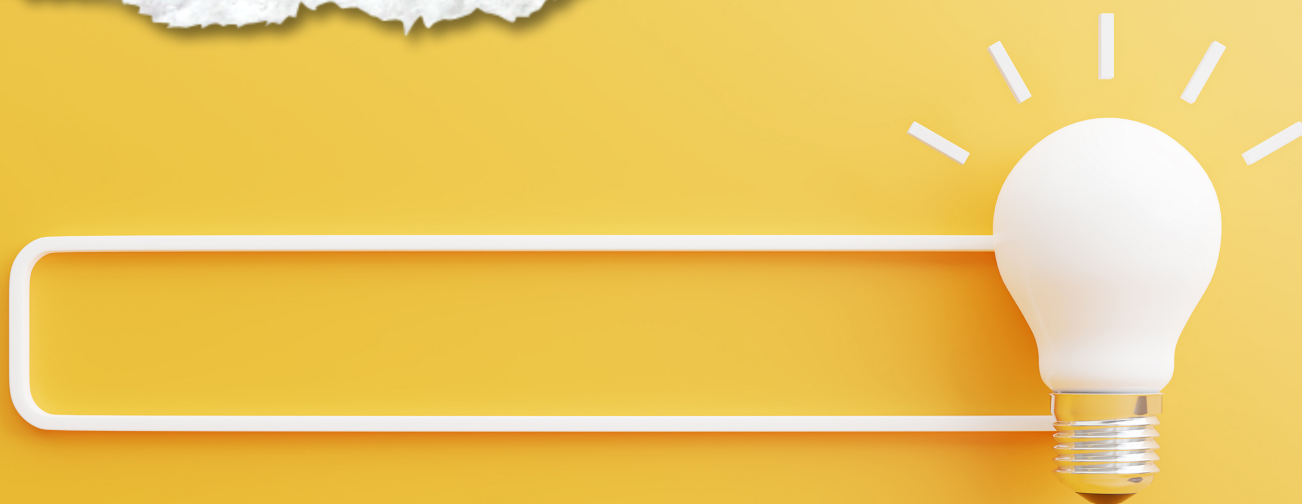




INSIDE THE "BLACK BOX"
AN INVESTIGATION INTO SEARCH ENGINE
RESULTS PAGE

JULIE VUILLERMOZ



INSIDE THE “BLACK BOX”

An investigation into search engine result page.

Julie M. Vuillermoz

Department of Communication Science

Faculty of Behavioural, Management & Social Sciences

University of Twente

2023-202000308. Bachelor Thesis

Dr. A.A.C.G. Van Der Graaf

July 1, 2024

Words count: 11031

ACKNOWLEDGMENT

Completing my bachelor's degree has been a remarkable journey, and I couldn't be more grateful for the decision to pivot my career and return to school to gain more knowledge and skills. Reflecting on the past three years, I find them incredibly insightful. I particularly enjoyed my third year, where I had the flexibility to focus on what truly interested me. Living in the Netherlands and completing my minor in Greece have broadened my horizons, exposing me to different cultures and ways of thinking. These experiences have taught me acceptance, adaptability, resilience and the importance of kindness. I couldn't have chosen a better place to pursue my bachelor's degree, and I am excited to continue my academic journey with a master's at the UT.

Regarding my bachelor's path, I am sincerely grateful to several persons whose support and guidance have been pivotal. First and foremost, I extend my appreciation to Shenja, who welcomed me into this enriching project. Her intellectual stimulation fuelled my curiosity and significantly enhanced my skills, playing an instrumental role throughout my studies, particularly during the completion of my thesis. The quality and depth of my thesis clearly reflect her invaluable mentorship along with the constructive feedback from Alex and Roel. I would also like to thank both of them for their assistance in refining my critical thinking and further developing my academic skills. Their unique perspectives and expertise have greatly enhanced my capabilities, each contributing in distinct ways to my approach and understanding.

On a personal note, the presence and encouragement of my partner have continually inspired me to pursue the best version of myself, filled with love and laughter. I am also grateful to my mother-in-law, whose her steadfast belief in me and her boundless kindness has been inspiring.

Thank you for being part of my journey.

Julia Vuilleumaz 

Table of Contents

ABSTRACT	1
1. INTRODUCTION	2
2. THEORETICAL FRAMEWORK	5
2.1. SERPS, FILTER BUBBLE AND ECHO CHAMBER	5
2.2. ENHANCING ALGORITHMIC FAIRNESS	6
2.3. EPISTEMIC EQUALITY	7
3. METHOD	9
3.1. ETHICAL CONSIDERATION	10
3.2. INSTRUMENT: BROWSER EXTENSION	10
3.3. PARTICIPANTS	11
3.5. DATA ANALYSIS	11
4. RESULTS	14
4.1. STEP 1: SCOPING.....	14
4.2. ANALYSIS FOR THE KEYWORD IMMIGRATION (PHASE 2-3)	15
4.3. ANALYSIS FOR THE KEYWORD SHELTER LOCATION (PHASE 2-3)	18
4.4. ANALYSIS FOR THE KEYWORD ASYLUM SEEKER (PHASE 2-3)	20
5. DISCUSSION	23
5.1. INFORMATION DIVERSITY.....	23
5.2. PERSONALISATION IMPACT	24
5.3. REFLEXION ON THE PROJECT'S INSTRUMENT	25
5.4 LIMITATIONS OF THIS STUDY.....	25
5.4 FURTHER RESEARCH	26
6. CONCLUSION	27
REFERENCES	28
APPENDIX	34

Abstract

Aim: The digital transformation of the media landscape has significantly impacted how the Dutch access information, largely through search engines operated by private companies. These engines use personalisation algorithms—often described as “black boxes”—to tailor content to individual preferences, which raises critical questions about the diversity of information accessed and its societal implications. This study seeks to understand the influence of search engine personalisation algorithms on information diversity in the Netherlands, exploring the extent to personalisation impact search results and their ranking mechanisms in term of information diversity. By investigating these epistemic rights, the study aims to highlight the importance of equal and fair access to diverse and accurate information, essential for informed knowledge acquisition. This research marks an initial step toward addressing these complex issues within the Dutch digital ecosystem.

Method: Data collection was conducted using a browser extension designed to capture search engine result page (SERPs) for 21 keywords across five distinct search engines. We employed an exploratory and quantitative analyses to identify patterns in the data. The analysis focuses on a single week, limiting it to one search per participant to prevent the duplication of searches by the same individuals. This approach was chosen to provide a more accurate representation of the diversity of information presented by different search engines under comparable conditions.

Result: The results indicate a low degree of information diversity, with relatively homogeneous search results, especially for Google. This suggests that there is a limited number of links, and the differences lie primarily in their ranking positions. The findings imply that the impact of personalisation algorithms is minimal. Additionally, the study determines “Gender” and “Political affiliation” as two dominant socio-demographic factors. While the accuracy of the models is low, more profound effect are expected with bigger sample. Further exploration of this aspect would be worthwhile.

Conclusion: These results indicate that the diversity of information within the Dutch digital ecosystem is notably limited, suggesting that individuals frequently encounter similar search results regardless of their demographic differences. From a societal viewpoint, this could suggest a form of epistemic equality; however, from an individual perspective, it restricts exposure to diverse viewpoints, funnelling users towards a mainstream narrative. A valuable direction for future research would be to delve into the specific types of content presented to users to better understand which information is deemed mainstream.

Keywords: search engine algorithm, personalisation, information diversity, epistemic rights, pluralism,

1. Introduction

As society rapidly transitioning into a digitally dominant era, the mechanisms through which individuals access and consume information have undergone significant transformations. The Netherlands exemplifies this shift, with an almost complete internet penetration rate of 99% (Kemp, 2024). Despite the high reliance on digital platforms, traditional news sources remain predominant; a notable 94% of Dutch citizens primarily rely on mainstream media outlets (NCTV, 2022), such as NPO and RTL (Swert et al., 2023), to stay informed. However, digital diversification is evident with 40% of the population also turning to search engines for news (NCTV, 2022). Notably, Google holds a dominant 91% share of this market, underscoring its significant influence in the online information ecosystem (Bianchi, 2024b). This reliance on digital news is paralleled by increasing concerns over information integrity. Approximately 47% of Dutch individuals express apprehensions regarding disinformation online (NCTV, 2022; Swert et al., 2023), signalling a growing mistrust in the digital discourse. This sentiment is starkly illustrated in a CBS's survey (2024), which found that the Netherlands had the highest incidence among the 27 EU member states of individuals aged 16 to 74 encountering what they perceived as false or dubious information online — reaching an alarming rate of 71%. Furthermore, a significant portion of the population, 69%, believe that disinformation is a primary cause of more societal division (NCTV, 2022). This prevalence of scepticism and concern over disinformation in a highly digital society such as the Netherlands mirrors global challenges related to digital information access and manipulation.

The substantial influence of search engines in shaping public discourse in the Netherlands is indicative of a broader, international phenomenon, where entities like Google significantly impact information consumption worldwide. As a matter of facts, Google commands a staggering 81% of the global search engine desktop market share (Bianchi, 2024a), orchestrating approximately 8.5 billion searches daily (Flensted, 2024), showcasing its unparalleled influence in the digital information sphere. Despite Google's dominance, Bing has shown remarkable growth, increasing its market share from 4.5% in 2015 to 10.5% in 2024 (Bianchi, 2024a), conducting around 900 million searches each day (Ch, 2024). Furthermore, the exploration of search query durations reveals a notably swift average of 53 seconds on Google (Lindner, 2024) with only 9% of the users navigating to the bottom of the first page (Dean, 2023) and 0.44% to the second-page results (Golebiewski & Boyd, 2019; Lindner, 2024). Acknowledging that, Google in 2016 introduce the direct answer snippet (Strzelecki & Rutecka, 2020), delivering concise answers directly in search results (Wu et al., 2020). It significantly enhances the user experience by improving the process of search (quick, and efficient) and eliminating the need for further website navigation (Tucker & Edwards, 2021). The alignment between user search intent and expected results bears economic significance for search engines, notably through the monetisation of user data and the promotion of sponsored content (Tucker & Edwards, 2021; Varian, 2006). This dynamic, combining personalisation, optimisation, and sponsored content, can be seen to shape ranking algorithms and have raised a series of concerns about information accessibility and equality.

Search engine and recommendation systems have been proven to be equipped with algorithmic biases, including those of race, culture, and gender (Noble, 2018), affecting the diversity and fairness of search results. Building upon similar considerations, language bias has been highlighted in Google searches to significantly affect the results obtained (Luo et al., 2023). This seems to align search results with the cultural context of the languages used, enhancing user satisfaction by better meeting their

cultural expectations. Lastly, human bias, including selective exposure and confirmation bias, significantly influences online behaviour (Saetra, 2019; Slechten et al., 2021), where individuals tend to seek information that aligns with their existing beliefs. Search engines are said to amplify this tendency by personalising search results based on past user behaviour, reinforcing biases and enclosing users within so-called echo chambers (Sunstein, 2001). In these echo chambers, information mirrors users' beliefs, attitudes, and viewpoints (Aguado & Hermida, 2022; Arguedas et al., 2022; Fletcher & Nielsen, 2018; Kolic et al., 2022; Mahmoudi et al., 2024). Personalisation also creates a feedback loop, also known as a filter bubble (Pariser, 2011), an epistemic structure (Furman, 2023; Nguyen, 2020), that limit exposure to a broader range of ideas, reinforcing existing beliefs and potentially narrowing understanding of the world (Abul-Fottouh et al., 2020; Yang et al., 2023). This "personalised universe of information" (Pariser, 2015) may distort public opinion by isolating individuals from broader discussions (Aguado & Hermida, 2022), reducing open-mindedness and contributing to polarisation (Einav et al., 2022). Such dynamics increase divisions in beliefs, opinions, and interactions online (Aguado & Hermida, 2022; Arora et al., 2022; Cano Macias & Ruiz Vera, 2024; Interian et al., 2023; Jones-Jang & Chung, 2022; Valensise et al., 2023; Y. Wu et al., 2023; Yi & Patterson, 2020). In this view, search engine algorithms and recommendation systems often are said to contribute to the amplification of echo chambers and filter bubbles, reinforcing existing viewpoints and isolating users from contrasting perspectives. However, this landscape also presents a paradoxical opportunity for enhancing diversity and equality. By consciously diversifying the algorithms and reducing biases in recommendation systems, platforms could expose individual user to a wider array of content and perspective, potentially enriching public discourse and fostering a more inclusive digital environment. This approach supports epistemic rights, ensuring that users encounter a broad spectrum of viewpoints and information, which is essential for a well-informed and critically thinking society. Additionally, it upholds the rights of individuals to equally access information sources, thus decentralising knowledge power and promoting a more equitable distribution of information. Therefore, this underscores the need for special care in managing digital information gatekeepers to ensure diversity and equality in the presented results.

The phenomenon of polarisation is comprehensively analysed in the field of news through various methodologies: sentiment analysis (Alam et al., 2022; Ludwig et al., 2023), content analysis to assess the effects of personalisation on news content diversity (Evans et al., 2023; Haim et al., 2018) or cluster analysis to analyse fragmented networks (Gaol et al., 2020) integrating linguistic parameter to examine how misinformation contributes to polarisation (Ruffo et al., 2023). Moreover, polarisation is thoroughly examined in political discourse, analysing aspects such as political attitudes (Feezell et al., 2021), the nexus of radicalism and political violence (Burton, 2023), the impact on democratic processes (Cho et al., 2020) as well as belief reinforcement or information diversity (Courtois et al., 2018; Dylko et al., 2017). Although these studies provide detailed insights into the mechanisms of polarisation within news and political discourse, they primarily focus on theoretical and momentary systemic outcomes rather than addressing the interplay between personalisation and information flow, particularly in relation to epistemic rights in everyday searches. This gap highlights the need for further research into the relationship between information diversity and the influence of algorithm-driven search engines on personal cognition and daily information processing. Although this study does not investigate the specifics of these algorithms or how users cognitively process the information they encounter, it does analyse the outcomes these search engines present, particularly within the Dutch cultural context. By auditing various search engines using a browser extension and analysing quantitatively the search result within a specific cultural context, this research aims to enhance the

understanding of digital information diversity and impact of personalisation. The study intends to shed light on the dynamic interplay between search results, ranking algorithms, and information diversity by addressing the following research question:

To what extent do search engine result pages and its ranking algorithm contribute to shaping the information diversity within the digital search ecosystem in The Netherlands?

This study positions itself within the broader framework of epistemic rights, a concept that underscores the critical importance of equitable access to all relevant and accurate information, and its profound impact on knowledge acquisition (Nieminen, 2024). As an initial exploration, this study aims to unravel the complexities of how personalisation in information dissemination affects these rights.

The structure of this paper is as follows: It begins with an exploration of algorithms and Search Engine Result Page (SERPs), focusing on its ranking process, links diversity, and a review of prior research conducted on these topics. This is followed by a detailed description of the study's methodology, data used and each step of the analysis. Subsequently, the results section presents the findings, which are then thoroughly examined in the concluding discussion. The discussion interprets the findings but also acknowledges the study's limitations and proposes avenues for future research.

2. Theoretical framework

The theoretical framework of this study delves into the nuanced aspects of SERPs, ranking system, and information diversity. It further explores the overarching theme of digital epistemic rights—specifically, the equitable accessibility of information and its implications. This research serves as an initial step into addressing these significant challenges. Although biases related to algorithms, language, and human factors have been extensively studied previously, this study shifts focus towards the outcomes—what is accessible to users—rather than the underlying processes (i.e., the algorithms). The objective here is to uncover how these elements influence the accessibility and visibility of information online, thereby offering deeper insights into the digital mechanisms that govern our access to knowledge.

2.1. SERPs, filter bubble and echo chamber

Recent studies have extensively explored search engine and recommendation algorithms to grasp their impact on information dissemination (Boeker & Urman, 2022; Rowland et al., 2023). These studies specifically examined the role of algorithmic bias in content diversity (Abul-Fottouh et al., 2020) and the effects of personalisation on user experience (Bastian et al., 2019; Chaney et al., 2018).

The private ownership of most search engines complicates data collection and analysis, prompting the focus of scrutiny on their results. SERPs have typically been designed to display ten links, ranked by relevance to the user's query. Each entry includes a clickable title, the document's URL, and a brief summary or snippet (Wu et al., 2020). The ranking of these links was revolutionised by the introduction of the PageRank algorithm by Larry Page, which assesses the importance of web pages based on the number and quality of links to them (Rogers, 2002). This algorithm does not consider the individual user's interests or search history, it is purely based on the interconnectedness and perceived importance of web pages. Unlike PageRank, which is the same for everyone who performs the same search, personalisation algorithm tailors search results to the individual user based on their unique behaviour, preferences, and past interactions with the search engine. In other words, personalisation algorithms adjust the content displayed to match individual user preferences, amplifying filter bubble and echo chamber phenomena (Eg et al., 2023; Graham, 2022). Filter bubbles manifest mainly digitally, through algorithmic filtering mechanisms, but also in physical settings influenced by homophily—the tendency for individuals with similar beliefs, behaviours, and habits to group together (Abul-Fottouh et al., 2020). These algorithmic filtering mechanisms, particularly in search engines, aim to deliver highly personalised and pertinent results, tailoring information access based on user preferences and historical interactions. The issue with filter bubbles is that “they are invisible, and people do not realize that they are seeing something different than anyone else” (Lunardi et al., 2020, p.3).

Similarly, the concept of the echo chamber refers to environments, especially in online communities, where individuals are exposed primarily or exclusively to opinions and information that mirror and reinforce their own beliefs, attitudes, and viewpoints (Aguado & Hermida, 2022; Arguedas et al., 2022; Fletcher & Nielsen, 2018; Kolic et al., 2022; Mahmoudi et al., 2024). This phenomenon limits exposure to diverse perspectives, leading to a reinforcement of existing biases.

These issues highlight significant concerns regarding the right to equal access to information and safeguarding epistemic rights concerning knowledge (Napoli, 2024). UNESCO (2024), in its recent report, emphasized that tackling these challenges have been a priority for the past three decades,

aligning with the United Nations Sustainable Development Goal 16.10.2, which advocates for the necessity of pluralism to sustain a functioning democracy (Coeckelbergh, 2023). It underscores a global commitment to fostering diverse public discourse, essential for democratic health.

Focusing more narrowly however, the factors influencing search results and their rankings remain unclear. Robertson et al. (2018) who researched the bias in search engine rankings in voting intention in eight experiments involving 8000 participants, found negligible or non-significant difference between the SERPs in standard and incognito windows. This study suggests that, for example, Google's personalisation minimally impacts the search results users receive, indicating a uniform access to information. However, this result is contradicted by findings that user login status, IP address, browsing history (Yang et al., 2023), and location (Ashokan & Haas, 2021; Rovira et al., 2021) do influence search results (Kliman-Silver et al., 2015). The complexity of isolating these factors means their individual effects are often intertwined, making it difficult to fully understand the type and ranking of displayed information. Additionally, evidence of algorithmic bias suggests that search result rankings disadvantage underrepresented minorities (Cui et al., 2022) indicating that current search engines may not provide fair and equal information dissemination.

While it is very challenging to influence search engines directly and access to their mechanism, as they remain private "black boxes," it is crucial to focus on offering fair and equal results to minimize or prevent the seemingly increasing societal division and polarisation.

2.2. Enhancing Algorithmic Fairness

In an earnest attempt to address those issues, Lunardi et al. (2020) and Ping et al. (2024) highlighted three quality dimensions that extend beyond mere accuracy to foster greater fairness in algorithms: diversity, novelty, and serendipity. While diversity refers to the inclusion of a wide range of content from various sources, viewpoints, or categories within search results, avoiding the risk of overrepresentation of any single viewpoint, source, or category; novelty relates to the freshness or uniqueness of the content provided in search results. An algorithm that prioritizes novelty seeks to present new, original, or previously unseen content to users, rather than repeatedly showing the same or highly similar information.

The concept of serendipity is defined as the fortuitous encounter with information while in pursuit of something else (Fletcher & Nielsen, 2018; Reviglio, 2019). This concept highlights the role of search engines in facilitating unexpected but valuable discoveries, enriching user experience by introducing them to content they were not initially seeking but find intriguing or useful. Reviglio (2019) argues that facilitating serendipity could serve as a countermeasure to the constriction of information diversity seen in filter bubbles and echo chambers by broadening the spectrum of information presented to users. The conundrum lies in achieving an ideal balance between personalisation, which delivers content aligned with the user's explicit interests, and serendipity (Huang et al., 2018; Lee, 2020). Central to the nature of serendipity are its defining characteristics of interestingness and unexpectedness (De Gemmis et al., 2015). Furthermore, Reviglio (2019) points out a fundamental contradiction in the effort to the act of deliberately creating serendipitous encounters, may inherently compromise the very principle of serendipity, suggesting that authentic serendipity cannot be fully automated. In response to this challenge, Kotkov et al. (2020) introduced a serendipity-oriented re-ranking algorithm, referred to as the Serendipity-Oriented Greedy Algorithm (SOG). However, they observed a trade-off where an increase in variety leads to a decrease in the accuracy of the search results.

An alternative approach involves presenting multiple perspectives of the same information to make readers aware of diverse viewpoints. Einav et al. (2022) in their study conducted an experiment in 2017, leading to the creation of "The Perspective." This innovative platform was conceived to counteract the effects of filter bubbles by presenting both sides of various issues. For instance, when confronted with a question such as "Should the US intervene militarily in foreign conflicts?" the website provided arguments both in favour of and against the statement. The outcome of this experiment revealed that reading articles from this website not only increased open-mindedness but also decreased opinion polarisation. Such findings shed light on the potential influence of social desirability bias or demand characteristics on research outcomes. In the same line, Epstein et al. (2017) conducted a political study with 3,600 participants to mitigate the Search Engine Manipulation Effect (SEME) using a fabricated search engine, Kadoodle, which notified users about ranking biases towards political candidates. They tested three scenarios: no alert, low alert, and high alert. The findings revealed that a low alert reduced vote attitude shifts by 16.9%, while a high alert achieved a 25.2% reduction. These results emphasized the critical role of algorithmic transparency and the necessity for user awareness regarding search result biases. Nonetheless, the proprietary nature of search engines poses challenges in adopting such transparency and bias-counteracting measures to address algorithm fairness and equal epistemic right.

2.3. Epistemic equality

As previously mentioned, various factors, both algorithmic and human, influence the content and ranking of search results which represent merely a surface manifestation of a larger and more complex issue. Trinchini & Baggio (2023) noted that the COVID-19 pandemic significantly impacted how users engage with information and communication technology, along with the ethical implications arising from altered perceptions of reality through these technologies. Moreover, misinformation has created numerous distortions of reality, with technology facilitating the spread of these distortions (Ruffo et al., 2023). This issue is particularly pronounced among younger demographics, who frequently encounter and propagate misinformation through social media platforms. For instance, one-third of adults aged 18-29 nowadays regularly access news via TikTok (Matsa, 2023), reflecting a deep reliance on social media for information seeking. This dependence makes them especially vulnerable to encountering and spreading misinformation. Addressing this issue, Barack Obama emphasised in 2020 the importance of discerning truth in the digital age: "If we do not have the capacity to distinguish what's true from what's false, then by definition the marketplace of ideas doesn't work. And by definition our democracy doesn't work. We are entering into an epistemological crisis." (Valaskivi & Robertson, 2022, p.1).

Epistemology concerns the nature of knowledge, including beliefs about what constitutes knowledge and how it is constructed, focusing on rationality, pluralism, and autonomy (Hoggan-Kloubert & Hoggan, 2023). The formation of beliefs is closely linked to the type of information individuals encounter, its accuracy, and their capacity to evaluate its relevance. Figà Talamanca & Arfini (2022) argue that how information is presented can influence perception more profoundly than the information's content itself. This situation increases the entrapment of individuals within online echo chambers and epistemic bubbles, which carry severe consequences. Societally, these structures can contribute to group polarisation and jeopardise democratic processes; on a personal level, they are said to reinforce cognitive biases, reduce open-mindedness, and impair critical-thinking skills, fostering intellectual rigidity. Thus, understanding the principle of epistemic equality within the online information ecosystem is crucial. This understanding is not only essential for refining search engine

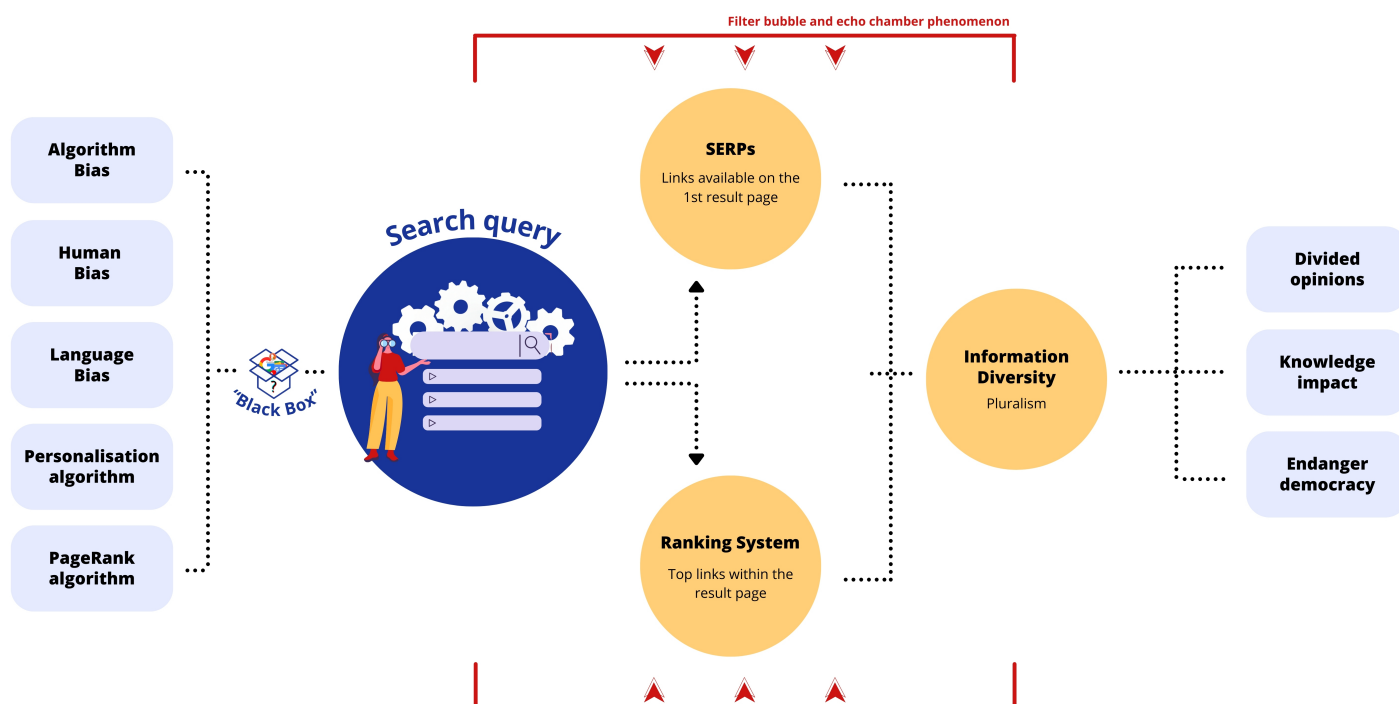
algorithms but also for promoting digital sustainability and combating information inequality (Trinchini & Baggio, 2023). In simpler terms, the role of daily searches and the information accessed through search engines are said to serve as significant epistemic agents, shaping human belief formation and affecting their comprehension of reality and truth.

In conclusion, online information ecosystem and more precisely the role of search engines therein in ensuring equal and diverse access to information remain poorly understood. Studies suggested that some underrepresented groups (Miah, 2024; Valdez & Javier, 2020), such as individuals with low digital literacy and those from economically disadvantaged backgrounds, face significant barriers that impact their access to technology and their digital skills, leading to an uneven distribution of information and knowledge. This disparity not only limits their ability to access diverse information but also affects their capability to critically assess and verify the truthfulness of that information, thereby increasing their vulnerability to misinformation.

In this view, this study evaluates the distribution of links within as well as across search engines to assess the diversity of links and identify any patterns across platforms. Additionally, we are examining the key socio-demographic factors influencing the SERPs. Ensuring fair and accurate information flow through search engines is crucial to prevent discrimination and ensure epistemic right. By analysing SERPs with a focus on ranking mechanisms and links' availability, illustrated in yellow in Figure 1, this study aims to understand information diversity from a communication perspective.

Figure 1

Epistemic rights framework

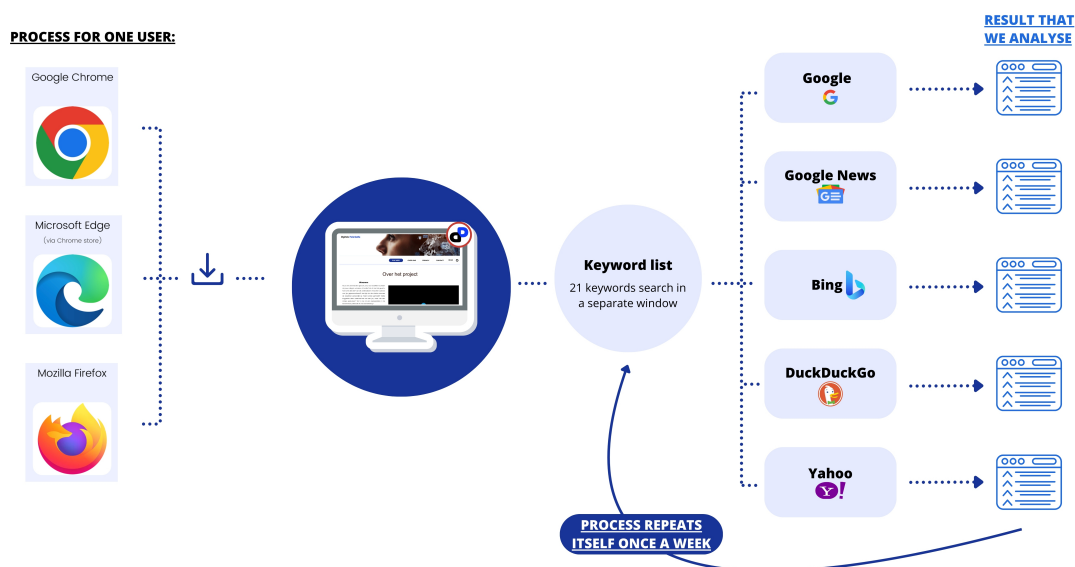


3. Method

This chapter provides a comprehensive examination of the methodology employed in this study, introducing the browser extension used and detailing its functionalities. Additionally, it outlines the profile of the participants involved in the study, including their demographic characteristics and the selection criteria used to include them in the research. The chapter further elaborates on the various analytical techniques applied in this study, explaining the statistical methods and data analysis procedures that were utilized to interpret the results and draw conclusions. Furthermore the methodology is clearly and comprehensively described, allowing other researchers to understand, replicate, or build upon the study.

This study is part of the "Digitale Polarisatie" project that was initiated in 2021. Its objective is to explore the diversity in the information presented in search engine results. To facilitate this information retrieval, the project has developed a browser extension that collects data from the SERPs of users (see Figure 2). The extension performs searches automatically once a week for each participant across multiple search engines, namely DuckDuckGo, Yahoo, Google, Bing, and Google News. It is important to note that these searches are conducted passively; users do not interact with them. Instead, searches are executed by means of opening a new window operated in the background, so to establish a controlled environment for systematic query generation. The methodology employed in this research is a systematic audit of the SERPs, utilising quantitative analysis to scrutinise the collected data. This study adopts an exploratory approach, aiming to identify and understand emerging patterns within the data. The decision to employ an exploratory methodology is particularly justified due to the absence of a predefined theoretical framework or hypothesis about the existing patterns of search engine results. Given the dynamic and often opaque algorithms that govern SERPs, this lack of predetermined expectations necessitates a flexible, discovery-oriented method that can adapt to the findings as they unfold. This approach is the most suitable for investigating complex systems where little is known about potential outcomes, making it an ideal choice in contexts where the phenomena under study are not yet well-defined.

Figure 2
Workflow of the Browser Extension



Note. This figure illustrates the operation of the browser extension associated with our project. The project's website, accessible digitalepolarisatie.nl provides comprehensive details on the project, including information on how to install the extension. Users can download the extension through major browsers such as Google Chrome, Microsoft Edge, and Firefox. Once installed, for each participant, the extension automatically executes searches once a week using 21 different keywords across five search engines: Google, Google News, Bing, DuckDuckGo, and Yahoo. For each search executed, the extension captures the first 10 links from the SERPs, in total 1050 links per participants, per search.

3.1. Ethical consideration

Prior to beginning the research, ethical approval was secured from the Ethics Committee BMS, with approval numbers 230687 for the investigation and 220261 for the browser extension. To comply with ethical standards, participants are required to give informed, voluntary consent and install the extension. All participants are adults over the age of 18 and reside in the Netherlands. The process of installing the extension involves a registration step where demographic information is collected, and consent forms are provided. Participants have the option to withdraw from the study at any time by disabling or uninstalling the extension.

3.2. Instrument: browser extension

The browser extension, developed by the BMS Lab from the University of Twente, serves as the primary data collection tool for this study. It is available for download on Google Chrome, Microsoft Edge, and Mozilla Firefox through the Digitale Polarisation Project's website (digitalepolarisatie.nl). Upon installation, users are prompted to answer demographic questions such as age, income, and postcode (detailed in Table 1 in appendix). By confirming these details on the pop-up page, users consent to the project's policy, allowing data collection to occur. To ensure privacy, all user data are anonymised, and the extension does not have any access to user's history.

The extension operates automatically once a week, running search queries in the background. It captures the top 10 links from the SERPs for each keyword and search engine, which are summarised in Table 2. The keywords are in Dutch since the project focus on Dutch culture and it allows to control human bias as well as language bias. This systematic approach enables the collection of consistent and comparative data across different search platforms.

Table 2

Search engines and keywords list

Search engine	Keywords
- Google	- Immigration
- Google News	- Asylum seekers
- Bing	- Shelter location
- DuckDuckGo	- Asylum seekers' centre
- Yahoo	- Asylum quota
	- Ter Apel
	- Climate
	- Energy transition
	- Agriculture
	- Nitrogen
	- Green energy
	- European politics
	- European elections
	- European Parliament

3.3. Participants

The data collection for this study began actively in October 2023 and has accumulated 233,288 searches to date, with each search generating a SERP consisting of a list of 10 links. Employing a bottom-up approach, the study concentrates on an in-depth analysis of SERPs from a specific week—named “week 2024-05-06,” spanning from May 6th to May 12th, 2024. This week was chosen for the highest amount of data across search engines. This focused analysis centres around three keywords: “Immigration,” “Asylum seekers,” and “Reception location.” These keywords were chosen for their significant relevance to Migration Policy, a current and complex topic that offers varied perspectives and potential more polarised outcome in content results. Each keyword represents a distinct aspect of migration policy, enabling a layered analysis of the topic.

The study involved 498 participants, recruited through diverse methods to ensure a broad and representative sample. These recruitment strategies included advertising a research extension on the project’s website, participating in panel discussions, distributing flyers with incentives like SONA points for students, conducting radio interviews, and issuing press releases. This comprehensive approach helped achieve widespread participant distribution across the country. During the week of 2024-05-06, the data contained 95 individuals. Due to data quality and technical limitations, the extension does not currently collect data every week from all participants, leading to 95 participants for that specific week instead of the maximum number of 498.

The data provides an overview of the 95 participant demographics. According to Table 3 in the appendix, the gender distribution among participants is uneven with 72% male and 28% female. Age demographics indicate that participants primarily fall within the 55 to 74 years age group, as shown in Table 4. Conversely, the age groups least represented are those aged 75 and older, and those between 25 to 44 years old. Geographically, a significant 52% of participants reside in the East region, including Overijssel, Gelderland, Flevoland, correlating with the project’s affiliation with the University of Twente, which is located in this area (Table 6). Approximately 20% of participants are from the North (Groningen, Friesland, Drenthe) and West (Utrecht, North Holland, South Holland, Zeeland), with only 1% from the South (North Brabant, Limburg). Notably, only one participant resides outside the Netherlands, as detailed in Table 5. Educationally, the most prevalent qualifications among participants are HBO and high school, representing 33% and 24% of the sample respectively, as per Table 7. There were no participants without any form of education. In terms of employment, 33% of participants are employed full-time, while 19% are either students or retirees, as reported in Table 8. The income distribution shows that the categories “less than €10,000” and “€20,001 to €30,000” are the most represented, comprising 28% and 24% of participants respectively, according to Table 9. The highest income bracket “over €100,001” has no representation. Political affiliations reveal that 24% of participants align with the VVD party, while a notable 21% chose not to disclose their political affiliation, indicating a significant reticence to share sensitive information (Table 10). Language preferences among participants show that 60% conduct their searches in Dutch and 34% in English, as documented in Table 11. The browser usage statistics highlight a strong preference for Chrome, used by 81% of participants, while Opera and Safari are not used at all (Table 12). Finally, in terms of search engine preferences, Google dominates with 93% usage, while DuckDuckGo accounts for 7%, as outlined in Table 13.

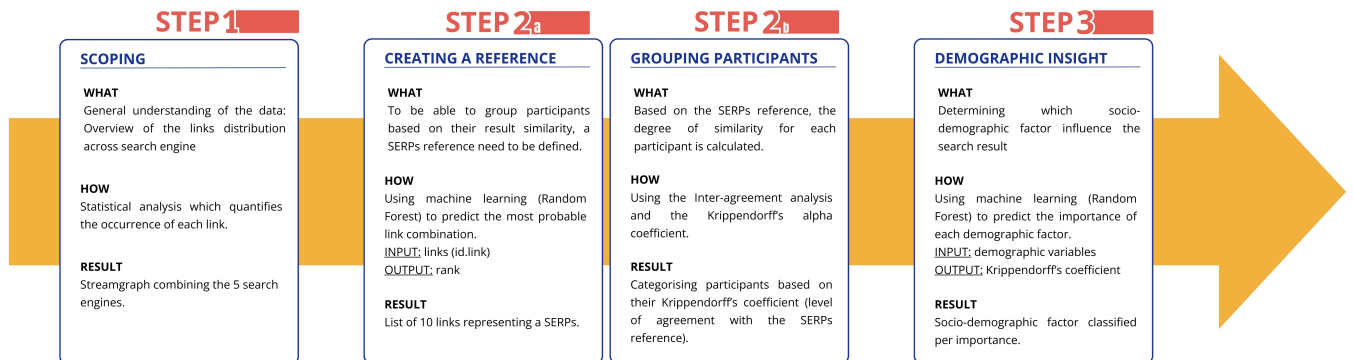
3.5. Data analysis

As outlined above, the analysis is quantitative and focused on the specific period of “week 2024-05-06.” The dataset includes 10 different variables and a total of 17,081 observations. The data

analysis methodology includes several steps designed to ascertain the diversity of SERPs and the socio-demographic factors influencing these results. This detailed process, is depicted and summarized in Figure 3.

Figure 3

Analysis process: the 3 steps



3.5.a. STEP 1: Scoping

The initial step of the analysis is designed to gain a comprehensive understanding of the data, focusing particularly on the distribution of links across various search engines. This involves conducting a detailed frequency analysis to explore how links are distributed by rank, keyword, and search engine. Each link is assigned a unique identifier based on its title and its website, which facilitates consistent tracking across different search platforms. To quantify the distribution, the occurrence of each link is counted, and transformed into percentages. These percentages are then visualized using a percentage streamgraph, providing a dynamic overview of how links are dispersed across the search engines. This visualisation highlights the frequency with which each link appears and underscores the diversity of links at each rank level.

3.5.b. STEP 2: Grouping participants (2.a + 2.b)

The second step of this study serves as a crucial preparatory step for Step 3, where the objective is to categorise users based on the similarity of their SERPs. Specifically, Step 2.a involves employing a machine learning approach—using the Random Forest algorithm—to identify the most common sequences of SERPs. This is a critical intermediary step designed to establish a reference link sequence, which serves as a benchmark for subsequent analysis. In Step 2.b, the analysis shifts to assessing the extent to which users' SERPs align with this established reference sequence. This assessment categorises users into classes based on their agreement with the reference, effectively grouping them by similarity in search results.

The use of machine learning in this phase is essential. While for keywords such as “Immigration,” where high instances of similar link sequences can be statistically identified with relative ease, the challenge increases with keywords that exhibit a high diversity in results. Employing a systematic machine learning approach allows the study to adapt (in the future) effectively to the diverse outcomes associated with the 21 different keywords analysed. This method improves user categorisation accuracy and ensures consistent and scalable handling of diverse and complex data sets across various search queries.

The selection of the Random Forest algorithm for this analysis is underpinned by its notable accuracy and robustness, characteristics derived from its ensemble learning technique (Grinsztajn et al., 2022). By constructing multiple decision trees during the training phase and synthesising their outcomes—typically through majority voting for classification tasks—the Random Forest model substantially reduces the risk of overfitting and boosts the model’s generalisability to novel datasets (Grinsztajn et al., 2022). This approach excels in managing extensive datasets characterized by complex structures, effectively accommodating missing values and sustaining high accuracy even amid considerable data noise. The inputs for this machine learning phase are the link identifiers, and the outputs are their corresponding ranks. This setup is ideal for handling big data scenarios, which are typical in this type of analysis. The process iteratively examines the three keywords across five different search engines, ensuring comprehensive coverage and robustness in the findings. This iterative process provides the likelihood of each link appearing at that particular rank.

Furthermore, the inter-agreement analysis in this phase leverages Krippendorff’s alpha coefficient to categorise users based on their conformity to the established SERPs reference. Krippendorff’s alpha is chosen for its flexibility and robustness, making it especially suitable for intercoder reliability testing in real-world datasets (Krippendorff, 2004). This coefficient is crucial as it serves as a foundational metric for the subsequent Step 3 of our study.

3.5.c. STEP 3: Demographic insight

The final step of the analysis uses Random Forest to determine which socio-demographic factors influence the SERPs. The analysis used a comprehensive set of 11 socio-demographic variables as inputs to the Random Forest algorithm. These variables included gender, age, residency status in the Netherlands, region of residence, educational background, employment status, income level, political affiliation, preferred language, most-used browser, and search engine. The output of this process was the Krippendorff’s alpha coefficient. One of the key strengths of this method is its ability to estimate the importance of each variable, which is crucial for discerning which socio-demographic factors most profoundly affect the search results. By leveraging the variable importance scores provided by Random Forest, this phase offers a detailed perspective on how search engines may personalise results based on user demographics.

The structured approach of this study analyses link distributions across different search engines and examines the impact of demographic factors. This provides valuable insights into information diversity and personalisation. The thorough analysis allows for multi-dimensional comparisons between search engines and across different keywords. By systematically dissecting the data, this study aims to identify patterns and variations in how information is ranked and displayed.

4. Results

4.1. STEP 1: Scoping

4.1.a. Keyword: Immigration

For the keyword “Immigration,” the analysis identified a total of 83 unique links across various search engines. The distribution of these links is as follows: Bing accounts for 32, DuckDuckGo for 28, Google for 26, Google News for 23, and Yahoo for 17 showing a low number of different links per search engine. This detailed breakdown is comprehensively listed in Table 14. Additionally, Figure 4 in appendix graphically illustrates the distribution of these links across the search engines, providing a visual comparison of how each platform prioritizes content related to “Immigration”. This statistical analysis reveals that certain links consistently dominate across multiple search engines. For instance, the link labelled “187” appears in 7.7% of search results, achieving notably high rankings: first on Bing and DuckDuckGo, third on Google, and seventh on Yahoo. Similarly, link “217” is also prevalent (see Table 15 for the specific domain of this link), appearing in 7.6% of search results, and is ranked second on Google, third on Bing and DuckDuckGo, and sixth on Yahoo. These findings highlight a trend of certain links consistently achieving high visibility across various search platforms.

4.1.b. Keyword: Shelter Location

For the keyword “Shelter Location,” the analysis identified a total of 106 unique links across different search engines, showing significant variation in link diversity. Bing displayed the most with 50 links, followed by DuckDuckGo with 45, Google with 33, and both Google News and Yahoo with 17 each. These details are documented in Table 14 and visually represented in Figure 5 of the appendix. Notably, three links appeared frequently across these platforms (see Table 15 for the specific domain of this link): Link 470 was found 5.9% of the time, ranking highest on DuckDuckGo at third place and appearing in multiple positions (2 to 5) on Bing and Google. Link 262 appeared 5.5% of the time, ranking first and second on Yahoo and Google, and similarly high on Bing (rank 1) and DuckDuckGo (rank 2 and 3). Finally, Link 364 appeared in 4.8% of cases, ranking highest on Bing and DuckDuckGo at first place, second on Google, and sixth on Yahoo.

4.1.c. Keyword: Asylum Seekers

For the keyword “Asylum Seekers,” a total of 218 unique links are catalogued across multiple search engines, marking the highest diversity of links among the three keywords. The distribution is as follows: Bing with 123 links, DuckDuckGo with 78, Google with 36, Google News with 42, and Yahoo with 16. The data are detailed in Table 14, with a visual representation in Figure 6, which illustrates the variations in content prioritisation among different search engines for this keyword. Among these, two links predominated (see Table 15 for the specific domain of this link): Link 97 appeared in 9.4% of search results, ranking within the top three positions on Yahoo and across all ten ranks on Google, DuckDuckGo, and Bing, particularly frequent at ranks one and two. The second prominent link, Link 54, appeared in 9.4% of search results, was mainly observed on Yahoo at ranks one and two, on Bing from ranks one to five, and on DuckDuckGo at ranks one, two, six, and eight.

Table 14*Unique links per search engines*

	Bing	DuckDuckGo	Google	Google News	Yahoo	TOTAL
Immigration	32	28	26	23	17	83
Shelter Location	50	45	33	17	17	106
Asylum Seeker	123	78	36	42	16	218

Table 15*Correspondent link number to link content*

Link number	Keyword	Domain	URL
187	Immigration	cbs.nl	https://www.cbs.nl/nl-nl/dossier/dossier-asiel-migratie-en-integratie/hoeveel-immigranten-komen-naar-nederland
217	Immigration	rijksoverheid.nl	https://www.rijksoverheid.nl/onderwerpen/immigratie-naar-nederland
470	Shelter Location	rijksoverheid.nl	https://www.rijksoverheid.nl/onderwerpen/asielbeleid/vraag-en-antwoord/soorten-opvang-asielzoekers
262	Shelter Location	coa.nl	https://www.coa.nl/nl/locatiezoeker
364	Shelter Location	coa.nl	https://www.coa.nl/nl/opvanglocaties-tijdens-de-asielprocedure
97	Asylum Seeker	coa.nl	https://www.coa.nl/nl
54	Asylum Seeker	unhcr.org	https://www.unhcr.org/nl/wie-we-zijn/wie-we-helpen/asielzoekers/

4.2. Analysis for the keyword Immigration (phase 2-3)

The results for Phases 2 and 3 of this study are organised and reported by individual keywords. This strategic choice facilitates a more straightforward comparison, as similar keywords often share links across different search engines. The presentation of the results follows a structured progression, beginning with the keyword “Immigration,” followed by “Shelter Location,” and concluding with “Asylum Seeker.”

4.2.a. STEP 2a: Creating a SERPs reference






The performance metrics of our Random Forest model, applied to data from various search engines, revealed significant differences in prediction accuracy, Out-of-Bag (OOB) error rates, statistical significance, and the Kappa statistic (Table 16). These metrics were computed to evaluate the model’s effectiveness across different platforms including Bing, DuckDuckGo, Google, Google News, and Yahoo. The model exhibited the highest accuracy and agreement (Kappa) on Google, with an accuracy of 83% and a Kappa statistic of 0.81, indicating almost perfect agreement. The lowest performance was observed on Yahoo, with an accuracy of 48% and a Kappa statistic of 0.51, indicating moderate agreement. All models achieved statistical significance with p-values < 2.2e-16, indicating that the models’ accuracies were significantly better than chance. OOB error rates varied across search engines, with Google showing the lowest OOB error rate at 15.77%, suggesting a better generalisability compared to others. Conversely, Yahoo exhibited the highest OOB error rate at 41.52%, indicating lower reliability in its predictions.

The predictions generated by the model are shown in Figure 7, providing a clear view of the most common link sequence prediction.

Table 16*Random Forest Model Performance - Immigration*

	OOB	Accuracy	P-value	Kappa
Bing	39	0.61	< 2.2e-16	0.57
DuckDuckGo	39.9	0.62	< 2.2e-16	0.58
Google	15.77	0.83	< 2.2e-16	0.81
Google News	31	0.78	< 2.2e-16	0.75
Yahoo	41.52	0.48	< 2.2e-16	0.51

Figure 7*Prediction of the most common link sequence (SERPs) - Immigration*

										
	ID.Title	Probability	ID.Title	Probability	ID.Title	Probability	ID.Title	Probability	ID.Title	Probability
RANK 1	187	0.99	187	0.98	231	0.99	87	0.99	221	0.99
RANK 2	456	0.99	456	0.98	217	0.99	242	0.98	192	0.99
RANK 3	217	0.99	217	0.99	187	0.99	276	0.98	456	0.96
RANK 4	206	0.97	206	0.99	218	0.99	110	0.96	226	0.92
RANK 5	84	0.90	210	0.81	206	0.98	147	0.96	84	0.88
RANK 6	6	0.79	433	0.32	434	0.95	275	0.97	217	0.91
RANK 7	402	0.93	402	0.71	203	0.95	479	0.99	187	0.95
RANK 8	401	0.92	401	0.94	214	0.97	453	0.97		
RANK 9	205	0.97	205	0.97	475	0.95	3	0.98		
RANK 10	434	0.94	434	0.93	118	0.95	277	0.99		

4.2.b. STEP 2b: Grouping participants

To group participant, the inter-rater agreement analysis was used across the various search engines, which is measured by Krippendorff's alpha and detailed in Table 17. Google and Google News demonstrated the highest levels of agreement, achieving perfect agreement ($\alpha = 1.0$) with 28 and 30 instances, respectively. In contrast, Bing and Yahoo showed no such instances. Moderate agreement levels were more frequently observed; DuckDuckGo reported the highest number of instances (23) in the alpha range >0.667 to ≤ 0.800 , suggesting a decent level of consistency among raters. The lowest category of agreement ($\alpha \leq 0.400$), indicative of poor consistency, was most prevalent in Bing with 27 instances and Yahoo with 15.

Table 17*Inter-agreement based on the number of users per class - Immigration*

	Bing	DuckDuckGo	Google	Google News	Yahoo
Alpha = 1.0	0	0	28	30	0
Alpha > 0.800	0	3	0	0	2
Alpha > 0.667 to ≤ 0.800	15	23	14	2	4
Alpha > 0.600 to ≤ 0.667	0	0	0	0	0
Alpha > 0.400 to ≤ 0.600	13	10	8	6	7
Alpha ≤ 0.400	27	19	3	13	15

4.2.c. STEP 3: Demographic insight

In the third phase, socio-demographic influence is assessed. Random Forest model was used to define the importance of each socio-demographic factors on search engine result for the keyword “Immigration”. Its performance metrics are summarised in Table 18. This part of the analysis focuses on four key performance indicators: the mean of squared residuals, percentage of variance explained, root mean square error (RMSE), and R-squared values, which collectively assess the accuracy and efficacy of the model.

The Random Forest model’s performance across various search engines, as summarised in Table 18, indicates generally low accuracy and effectiveness in predicting search engine outputs for the keyword “Immigration.” All search engines—Bing, DuckDuckGo, Google, Google News, and Yahoo—demonstrated suboptimal results with negative R-squared values, highlighting an inability to capture the variance of the dependent variable effectively. While some search engines such as Yahoo showed a lower RMSE values, indicating slightly better prediction errors, the overall negative trends in percentage of variance and R-squared across platforms confirm a consistent lack of model reliability.

These results across all search engines indicate that the Random Forest model does not perform well in accurately predicting the influence of socio-demographic factors on search engine outputs for the keyword “Immigration.” The key determinants predicted by the model are illustrated in Figure 8, providing a visual representation of the most important socio-demographic factors.

Table 18
Random Forest Model Performance - Immigration

	Mean of squared residuals	Percentage of variance	RMSE	R-squared
Bing	0.0038	-25.37	0.1001	-0.0190
DuckDuckGo	0.0361	-0.25	0.1604	-0.0718
Google	0.0045	-9.64	0.1115	-0.0301
Google News	0.0121	-9.7	0.1604	-0.0401
Yahoo	0.0047	-8.68	0.0942	-0.0423

Figure 8
Most important socio-demographic factors - Immigration



4.3. Analysis for the keyword Shelter Location (phase 2-3)

4.3.a. STEP 2a: Creating a SERPs reference

Regarding the performance metrics of the Random Forest Model for the keyword “Shelter Location” (Table 19), the data across various search engines demonstrate significant variations in predictive effectiveness. For Bing, the OOB error rate was 61.49%, with an accuracy rate of 43% and a Kappa statistic of 0.36, suggesting minimal agreement beyond chance. Similarly, DuckDuckGo showed an OOB error rate of 53.91%, an accuracy of 43%, and a Kappa of 0.36, indicating comparable performance levels with significant prediction errors. Google displayed slightly better performance with an OOB error rate of 56.83%, an accuracy of 45%, and a Kappa statistic of 0.39, suggesting a modest improvement in model agreement over Bing and DuckDuckGo. In contrast, the prediction of the Google News’ model outperformed the other search engines’ model significantly, with an OOB error rate of 32.5%, an accuracy of 67%, and a Kappa statistic of 0.64. These metrics indicate substantial agreement and higher reliability in predictions. Yahoo recorded an OOB error rate of 41.92%, an accuracy of 52%, and a Kappa of 0.44, denoting moderate agreement. Statistically, all models’ prediction achieved significant results with p-value < 2.2e-16. Overall, these results demonstrate the varied capabilities of the Random Forest model across different digital environments, with Google News showing the most reliable and accurate predictions.

The predictions generated by the model are illustrated in Figure 9, providing a visual representation of the most common link sequence prediction.

Table 19

Random Forest Model Performance – Shelter Location

	OOB	Accuracy	P-value	Kappa
Bing	61.49	0.43	< 2.2e-16	0.36
DuckDuckGo	53.91	0.43	< 2.2e-16	0.36
Google	56.83	0.45	< 2.2e-16	0.39
Google News	32.5	0.67	< 2.2e-16	0.64
Yahoo	41.92	0.52	< 2.2e-16	0.44

Figure 9

Prediction of the most common link sequence (SERPs) – Shelter Location



4.3.b. STEP 2b: Grouping participants

The inter-rater agreement for the keyword “Shelter Location,” is summarised in Table 20. The data shows a complete absence of perfect agreement ($\alpha = 1.0$) across all search engines. Similarly, no instances of strong agreement ($\alpha > 0.800$) were observed. Moderate agreement ($\alpha > 0.667$ to ≤ 0.800) was only recorded by Google News and Yahoo, each posting seven instances. The fair agreement category ($\alpha > 0.400$ to ≤ 0.600) showed somewhat more distribution, with Google News notably higher at twenty instances, indicating a moderate level of consistency, while other search engines like Bing, DuckDuckGo, Google, and Yahoo exhibited significantly fewer instances. The key findings are in the lowest category of agreement ($\alpha \leq 0.400$), where Bing, DuckDuckGo, and Google displayed notably high instances—42, 39, and 25 respectively. Conversely, Google News and Yahoo showed a relatively lower frequency of poor agreement with 7 and 19 instances, respectively.

Table 20

Inter-agreement based on the number of users per class – Shelter Location

	Bing	DuckDuckGo	Google	Google News	Yahoo
Alpha = 1.0	0	0	0	0	0
Alpha > 0.800	0	0	0	0	0
Alpha > 0.667 to ≤ 0.800	0	0	0	7	7
Alpha > 0.600 to ≤ 0.667	0	0	0	0	0
Alpha > 0.400 to ≤ 0.600	1	1	6	20	2
Alpha ≤ 0.400	42	39	25	7	19

4.3.c. STEP 3: Demographic insight

For Phase 3, focusing on the keyword “Shelter Location,” similar results with “Immigration” are observed, with consistently negative percentage variances across all search engines. For example, Google News showed a percentage variance of -1.31, indicating a general trend where the model failed to account for a significant proportion of the variability in the data. These metrics are summarised in Table 21.

Regarding the primary socio-demographic determinants influencing the outputs, the Random Forest model identified distinct factors for each search engine. For Bing, the most significant factors were Region and Gender; DuckDuckGo also highlighted Gender and Region as critical, albeit in a reversed order of prominence. In contrast, Google and Yahoo found Political affiliation and Income to be the main determinants. Google News differed from these patterns by identifying Age and Education as the primary factors affecting search results. A detailed list of all factors considered by the model is depicted in Figure 10.

Table 21

Random Forest Model Performance – Shelter Location

	Mean of squared residuals	Percentage of variance	RMSE	R-squared
Bing	0.0038	-9.36	0.1042	-0.0283
DuckDuckGo	0.0035	-6.94	0.0864	-0.0059
Google	0.0023	-3.81	0.0348	-0.1775
Google News	0.0009	-1.31	0.0180	-0.1374
Yahoo	0.0091	-5.83	0.0890	-0.0091

Figure 10

Most important socio-demographic factors - Shelter Location



4.4. Analysis for the keyword Asylum Seeker (phase 2-3)

4.4.a. STEP 2a: Creating a SERPs reference

Regarding the performance metrics of the Random Forest Model for the keyword “Asylum Seeker,” notable variations are observed across different search engines (Table 22). For Bing, the OOB error rate was reported at 55.89%, with an accuracy of 46% and a Kappa statistic of 0.40, indicating moderate agreement beyond chance. DuckDuckGo displayed a similar OOB error rate of 55.77%, but with a slightly lower accuracy of 43% and a Kappa of 0.37, suggesting less effective prediction capabilities. Google showed a comparable performance to DuckDuckGo with an OOB error rate of 55.58%, an accuracy of 43%, and a Kappa statistic of 0.37, further highlighting challenges in predictive accuracy among these platforms. Conversely, Google News demonstrated better model performance with an OOB error rate of 41.05%, an accuracy of 59%, and a Kappa statistic of 0.54. These metrics indicate a higher reliability and substantial agreement in its predictions. Yahoo also showed improved outcomes with an OOB error rate of 52.07%, an accuracy of 58%, and a Kappa of 0.39, marking it as moderately effective in comparison to Bing and DuckDuckGo, yet not as effective as Google News. Statistically, all models achieved significant results (p-values < 2.2e-16).

The predictions generated by the model are illustrated in Figure 11, providing a visual representation of the most common link sequence prediction.

Table 22

Random Forest Model Performance – Asylum Seeker

	OOB	Accuracy	P-value	Kappa
Bing	55.89	0.46	< 2.2e-16	0.40
DuckDuckGo	55.77	0.43	< 2.2e-16	0.37
Google	55.58	0.43	< 2.2e-16	0.37
Google News	41.05	0.59	< 2.2e-16	0.54
Yahoo	52.07	0.58	< 2.2e-16	0.39

Figure 11

Prediction of the most common link sequence (SERPs) – Asylum Seeker

RANK	Bing		DuckDuckGo		Google		Google News		Yahoo	
	ID.Title	Probability	ID.Title	Probability	ID.Title	Probability	ID.Title	Probability	ID.Title	Probability
RANK 1	97	1.00	97	0.99	424	0.92	56	1.00	54	0.96
RANK 2	483	0.81	80	0.93	56	0.98	245	0.99	97	0.95
RANK 3	70	0.99	262	0.86	302	0.76	153	0.99	44	0.55
RANK 4	44	1.00	478	0.93	21	0.95	75	0.97	44	0.43
RANK 5	1	0.99	450	0.95	154	0.97	69	0.96	466	0.90
RANK 6	4	0.99	457	0.96	462	0.96	432	0.94	459	0.97
RANK 7	463	0.98	40	1.00	63	0.99	281	0.90	40	0.99
RANK 8	105	0.82	430	0.99	267	0.95	377	0.88		
RANK 9	278	0.99	42	0.95	327	0.97	247	0.87		
RANK 10	47	0.99	437	0.99	51	0.60	274	0.43		

4.4.b. STEP 2b: Grouping participants

Table 23 summarises the inter-rater agreement for the keyword “Asylum Seeker” across various search engines, employing Krippendorff’s alpha to evaluate the consistency of classifications by different users. For all search engines, there were no instances of perfect agreement (Alpha = 1.0), indicating a complete absence of unanimous rater consensus. Similarly, except for Yahoo, which reported two instances, there were no occurrences of strong agreement (Alpha > 0.800) across the other platforms, demonstrating a general challenge in achieving high consistency levels. Moderate agreement (Alpha > 0.667 to ≤ 0.800) was only noted in Google News and Yahoo, with 4 and 3 instances, respectively. No instances of agreement (Alpha > 0.600 to ≤ 0.667) were recorded in any search engine, indicating an absence of this moderate level of agreement across the board. The category of fair agreement (Alpha > 0.400 to ≤ 0.600) saw some instances, with Yahoo showing six, DuckDuckGo three, and Google News one, suggesting that these platforms achieved a basic level of consistency in some cases. However, the most significant observation is the large number of instances with the lowest level of agreement (Alpha ≤ 0.400), with Bing reporting 47, DuckDuckGo 40, Google 48, Google News 36, and Yahoo 28. This predominance suggests significant discrepancies in rater classifications across all platforms for the keyword “Asylum Seeker.”

Table 23

Inter-agreement based on the number of users per class – Asylum Seeker

	Bing	DuckDuckGo	Google	Google News	Yahoo
Alpha = 1.0	0	0	0	0	0
Alpha > 0.800	0	0	0	0	2
Alpha > 0.667 to ≤ 0.800	0	0	0	4	3
Alpha > 0.600 to ≤ 0.667	0	0	0	0	0
Alpha > 0.400 to ≤ 0.600	0	3	0	1	6
Alpha ≤ 0.400	47	40	48	36	28

4.4.c. STEP 3: Demographic insight

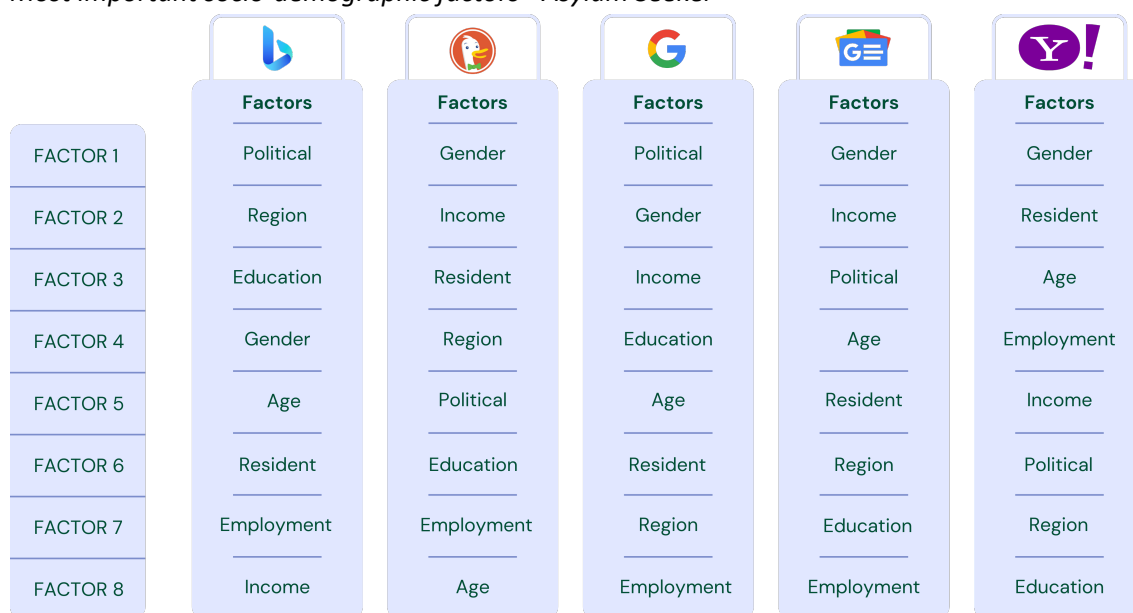
In the analysis of the Random Forest model’s performance for the keyword “Asylum Seeker,” the results indicated varied effectiveness across different search engines, as summarised in the key performance metrics in Table 24. For Bing, the model showed a relatively low mean of squared residuals at 0.0007 with a small positive percentage of variance at 1.51%, although the R-squared value ($R^2 = -0.01351$) indicated a poor model fit. DuckDuckGo and Google displayed more substantial challenges, with negative percentage variances of -10.01% and -2.97% respectively, and corresponding negative R-squared values, highlighting a decrease in predictive accuracy. Google News’ model exhibited the most significant discrepancies with a percentage of variance at -3.06% and a notably low R-squared ($R^2 = -0.4492$), pointing to a substantial deviation from actual outcomes. Yahoo’s model prediction also struggled, with a percentage of variance at -5.67% and an almost neutral R-squared ($R^2 = -0.0011$), indicating neither improvement nor deterioration from the model.

The analysis of socio-demographic determinants, shown in Figure 12, revealed the primary and secondary factors influencing model outputs for each search engine. Political affiliation emerged as the most important factor for Bing and Google, while gender was the key determinant for DuckDuckGo, Google News, and Yahoo. Secondary factors varied across platforms: income significantly impacted DuckDuckGo and Yahoo, gender was important for Google, region influenced Bing, and residency status was notable for Yahoo.

Table 24
Random Forest Model Performance – Asylum Seeker

	Mean of squared residuals	Percentage of variance	RMSE	R-squared
Bing	0.0007	1.51	0.0471	-0.01351
DuckDuckGo	0.0065	-10.01	0.0697	-0.0590
Google	0.0007	-2.97	0.0558	-0.0498
Google News	0.0167	-3.06	0.0765	-0.4492
Yahoo	0.0059	-5.67	0.1167	-0.0011

Figure 12
Most important socio-demographic factors - Asylum Seeker



5. Discussion

To address the research question, *“To what extent do search engine result pages and its ranking algorithm contribute to shaping the information diversity within the digital search ecosystem in The Netherlands?”*, this study conducts a comprehensive analysis of SERPs. This section interprets the results, discusses the implications and limitation of this study, offering insights into information in the Netherlands and epistemic rights.

5.1. Information diversity

The findings indicate a low diversity of links within SERPs, with more variation in ranking position than in the type of links displayed. For example, under the keyword “Immigration,” the number of links is limited, and the results are quite homogenous. In contrast, for “Asylum Seekers,” Bing (123 links) and DuckDuckGo (78 links) show significantly more results than Google (36 links) and Yahoo (16 links). This pattern suggests that Bing and DuckDuckGo might employ broader or more varied indexing strategies, potentially offering richer content diversity than Google and Yahoo. Additionally, Bing and DuckDuckGo consistently exhibit similar distribution patterns across different keywords. For instance, the most common SERPs for “Immigration” and the top links for “Asylum Seekers” are nearly identical between these two search engines. Conversely, Google, Google News, and Yahoo tend to present a more uniform pattern of results, often dominated by one or two links.

Delving deeper, this analysis also sheds light on the importance of ranking in SERPs. Statistically, the first link garners approximately 40% of all clicks, significantly higher than the 19% for the second link and 10% for the third (Chaffey, 2024), which underscore the critical role of ranking in influencing user engagement and visibility. In the data, it appears that specific sources such as Rijksoverheid, COA, CBS, and UNHCR are predominantly displayed across search engines, which could have epistemic consequences. These organisations, while authoritative, primarily present governmental and institutional viewpoints. Coeckelbergh (2023) argues that citizens require diverse knowledge to vote effectively and maintain democracy. However, if the mainstream information provided is from governmental and institutional sources, could this not skew the construction of knowledge?

Moreover, acquiring scientific knowledge requires time and effort that many search engine users may not expend, a concern tackled by Google’s forthcoming update to its “Web” platform (Orlowski, 2024). This new platform aims to revolutionise the search experience by offering users direct answers without traditional links, signifying a move towards even more controlled access to information and emphasising the vital need for reliable information gatekeepers. Furthermore, the recent advancements in large language models like GPT-4 are changing the nature of search from a traditional model to a conversational interface, which could increase selective exposure and opinion polarisation (Sharma et al., 2024). This inclination and reliance on AI-mediated information access raises concerns at multiple level, such as even more misinformation (Shin et al., 2024), potential linguistic homogenization affecting language richness (Creely, 2024), and issues related to privacy and ethics (X. Wu et al., 2024).

Additionally, there are reservations about the concentration of power within private entities such as Google, Bing, or OpenAI among others, which hold significant sway over the curation and dissemination of information. This power dynamic presents a paradox and reveals inherent limitations; while users appreciate the convenience of conversational and tailored information access, they may not fully grasp the profound implications of this exchange. This evolving landscape suggests a shift towards a reality where technology, not individual users, shapes the information environment,

tailoring individual perception of reality and granting disproportionate power and control to technology providers.

5.2. Personalisation impact

The uniformity observed in SERPs implies that personalization algorithms might have a limited impact, which could be related to the simplicity of the keywords used—often just one word. The impact of personalisation and the “filter bubble” effect become more pronounced when searches display a certain opinion orientation, influenced by the specific wording of the query (Gottron & Schwagereit, 2024). This study also highlights a lack of accuracy in models designed to predict how socio-demographic factors affect search outcomes. This could be due to inconsistent data or the possibility that these factors do not significantly impact the results. However, the latter seems to contradict with Google’s approach, which emphasises its relevancy focus (Sullivan, 2019), tracking decades of search behaviour to tailor content specifically for each user. The vast data repository Google has developed complicates comparisons with other search engines like Yahoo, Bing, and DuckDuckGo, which do not possess databases of comparable scale. This discrepancy likely influences the reliability of determining the impact of socio-demographic factors on search results. Therefore, while it is feasible to compare the outputs of different search engines among Dutch internet users, comparing their underlying processing mechanisms remains challenging.

This situation presents a broader societal dilemma regarding information dissemination. On one hand, the lack of diversity could be seen as promoting equal access to information, as all users receive a similarly narrow content range. On the other hand, it could be viewed as limiting the variety of accessible viewpoints, potentially moulding public opinion and promoting cultural homogeneity by sidelining lesser-known or alternative perspectives and reinforcing mainstream narratives. Ensuring accountability for fairness, pluralism, and equality remains a challenge, as the operations of search engines often remain “black boxes” with processes that are opaque and not readily understandable to users or regulators. These issues emphasize the urgent need for greater transparency and ethical considerations in the design and operation of search technologies, to ensure they serve the public good while respecting the diversity of user needs and perspectives. Such dynamics pose significant risks to democratic values, a concern echoed by global entities like the United Nations, which recognises the broader implications of these trends on democracy itself.

Lastly, this study also aimed to identify key socio-demographic factors influencing SERPs. Across all search engines and keywords, “Gender” emerged as the most dominant factor, appearing in the top two positions in 10 out of 15 instances. For keywords like “Shelter Location” and “Asylum Seeker,” Political Affiliation also seemed to play a critical role in shaping SERPs. While these insights are intriguing, the model’s lack of robustness limits the strength of the conclusions that can be drawn about socio-demographic influences from this dataset. The accuracy of the results was compromised due to data quality issues, primarily stemming from the low number of users at this stage, as Random Forest algorithms require larger datasets. However, the robustness of this method holds promise and could be effectively applied once the extension is fully functional. At present, it is challenging to draw definitive conclusions regarding socio-demographic factors, except that different factors seem to play a role in different search engines.

5.3. Limitation of the project's instrument

This project's methodology encompasses distinct advantages and certain limitations. On the one hand, the primary benefit is its systematic nature, allowing for consistent comparisons across different search engines with minimal bias. On the downside, this approach does not consider the variability in individual search behaviours, which significantly influences the search results. Nonetheless, certain factors need to be controlled to achieve meaningful results, and keyword selection was one such controlled factor in this experiment. Another significant advantage is that the extension collects a substantial amount of data, which enables a broad range of analysis using the same dataset. This extensive data collection enhances the robustness and depth of the research findings.

On the other hand, this project's methodology also presents several limitations. Firstly, the technology behind the browser extension is susceptible to issues, including bugs that have surfaced following updates to Google or participant closing their laptop during the search's run, temporarily disrupting data collection. Each time a search engine updates its policies, the extension must also be updated to ensure effective data gathering. Additionally, the success of our data collection heavily relies on the accurate functioning of the extension and the developer's ability to address these challenges promptly. Thirdly, this study leans on a data collection which is dependent on participants voluntarily allowing the extension to run on their laptops for a designated period. This dependence introduces challenges related to participant willingness and concerns about privacy. Participants' apprehensions about how the extension functions can make it difficult to recruit and retain a large and diverse group, thus impacting the breadth and depth of data collected. Lastly, the extension was designed to operate exclusively on laptops and desktops, which poses challenges for recruiting participants and may affect the realism of the experiment since most of the search are operates on phones. Most search queries today are conducted on mobile devices. However, developing the extension for mobile use was not feasible without creating a separate app, which presents its own set of challenges and complexities.

5.4 Limitations of this study

The study encountered several significant limitations, primarily related to data constraints. First, there was a limited amount of data available per week, which posed challenges in selecting a week that provided the most comprehensive data for each keyword and search engine. For the second machine learning analysis, the pool of users was quite low which is one of the reasons of the low results reliability for phase 3. This limitation was compounded by issues related to the consistency of the data. The reliability of the browser extension used for data collection was not fully realized, meaning that ongoing updates—ranging from keyword modifications to technical and logistical adjustments—potentially compromised the consistency and comparability of the data over time. Regarding the data, the gender spread was also uneven which could include bias in the results.

A second major limitation involved the complexity of the data itself. Ideally, the study aimed to conduct a detailed comparative analysis for each keyword (21) across different search engines (5), potentially leading to a meta-analysis of the SERPs. However, such a comprehensive comparison proved to be exceedingly challenging. The process of collecting, analysing, and reporting data across multiple variables was complex, making it difficult to draw clear conclusions from the meta-analysis. This limitation underscores the need for more robust data collection tools and methods that can handle the dynamic and multifaceted nature of search engine data effectively.

5.4 Further research

In considering future research directions, several enhancements and expansions to the current study could be pursued to deepen the understanding of information access and diversity. Firstly, including a more diverse dataset that reflects a wider range of backgrounds could offer a more complete view of how different populations interact with and are served by search engines. Additionally, instead of utilising a Random Forest model as was done in step 2.a, exploring the Dynamic Markov Model could offer significant advantages (Chen et al., 2021). This model, by taking into account the sequences of links rather than evaluating individual links, could enhance the flexibility of the analysis and potentially expand the data pool available for phase 3.

Further methodological improvements could include timeline comparisons to track changes over time and thematic comparisons to discern patterns across different search queries. Conducting a detailed content analysis of the titles and descriptions returned by search queries would also provide deeper insights into how content is being framed and presented by search engines. Additionally, examining the issue from a cultural perspective could reveal the societal impact of the displayed information and its influence on knowledge acquisition. These directions promise to enrich the scholarly understanding of digital information landscapes and enhance practical strategies for achieving more equitable and accurate search engine results.

6. Conclusion

In conclusion, digitalisation has drastically changed the media landscape, resulting in a growing dependence on search engines for information consumption. This shift raises concerns about the equality of information accessibility and the diversity of perspectives offered. This study offers an initial glimpse into the vast potential of the Digitale Polarizatie Project. It aims to explore the diversity of information and analyse the similarity of SERPs among users for three specific keywords across five different search engines, identifying factors that influence information dissemination.

Using an exploratory quantitative approach, the study progresses through four stages, from statistical observations to the application of machine learning methodologies. The primary finding reveals that while individuals receive similar search results, the nuance lies onto the ranking position of the links. While Google, Google News and Yahoo presents a concerning uniformity, Bing and DuckDuckGo display a higher degree of links' diversity. In other words, such patterns highlight the critical role of ranking algorithms in influencing content visibility, prompting questions about the range of accessible perspectives. Moreover, the upcoming changes in search technologies, including Google's new "Web" platform and the rise of large language models, signal a shift towards more AI-mediated, conversational information retrieval. This evolution could potentially exacerbate the issues observed in traditional search like misinformation, selective exposure, and opinion polarisation, further complicating the landscape of digital information access.

This research emphasises the importance of improving transparency, fairness, and diversity in search technologies to better serve the public good and meet the varied needs of users. As the digital search ecosystem continues to evolve it is crucial for stakeholders—policymakers, technology developers, and civil society—to collaborate in addressing these challenges. By doing so, they can safeguard democratic values and promote a more inclusive and informed public sphere. Hence, the findings from this study offer a preliminary look at the digital information landscape, encouraging further exploration and improvement.

References

- Abul-Fottouh, D., Song, M. Y., & Gruzd, A. (2020). Examining algorithmic biases in YouTube's recommendations of vaccine videos. *International Journal of Medical Informatics*, 140. <https://doi.org/10.1016/j.ijmedinf.2020.104175>
- Aguado, J. M., & Hermida, A. (2022). *Hate Speech and Polarization in Participatory Society*.
- Alam, M., Iana, A., Grote, A., Ludwig, K., Müller, P., & Paulheim, H. (2022). Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection. *WWW 2022 - Companion Proceedings of the Web Conference 2022*, 448–457. <https://doi.org/10.1145/3487553.3524674>
- Arguedas, A. R., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2022). *Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review*.
- Arora, S. D., Singh, G. P., Chakraborty, A., & Maity, M. (2022). Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change*, 183. <https://doi.org/10.1016/j.techfore.2022.121942>
- Ashokan, A., & Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing and Management*, 58(5). <https://doi.org/10.1016/j.ipm.2021.102646>
- Bastian, M., Makhortykh, M., & Dobber, T. (2019). News personalization for peace: how algorithmic recommendations can impact conflict coverage. *International Journal of Conflict Management*, 30(3), 309–328. <https://doi.org/10.1108/IJCM-02-2019-0032>
- Bianchi, T. (2024a). *Global search engine desktop market share*. Statista. <https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/>
- Bianchi, T. (2024b). *Netherlands search engines market share 2024*. Statista. <https://www.statista.com/statistics/688737/market-shares-of-search-engines-in-the-netherlands/>
- Boeker, M., & Urman, A. (2022). An Empirical Investigation of Personalization Factors on TikTok. *WWW 2022 - Proceedings of the ACM Web Conference 2022*, 2298–2309. <https://doi.org/10.1145/3485447.3512102>
- Burton, J. (2023). Algorithmic extremism? The securitization of artificial intelligence (AI) and its impact on radicalism, polarization and political violence. *Technology in Society*, 75. <https://doi.org/10.1016/j.techsoc.2023.102262>
- Cano Macias, R., & Ruiz Vera, J. M. (2024). Dynamics of opinion polarization in a population. *Mathematical Social Sciences*, 128, 31–40. <https://doi.org/10.1016/j.mathsocsci.2024.01.009>
- CBS. (2024). *More people doubt the accuracy of information seen online*. <https://www.cbs.nl/en-gb/news/2024/15/more-people-doubt-the-accuracy-of-information-seen-online>
- Ch, D. (2024). *Bing Users and Growth Statistics*. SignHouse. <https://www.usesignhouse.com/blog/bing-stats>
- Chaffey, D. (2024). *2024 comparison of Google organic clickthrough rates (SEO CTR) by ranking position*. Smart Insights. <https://www.smartinsights.com/search-engine-optimisation-seo/seo-analytics/comparison-of-google-clickthrough-rates-by-position/>
- Chaney, A. J. B., Stewart, B. M., & Engelhardt, B. E. (2018). *How algorithmic confounding in recommendation systems increases homogeneity and decreases utility*. 224–232. <https://doi.org/10.1145/3240323.3240370>

- Chen, R., Sun, H., Chen, L., Zhang, J., & Wang, S. (2021). *Dynamic order Markov model for categorical sequence clustering*. <https://doi.org/10.1186/s40537-021-00547-2>
- Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization. *Journal of Broadcasting and Electronic Media*, 64(2), 150–172. <https://doi.org/10.1080/08838151.2020.1757365>
- Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 3(4), 1341–1350. <https://doi.org/10.1007/s43681-022-00239-4>
- Courtois, C., Slechten, L., & Coenen, L. (2018). Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics and Informatics*, 35(7), 2006–2015. <https://doi.org/10.1016/j.tele.2018.07.004>
- Creely, E. (2024). Exploring the Role of Generative AI in Enhancing Language Learning: Opportunities and Challenges. *International Journal of Changes in Education*. <https://doi.org/10.47852/bonviewijce42022495>
- Cui, M., Mariani, M. S., & Medo, M. (2022). Algorithmic bias amplification via temporal effects: The case of PageRank in evolving networks. *Communications in Nonlinear Science and Numerical Simulation*, 104. <https://doi.org/10.1016/j.cnsns.2021.106029>
- De Gemmis, M., Lops, P., Semeraro, G., & Musto, C. (2015). An investigation on the serendipity problem in recommender systems. *Information Processing & Management*, 51(5), 695–717. <https://doi.org/10.1016/J.IPM.2015.06.008>
- Dean, B. (2023). *How People Use Google Search*. Backlinko. <https://backlinko.com/google-user-behavior>
- Dylko, I., Dolgov, I., Hoffman, W., Eckhart, N., Molina, M., & Aaziz, O. (2017). The dark side of technology: An experimental investigation of the influence of customizability technology on online political selective exposure. *Computers in Human Behavior*, 73, 181–190. <https://doi.org/10.1016/j.chb.2017.03.031>
- Eg, R., Demirkol Tønnesen, Ö., & Tennfjord, M. K. (2023). A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. In *Computers in Human Behavior Reports* (Vol. 9). Elsevier B.V. <https://doi.org/10.1016/j.chbr.2022.100253>
- Einav, G., Allen, O., Gur, T., Maaravi, Y., & Ravner, D. (2022). Bursting filter bubbles in a digital age: Opening minds and reducing opinion polarization through digital platforms. *Technology in Society*, 71. <https://doi.org/10.1016/j.techsoc.2022.102136>
- Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017). Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW). <https://doi.org/10.1145/3134677>
- Evans, R., Jackson, D., & Murphy, J. (2023). Google News and Machine Gatekeepers: Algorithmic Personalisation and News Diversity in Online News Search. *Digital Journalism*, 11(9), 1682–1700. <https://doi.org/10.1080/21670811.2022.2055596>
- Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior*, 116. <https://doi.org/10.1016/j.chb.2020.106626>

- Figà Talamanca, G., & Arfini, S. (2022). Through the Newsfeed Glass: Rethinking Filter Bubbles and Echo Chambers. *Philosophy and Technology*, 35(1). <https://doi.org/10.1007/s13347-021-00494-z>
- Flensted, T. (2024). *How Many People Use Google? Statistics & Facts*. SEO.AI. <https://seo.ai/blog/how-many-people-use-google>
- Fletcher, R., & Nielsen, R. K. (2018). Automated Serendipity: The effect of using search engines on news repertoire balance and diversity. *Digital Journalism*, 6(8), 976–989. <https://doi.org/10.1080/21670811.2018.1502045>
- Furman, K. (2023). Epistemic Bunkers. *Social Epistemology*, 37(2), 197–207. <https://doi.org/10.1080/02691728.2022.2122756>
- Gaol, F. L., Maulana, A., & Matsuo, T. (2020). News consumption patterns on Twitter: fragmentation study on the online news media network. *Heliyon*, 6(10). <https://doi.org/10.1016/j.heliyon.2020.e05169>
- Golebiewski, M., & Boyd, D. (2019). *Data Voids: Where Missing Data Can Easily Be Exploited*. <https://www.researchgate.net/publication/356909935>
- Gottron, T., & Schwagereit, F. (2024). *The Impact of the Filter Bubble-A Simulation Based Framework for Measuring Personalisation Macro Effects in Online Communities*. <http://edudemic.com/2012/12/>
- Graham, R. (2022). *Investigating Google's Search Engine*. Bloomsbury Publishing.
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on typical tabular data?*
- Haim, M., Graefe, A., & Brosius, H. B. (2018). Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News. *Digital Journalism*, 6(3), 330–343. <https://doi.org/10.1080/21670811.2017.1338145>
- Hoggan-Kloubert, T., & Hoggan, C. (2023). Post-Truth as an Epistemic Crisis: The Need for Rationality, Autonomy, and Pluralism. *Adult Education Quarterly*, 73(1), 3–20. <https://doi.org/10.1177/07417136221080424>
- Huang, J., Ding, S., Wang, H., & Liu, T. (2018). Learning to Recommend Related Entities With Serendipity for Web Search Users. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, 17. <https://doi.org/10.1145/3185663>
- Interian, R., G. Marzo, R., Mendoza, I., & Ribeiro, C. C. (2023). Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods. *International Transactions in Operational Research*, 30(6), 3122–3158. <https://doi.org/10.1111/itor.13224>
- Jones-Jang, S. M., & Chung, M. (2022). Can we blame social media for polarization? Counter-evidence against filter bubble claims during the COVID-19 pandemic. *New Media and Society*. <https://doi.org/10.1177/14614448221099591>
- Kemp, S. (2024). *Digital 2024: The Netherlands*. DataReportal. <https://datareportal.com/reports/digital-2024-netherlands>
- Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, location, location: The impact of geolocation on web search personalization. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, 2015-October*, 121–127. <https://doi.org/10.1145/2815675.2815714>
- Kolic, B., Aguirre-López, F., Hernández-Williams, S., & Garduño-Hernández, G. (2022). *Quantifying the structure of controversial discussions with unsupervised methods: a look into the Twitter climate change conversation*. <http://arxiv.org/abs/2206.14501>

- Kotkov, D., Veijalainen, J., Wang, S., & Veijalainen, J. (2020). How does serendipity affect diversity in recommender systems? A serendipity-oriented greedy algorithm. *Computing*, *102*, 393–411. <https://doi.org/10.1007/s00607-018-0687-5>
- Krippendorff, K. (2004). *Reliability in Content Analysis: Some Common Misconceptions and Recommendations*. <http://repository.upenn.edu/ascpapers/242>
- Lee, Y. (2020). Serendipity adjustable application recommendation via joint disentangled recurrent variational auto-encoder. *Electronic Commerce Research and Applications*, *44*. <https://doi.org/10.1016/j.elerap.2020.101017>
- Lindner, J. (2024). *Must-Know Google Search Statistics*. Gitnux. <https://gitnux.org/google-search-statistics/>
- Ludwig, K., Grote, A., Iana, A., Alam, M., Paulheim, H., Sack, H., Weinhardt, C., & Müller, P. (2023). Divided by the Algorithm? The (Limited) Effects of Content- and Sentiment-Based News Recommendation on Affective, Ideological, and Perceived Polarization. *Social Science Computer Review*, *41*(6), 2188–2210. <https://doi.org/10.1177/08944393221149290>
- Lunardi, G. M., Machado, G. M., Maran, V., & de Oliveira, J. P. M. (2020). A metric for Filter Bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing Journal*, *97*. <https://doi.org/10.1016/j.asoc.2020.106771>
- Luo, Q., Puett, M. J., & Smith, M. D. (2023). *A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and YouTube*.
- Mahmoudi, A., Jemielniak, D., & Ciechanowski, L. (2024). Echo Chambers in Online Social Networks: A Systematic Literature Review. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2024.3353054>
- Matsa, K. (2023). *More Americans are getting news on TikTok, in contrast with most other social media sites*. Pew Research Center. <https://www.pewresearch.org/short-reads/2023/11/15/more-americans-are-getting-news-on-tiktok-bucking-the-trend-seen-on-most-other-social-media-sites/>
- Miah, M. (2024). *Digital Inequality: The Digital Divide and Educational Outcomes*. <https://www.researchgate.net/publication/379258768>
- Napoli, P. M. (2024). Epistemic Rights, Information Inequalities, and Public Policy. In *Global Transformations in Media and Communication Research: Vol. Part F2069* (pp. 47–62). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-45976-4_4
- NCTV. (2022). *Public survey on disinformation Outcomes and initial communication insights*.
- Nguyen, C. T. (2020). Echo Chambers and Epistemic Bubbles. In *Episteme* (Vol. 17, Issue 2, pp. 141–161). Cambridge University Press. <https://doi.org/10.1017/epi.2018.32>
- Nieminen, H. (2024). Why We Need Epistemic Rights. In *Global Transformations in Media and Communication Research: Vol. Part F2069* (pp. 11–28). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-45976-4_2
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.
- Orlowski, A. (2024). *Google declares the end of the World Wide Web*. UnHerd. <https://unherd.com/newsroom/google-declares-the-end-of-the-world-wide-web/>
- Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You* (Penguin).
- Pariser, E. (2015). *Did Facebook's Big Study Kill My Filter Bubble Thesis?* . Wired. <https://www.wired.com/2015/05/did-facebooks-big-study-kill-my-filter-bubble-thesis/>

- Ping, Y., Li, Y., & Zhu, J. (2024). Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement. *Springer*.
<https://doi.org/10.1007/s10660-024-09813-w>
- Reviglio, U. (2019). Serendipity as an emerging design principle of the infosphere: challenges and opportunities. *Ethics and Information Technology*, 21(2), 151–166.
<https://doi.org/10.1007/s10676-018-9496-y>
- Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within Google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW). <https://doi.org/10.1145/3274417>
- Rogers, I. (2002). *The Google Pagerank Algorithm and How It Works*. www.searchenginesystems.net
- Rovira, C., Codina, L., & Lopezosa, C. (2021). Language bias in the google scholar ranking algorithm. *Future Internet*, 13(2), 1–17. <https://doi.org/10.3390/fi13020031>
- Rowland, J., López-Asensio, S., Bagci, A., Delicado, A., & Prades, A. (2023). Shaping information and knowledge on climate change technologies: A cross-country qualitative analysis of carbon capture and storage results on Google search. *Journal of the Association for Information Science and Technology*. <https://doi.org/10.1002/asi.24828>
- Ruffo, G., Semeraro, A., Giachanou, A., & Rosso, P. (2023). Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. In *Computer Science Review* (Vol. 47). Elsevier Ireland Ltd. <https://doi.org/10.1016/j.cosrev.2022.100531>
- Saetra, H. S. (2019). *The tyranny of perceived opinion: Freedom and information in the era of big data*. <https://doi.org/10.1016/j.techsoc.2019.101155>
- Sharma, N., Liao, Q. V., & Xiao, Z. (2024). *Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking*. 17(24). <https://doi.org/10.1145/3613904.3642459>
- Shin, D., Koerber, A., & Lim, J. S. (2024). Impact of misinformation from generative AI on user information processing: How people understand misinformation from generative AI. *New Media and Society*.
https://doi.org/10.1177/14614448241234040/ASSET/IMAGES/10.1177_14614448241234040-IMG3.PNG
- Slechten, L., Courtois, C., Coenen, L., & Zaman, B. (2021). Adapting the Selective Exposure Perspective to Algorithmically Governed Platforms: The Case of Google Search. *Sage*.
<https://doi.org/10.1177/00936502211012154>
- Strzelecki, A., & Rutecka, P. (2020). Direct Answers in Google Search Results. *IEEE Access*.
<https://doi.org/10.1109/ACCESS.2020.2999160>
- Sullivan, D. (2019). *How we keep Search relevant and useful*. Google Blog.
<https://blog.google/products/search/how-we-keep-google-search-relevant-and-useful/>
- Sunstein, C. (2001). *Republic.com*.
- Swert, D., Schuck, K. ;, Boukes, A. ;, Dekker, M. ;, & Helwegen, N. ; (2023). *Monitoring media pluralism in the digital era : Application of the Media Pluralism Monitor In the European Union, Albania, Montenegro, Republic of North Macedonia, Serbia and Turkey in the year 2022 Country report: The Netherlands*. <https://doi.org/10.2870/884842>
- Trinchini, L., & Baggio, R. (2023). *Digital sustainability: Ethics, epistemology, complexity and modelling*.
- Tucker, V. M., & Edwards, S. L. (2021). Search evolution for ease and speed: A call to action for what's been lost. *Journal of Librarianship and Information Science*, 53(4), 668–685.
<https://doi.org/10.1177/0961000620980827>

- UNESCO. (2024). *The Need to Accelerate Worldwide Progress*. <http://en.unesco.org/open-access/terms-use-ccbysa-en>
- Valaskivi, K., & Robertson, D. G. (2022). Introduction: epistemic contestations in the hybrid media environment. *Popular Communication, 20*(3), 153–161. <https://doi.org/10.1080/15405702.2022.2057998>
- Valdez, V. B., & Javier, S. P. (2020). *Digital Divide: From a Peripheral to a Core Issue for all SDGs*. 1–14. https://doi.org/10.1007/978-3-319-71060-0_107-1
- Valensise, C. M., Cinelli, M., & Quattrociocchi, W. (2023). The drivers of online polarization: Fitting models to data. *Information Sciences, 642*. <https://doi.org/10.1016/j.ins.2023.119152>
- Varian, H. R. (2006). *The Economics of Internet Search*.
- Wu, X., Duan, R., & Ni, J. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence, 2*(2), 102–115. <https://doi.org/10.1016/j.jiixd.2023.10.007>
- Wu, Y., Li, L., Yu, Q., Gan, J., & Zhang, Y. (2023). Strategies for reducing polarization in social networks. *Chaos, Solitons and Fractals, 167*. <https://doi.org/10.1016/j.chaos.2022.113095>
- Wu, Z., Sanderson, M., Cambazoglu, B., Croft, B., & Scholer, F. (2020). Providing Direct Answers in Search Results: A Study of User Behavior. *International Conference on Information and Knowledge Management, Proceedings, 1635–1644*. <https://doi.org/10.1145/3340531.3412017>
- Yang, C., Xu, X., Nunes, B. P., & Siqueira, S. W. M. (2023). Bubbles bursting: Investigating and measuring the personalisation of social media searches. *Telematics and Informatics, 82*. <https://doi.org/10.1016/j.tele.2023.101999>
- Yi, Y., & Patterson, S. (2020). Disagreement and polarization in two-party social networks. *IFAC-PapersOnLine, 53*(2), 2568–2575. <https://doi.org/10.1016/j.ifacol.2020.12.302>

Appendix

Table 1

Demographic pop-up

Questions	Possible answers	
Are you residing in the Netherlands?	- Yes	- No
What is your sex?	- Male - Female	- Other - Rather not say
What is your highest level of education?	- No education - High school (VMBO, HAVO, VWO) - Middle professional education (MBO)	- Higher professional education (HBO) - University
Which political party do you prefer?	- VVD - D66 - PVV - CDA - SP - PvdA - Groenlinks - FVD - Partij voor de Dieren - ChristenUnie	- Volt - JA21 - SGP - DENK - 50PLUS - BBB - BIJ1 - Other - Rather not say
In which language do you perform search queries? (multiple answers possible)	- Dutch - English - German - French	- Spanish - Italian - Rather not say
What is your postcode? * (only numbers)	"xxxx"	
What is your age?	- 16-24 - 25-34 - 35-44 - 45-54	- 55-64 - 65-74 - 75+ - Rather not say
What is your personal annual net income?	- Less than 10.000 euro - 10.000 to 20.000 euro - 20.001 to 30.000 euro - 30.001 to 40.000 euro	- 40.001 to 50.000 euro - 50.001 to 100.000 euro - 100.001 or more - Rather not say
What is your current employment situation?	- Full-time employment - Part-time employment - Unemployed - Retired	- Self-employed - Student - Rather not to say
Which (social) media channels do you use for news and information? (multiple answers possible)	- TV - The Newspaper - News websites - YouTube - Facebook - Instagram	- LinkedIn - Twitter - Telegram - Reddit - Radio - Other

	- WhatsApp	- Rather not say
On which browser do you mainly perform search queries? (Multiple answers possible)	- Firefox	- Opera
	- Chrome	- Safari
	- Microsoft Edge	
On which search engine do you mainly perform search queries? (Multiple answers possible)	- Google	- StartPage
	- DuckDuckGo	- Ecosia
	- Bing	- Others
	- Yahoo	- Rather not say

Table 3*Demographic: Gender*

	Man	Women	Other	Unselected
N	42	16	0	0
Percentage	72.41	27.59	0	0

Table 4*Demographic: Age*

	16-24	25-34	35-44	45-54	55-64	65-74	75+	Unselected
N	11	5	5	8	19	7	3	0
Percentage	18.97	8.62	8.62	13.79	32.76	12.07	5.17	0

Table 5*Demographic: Resident*

	Yes	No
N	57	1
Percentage	98.28	1.72

Table 6*Demographic: Region*

	East	North	South	West	Unselected
N	34	12	1	11	0
Percentage	52.62	20.69	1.72	18.96	0

Table 7*Demographic: Education*

	No education	High School	MBO	HBO	University	Unselected
N	0	14	13	19	12	0
Percentage	0	24.14	22.41	32.78	20.69	0

Table 8*Demographic: Employment*

	Full-time employment	Part-time employment	Unemployed	Self-employed	Student	Retired	Unselected
N	19	9	3	5	11	11	0
Percentage	32.76	15.52	5.17	8.62	18.97	18.97	0

Table 9*Demographic: Income*

	<10000	10000-20000	20001-30000	30001-40000	40001-50000	50001-100000	>10001	Unselected
N	16	5	14	8	10	5	0	0
Percentage	27.59	8.62	24.14	13.79	17.24	8.62	0	0

Table 10*Demographic: Political affiliation*

	VVD	D66	PVV	CDA	SP	PvdA	Groenlinks	FVD	Partij voor de Dieren	ChristenUnie
N	14	3	5	1	8	0	2	0	0	1
Percentage	24.1	5.1	8.6	1.7	13.8	0	3.4	0	0	1.7
	Volt	JA21	SGP	DENK	50PLUS	BBB	BIJ1	Others	Rather not say	
N	3	0	4	0	0	0	0	0	12	
Percentage	5.2	0	6.9	0	0	0	0	0	20.7	

Table 11*Demographic: Language*

	Dutch	English	German	French	Spanish	Italian	Unselected
N	35	20	2	1	0	0	0
Percentage	60.34	34.48	3.45	1.72	0	0	0

Table 12*Demographic: browser most used*

	Firefox	Chrome	Microsoft Edge	Opera	Safari
N	8	47	3	0	0
Percentage	13.79	81.03	5.17	0	0

Table 13*Demographic: search engine most used*

	Google	DuckDuckGo	Bing	Yahoo	StartPage	Ecosia	Others
N	54	4	0	0	0	0	0
Percentage	93.10	6.90	0	0	0	0	0

Figure 4
Link Distribution for Immigration

Percentage Distribution for Immigrate

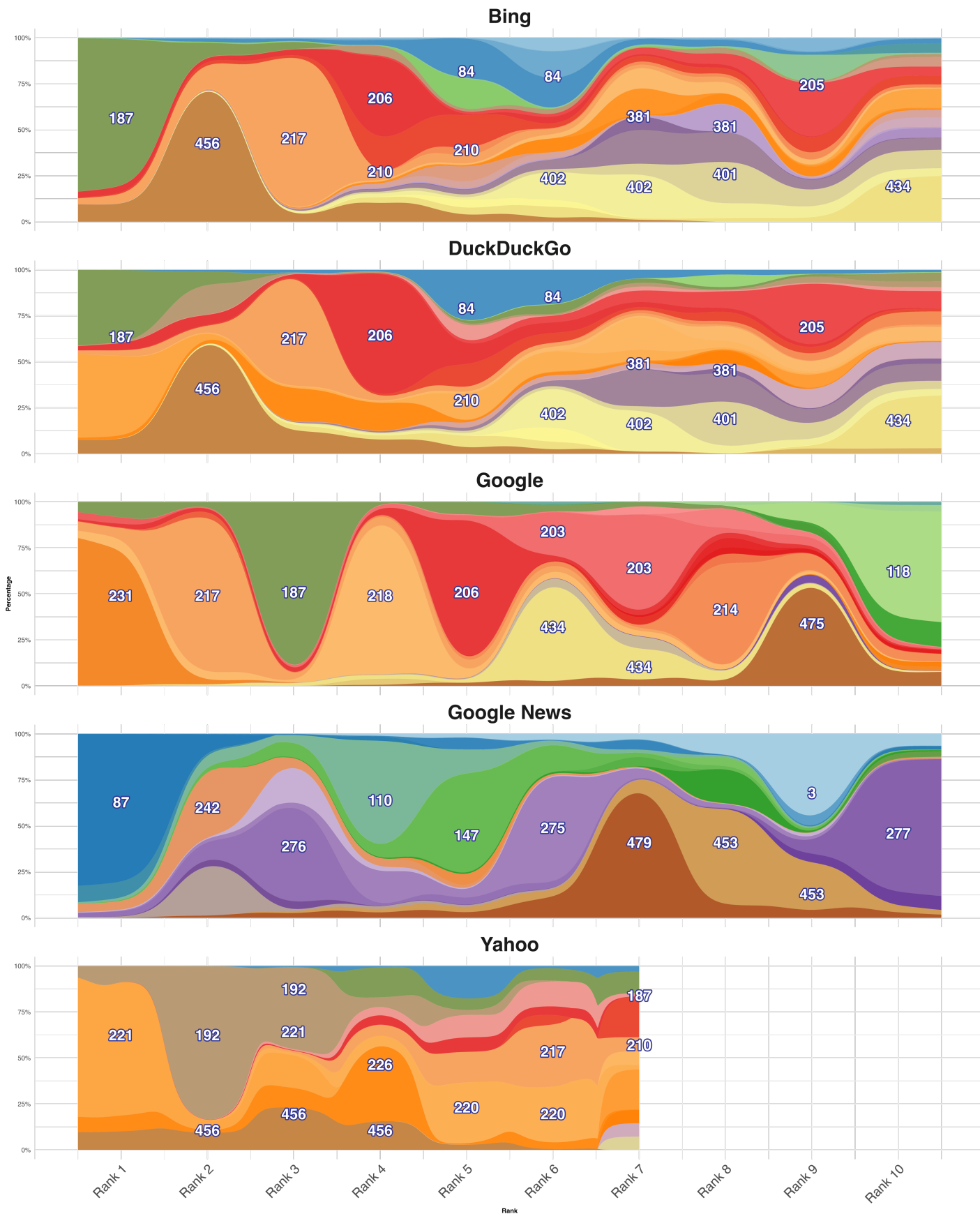


Figure 5
Link Distribution for Shelter Location

Percentage Distribution for Opvanglocatie

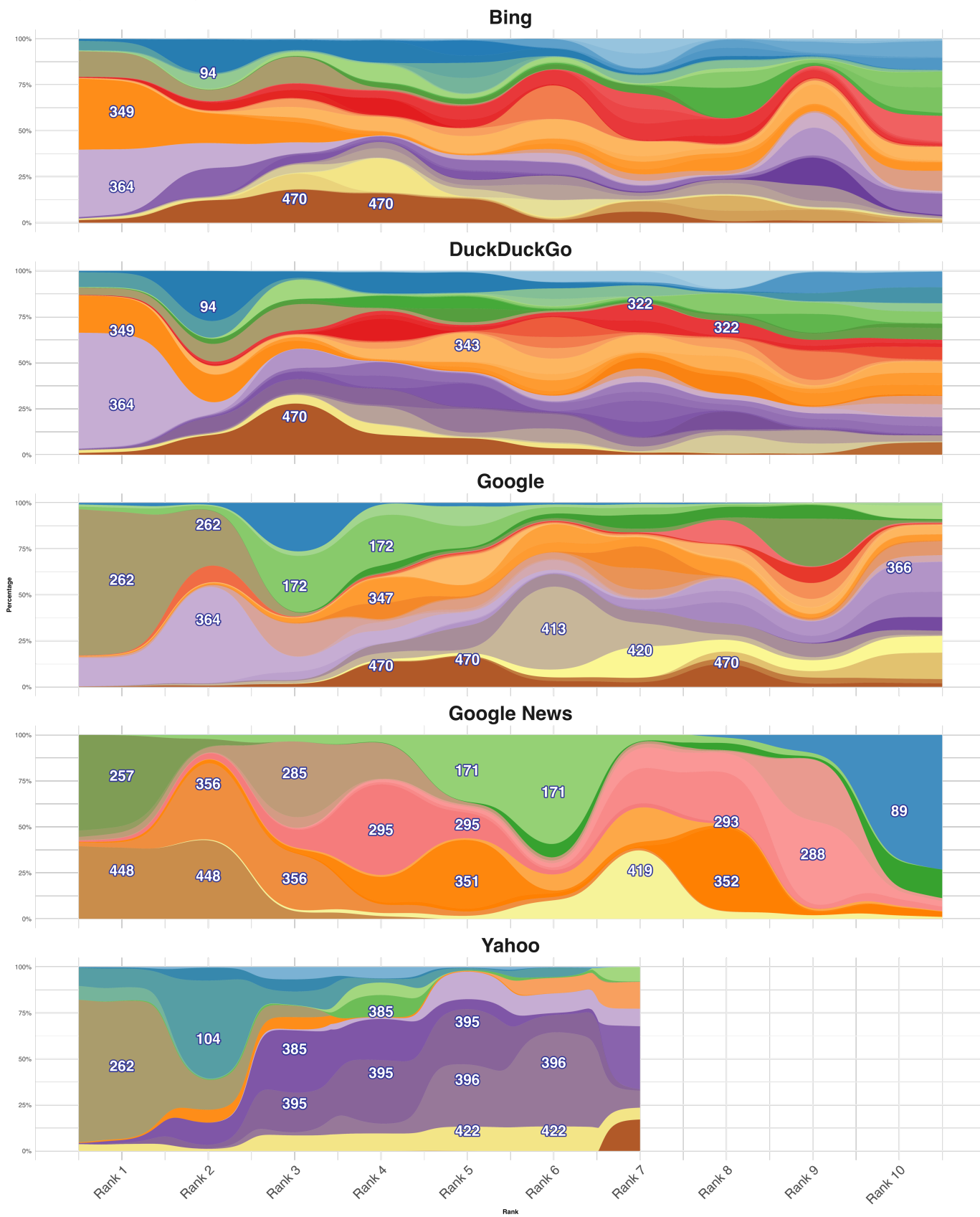


Figure 6
Link Distribution for Asylum Seeker

Percentage Distribution for Asielzoekers

