9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Googling Politics? The Computational Identification of Political and News-Related Searches from Web Browsing Data

Keywords: political communication, search engines, news consumption, browsing data, automated content analysis

Extended Abstract

Today's media landscape offers countless sources of information needed for well-informed political decision-making. Not only have search engines emerged as important access points for news [6], search results can influence voter preferences [3] and can contain misinformation [7]. This highlights the democratic implications of search engine use. Yet, there is limited understanding of how search engines fit into the public's information routines: Who is using search to stay informed about politics and current events – and how frequently? While some prior studies have produced valuable insights [5, 2], this research has been held back by a lack of a scalable approach to classify political and news-related (PNR) searches in browsing data. The large volume of browsing, the brevity of the search queries, and context-sensitivity of what should be PNR complicate this task.

This paper aims to fill this gap by comparing five computational methods for distinguishing PNR searches. We outline the limitations and advantages of each method, both in terms of theoretical implications (e.g., missing relevant components of PNR search, context-sensitivity) and practical use (e.g., computational requirements, supplementary data), and propose the most suitable approach for this task. Finally, we demonstrate the practical application of this technique by answering questions about the frequency, search strategies and individual differences in PNR search behaviour. To this end, we use Dutch browsing history data coupled with self-reported characteristics ($N_{users} = 359, N_{records} = 9.8M, N_{searches} = 700k$). Additionally, we use 35.5k manually annotated searches, which passed extensive validation by multiple researchers, and a corpus of news covering the entire browsing period.

We will evaluate the following approaches: a browsing sequences approach, dictionary approach, bag-of-words supervised machine learning (SML), BERT-based classification, and a context-sensitive approach.

Building on news referral research [6], we can use the **sequence of browsing** to label searches as PNR if they result in a news website visit. Although it offers a straightforward and scalable approach, it misses two other plausible outcomes of PNR search. First, it may *underestimate* PNR search as modern search result pages increasingly offer features aimed at providing direct answers (e.g., Knowledge Panel) , which reduces the need to click on search results. Second, this approach may lead to *mislabelling* because PNR search can lead to background information websites (e.g., Wikipedia) or lesser known information channels (e.g., blogs, alternative news websites) excluded from the list of news websites, and news websites offer non-PNR content (e.g., weather, entertainment) which can also be accessed via search.

This underscores the need for an approach based on the search terms' content, instead of inferring meaning from potential subsequent activity. A commonly used **dictionary approach** allows for transparency and low computational requirements, but creating an exhaustive dictionary for a complex topic is not straightforward and off-the-shelve validated dictionaries for

PNR content are not readily available. They are also often outperformed by more sophisticated SML methods [1] that are able to classify data in an inductive manner. In this category, a **BERT-based classifier** built on top of a large language model often outperforms **bag-of-words classifiers** [4] because they can classify words not in the training data, which is particularly useful for short texts. Yet, SML requires annotated data which is costly to obtain, especially for minority class situations, like PNR search. Furthermore, BERT models have high computational requirements and typically require GPU access for fine-tuning.

Another limitation of these approaches, is that they are mainly able to detect a core set of words that are always PNR (e.g., *parliament*), but what is news-related is highly context-dependent (e.g., *Bucha* was once just a place in Ukraine, but now has a different meaning). We develop such a **context-sensitive approach** by mapping searches to a moving window on a corpus of news, which means that searches containing words that are important during a news event will be scored higher. This approach will be used as a context-sensitive layer on top of another approach, but at the price of extensive supplementary data.

Preliminary results for two of the five approaches (i.e., sequence-based and BERT-based) are shown in Table 1. A sequences-based approach does not capture PNR search accurately, despite using an exceptionally extensive list of (institutionalised and alternative) news and background information websites ($n = 472$). We flag between 66% and 75% searches as PNR when they are not. The low recall signals that we miss many PNR searches, which is especially the case if one only relies to visits to news websites, which is what has been done to date [6]. As expected, we observe a clear improvement in performance for BERT-based classification, particularly when accounting for class imbalance and using a Dutch-specific BERT model. Methodologically, these preliminary results confirm our expectation that we need approaches that work with the search term's content. Substantially, this may indicate that those who use search to inform themselves about politics and current events, in many cases do not end up on news websites. These findings also suggest that to move beyond a partial understanding of PNR search behaviour, extensive supplementary data, such as annotations, are necessary.

References

[1] Wouter van Atteveldt, Mariken A. C. G. van der Velden, and Mark Boukes. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms". In: *Communication Methods and Measures* 15.2 (Apr. 2021), pp. 121–140.

[2] Sina Blassnig et al. "Googling Referendum Campaigns: Analyzing Online Search Patterns Regarding Swiss Direct-Democratic Votes". In: *Media and Communication* 11.1 (Jan. 2023), pp. 19–30.

[3] Robert Epstein and Ronald E. Robertson. "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections". In: *Proceedings of the National Academy of Sciences* 112.33 (Aug. 2015), E4512–E4521.

[4] Zilin Lin et al. "Beyond Discrete Genres: Mapping News Items onto a Multidimensional Framework of Genre Cues". In: *arXiv preprint arXiv:2212.04185* (2022).

[5] Ericka Menchen-Trevino et al. "Searching for politics: Using real-world web search behavior and surveys to see political information searching in context". In: *The Information Society* 0.0 (Dec. 2022), pp. 1–14.

[6] Judith Möller et al. "Explaining Online News Engagement Based on Browsing Behavior: Creatures of Habit?:" in: *Social Science Computer Review* (Feb. 2019).

[7] Aleksandra Urman et al. "Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results". In: *Telematics and Informatics* 72 (Aug. 2022), p. 101860.

Table 1: Model performance

Model	Precision	Recall	F1	N
Browsing-sequence-based classification				
institutionalized news	.31	.13	.18	1070
inst. and alternative news	.31	.13	.18	1070
inst. and alt. news and background information	.26	.32	.28	1070
BERT-based classification				
Dutch BERT, unbalanced class weights				
batch size 16, epochs 2, learning rate 2e-5	.77	.52	.62	541
batch size 16, epochs 3, learning rate 2e-5	.74	.55	.63	541
batch size 16, epochs 4, learning rate 2e-5	.76	.55	.64	541
batch size 32, epochs 2, learning rate 2e-5	.76	.53	.63	541
batch size 32, epochs 3, learning rate 2e-5	.71	.61	.66	541
batch size 32, epochs 4, learning rate 2e-5	.72	.59	.65	541
Dutch BERT, balanced class weights				
batch size 16, epochs 2, learning rate 2e-5	.73	.55	.63	541
batch size 16, epochs 3, learning rate 2e-5	.71	.59	.65	541
batch size 16, epochs 4, learning rate 2e-5	.73	.56	.64	541
batch size 32, epochs 2, learning rate 2e-5	.54	.68	.60	541
batch size 32, epochs 3, learning rate 2e-5	.60	.65	.63	541
batch size 32, epochs 4, learning rate 2e-5	.68	.59	.63	541
Multilingual BERT, unbalanced class weights				
batch size 16, epochs 2, learning rate 2e-5	.79	.49	.60	541
batch size 16, epochs 3, learning rate 2e-5	.77	.57	.65	541
batch size 16, epochs 4, learning rate 2e-5	.73	.59	.65	541
batch size 32, epochs 2, learning rate 2e-5	.77	.54	.64	541
batch size 32, epochs 3, learning rate 2e-5	.76	.61	.68	541
batch size 32, epochs 4, learning rate 2e-5	.71	.63	.67	541

Note: Due to the class imbalance, performance scores for predicting non-PNR are uninformative and therefore not displayed. The N is unequal between approaches due to the train-validation-test-split for fine tuning BERT-based classifiers.