of Large Data Sets. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 4039–4048.

King, N., Brooks, J., & Tabari, S. (2018). Template analysis in business and management research. In M. Ciesielska & D. Jemielniak (eds.), *Qualitative methodologies in organization studies* (pp. 179–206). Springer.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: a survey. *Information*, **10**(4), 150.

Lapesa, G., Blessing, A., Blokker, N., Dayanik, E., Haunss, S., Kuhn, J., & Padó, S. (2020). Analysis of political debates through newspaper reports: methods and outcomes. *Datenbank-Spektrum*, **20**, 143–153.

Laurer, M., Van 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.

Leifeld, P. (2013). Reconceptualizing major policy change in the advocacy coalition framework: a discourse network analysis of German pension politics. *Policy studies journal*, **41**(1), 169–198.

Leifeld, P. (2016). *Policy debates as dynamic networks: German pension politics and privatization discourse* (Vol. **29**). Campus Verlag.

Li, J., Dada, A., Puladi, B., Kleesiek, J., & Egger, J. (2024). ChatGPT in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, **245**, 1–12.

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, **31**(3), 366–379.

Markard, J., Rinscheid, A., & Widdel, L. (2021). Analyzing transitions through the lens of discourse networks: coal phase-out in Germany. *Environmental Innovation and Societal Transitions*, **40**, 315–331.

McGee, R. W. (2023). Is chat gpt biased against conservatives? An empirical study. *An Empirical Study (February 15, 2023)*.

Miller, B., Linder, F., & Mebane, W. R. (2020). Active learning approaches for labeling text: review and assessment of the performance of active learning approaches. *Political Analysis*, **28**(4), 532–551.

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. *International Journal of Qualitative Methods*, **22**, Forthcoming.

Nguyen-Trung, K. (2024). ChatGPT in thematic analysis: can AI become a research assistant in qualitative research? *OSF Preprint*. https://doi.org/10.31219/osf.io/vefwc.

OpenAI (2024). ChatGPT Technical Report. https://doi.org/10.48550/arXiv.2303.08774.

Petchler, R., & González-Bailon, S. (2015). Automated content analysis of online political communication. In *Handbook of digital politics* (pp. 433–450). Edward Elgar Publishing.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Prakash, N., Wang, H., Hoang, N. K., Hee, M. S., & Lee, R. K. W. (2023). PromptMTopic: unsupervised multimodal topic modeling of memes using large language models. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23). Association for Computing Machinery, New York, NY, USA (pp. 621–631).

Prior, L., Hughes, D., & Peckham, S. (2012). The discursive turn in policy analysis and the validation of policy stories. *Journal of Social Policy*, **41**(2), 271–289.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, **28**(1), 112–133.

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, **12**(3), 148.

Schmidt, V. A. (2008). Discursive institutionalism: the explanatory power of ideas and discourse. *Annual Review of Political Science*, **11**, 303–326.

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, **631**(8022), 755–759.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, **52**(3), 705–722.

Symeonidis, V., Francesconi, D., & Agostini, E. (2021). The EU's education policy response to the Covid-19 pandemic: a discourse and content analysis. *CEPS Journal*, **11**(Special Issue), 89–115.

Theiss, M. (2023). How does the content of deservingness criteria differ for more and less deserving target groups? An analysis of polish online debates on refugees and families with children. *Journal of Social Policy*, **52**(4), 962–980.

Thomas, P., Spielman, S., Craswell, N., & Mitra, B. (2023). Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*.

Towns, W., & Henstra, D. (2018). Federal policy ideas and involvement in Canadian urban transit, 2002–2017. *Canadian Public Administration*, **61**(1), 65–90.

Turobov, A., Coyle, D., & Harding, V. (2024). Using ChatGPT for thematic analysis. *arXiv preprint arXiv:2405.08828*.

Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). Want to reduce labeling cost? GPT-3 can help. *arXiv preprint arXiv:2108.13487*.

Widmann, T., & Wich, M. (2023). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*, **31**(4), 626–641.

Willems, E., & Beyers, J. (2023). Public support and advocacy success across the legislative process. *European Journal of Political Research*, **62**(2), 397–421.

Wu, Y., & Hu, G. (2023). Exploring prompt engineering with GPT language models for document-level machine translation: insights and findings. In Proceedings of the Eighth Conference on Machine Translation, Singapore. Association for Computational Linguistics. (pp. 166–169).

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

Cite this article: Gielens, E., Sowula, J., and Leifeld, P. (2026) Goodbye human annotators? Content analysis of social policy debates using ChatGPT. *Journal of Social Policy* **55**, 385–404. https://doi.org/10.1017/S0047279424000382

https://doi.org/10.1017/S0047279424000382 Published online by Cambridge University Press