Googling politics? Comparing five computational methods to identify political and news-related searches from web browser histories

van Hoof, M.; Trilling, D.; Meppelink, C.; Möller, J.; Loecherbach, F.

DOI
10.1080/19312458.2024.2363776

Publication date
2025

Document Version
Final published version

Published in
Communication Methods and Measures

COMMUNICATION METHODS AND MEASURES
2025, VOL. 19, NO. 1, 63–89
https://doi.org/10.1080/19312458.2024.2363776

Routledge
Taylor & Francis Group

Googling Politics? Comparing Five Computational Methods to Identify Political and News-related Searches from Web Browser Histories

Marieke van Hoof[a], Damian Trilling[b], Corine Meppelink[a], Judith Möller[c], and Felicia Loecherbach[a]

[a]Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, Netherlands; [b]Language, Literature and Communication, Vrije Universiteit Amsterdam, Amsterdam, Netherlands; [c]Communication Science, University of Hamburg – Leibniz Institute for Media Research,Hans Bredow Institute, Hamburg, Germany

ABSTRACT

Search engines play a crucial role in today's information environment. Yet, political and news-related (PNR) search engine use remains understudied, mainly due to the lack of suitable measurement methods to identify PNR searches. Existing research focuses on specific events, topics, or news articles, neglecting the broader scope of PNR search. Furthermore, self-reporting issues have led researchers to use browsing history data, but scalable methods for analyzing such data are limited. This paper addresses these gaps by comparing five computational methods to identify PNR searches in browsing data, including browsing sequences, context-enhanced dictionary, Traditional Supervised Machine Learning (SML), Transformer-based SML, and zero-shot classification. Using Dutch Google searches as a test case, we use Dutch browsing history data obtained via data donations in May 2022 linked to surveys (N_{users} = 315; $N_{records}$ = 9,868,209; $N_{searches}$ = 697,359), along with 35.5k manually annotated search terms. The findings highlight substantial variation in accuracy, with some methods being more suited for narrower topics. We recommend a two-step approach, applying zero-shot classification followed by human evaluation. This methodology can inform future empirical research on PNR search engine use.

Introduction

Healthy democracies rely on an informed public that is able to make informed political decisions (Helberger et al., 2018). In today's media landscape, there are countless options for obtaining information, including social media, news aggregators, and search engines. Search engines, in particular, are known to play a crucial role in shaping the public's access to news (Arendt & Fawzi, 2019; J. Möller et al., 2020; Wojcieszak et al., 2022). Furthermore, search results can impact voter preferences (Epstein & Robertson, 2015) and are widely trusted (Dutton et al., 2017; Haas & Unkel, 2017; Pan et al., 2007; Unkel & Haas, 2017). However, public and academic concerns about the impact of search engines on the representation of information (Pradel, 2021; Ulloa et al., 2022; Urman & Makhortykh, 2022; Urman, Makhortykh, & Ulloa, 2022), the spread of misinformation (Urman, Makhortykh, Ulloa, & Kulshrestha, 2022) and the creation of filter bubbles (Pariser, 2011), call into question the extent to which this trust is justified. These concerns become more pressing given

CONTACT Marieke van Hoof m.vanhoof@uva.nl Amsterdam School of Communication Research, University of Amsterdam, P.O. Box 15791, Amsterdam 1001 NG, Netherlands

This article has been republished with minor changes. These changes do not impact the academic content of the article.

developments set to revolutionize information consumption, like AI-generated answers on Google's search result pages (Google, 2023). In the near future, search engines will not only be gateways to other websites but also primary information sources.

Consequently, we need to understand how, how frequently, and who is using search engines to inform themselves about politics and current affairs. Yet, our current knowledge of the use of search engines for political and news-related (PNR) information-seeking is limited to specific political events (e.g., elections, referendums) (Blassnig et al., 2023; Trielli & Diakopoulos, 2022) and issues (e.g., climate, immigration) (van Hoof et al., 2022; Wang & Jaidka, 2024), and news articles selected via search results (e.g., J. Möller et al., 2020; Robertson et al., 2023; Vermeer et al., 2020; Wojcieszak et al., 2022). While considerable research on key concepts like political interest, knowledge and news reading exists, an equivalent broad conceptualization for PNR search remains largely missing. Existing research with a broader perspective is confined to a small set of respondents from the United States (Menchen-Trevino et al., 2023).

Furthermore, to combat self-reporting issues in our complex and fast-paced information environment, researchers have begun to use digital traces, such as browsing history data, offering an ecologically valid account of search behavior. Nonetheless, limited work has proposed methodologies for studying such data. Browsing data has yielded valuable in-depth insights into PNR search behavior, as demonstrated by Menchen-Trevino et al. (2023). However, their manual approach, while able to cover PNR search more comprehensively, struggles with scalability when applied to larger populations or longer time periods.

Studies on specific cases or small sets of individuals offer valuable insights, but to fully understand the role of search engines in the public's political information and news routines, we require a method to distinguishing PNR searches from unrelated ones in search history data in a comprehensive and scalable manner. By developing such automated methods and discussing their conceptual and methodological advantages and limitations, we hope to foster empirical research that is able to answer questions like those posed above. We conduct a comparison of five computational methods, i.e. browsing sequences, context-enhanced dictionary, traditional supervised machine learning, transformer-based supervised machine learning, and zero-shot classification.

We use browsing history data obtained in May 2022 via data donations linked to surveys with sociodemographic, political and news-related characteristics ($N_{users} = 315$; $N_{records} = 9{,}868{,}209$; $N_{searches} = 697{,}359$) as well as a sample of 35.5k manually annotated search terms. We use Dutch Google searches about politics and news as a test case, but our insights should apply across multiple languages and search engines.

Theoretical framework and related research

We define PNR (i.e., politics and news-related) search as seeking information contributing to opinion formation on political and societal topics, which aligns with the conceptualization of "political searches" by Menchen-Trevino et al. (2023) that refers to searches for issues of public concern. This is an intentionally broad conceptualization, which includes searches for political entities (e.g., political parties, institutions, politicians), policy areas (e.g., taxes, education, health care), societal topics (e.g., climate, immigration, racism), but also news events of political or societal relevance. It excludes, for instance, searches for sports, entertainment and accidents (see Appendix D for details). This definition distinguishes searches of *public* concern (e.g., "news unemployment benefits") from queries for issues of *personal* concern (e.g., "where to apply for unemployment benefits").

Furthermore, by definition, PNR search includes information-seeking for topics that are relatively stable over time, such as political entities, societal topics, and policy areas (e.g., "when are the next elections," "immigration"), but also refers to searches that are only, or particularly relevant at the time of searching. News-related search terms are influenced by situational factors that trigger an information need, such as key mediated political events (Arendt & Fawzi, 2019; Trevisan et al., 2018). Such context-dependent search terms can only be understood in tandem with the political and news context

in which they are used. For example, a search for a foreign city would evolve from a personal search to a PNR search after the city is invaded by another country. The same applies for short-term news events, such as a search for a relatively unknown village where a political party celebrated their election victory, the search would be considered PNR only during the election. Consequently, methods identifying PNR search queries should ideally consider timing in their application.

Classifying search queries presents a unique challenge compared to other textual content analyzed in communication science, such as news articles or social media posts. Search terms typically consist of a few words only, lack regular sentence structure (keyword-like), and lack contextualizing metadata. As a result, search queries can be *ambiguous*, where it may be unclear solely from the text whether the query was aimed to retrieve PNR information (e.g. "unemployment benefits"). This ambiguity poses a challenge for human annotation, but also for automated procedures, as such distinction is often expressed in subtle textual nuances (if discernible at all). For instance, "vaccination rates 2021" and "vaccination location amsterdam" share keywords, but their different goals are clear. Note that there is a distinction between how we believe PNR search is best conceptualized (i.e., as a characteristic of the user), and how in many study designs, including ours, we are able to methodologically approach it (i.e., as a characteristic of a search query, without knowing the true motivation behind the query).

Furthermore, the vast volume of online searches and the infrequency of PNR search render fully manual methods impossible if one wants to scale this research beyond small case studies (Menchen-Trevino et al., 2023). Hence, an automated approach is needed. Automated methods vary in resource requirements, including the need for additional data sources (i.e. annotated data, news content, human validation) and computing power (i.e. requiring GPUs). The feasibility and reproducibility of resource-intensive approaches may be limited, especially when applied across different contexts, time periods and languages. This is particularly relevant given the growing need for comparative and longitudinal research in communication science.

In the following sections, we present an overview of the possible approaches for identifying PNR searches and discuss their (dis)advantages.

Browsing sequences

Browsing data has been used in news referrals research by utilizing the sequence of browsing to identify how people access news websites, including through search engines (Cardenal et al., 2022; Guess et al., 2020; J. Möller et al., 2020; Vermeer et al., 2020; Wojcieszak et al., 2022). The news websites are typically identified with a list of news sources. Similarly, we can use browsing patterns to label searches as PNR if the search led to the selection of a news website. However, extending this technique to distinguish PNR searches hinges on the assumption that the majority of such searches lead to a visit to a well-known news outlet and therefore misses other plausible outcomes.

Firstly, search engines have evolved beyond being a mere gateway to other websites, as they now include features aimed at providing direct answers as standard parts of the search result page (Oliveira & Lopes, 2023). One such feature, the Knowledge Panel, appears in 69% of political searches in the US (Robertson et al., 2018). Leading search engines Google and Bing accelerated this development by introducing AI-generated answers and chatbot functionalities (e.g., Google's Gemini, Microsoft's Copilot). As a result, users may not always select a website, including for political search queries (Blassnig et al., 2023; Menchen-Trevino et al., 2023), possibly because the search result page informed them sufficiently.

Secondly, if users do select a website, they may select websites that are excluded from the pre-defined list of news(−related) websites, such as niche alternative and local news websites or blogs. Most studies comprise such a list of larger and national (news) outlets, using Alexa top lists, annotating top domains in their sample or different quantiles of the data (see, e.g. Stier, Kirkizh, et al., 2020; Wojcieszak et al., 2022). This is informative to calculate a share of consumed news in browsing histories, since the distributions of website visits follow the power law and relatively few websites are missed. Yet, for sequence-based methods missing out on many news-related domains reduces

accuracy. Furthermore, this approach may incorrectly label search terms as PNR because news websites also contain content on sports, entertainment and the weather. In this vein, Reiss (2023a) has shown that this list-based approach incorrectly labels non-news content as news-related, and, to a lesser extent, misses news(related) content on non-news websites. Additionally, these lists are slow to adapt to new sources of information coming up over time and remain in constant need of updating.

Using the sequence of browsing to label search queries is straightforward, scalable, and requires few additional resources. Notably, it avoids issues with the ambiguity and context-sensitivity of search queries since it infers meaning based on browsing patterns. This, however, likely comes at the expense of a crude operationalization of PNR search: it may *underestimate* PNR search and be *biased* toward well-known websites. Furthermore, establishing a referral is challenging due to non-linear browsing (e.g., multi-tabbing), resulting in disconnected browsing patterns. In this emerging field, there is no single validated approach; approaches vary, for instance, in the time thresholds between website visits, and in their use of URL parameter information (e.g., Wojcieszak et al., 2022).

Context-enhanced dictionary

For the aforementioned factors, it is useful to consider content-based approaches. Dictionary approaches are commonly used for text classification tasks. This approach involves creating a list of words that serve as indicators of a PNR search, with the list typically being generated by researchers. A search is considered PNR if the search query matches with words in the dictionary.

Although dictionary approaches have benefits, such as their simplicity, transparency and limited resource requirements, it is not straightforward to apply them to PNR search terms. Differentiating between personal and PNR queries requires an understanding of the context, meaning that simple textual overlap may not be enough. Additionally, it can be challenging to create an exhaustive dictionary that covers all aspects of a concept. Therefore, dictionaries are usually designed for specific cases, such as searches for a COVID-19 referendum (Blassnig et al., 2023), but fail to capture broader concepts (see Van Atteveldt et al., 2022) like PNR search. Specifically for search, queries are short and may not include variations of similar words like news content would do, which can make it easier to miss with a dictionary. As different users use distinct search terms to express comparable information needs (van Hoof et al., 2022), dictionary approaches may fail to capture searches conducted by particular groups of people.

Moreover, dictionary approaches seem reasonable for identifying a stable set of search terms that are always related to politics and news, but capturing context-dependent search for current events and politics is more challenging. One solution is to complement a base dictionary with a rolling set of dictionaries consisting of entities (e.g., people, locations, . . .) mentioned in news content at the time of searching. We call this combination of dictionaries the context-enhanced dictionary approach. While we cannot determine if news-related searches are directly *influenced* by news media, it is theoretically plausible that news content and search terms at least share keywords about the key elements of current events. This advantage, however, necessitates a news corpus that covers the entire search data, which may not be feasible to obtain, especially for longer time periods and multiple languages.

Traditional supervised machine learning

Importantly, dictionary approaches have often been outperformed by more sophisticated methods based on Supervised Machine Learning (SML) (e.g., Kroon et al., 2022; Lind et al., 2021; Van Atteveldt et al., 2021). Unlike dictionaries, SML approaches take annotated documents as examples from which patterns are learned and subsequently applied to new, unseen data. In doing so, it takes a bottom-up approach and can identify patterns that researchers may not be able to capture. It should therefore better recognize the (subtle) differences between queries for personal and PNR information, and the varied language used by searchers.

Despite SML approaches typically being a significant improvement over dictionary approaches, they have drawbacks. Their effectiveness depends on the annotated data they have been trained on to be large and diverse enough to capture all aspects of PNR search. This is a considerable challenge given the sparseness of search terms and the breadth of the concept. Additionally, PNR information-seeking is infrequent, making it difficult to obtain a large sample size for annotation while still finding enough useful queries. Furthermore, as SML is only able to capture what is in the training data, it is also less adaptable to other contexts (e.g., time periods and languages). Traditional SML techniques also consider documents as collections of words and therefore disregards informative features like word order and semantics.

Transformer-based supervised machine learning

To overcome drawbacks of dictionary and traditional SML, researchers have recently successfully applied sophisticated Large Language Models (LLMs), many of which are built on the Transformer architecture, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and GPT-3 (Brown et al., 2020). LLM methods have shown to outperform other methods for classifying related concepts like news genres in news articles (Lin et al., 2023), sentiment in news headlines (Van Atteveldt et al., 2021), and PNR content in web content (Makhortykh et al., 2022) and website titles (Wojcieszak et al., 2024).

LLMs are pre-trained on large collections of text to learn relationships between words and their surrounding context, allowing them to obtain a general understanding of language. One of the most popular LLMs is BERT, which is a language model developed by Google (Devlin et al., 2018). By fine-tuning a BERT model with a much smaller set of annotated examples compared to Traditional SML, it updates its parameters to be better able to understand PNR search terms while retaining its general understanding of language. This enables a transformer-based classifier to understand the semantic meaning of search terms, even when these terms are not represented in the annotations. For instance, even if the training data does not include a search for "*Joe Biden*," these models may still be able to understand its political meaning. While sparse search terms may still pose some challenge for transformer-based SML, LLM approaches are expected to better understand searcher's intent expressed in (subtle) textual differences, if such intent is evident from the wording.

This increase in performance requires sacrificing simplicity and transparency, as well as needs large computational resources, typically requiring GPU access for fine-tuning. Despite the ability of LLMs to understand a wide range of words, the data they are trained on is critical, as they cannot recognize PNR concepts emerged beyond their training date. Moreover, if the pre-training data is biased toward certain topics or perspectives (Bender et al., 2021), the classifier may be less accurate at identifying PNR searches outside of those biases.

Furthermore, obtaining the amount of annotated data necessary for additional training is a challenge for both types of SML, significantly reducing their applicability, particularly for applications across languages (Lind et al., 2021) and time periods.

Zero-shot classification

The recent availability of powerful LLMs like OpenAI's GPT-3.5 and GPT-4 (Brown et al., 2020) has made it easy for researchers to use zero-shot classification to circumvent the need for annotated data. Instead of fine-tuning a pre-trained LLM using a SML approach, this method involves directly instructing the model to classify texts without fine-tuning. While it shares disadvantages with SML, zero-shot classification's major advantage is that it, in theory, eliminates the need for annotations, which facilitates comparisons across languages (Kuzman et al., 2023b). Recent papers and preprints demonstrated the accuracy of zero-shot text

classification (using GPT-3.5 or GPT-4) for complex topics like hate speech (Huang et al., 2023), genre identification (Kuzman et al., 2023b), fact-checking (Hoes et al., 2023), and political affiliation (Törnberg, 2023), in some cases outperforming (trained) human coders (Gilardi et al., 2023; Huang et al., 2023). Notably, such models are designed for text generation and are trained to understand prompts, making them potentially very suitable to infer user intent in cases such as search queries. We can also leverage the opportunity for elaborate prompting to instruct models with dates to contextualize the search term within a given time frame, thereby capturing the evolving nature of news-related terms. Finally, LLMs' ability to explain their own reasoning (Kojima et al., 2022), and thereby enlightening its opaque decision-making, sets it apart from other machine learning methods.

Significant drawbacks of frequently-used commercial models like GPT-3.5 or GPT-4 are issues of reproducibility and transparency. For instance, when using GPT models, we are dependent on OpenAI for updates and changes to their models and API, limiting researchers' control over the scientific process – for this purpose open-weight models are preferred.

Additionally, data potentially containing privacy sensitive information should (or, in some legal context cannot) be sent to commercial AIs but should instead be run locally. Moreover, the use of the newest models especially with larger datasets can be resource intensive (both financially and environmentally). Regardless of the specific model of choice, even small variations in prompting can lead to substantially different outcomes (Reiss, 2023b), so substantial prompt engineering is necessary when applying this approach. While being zero-shot in terms of training, manual annotations will still be necessary to validate the model (e.g., Reiss, 2023b; Törnberg, 2023).

Data

We use Chrome browsing history data obtained via data donations in May 2022, which are linked to self-reports on socio-demographic, political and news-related characteristics (N_{users} = 340, $N_{records}$ = 13,093,451) (Welbers et al., in press). After completing the survey, respondents were asked to request their Chrome browser history using the Google Takeout service.[1] They could then use a designated application to review their data, remove specific parts of the data if desired, and upload their data to the database. The browsing data includes the full URL, website title, timestamp and transition type (i.e. website accessed via a link, typed in, bookmarked) of all logged-in activity on Chrome desktop and mobile browsers and the Google app. The Ethical Board of the University of Amsterdam has approved of this study.

Preprocessing steps and sample selection

Browser history data contain records that are not "meaningful" website visits (e.g., automatic redirects, page refreshes, pop ups, etc.). Besides affecting the total level browsing activity, these "meaningless" visits greatly affect the frequency of visits to certain domains, as the production of these "meaningless" visits differs between domains. Hence, we remove records from (sub)domains that did not return a valid DNS setting, were automatically redirected, or returned extremely limited data (Faroughi et al., 2021). The categorization of the top domains was manually checked and changed if necessary. Similar to Guess (2021) and Robertson et al. (2023), we also remove sequential duplicates, which we operationalized as exact URL matches within a 5-second time window. We also restrict the sample to users who used Google Search at least once. These strategies reduce the number of records by 35% (N_{users} = 315, $N_{records}$ = 9, 868,209).

[1] https://takeout.google.com/settings/takeout

Search queries

To measure whether a record was a search, we first detected whether the URL contained *google.nl/search* or *google.com/search*,[2] and extracted the search type.[3] We then classified the record as search if the search query parameter q is present, and it is a text search type, to exclude records other than regular Google searches (such as image search, clicks on images, etc.). 7.1% of all records are Google searches ($N_{searches} = 697{,}359$). The search query's textual content was subsequently extracted from the URL's q parameter.

Searches were manually annotated by six trained annotators using a codebook (see Appendix D). The annotators participated in a joint coder training where they were trained on the codebook and coding procedure. Subsequently, any disagreements in a training sample of $n = 200$ were reviewed and discussed with the first author, resulting in improvements to the codebook. For each search, the coders were shown the search query, date of search, and up to three visited websites following the search. Additionally, the annotators were able to review a Google Search result page containing websites with a publishing date within a 7-day time frame of the search query's date. These websites can inform the interpretation of unfamiliar, potentially context-dependent, search queries, such as those related to news events. To construct the annotation sample, we used two different sampling approaches. We first applied a weighted random sampling method, where searches from users with high volumes of browsing had a lower probability of being selected. To increase the potential share of PNR queries, we then additionally used a sampling method that sampled queries leading to PNR websites. Each of these methods contributed roughly half of the final annotation sample. For more details on the annotation sampling procedure, please refer to 703; 750), which, given duplicate queries, corresponds to 118.6k searches in the browsing data.

Each of these methods contributed roughly half of the final annotation sample. For more details on the annotation sampling procedure, please refer to Appendix E. $n = 35,453$ unique search queries were annotated in this procedure ($n_{P\ NR} = 2,703$; $n_{non-P\ NR} = 32,750$), which, given duplicate queries, corresponds to 118.6k searches in the browsing data.

We ensured the annotations' quality by proceeding in two steps. First, we calculated the inter-coder reliability on a random sample of 879 search queries between all coders, which resulted in a Krippendorff's α of .65. Some disagreement is expected given the complex coding task (i.e., sparse search terms, complex concept). None of the coders performed particularly better or worse than the others (excluding a coder results in a maximum of .03 improvement). Second, we subjected *all* 35k annotations to an additional round of coding by a coder with expertise in the subject, effectively coding all units at least twice. The expert coder reviewed the annotations made by the six other coders, and disagreed with 775 annotations. The disagreements were checked by the first author who made the final decision. We changed the annotations of 578 queries (1.7%), mainly involving non-PNR queries reclassified as PNR by the expert coder. This process addressed issues stemming from a lack of knowledge about political and news events, which is also described by earlier work with a highly similar coding task (Menchen-Trevino et al., 2023).

We distinguish two sources of potential error: misjudgments of annotators and insufficient information in (short) search queries. We are confident that our approach significantly mitigates the first source of error. We revisit the second source of error in the discussion.

Sample description

Table 1 provides an overview of the characteristics of the survey and browsing data after preprocessing and selection. The browsing data covers the period from April 12, 2021 to May 31, 2022. Women and lower educated individuals are underrepresented in our sample, and the respondents self-report a relatively high interest in news. The browsing characteristics show significant individual-level variation

[2]We focus on Google, the dominant search engine in the Netherlands, with 98% of searches being Google Search compared to 1.4% DuckDuckGo and < 0.01% Yahoo, Bing and Baidu, and 0% Yandex, based on our sample. Furthermore, we focus on Google Search, and exclude other Google components (e.g., Google News). Yet, applying the proposed approaches should be similar for different search engines and Google components.

[3]Text search 81.6%, image 15.4%, shopping 1.1%, and local, video, news, and books < 1%.

Table 1. Sample characteristics (N_{users} = 315, $N_{records}$ = 9,868,209).

	M /%	Median	SD	Min	Max	*M* /% 2022 Census
Survey data						
Age	47.43	47	16.26	18	80	42
Female	37					50
Education						
low	7					26
medium	25					38
high	65					36
unknown	3					
Political interest (1 = low, 7 = high)	4.48	5	1.73	1	7	
Political position (0 = left, 10 = right)	4.72	5	2.36	0	10	
Self-reported news use: *I follow the news…*						
every day closely	41					
every day briefly	42					
a few times per week	10					
a few times per month	3					
less often	3					
never	2					
Browsing data						
Browsing duration (in days)	253.70	362	137.69	1	370	
Website visits	31,327.65	17,745	42,137.80	13	394,578	
Average website visits per day	127.67	97.01	126.47	0.07	1081.04	
Average searches per day	8.76	5.59	11.23	0.0	125.10	
Average news website visits per day	5.47	1.27	11.16	0.0	79.41	
Average background information website visits per day	0.72	0.28	1.35	0.0	9.75	

Note. Age is unknown for 11 respondents. Percentage instead of mean presented for categorical variables. Census data from Statistics Netherlands 2022.

in browsing duration and number of records. Obtaining a representative sample is a challenge when working with digital trace data (Stier, Breuer, et al., 2020). A biased sample could affect follow-up analyses because information-seeking for PNR information may be related to these characteristics (e.g. political interest, Menchen-Trevino et al., 2023). While we publish our models open source, the proposed approaches are generally meant to be applied to one's own data set, which should be more representative to make valid conclusions about PNR search behavior. As the primary aim of this paper is to classify search terms and the number of records highly impacts whose searches would be included in the annotations, we addressed this bias in constructing the annotation sample (see above).

Methods

The code and models developed for this project are available at Github.[4]

Browsing sequences

The browsing sequences method consists of three consecutive steps. The flowchart depicted in Figure 1 describes the first two steps. First, we determine whether a search led to the selection of a search result, a new search (e.g., search query refinement), or no selection. This is indicated as the initial step in the flow chart in Figure 1. Building on news referrals research (e.g., J. Möller et al., 2020; Wojcieszak et al., 2022), to identify whether a search result was selected, we check if the immediately following record is accessed through a link within 30 seconds after a search.[5] If the next record is another search, but not simply a reload, it indicates

[4]https://github.com/mariekevh/GooglingPolitics
[5]A 5-minute threshold yields similar results.

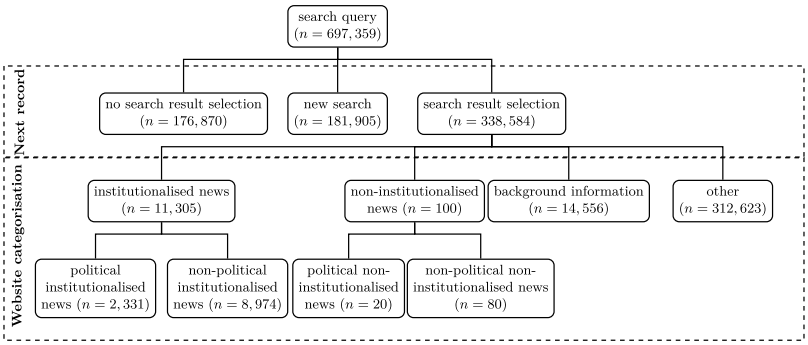

Figure 1. Flowchart for browsing sequences approach. *Note*. n refers to the number of records.

a new search. If the next record is not accessed via a link, not a new search or not within 30 seconds, it means no website was selected.

Second, moving on to the next step in Figure 1, when a search result is selected, we categorize this record using a combination of domain-level and content-level classification. We first use a list of domain categorizations from Loecherbach (2023) to classify the selected domains as institutionalized news (i.e., professionally and editorially organized) ($n_{websites} = 373$), non-institutionalized news ($n_{websites} = 54$), or background information (about public events and figures, e.g. Wikipedia) ($n_{websites} = 306$), which is able to categorize 76% of all records. To address the issue of non-news content on news websites (Reiss, 2023a), we then apply a political content classifier created by Wojcieszak et al. (2024) to the news website titles to distinguish political and news-related website records from other types of records.

Third, we label a search as PNR based on the selected website's category. We consider six different options, ranging from narrower (e.g., political institutionalized news) to broader (e.g., including background information websites) definitions of PNR search. We classify a search as PNR if the selected website is (1) institutionalized news, (2) political institutionalized news, (3) institutionalized and non-institutionalized news, (4) political institutionalized and non-institutionalized news, (5) institutionalized and non-institutionalized news or background information, or (6) political institutionalized and non-institutionalized news or background information website.

Context-enhanced dictionary

The context-enhanced dictionary approach consists of base dictionaries meant to capture context-independent search terms, enriched with a rolling set of dictionaries capturing context-dependent search terms containing entities that appeared in the news.

There is not one off-the-shelf base dictionary that will perfectly capture (Dutch) PNR information, especially not for user-generated language. However, one promising option is the Dutch-language Lexicoder Topic Dictionaries by Albaugh et al. (2013), which is a validated dictionary aimed at capturing policy topics in news content, legislative debates and policy documents ($n = 1,\ 424$) (*Policy Agendas*). Additionally, we created a list of nationally and locally represented political parties ($n = 66$) and members of government and house of representatives ($n = 180$) (*Politics*), and a list of the largest news outlets in the Netherlands (i.e., ≥100 visits in our data) and the term *news* ($n = 198$) (*News*). These lists include abbreviations (e.g., FvD/Forum voor Democratie). The search term was labeled as PNR if at least one term from the dictionary is present in the search term. Following the approach by Kroon et al. (2022), we removed punctuation, lower-cased the words, and stemmed (i.e., reduced to its root form) the search terms to match the Policy Agendas dictionary. We did not stem the search terms to match the News and Politics dictionaries.

To construct our rolling dictionaries, we collected a data set of news content from the six largest Dutch newspapers[6] ($n = 296{,}355$) spanning from March 1, 2021 to June 30, 2022 via the Amsterdam Content Analysis Tool (AmCAT).[7] Each article included its full text, headlines, publisher, date and URL. We label a search query as PNR if it matched with news headlines, which contains the key information regarding the news article's content. Specifically, we consider news articles published one day before and one day after the search. To this end, we apply fuzzy matching based on the Levenshtein Distance on the character-level (0 = no similarity, 100 = exact overlap), which is able to detect if two texts are similar beyond an exact match. A similarity threshold of ≥ 80 was applied to identify matches since the performance did not substantially improve for higher thresholds. Prior to matching, we preprocessed the query and headlines by lowercasing and removing stop words, punctuation and digits. From the headlines, we extracted named entities (e.g., locations, people) and/or nouns (both versions are presented in Table 3).

Finally, a search term is labeled as PNR if either the base or rolling dictionaries identified it as such. We evaluate all dictionaries separately as well as in combination. Note that the same query may be labeled differently on different dates because this approach is applied to a combination of query and date.

Traditional supervised machine learning

We split the annotated data into a training and test set (8:2 ratio), and trained six SML models using 5-fold cross-validation: Naive Bayes, Logistic Regression and Linear Support Vector Machine (SVM) with two vectorizers (CountVectorizer, TfidfVectorizer) (for details see Appendix A). We used a grid search to find the best configuration for each model. Specifically, we experimented with pre-processing parameters (n-grams, minimum and maximum document frequency, lowercasing and removal of punctuation and digits, stopword removal, and stemming), introducing class weights to account for class imbalance, and C for Linear SVM specifically. The best model was selected based on the highest f1 score for the positive class (i.e., PNR), thereby aiming for a balance between precision and recall. The performance metrics presented in Table 3 are the results on the test set after (hyper) parameter optimization.

Transformer-based supervised machine learning

Using the same train-test split, we further split the training set into a training and validation set (8:2 ratio) to fine-tune four pre-trained BERT models (for details see Appendix A). Three Dutch-specific models, i.e. *BERTje* (De Vries et al., 2019) and the state-of-the-art *robBERT* model (Delobelle et al., 2020) trained on diverse texts up to 2019, as well as the more up-to-date *robBERT 2022* model (Delobelle et al., 2022) with training data up to January 2022. We also considered a *multilingual BERT* model (Devlin et al., 2018) trained on Wikipedia to account for multilingual search.[8] We experimented with the (hyper)parameters batch size, number of epochs and learning rate (recommended by Devlin et al., 2018), as well as number of warm up steps and class weights. As BERT models benefit from contextual data, we did not experiment with stop word removal but apply lowercasing and removal of punctuation and digits. Again, the best model was selected based on the highest f1-score for the positive class. The performance presented in Table 3 are the results of the four models on the test set after (hyper)parameter optimization.

[6]De Telegraaf ($n = 95{,}100$), NRC Handelsblad ($n = 50{,}152$), De Volkskrant ($n = 49{,}194$), Algemeen Dagblad ($n = 44{,}956$), Trouw ($n = 43{,}944$) and Financieel Dagblad ($n = 13{,}009$).
[7]https://github.com/amcat
[8]36.7% of the annotated PNR search terms are not in Dutch.

Zero-shot classification

In a zero-shot fashion, we collected annotations for a short definition and longer definition adapted for LLM prompting from the codebook. Where possible (given the training date threshold), we also collect annotations with date indications (month, year) added to these instructions. Following Kojima et al. (2022), we used a two-step prompt. First, we asked for reasoning by prompting "*Is this search query '**{search query}**' political or news-related when searched in the Netherlands (yes or no)? Political or news-related search queries are defined as seeking information contributing to opinion formation on political and societal topics. Give your reasoning.*" (reasoning extraction). Then, we asked for a binary label based on its reasoning by prompting "*Therefore the answer (yes or no) is*" (answer extraction). We used English prompts instead of Dutch, diverging from the original codebook, because zero-shot classifications have yielded substantially better results when instructed in English (Kuzman et al., 2023a). An example of this approach is presented in Table 2 (for details on instructions, see Appendix B).

We tested two models using a zero-shot approach: GPT-3.5 and Llama 3 with 8 billion parameters (Meta, 2024). First, we used OpenAI's "gpt-3.5-turbo" model (Brown et al., 2020) and chat completion API to collect annotations between March 31 and April 4, 2023.

The GPT-3.5 model's training data at this time cut off in September 2021, preventing us from collecting date-specific annotations for search queries after that time. Details on date-specific indications can be found in Appendix B. We set the temperature parameter to 0.2 for all classifications to make the output more robust (Gilardi et al., 2023; Reiss, 2023b). Despite the two-step approach, GPT-3.5 failed to label some search terms, which we joined with the unrelated category when parsing the final label from the output. As a robustness check, we collected these classification for the shorter definition again, which produced similar results at this temperature setting. However, minor adjustments to the (longer) instructions yielded slightly different results (Reiss, 2023b), which we discuss in the conclusion section.

Second, we ran Meta's Llama 3 model with 8 billion parameters via Python locally on a laptop (without a GPU), implemented by Ollama (Ollama, 2024). The cutoff of this version of Llama 3 is March 2023, therefore all date-specific queries could be annotated. The temperature setting and two-step approach were identical with the GPT-3.5 annotations. For Llama 3, all search terms were labeled on the first try, eliminating the need for an unrelated category.

Evaluation of performance

We report precision, recall and f1-scores for all approaches. F1-scores and recall are our primary evaluation metrics. We chose f1-scores because, in principle, we aim to avoid both false positives (incorrectly predicting a search as PNR, leading to lower precision) and false negatives (predicting a search as unrelated when it is PNR, leading to lower recall). Yet, we prioritize recall over precision.

Table 2. Example of input and output of zero-shot classification.

Step		Content
Reasoning extraction	Prompt	Is the search query "toeslagen belastingdienst" political or news-related when searched in August 2021 in the Netherlands (yes or no)? Political or news-related search queries are defined as seeking information contributing to opinion formation on political and societal topics. Give your reasoning.
	Answer	Yes, the search query "toeslagen belastingdienst" is political when searched in August 2021 in the Netherlands. This is because the topic of "toeslagen" has been a major political issue in the Netherlands for several years, with the government facing criticism for its handling of the system and the impact it has had on citizens. The search query is likely to be related to news and information about the ongoing political discussions and developments surrounding the issue. Therefore, it can be considered a political search query as it seeks information contributing to opinion formation on a political and societal topic.
Answer extraction	Prompt	Therefore, the answer (yes or no) is
	Answer	Yes.

Note. Example of GPT-3.5 output with short definition instructions with date indication.

Methods scoring high on recall, but slightly lower on precision, allow for filtering out searches which are *potentially* PNR. This allows for a reasonable alternative approach: first applying the method to narrow down the set of search terms, and secondly, manually evaluating this substantially reduced set on precision.

Furthermore, it is preferable to use all available search queries for evaluation, which is less for those methods that require a training set and more for those that do not. However, to ensure a fair comparison, we aim to evaluate all methods on the *same* test set ($n = 7{,}901$) derived from the train-test split. We focus on the $n = 541$ PNR test set queries (as identified by human coders), as our primary aim is to identify PNR queries. Performance scores for the majority class (i.e., non-PNR) are affected by class imbalance and are therefore not included. A flowchart that details the sampling steps taken is presented in Appendix E.

The test sizes for some methods in Table 3 deviate from $n = 541$ for two reasons. Firstly, the conceptual distinction between methods that do and do not consider context results in differences in the applied level of these methods, and consequently, the test size. The browsing sequences approach

Table 3. Performance on the test set of all models per approach.

model	precision	recall	f1-score	*N*
Browsing sequences				
Institutionalized news	0.20	0.06	0.09	271
Political inst. news	0.33	0.01	0.03	271
Inst. and non-inst. news	0.21	0.06	0.10	271
Pol. inst. and non-inst.	0.33	0.01	0.03	271
Inst. and non-inst. news and background information	0.23	0.21	0.22	271
Pol. inst. and non-inst. news and background information	0.25	0.16	0.20	271
Context-enhanced dictionary				
Base dictionaries				
Politics	0.79	0.03	0.05	541
News	0.61	0.13	0.21	541
Policy Agendas	0.14	0.09	0.11	541
Politics + News	0.64	0.16	0.25	541
Policy Agendas + News	0.25	0.22	0.23	541
Policy Agendas + Politics	0.17	0.12	0.14	541
Policy Agendas + Politics + News	0.27	0.24	0.25	541
Rolling news dictionaries				
Entities	0.21	0.28	0.24	1132
Entities + Nouns	0.16	0.38	0.22	1132
Combined				
Best base dictionary + Entities	0.32	0.65	0.43	1132
Best base dictionary + Entities + Nouns	0.25	0.73	0.37	1132
Traditional Supervised Machine Learning				
Logistic Regression with Count	0.69	0.58	0.63	541
Logistic Regression with Tfidf	0.55	0.66	0.60	541
Naive Bayes with Count	0.83	0.41	0.55	541
Naive Bayes with Tfidf	0.90	0.28	0.43	541
Linear Support Vector Machine with Count	0.75	0.55	0.63	541
Linear Support Vector Machine with Tfidf	0.66	0.64	0.65	541
Transformer-based Supervised Machine Learning				
BERTje	0.68	0.57	0.62	541
robBERT	0.68	0.67	0.67	541
robBERT 2022	0.74	0.61	0.67	541
Multilingual BERT	0.74	0.58	0.65	541
Zero-shot classification				
GPT-3.5 with short definition	0.43	0.78	0.55	541
GPT-3.5 with long definition	0.43	0.82	0.57	541
Llama 3 with long definition	0.16	0.89	0.26	541
Llama 3 with short definition	0.16	0.87	0.27	541
Llama 3 with long definition & date indication	0.13	0.91	0.23	541
Llama 3 with short definition & date indication	0.44	0.53	0.48	541

Note. Best performing model per approach in bold, based on f1-score. The test size differs between methods due to the conceptual distinction between methods that do and do not consider context, and due to oversampling in the annotation set (see Methods section for details).

operates at the level of the combination of search query and selected website because this method uses the latter as an indicator for PNR. Consequently, identical search queries may be labeled differently based on the user's website selection following the search. Similarly, context-enhanced dictionaries are applied to combinations of queries and dates, meaning that the classification of identical search queries may vary depending on the search date. In contrast, traditional SML, transformer-based SML and zero-shot classification operate at the level of a search query, consistently assigning the same label to queries regardless of its associated date or website. This difference in applied levels affects the test size for evaluation because some queries appear multiple times for browsing sequences and context-enhanced dictionaries, but with different dates or selected websites.

Secondly, note that during the construction of the annotation sample, we deliberately oversampled search queries leading to news websites (see Data section). Hence, evaluating the performance of the browsing sequences approach on these queries may result in inflated performance scores. To address this, these specific search queries were excluded from the test set for the browsing sequences approach specifically.

Finally, we also manually inspect some of the results of the best performing models to get a clearer picture of the (mis)classifications. See Table C1 for examples. The examples highlighted in the Results section are translations from Dutch.

Results

Browsing sequences

According to Table 3, all sequence-based approaches fail to accurately capture PNR searches, despite using an exceptionally extensive list of news and background information websites. Around 67–80% of search terms identified as PNR by these models is in fact unrelated, as indicated by the low precision of .20 to .33. Moreover, by far most PNR searches are missed, resulting in low recall of .16 to .21, particularly when relying solely on news website visits (.06 to .01). Narrowing down the categorization to political news websites does result in the expected increase in performance. The results indicate that there are indeed some (but not many) visits to news websites are unrelated to politics and news (increased precision), but searches identified as PNR do lead to websites that do not conform to this strict definition (decreased recall). Lower recall could also be explained by the political classifier not correctly recognizing political headlines after its training date.

An inspection of the classifications of the best performing model reveals that correct classifications contained explicit references to news websites or news events, but these type of search terms are also often misclassified. Moreover, many search terms incorrectly labeled as PNR are clearly unrelated to politics or news (e.g., shopping).

It is worth mentioning that using this approach leads to the conclusion that merely 3.7% of searches in the entire data set lead to a news or background information website, which aligns with findings of similar research (Menchen-Trevino et al., 2023). 45% of all searches lead to websites other than news (−related) websites (most often shopping, Facebook and other Google services). Yet, 26% of searches result in a new search and 25% lead nowhere, which makes these searches challenging to classify using a browsing sequences approach.

Context-enhanced dictionary

The results in Table 3 present the base and rolling dictionaries separately and in combination. Using the best performing base and rolling dictionaries in isolation barely outperform the best sequence-based approach, as indicated by f1-scores of .25 and .22, respectively. The base dictionaries on their own work relatively well for detecting search terms that explicitly mention news outlets or politics (e.g., *prime minister*). Enhancing the base dictionaries with news content allows the dictionary

approach, which typically does not perform well for broad concepts (Kroon et al., 2022; Van Atteveldt et al., 2021), to be competitive in terms of recall (.73), without reducing its precision (.25). It additionally labels some searches for news events (e.g., *truth social*), including COVID19-related queries that were not represented in the Policy Agendas data set from 2013. Furthermore, using nouns in addition to entities appearing in the news presents a trade-off between precision (−.07) and recall (+.08). Since the performance of both models is sub-optimal given common standards, this method would require human validation. In that case, we would opt for the model with higher recall, including both nouns and entities.

Furthermore, dictionary approaches do not consider the surrounding context, leading to searches of *personal* concern being mislabeled as PNR due to the presence of words that indicate PNR in other contexts, e.g. *reckless driving* ***police*** *complaint* (matching words in bold). The dictionary-based approach, particularly the base dictionaries, also fail to identify PNR queries containing synonyms or spelling errors.

Traditional supervised machine learning

Traditional SML approaches show a substantial improvement compared to the sequences (f1-score +.45) and dictionary methods (f1-score +.28), with Linear SVM with a tf·idf vectorizer achieving the best f1-score (.65) and balance between precision and recall. Yet, even the best performing traditional SML model faces challenges similar to the dictionary approach. While explicit references to news outlets and politics lead to correct classifications, the best model incorrectly predicts search queries compromised of political or news-related names, spelling errors, or news-related topics that were not (adequately) represented in the training data as unrelated, as well as mislabels searches because it considers search terms as a bag-of-words. This model correctly classifies many COVID19-related search terms due to the high frequency of COVID-19 in the training data. However, this prevalence may also hinder the model's ability to learn other PNR features accurately, resulting in a lower performance overall. Moreover, it reduces the fit of a model based on these training data on future data.

Transformer-based supervised machine learning

The robBERT model outperforms other transformer-based SML models in terms of f1-score and balance between precision and recall. Surprisingly, transformer-based SML shows only a marginal improvement compared to the traditional SML, despite being based on a LLM, as indicated by a .02 difference in f1-scores. Manual inspection of the misclassifications reveals that this model struggles to recognize names and spelling errors. It also faces difficulty in labeling news-related search terms, either due to their time-dependency or emergence after the robBERT's training data ended in 2019. For instance, the model incorrectly predicts a search for a Dutch political talk show as unrelated, likely because it only began airing in 2021. Although the updated robBERT 2022 model avoids such issues, it overall does not perform better in this case. Additionally, the robBERT model makes incorrect predictions for search terms that are of personal concern yet contain references that would be considered PNR in other contexts (e.g., *how to apply for social housing*) and struggles with (shorter) search queries lacking contextualizing information. Furthermore, despite considerable effort to ensure the annotations' quality, robBERT retrieved some PNR search terms that the annotators missed (e.g., *hourly cost government aircraft*).

Zero-shot classification

Zero-shot classification using GPT-3.5 outperforms that with Llama 3 based on f1-scores. Providing GPT-3.5 with a longer definition yields a higher f1-score compared to a shorter one. The best performing model (f1-score .57) correctly understands names, media outlets, spelling errors and various political and news topics as PNR, but stretches the definition too far, resulting in low precision. For instance, "*The search query 'Italy' is news-related [. . .] as it is likely seeking information about current events and news related to Italy.*" It also occasionally categorizes sports or entertainment-related queries as PNR despite the instructions clearly

stating to *exclude* these. These mistakes indicate a misunderstanding, or even ignoring, of (part of) the instructions that were clear to human coders. Similar to Transformer-based SML, GPT-3.5's false positives are also sometimes due to annotators incorrectly labeling a search term as unrelated. For instance, human coders missed a query about the relationship status of a local Dutch politician whose name GPT-3.5 was able to recognize as political.

Similar to the context-enhanced dictionary method, GPT-3.5 achieves a high level of recall (.82) at the expense of precision (.43). While GPT-3.5's precision is notably higher than the context-enhanced dictionary (+.18), both are not up to par with common standards and would benefit from validation after classification. Some issues remain regarding the disagreement between personal and PNR search terms, and GPT-3.5 sometimes changes its judgment between reasoning and answer extraction.

Discussion

Search engines playing an important role in the current information environment. However, we had yet to find a method to identify political and news-related searches, while taking all relevant aspects of this concept into account. We set out to map how five potential approaches perform in order to provide recommendations to researchers using such browsing history data. We find that PNR search terms are complex to classify because of the comprehensive concept and the lack of information in search query content.

Our findings suggests that relying solely on browsing sequences is insufficient as a method to measure the use of web search for PNR information. This signals that individuals who seek information on politics and current events through search engines in many cases do not visit (large) news or background information websites like Wikipedia, if any website for that matter, even though it is theoretically plausible. This outcome may be partly attributed to measurement issues caused by web browser activities like multi-tabbing or reloading, which disrupts the tracking of browsing patterns. Future research should aim to develop valid methods to address these issues. Yet, the analysis of browsing patterns also indicated that a large share of searches resulted in the selection of smaller or news-unrelated websites that were not captured by our list of (news) websites, new searches (e.g., refinement of search queries), or no visit at all. Within our sample, we observed that one in four PNR searches ended at the search result page, which echoes the findings of previous studies (Blassnig et al., 2023; Menchen-Trevino et al., 2023). Individuals can increasingly find answers to PNR queries directly on the search result page due to information-rich features (Oliveira & Lopes, 2023), some of which lower click-through-rates for regular search results (Gleason et al., 2023). As these features become more common, methods that rely on clicks could become (even) less suitable for identifying PNR searches in browsing history data. An alternative approach is to classify the search query's textual content, as done by the other four methods in this paper.

Supervised methods, both traditional and transformer-based, are best at striking a balance between identifying relevant queries and doing so with some precision. However, these approaches still leave room for improvement. Despite using a substantial sample of annotated search terms, it is likely that these models will improve by further training them on more (diverse) annotation data, allowing the model to recognize a wider range of topics as PNR, compensating for the sparsity of search query content. Nevertheless, further increasing the number of annotations, even with techniques like active learning, may not be the most rewarding because annotations would need to be updated to accommodate emerging topics. Additionally, it is surprising that transformer-based SML, which is specifically trained to understand the meaning of words in context, did not lead to a significant improvement in distinguishing PNR queries from personal queries using similar words in cases where this would be clear to human annotators. We suspect that the limited input text does not provide sufficient context for BERT models to make accurate predictions.

It becomes clear that utilizing the sequence of browsing is useful to understand how people access a selection of (news) websites, and SML is useful for more narrow, targeted topics that can be more easily captured in the training data and do not need updating as much, because this method can strike a balance between precision and recall. Future research could compare the performance of our general classifiers with fine-tuned classifiers designed to identify searches related to specific topics.

Given the scope of PNR search, which could encompass any information that can aid political opinion formation, it is important to prioritize recall. We show that enhancing a dictionary with context-sensitive information based on news content substantially improves its ability to recall PNR search terms. Context-enhanced dictionaries perform comparatively well in terms of recall, but lose out to zero-shot classification in terms precision. We acknowledge, however, that its performance could be improved if we improved this method on several points, such as the base dictionary's age and application to a context for which it was not originally designed, and the expansion of the rolling dictionary with a filter for politically relevant news. Nonetheless, we do not expect it to outperform zero-shot classification nor be more useful in terms of the necessary preparation to employ it (i.e., dictionary construction).

The biggest advantage of zero-shot classification is its ability to recognize a wide range of PNR topics without the need for training, which opens opportunities for applications in multilingual and large-scale settings. The two-step prompting approach where the model provides an explanation, and subsequently makes a decision based on its own reasoning, alleviates some of the sparsity of search terms by adding contextualizing information from the knowledge in its LLM. Yet, GPT-3.5 in a zero-shot setting appears *too lenient* its classifications by making many false positives that would be avoided by human coders, and to a lesser extent, the SML models. Hence, these mistakes require a human to make the final decision. Given that this is an area of rapid change, both for commercial and open-weight LLMs, where earlier versions of models quickly become outdated, we recognize that the performance of the zero-shot approach may improve using a more powerful and up-to-date models (e.g., GPT-4). We encourage testing other, preferably more open models, in future research. At the same time, the release of new models is a moving target – any specific release will be outdated soon. Hence, our results should not so much be seen as a definitive verdict on which exact LLM to use, but more as an indication of their potential. Ultimately, researchers are advised to make an informed decision among recent releases available at the time of their study, taking criteria such as reproducibility, openness, and resource requirements into account.

We recommend future research studying PNR search to leverage zero-shot classification using LLMs to filter search terms that are *potentially* related to politics or news, significantly reducing the number of queries human annotators need to assess (i.e., 85.5% reduction in our test set using GPT-3.5). This sample of search queries should subsequently be evaluated critically by trained human coders. Identifying PNR search queries is only one example where the classification of very short textual content into relatively generic concepts is an issue. Similar considerations apply to tasks such as categorizing potentially short social media comments as (ir)relevant (e.g. Youtube, A. M. Möller et al., 2024). In such cases as well, we suggest employing methods with high recall and complementing them with human evaluation.

Nevertheless, we need to stress a few caveats of human annotation of search terms. In this study, we used fully manual annotations as the golden standard and for training SML models. The lower inter-coder reliability observed may be attributed to two sources of error: misjudgments by coders and a lack of information in the search query. We believe that the first source of error is largely addressed by subjecting all search terms to an additional round of expert coding. In future research, the explanations generated by the LLM during annotation, which typically include explanations of key entities in the query, can assist coders in their evaluation when approached critically. Yet, the second source of error remains a shortcoming of the (human) annotation of short search terms. Despite employing various strategies to provide context to coders (e.g., search date, sequence of visited websites), a certain level of disagreement persists due to limited information in short search terms. The extent to which

ambiguous search queries pose an issue to the identification of them remains an open question. The level of ambiguity of search terms may vary among different sub(types) of PNR queries. For instance, a potential (sub)type may consist of searching for governmental services (e.g., childcare facilities), which can happen for both personal and political reasons. Future research should therefore develop a more nuanced typology of PNR queries and assess the level of ambiguity within each category.

Furthermore, truly grasping the motivation behind short search terms, especially ambiguous ones, requires different research designs, such as surveying respondents about their queries. Future research could leverage the data donation process by enabling respondents to annotate their own search terms. However, due to the infrequent occurrence of PNR queries presenting participants with a random sample of their searches likely yields unsatisfactory results. A multi-step process is therefore likely needed to make this feasible. Employing zero-shot classification as an initial filtering step to obtain a more balanced (and smaller) sample of queries for annotation could be a promising path forward, similar to oversampling rare events for training classifiers.

We also stress a few considerations of using zero-shot classification that future researchers need to be aware of. The use of commercial models like GPT comes with issues of reproducibility (as the models underlying APIs are constantly being adapted and replaced by newer ones), costs (both for querying the API but also environmental costs of each query run), and data security (annotation of privacy sensitive material, especially problematic under GDPR ruling) (Spirling, 2023). Although currently a fully locally running LLM with open weights appeared to still be lacking behind the commercial solution, we are confident that the fast-paced development in this sector will soon produce competitive solutions. This would allow the wider research community to make use of LLMs for different types of tasks without needing to have access to competitive computing infrastructure. Regardless of the model used, prompt engineering is necessary for LLMs to function effectively as slightly different prompts, that would go unnoticed by humans, can lead to different results (see also Reiss, 2023b).

Conclusion

The role of search engines in shaping our political information ecosystem is increasingly recognized. However, our understanding of the role of search engines in political information diets is limited to small samples (Menchen-Trevino et al., 2023), specific topics (Blassnig et al., 2023), or news content only (e.g., Robertson et al., 2023; Wojcieszak et al., 2022). While recent advances in the collection of digital traces, such as browsing data, have opened up new possibilities to study search behavior in an ecologically valid way, this has not yet been leveraged to study political and news-related web search comprehensively and at a larger scale. In the current paper, we explored methodological solutions to identify PNR searches – a first step for empirical research on political and news-related information-seeking via search engines. Our findings show substantial variation in the accuracy of the methods employed, with some methods being more suited to more narrow topics, and therefore leading to substantially different conclusions if used. Specifically, we recommend a two-step approach leveraging LLMs for zero-shot classification and human evaluation to capture PNR search behavior most accurately.

The use of search engines for politics and news-related information-seeking will remain a relatively small part of the public's online information diet, no matter how accurately it is measured. Yet, due to the active and engaged nature of this type of political information consumption and its potential impact on political decision-making (Epstein & Robertson, 2015), its role should not be overlooked. Our methodological approach can help answer important questions about the frequency (e.g., relative to other information channels), search strategies (e.g., confirmation bias, selective exposure), and how individual-level characteristics moderate these searches (e.g., political knowledge, digital literacy),

which have not been addressed to date. This information is necessary to contextualize existing findings about the democratic implications of political information delivered by search engines (Nechushtai et al., 2023; Puschmann, 2019; Urman, Makhortykh, Ulloa, & Kulshrestha, 2022).

Acknowledgments

This study was supported by the Amsterdam School of Communication Research, and its Digital Communication Methods Lab at the University of Amsterdam. This work was in part carried out on the Dutch national e-infrastructure with the support of SURF (Grant No. EINF-5514). We thank our research assistants and Roeland Dubel for their coding assistance, and Anne Kroon for help with transformer-based SML.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by SURF and the Amsterdam School of Communication Research, University of Amsterdam.

Notes on contributors

Marieke van Hoof is a postdoctoral researcher in political communication at the Amsterdam School of Communication Research, University of Amsterdam. In her research, she explores how digital media, algorithms and AI shape the use of political information.

Damian Trilling (PhD, University of Amsterdam) is a professor of journalism studies at the Department of Language, Literature and Communication, VU Amsterdam. He is interested in news use and dissemination and in the adoption and development of computational methods.

Corine Meppelink (PhD, University of Amsterdam) is an assistant professor at the Amsterdam School of Communication Research, University of Amsterdam. Her research focuses on how people access, understand, and evaluate digital information and inequalities caused by literacy differences such as health literacy and digital literacy.

Judith Möller (PhD, University of Amsterdam) is a professor of media use and effects at the University of Hamburg and the Leibniz Institute of Media Research|Hans Bredow Institute. In her research, she focuses on the effects of political communication, in particular, social media and digital media.

Felicia Loecherbach (PhD, VU Amsterdam) is an assistant professor at the Amsterdam School of Communication Research, University of Amsterdam. Her research focuses on news diversity and how it is impacted by selection and personalization as well as the collection and modeling of digital trace data.

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 Conference*, June 27–29, Antwerp.

Arendt, F., & Fawzi, N. (2019). Googling for trump: Investigating online information seeking during the 2016 US presidential election. *Information, Communication & Society*, *22*(13), 1945–1955. https://doi.org/10.1080/1369118X.2018.1473459

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Blassnig, S., Mitova, E., Pfiffner, N., & Reiss, M. V. (2023). Googling referendum campaigns: analyzing online search patterns regarding Swiss direct-democratic votes. *Media and Communication*, *11*(1), 19–30. https://doi.org/10.17645/mac.v11i1.6030

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). *Language models are few-shot learners*. https://doi.org/10.48550/arXiv.2005.14165

Cardenal, A. S., Victoria-Mas, M., Majó-Vázquez, S., & Lacasa-Mas, I. (2022). Assessing the validity of survey measures for news exposure through digital footprints: Evidence from Spain and the UK. *Political Communication*, *39*(5), 634–651. https://doi.org/10.1080/10584609.2022.2090038

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based language model. *Findings of the Association for Computational Linguistics: EMNLP*, *2020*, 3255–3265. https://doi.org/10.18653/v1/2020.findings-emnlp.292

Delobelle, P., Winters, T., & Berendt, B. (2022). Robbert-2022: Updating a Dutch language model to account for evolving language use. *arXiv preprint arXiv:2211.08192*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). Bertje: A Dutch Bert model. *arXiv preprint arXiv:1912.09582*.

Dutton, W. H., Reisdorf, B. C., Dubois, E., & Blank, G. (2017). Search and politics: The uses and impacts of search in Britain, France, Germany, Italy, Poland, Spain, and the United States. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2960697

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, *112*(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

Faroughi, A., Morichetta, A., Vassio, L., Figueiredo, F., Mellia, M., & Javidan, R. (2021). Towards website domain name classification using graph based semi-supervised learning. *Computer Networks*, *188*, 107865. https://doi.org/10.1016/j.comnet.2021.107865

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. https://doi.org/10.1073/pnas.2305016120

Gleason, J., Hu, D., Robertson, R. E., & Wilson, C. (2023). Google the gatekeeper: How search components affect clicks and attention. *Proceedings of the International AAAI Conference on Web & Social Media*, *17*, 245–256. https://ojs.aaai.org/index.php/ICWSM/article/view/22142

Google. (2023). *Supercharging search with generative AI*. Retrieved May 10, 2023, from https://blog.google/products/search/generative-ai-search/

Guess, A. M. (2021). (Almost) everything in moderation: New evidence on Americans' online media diets. *American Journal of Political Science*, *65*(4), 1007–1022. https://doi.org/10.1111/ajps.12589

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Haas, A., & Unkel, J. (2017). Ranking versus reputation: Perception and effects of search result credibility. *Behaviour & Information Technology*, *36*(12), 1285–1298. https://doi.org/10.1080/0144929X.2017.1381166

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

Hoes, S., Altay, E., & Bermeo, J. (2023). Tokenization of social media engagements increases the sharing of false (and other) news but penalization moderates it. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-40716-2

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *Companion Proceedings of the ACM Web Conference 2023* (pp. 294–297). https://doi.org/10.1145/3543873.3587368

Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Kroon, A. C., van der Meer, T., & Vliegenthart, R. (2022). Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research*, *4*(2), 528–570. https://doi.org/10.5117/CCR2022.2.006.KROO

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023a). Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, *5*(3), 1149–1175. https://doi.org/10.3390/make5030059

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023b). *ChatGPT: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification*. Retrieved March 29, 2023, from http://arxiv.org/abs/2303.03953

Lind, F., Heidenreich, T., Kralj, C., & Boomgaarden, H. G. (2021). Greasing the wheels for comparative communication research. *Computational Communication Research*, *3*(3), 1–30. https://doi.org/10.5117/CCR2021.3.001.LIND

Lin, Z., Welbers, K., Vermeer, S., & Trilling, D. (2023). Beyond discrete genres: Mapping news items onto a multidimensional framework of genre cues. *Proceedings of the International AAAI Conference on Web & Social Media*, *17*, 542–553. https://doi.org/10.1609/icwsm.v17i1.22167

Loecherbach, F. (2023). *Diversity of news consumption in a digital information environment* [Doctoral dissertation]. Vrije Universiteit Amsterdam.

Makhortykh, M., de León, E., Urman, A., Christner, C., Sydorova, M., Adam, S., Maier, M., & Gil-Lopez, T. (2022). Panning for gold: Lessons learned from the platform-agnostic automated detection of political content in textual data. *arXiv preprint arXiv:2207.00489*.

Menchen-Trevino, E., Struett, T., Weeks, B. E., & Wojcieszak, M. (2023). Searching for politics: Using real-world web search behavior and surveys to see political information searching in context. *The Information Society*, *39*(2), 98–111. https://doi.org/10.1080/01972243.2022.2152915

Meta, A. I. (2024). *Introducing meta llama 3: The most capable openly available llm to date*. Retrieved April 18, 2024, from https://ai.meta.com/blog/meta-llama-3/

Möller, J., Van De Velde, R. N., Merten, L., & Puschmann, C. (2020). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, *38*(5), 616–632. https://doi.org/10.1177/0894439319828012

Möller, A. M., Vermeer, S. A. M., & Baumgartner, S. E. (2024). Cutting through the comment chaos: A supervised machine learning approach to identifying relevant YouTube comments. *Social Science Computer Review*, *42*(1), 162–185. https://doi.org/10.1177/08944393231173895

Nechushtai, E., Zamith, R., & Lewis, S. C. (2023). More of the same? Homogenization in news recommendations when users search on google, youtube, Facebook, and twitter. *Mass Communication and Society*, 1–27. https://doi.org/10.1080/15205436.2023.2173609

Oliveira, B., & Lopes, C. T. (2023). The evolution of web search user interfaces – an archaeological analysis of google search engine result pages. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (pp. 55–68). https://doi.org/10.1145/3576840.3578320

Ollama. (2024). *Ollama llama3*. Retrieved May 19, 2024, from https://ollama.com/library/llama3

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, *12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin UK.

Pradel, F. (2021). Biased representation of politicians in google and Wikipedia search? The joint effect of party identity, gender identity and elections. *Political Communication*, *38*(4), 447–478. https://doi.org/10.1080/10584609.2020.1793846

Puschmann, C. (2019). Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism*, *7*(6), 824–843. https://doi.org/10.1080/21670811.2018.1539626

Reiss, M. (2023a). Dissecting non-use of online news – systematic evidence from combining tracking and automated text classification. *Digital Journalism*, *11*(2), 363–383. https://doi.org/10.1080/21670811.2022.2105243

Reiss, M. (2023b). Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. *OSF Preprints*. https://doi.org/10.31219/osf.io/rvy5p

Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., & Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, *618*(7964), 324–348. https://doi.org/10.1038/s41586-023-06078-5

Robertson, R. E., Lazer, D., & Wilson, C. (2018). Auditing the personalization and composition of politically-related search engine results pages. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW*, *18*, 955–965. https://doi.org/10.1145/3178876.3186143

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, *616*(7957), 413. https://doi.org/10.1038/d41586-023-01295-4

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stier, S., Kirkizh, N., Froio, C., & Schroeder, R. (2020). Populist attitudes and selective exposure to online news: A cross-country analysis combining web tracking and surveys. *The International Journal of Press/politics*, *25*(3), 426–446. https://doi.org/10.1177/1940161220907018

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Trevisan, F., Hoskins, A., Oates, S., & Mahlouly, D. (2018). The google voter: Search engines and elections in the new media ecology. *Information, Communication & Society*, *21*(1), 111–128. https://doi.org/10.1080/1369118X.2016.1261171

Trielli, D., & Diakopoulos, N. (2022). Partisan search behavior and google results in the 2018 U.S. midterm elections. *Information, Communication & Society*, *25*(1), 145–161. https://doi.org/10.1080/1369118X.2020.1764605

Ulloa, R., Richter, A. C., Makhortykh, M., Urman, A., & Kacperski, C. S. (2022). Representativeness and face-ism: Gender bias in image search. *New Media & Society*, *26*(6), 3541–3567. https://doi.org/10.1177/14614448221100699

Unkel, J., & Haas, A. (2017). The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology*, *68*(8), 1850–1862. https://doi.org/10.1002/asi.23820

Urman, A., & Makhortykh, M. (2022). "Foreign beauties want to meet you": The sexualization of women in google's organic and sponsored text search results. *New Media & Society*, *26*(5), 2932–2953. https://doi.org/10.1177/14614448221099536

Urman, A., Makhortykh, M., & Ulloa, R. (2022). The matter of chance: Auditing web search results related to the 2020 U.S. Presidential primary elections across six search engines. *Social Science Computer Review*, *40*(5), 1323–1339. https://doi.org/10.1177/08944393211006863

Urman, A., Makhortykh, M., Ulloa, R., & Kulshrestha, J. (2022). Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and Informatics*, *72*, 101860. https://doi.org/10.1016/j.tele.2022.101860

Van Atteveldt, W., Trilling, D., & Calderón, C. A. (2022). *Computational analysis of communication*. Wiley Blackwell.

Van Atteveldt, W., Van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

van Hoof, M., Meppelink, C. S., Moeller, J., & Trilling, D. (2022). Searching differently? How political attitudes impact search queries about political issues. *New Media & Society*, 14614448221104405. https://doi.org/10.1177/14614448221104405

Vermeer, S., Trilling, D., Kruikemeier, S., & De Vreese, C. (2020). Online news user journeys: The role of social media, news websites, and topics. *Digital Journalism*, *8*(9), 1114–1141. https://doi.org/10.1080/21670811.2020.1767509

Wang, Y., & Jaidka, K. (2024). Confirmation bias in seeking climate information: Employing relative search volume to predict partisan climate opinions. *Social Science Computer Review*, *42*(1), 4–24. https://doi.org/10.1177/08944393231160963

Welbers, K., Loecherbach, F., Lin, Z., & Trilling, D. (in press). Anything you would like to share? evaluating a data donation application in a survey and field study. *Computational Communication Research*.

Wojcieszak, M., Menchen-Trevino, E., Clemm Von Hohenberg, B., De Leeuw, S., Gonçalves, J., Davidson, S., & Gonçalves, A. (2024). Non-news websites expose people to more political content than news websites: Evidence from browsing data in three countries. *Political Communication*, *41*(1), 129–151. https://doi.org/10.1080/10584609.2023.2238641

Wojcieszak, M., Menchen-Trevino, E., Goncalves, J. F. F., & Weeks, B. (2022). Avenues to news and diverse news exposure online: Comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks. *The International Journal of Press/politics*, *27*(4), 860–886. https://doi.org/10.1177/19401612211009160

Appendices

Appendix A: Details on Traditional and Transformer-based Supervised Machine Learning

Traditional Supervised Machine Learning

We trained three Supervised Machine Learning (SML) models, i.e. Logistic Regression, Naive Bayes, Linear Support Vector Machine (SVM), each with two vectorizers, i.e. CountVectorizer, TfidfVectorizer. We experiment with the following parameters per model-vectorizer combination in a grid search:

- **Preprocessing parameters**: n-grams (unigrams, unigrams+bigrams), min document frequency (1, 5), max document frequency (0.5, 1), preprocessing,[9] stopword removal, stemming
- **Training parameters**: class weights (balanced, proportional, none)
- **Hyper parameters (for Linear SVM)**: C (0.1, 1, 100)

We select the best model based on the f1-score for the positive class (i.e., PNR) on the held-out test set. We opt for this strategy since prediction of the negative class (i.e., unrelated to politics and current events) is both easy to learn the classifier due to class imbalance and not as informative to us in the context of this paper. The full model performance is presented in Table A1. The configurations for each model as presented in Table 3 are as follows:

- **Logistic Regression with Tfidf**: balanced class weights, min document frequency 1, max document frequency 0.5, preprocessing, unigrams, stopword removal
- **Logistic Regression with Count**: balanced class weights, min document frequency 1, max document frequency 0.5, unigrams+bigrams, stopword removal
- **Naive Bayes with Tfidf**: min document frequency 5, max document frequency 0.5, preprocessing, unigrams +bigrams, stopword removal
- **Naive Bayes with Count**: min document frequency 5, max document frequency 0.5, preprocessing, unigrams, stopword removal
- **Linear SVM with Count**: min document frequency 1, max document frequency 0.5, unigrams, C 1
- **Linear SVM with Tfidf**: balanced class weights, min document frequency 1, max document frequency 0.5, unigrams +bigrams, stopword removal, C 1

Table A1. Full performance: Traditional and Transformer-based SML.

model	label	precision	recall	f1-score	macro f1-score	N
Traditional Supervised Machine Learning						
Logistic Regression with Count	0	0.97	0.98	0.97	0.80	6550
	1	0.69	0.58	0.63		541
Logistic Regression with Tfidf	0	0.97	0.96	0.96	0.78	6550
	1	0.55	0.66	0.60		541
Naive Bayes with Count	0	0.95	0.99	0.97	0.76	6550
	1	0.83	0.41	0.55		541
Naive Bayes with Tfidf	0	0.94	1.00	0.97	0.70	6550
	1	0.90	0.28	0.43		541
Linear SVM with Count	0	0.96	0.98	0.97	0.80	6550
	1	0.75	0.55	0.63		541
Linear SVM with Tfidf	0	0.97	0.97	0.97	0.81	6550
	1	0.66	0.64	0.65		541
Transformer-based Supervised Machine Learning						
BERTje	0	0.96	0.98	0.97	0.80	6550
	1	0.68	0.57	0.62		541
robBERT	0	0.97	0.97	0.97	0.82	6550
	1	0.68	0.67	0.67		541
robBERT 2022	0	0.97	0.98	0.98	0.82	6550
	1	0.74	0.61	0.67		541
Multilingual BERT	0	0.97	0.98	0.97	0.81	6550
	1	0.74	0.58	0.65		541

[9]lowercasing and removal of punctuation and numbers.

Transformer-based Supervised Machine Learning

We fine-tuned four pre-trained models, i.e. BERTje (De Vries et al., 2019), robBERT (Delobelle et al., 2020), robBERT 2022 (Delobelle et al., 2022), and multilingual BERT (Devlin et al., 2018). During fine-tuning, we experimented with the following (hyper)parameters:

- **Training parameters**: class weights (balanced, proportional, none)
- **Hyper parameters**: batch size (16, 32), learning rate (5e-5, 3e-5, 2e-5), number of epochs (2, 3, 4), number of warm up steps (0, 1, 1000)

We did not apply much preprocessing (other than the removal of punctuation and digits and lower casing) because BERT models can benefit from contextual information. Note that not all combinations are tested (due to computational resources), but some were excluded based on performance in prior iterations. The best performing model for each pre-trained BERT model is selected based on the f1-score for the positive class (i.e., PNR) on the held-out test set, as presented in Table 3. The full model performance is presented in Table A1. The configurations for each model as presented in Table 3 are as follows:

- **BERTje**: balanced class weights, number of epochs 3, learning rate 5e-5, batch size 16, number of warm up steps 0
- **robBERT**: balanced class weights, number of epochs 3, learning rate 5e-5, batch size 16,number of warm up steps 1 number of warm up steps 1

- **robBERT 2022**: number of epochs 3, learning rate 2e-5, batch size 16, number of warm up steps 1
- **Multilingual BERT**: balanced class weights, number of epochs 3, learning rate 5e-5, batch size 16, number of warm up steps 1

Appendix B: Details on zero-shot classification Reasoning extraction prompt variations

The following four reasoning extraction prompt variations are used to instruct GPT-3.5 and Llama 3 (8b) in a zero-shot setting.

1. Short definition without date indication

Input variables: search query

Prompt: Is the search query "{search query}" political or news-related when searched in the Netherlands (yes or no)? Political or news-related search queries are defined as seeking information contributing to opinion formation on political and societal topics. Give your reasoning.

2. **Short definition with date indication**

Input variables: search query, date (month, year)

Prompt: Is the search query "{search query}" political or news-related when searched in {date} in the Netherlands (yes or no)? Political or news-related search queries are defined as seeking information contributing to opinion formation on political and societal topics. Give your reasoning.

3. Long definition without date indication

Input variables: search query

Prompt: Is the search query "{search query}" political or news-related when searched in the Netherlands (yes or no)? Answer yes or no for whether the search query is political or news-related. Give your reasoning. Political or news-related search queries are defined as seeking information contributing to opinion formation on political and societal topics. This includes search queries about (international and national) political actors (e.g., political parties, politicians), elections, policy, political events (e.g., statements from political actors), news media (e.g., nos.nl, RTL Nieuws, NOS journaal vandaag) talk shows or programs that focus on societal themes (e.g., Op1, Boos, Zembla) or figures in these media (e.g., Tim Hofman). It also includes search queries seeking out general information or news about societal themes (e.g., climate change, immigration, COVID-19, LGBT+, racism, crime, economy, war, etc.), but excludes those about practical information about these themes (e.g., checking pension benefits, getting vaccinated). In cases where it is unclear whether the search term is seeking general information or news about societal themes or practical information, follow the following rule: If the searcher's intention can be interpreted as interested in finding news or information about societal themes as well as practical, then answer yes (e.g., wait time for booster shot, easing of restrictions in France). If the search term can only be interpreted as seeking practical information, then answer no (e.g., vaccination line for Jansen, I want to get vaccinated). A political or news-related search query can also be related to current events about political or societal themes (e.g., COVID-19, train strikes, The Voice scandal). Political and news-related search queries are not about (natural) disasters (e.g., earthquakes, accidents), entertainment news (e.g., celebrities, fashion, gadgets, food), sports, culture (e.g., music radio, festivals), unless when they concern policy related to these themes. A search term is never political or news-related when it concerns, for example, practical information (e.g., how long can you wear contact lenses, temperature tomorrow), shopping (e.g. IKEA Malm) or health.

4. Long definition with date indication

Input variables: search query, date (month, year)

Prompt: Is the search query "'{search query}" political or news-related when searched in {date} in the Netherlands (yes or no)? Answer yes or no for whether the search query is political or news-related. Give your reasoning. Political or news-related search queries are defined as seeking information contributing to opinion formation on political and societal topics. This includes search queries about (international and national) political actors (e.g., political parties, politicians), elections, policy, political events (e.g., statements from political actors), news media (e.g., nos.nl, RTL Nieuws, NOS journaal vandaag) talk shows or programs that focus on societal themes (e.g., Op1, Boos, Zembla) or figures in these media (e.g., Tim Hofman). It also includes search queries seeking out general information or news about societal themes (e.g., climate change, immigration, COVID-19, LGBT+, racism, crime, economy, war, etc.), but excludes those about practical information about these themes (e.g., checking pension benefits, getting vaccinated). In cases where it is unclear whether the search term is seeking general information or news about societal themes or practical information, follow the following rule: If the searcher's intention can be interpreted as interested in finding news or information about societal themes as well as practical, then answer yes (e.g., wait time for booster shot, easing of restrictions in France). If the search term can only be interpreted as seeking practical information, then answer no (e.g., vaccination line for Jansen, I want to get vaccinated). A political or news-related search query can also be related to current events about political or societal themes (e.g., COVID-19, train strikes, The Voice scandal). Political and news-related search queries are not about (natural) disasters (e.g., earthquakes, accidents), entertainment news (e.g., celebrities, fashion, gadgets, food), sports, culture (e.g., music radio, festivals), unless when they concern policy related to these themes. A search term is never political or news-related when it concerns, for example, practical information (e.g., how long can you wear contact lenses, temperature tomorrow), shopping (e.g. IKEA Malm) or health.

Context-dependent LLM results

In theory, adding a date can help LLMs to situate the search term in a given context. For instance, for the query "sister Holleeder stabbed" GPT-3.5 returned "*the query refers to a recent event where the sister of a notorious Dutch criminal was stabbed, which is a current event related to crime and safety in society*." Selecting only queries searched before GPT-3.5's training threshold of September 2021 and comparing the performance between instructions with and without a date indication, reveals that such date indication does not result in substantially improved results, and even lowers the f1-score when defining the concept more explicitly (see Table B1). Note that we could not fully explore this because most search terms were searched after GPT-3.5's training threshold. For Llama 3, the model training threshold is March 2023, therefore the whole set could be annotated with the date indication. Interestingly, the results (see Table 3) show that with the long prompt adding the date did not lead to a substantial improvement in performance, however for the short prompt precision increased substantially (at the expense of recall). The interpretation of this remains speculative with only this one task to go from, but it does indicate that with only little instructions given the additional contextual information improved the correct identification of PNR search terms.Future research should investigate whether contextualizing content in this or a similar manner improves the performance.

Table B1. Performance of GPT-3.5 instructions with and without date indications on pre-training date threshold search queries.

model	precision	recall	f1-score	N
Short definition	0.53	0.63	0.58	150
Short definition & date indication	0.44	0.75	0.55	150
Long definition	0.41	0.85	0.55	150
Long definition & date indication	0.45	0.83	0.58	150

Appendix C: Examples classifications

We recommend to always inspect some of the classifications manually, in order to get a clearer picture of where different methods get it wrong. Table C1 shows some examples based on a random sample of the search terms labeled as PNR and non-PNR.

Table C1. Classifications of methods based on a random sample of five PNR and five non-PNR search terms.

Search query	Annotation	Browsing seq	Cont. dictionary	Trad. SML	Transf. SML	Zero-shot
delays ns	1	0/1	1/1	0	0	1
adverse effects incorrect use face mask	1	2/3	1/1	0	1	1
testing flying back to the Netherlands	1	1/1	1	1	1	1
guinea pig corona	1	0/2	1/1	1	1	1
baudet receives money from russia[a]	1	1/2	1/1	1	1	1
jan host[b]	0	0/1	0/1	1	0	0
why was a clockwork orange banned	0	0/2	0/1	0	0	1
current programs	0	1/3	0/1	0	0	1
complaint ns[c]	0	0/4	2/2	0	0	1
corona check app qrcode	0	0/7	1/1	1	1	1

Note. Recall that methods are applied to different levels: content-based methods (i.e. Traditional SML, Transformer-based SML and Zero-shot classification) operate at the level of a search term, the context-enhanced dictionary method combines query and date, and the browsing sequences method combines query and selected website. Hence, the latter two methods are presented as the share of search terms that are predicted as PNR (i.e. *frequency labelled PNR /frequency*).
[a]Baudet is a Dutch politician. [b] Host in the context of a television program. [c] NS is the Dutch railway operator.

Appendix D: Codebook Political and News(-related) Search Queries

Find below the codebook section used for the manual annotation of political and news-related search queries. Note that the codebook is translated from Dutch for the purpose of this appendix. For the full untranslated codebook, see https://github.com/mariekevh/GooglingPolitics.

Coding Political and News-related Queries

You will answer some questions about search terms. These search terms are derived from browsing histories. Based on your answers, we can categorize search terms to understand how people search for information online about political and societal topics.

You will be presented with a search term, the website(s) the respondent visited subsequently, and the date of the search. If no websites are listed, it means the respondent either did not visit any website or the website contained sensitive information. For each search term, first determine if it is related to politics or news. If yes, answer two follow-up questions. If not, stop coding and move on to the next search term. Qualtrics will ensure you see the appropriate questions. For each question, you can indicate "I don't know" and provide an explanation for your uncertainty. Another coder will evaluate these search terms later. It is possible that after answering the questions, you realize you would have liked to give a different answer. At the end of the questionnaire, there is space to leave comments. You can explain what went wrong during the question-answering process and how it should have been done differently. Follow these steps for each search term:

(1) Answer the questions based solely on the search term.
(2) If the answer is not immediately clear, you may consult the visited website(s). NOTE: Try not to be too influenced by the website! Even if a respondent visited an unrelated website or did not visit any website, the search term may still be relevant.

- For example: It may not be immediately clear that the search term "Sterrebos" is related to politics or news. However, based on the fact that the respondent visited the news website *nos.nl*, you can infer that this search term is related to news or politics. The search term refers to an environmental action in the Sterrebos.

3. If you do not know what the search term means, still have doubts, or the search term contains a name you are unfamiliar with, you should Google the search term. You can do this by clicking on the provided link. It will open a new

window with the search results for the search term within the original time frame (3 days before and after). Consult the first search results page to answer the questions.

Q1 - Is the search term **related to politics or news**?

- This means that the intention of the search query was to find information that could help form an opinion on political and societal topics.
- This includes references to (international and national)...
- Political actors (such as political parties, politicians (e.g., *Justin Trudeau*, *Putin*))
- Elections
- Policies
- Political events (such as statements by political actors)
- News media (such as nos.nl, talk shows, or programs focusing on societal issues (e.g., *Op1*, *Boos*, *Zembla*) (e.g., *RTL Nieuws*, *NOS Journaal Vandaag*), or individuals in these media (e.g., *Tim Hofman*))
- Societal topics (such as climate, immigration, COVID-19, LGBT+, racism, crime, economy, war, ...)

*Excluding: search queries for practical information related to these topics

(e.g., buying Volkswagen shares, selling a house without a real estate agent, checking pension amount, I want to get vaccinated).

*In case it is unclear whether the search query is seeking information/news about societal topics or practical information, follow the following guideline. If the intention can be interpreted as an interest in finding news or information about societal topics in addition to practical purposes, categorize it as related (e.g., *wait time for booster shot, easing of restrictions in France, new COVID-19 measures*). If the search query can only be understood as seeking practical information, categorize it as unrelated (e.g., *vaccination hotline for Jansen, I want to get vaccinated*).

- Current affairs on political or societal topics (e.g., *COVID-19 figures on January 9, 2022, train strike, The Voice scandal, Brexit river pollution*)

*This includes current affairs related to the aforementioned societal topics.
*This does not include searches for current affairs regarding (natural) disasters (such as earthquakes, accidents), entertainment news (such as celebrities, fashion, gadgets, food), sports, culture (such as music, radio, festivals), unless it pertains to policies related to these topics.

- Answer "no" if the search query is about:

*Practical information (e.g., *how long can you wear contact lenses, tomorrow's temperature*)
*Shopping (e.g., *IKEA Malm*)
*Entertainment (e.g., *The Voice of Holland, YouTube Slam FM*)
*Health (e.g., *menopause)*

Appendix E: Sampling procedure

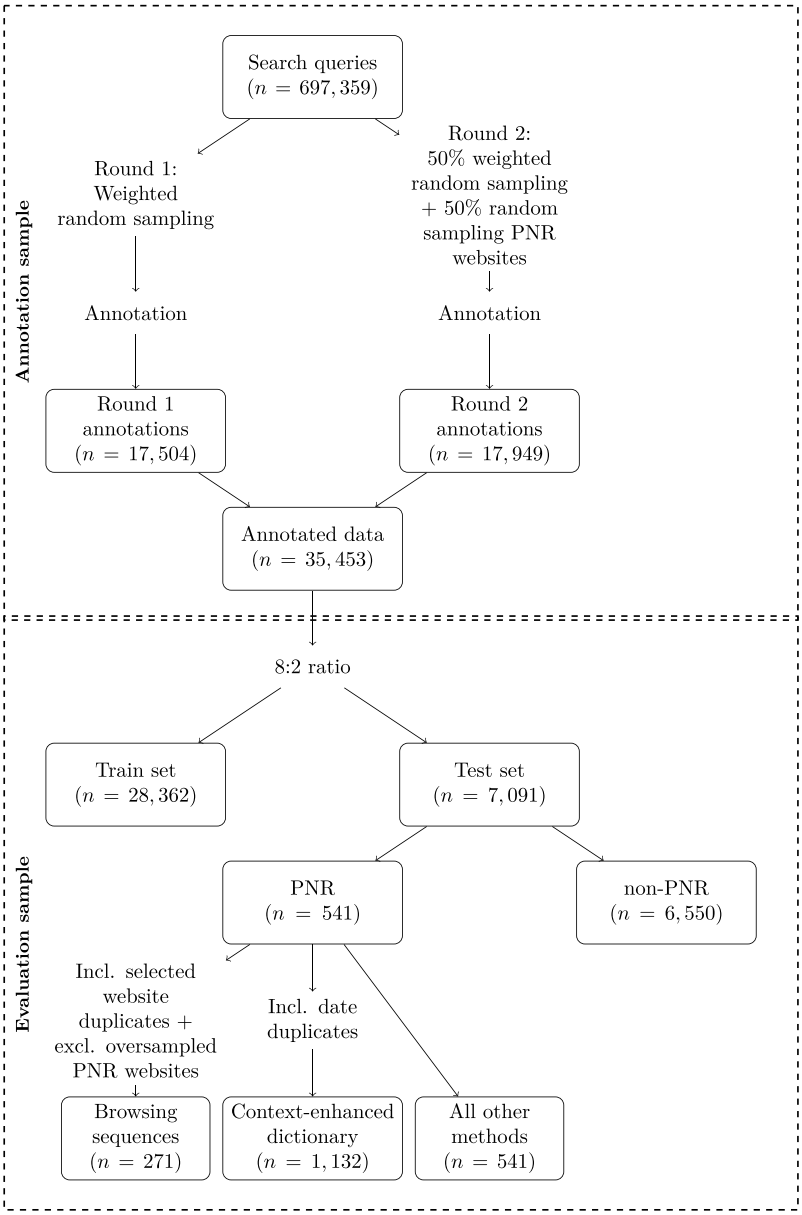

Figure E1. Steps of sampling procedure.
Note. The annotation sampling procedure involved two consecutive rounds. In round 1, we used weighted random sampling of search queries, giving searches from users with high volumes of browsing a lower probability of being selected. Due to the extreme class imbalance in round 1 (only ~3.5% PNR according to coders), the approach for round 2 was adjusted. In round 2, 50% of the sample was selected using the same weighted random sampling approach from round 1, while the other 50% was a random sample of queries that led to PNR websites.