UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

# INVESTIGATING PERSONALIZATION IN TWITTER SEARCH THROUGH AN EXPERIMENT ON POLARIZED ACCOUNTS AND OPINIONS ON THE BRAZILIAN SOCIAL WELFARE REFORM

Jônatas Castro dos Santos

**Orientadores**

Sean Wolfgand Matsui Siqueira

Bernardo Pereira Nunes

RIO DE JANEIRO, RJ - BRASIL

OUTUBRO DE 2020

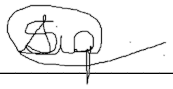
INVESTIGATING PERSONALIZATION IN TWITTER SEARCH THROUGH AN EXPERIMENT ON POLARIZED ACCOUNTS AND OPINIONS ON THE BRAZILIAN SOCIAL WELFARE REFORM

Jônatas Castro dos Santos

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Em conformidade com a Resolução nº 5.257 de 25/03/2020, esta ata vai somente por mim assinada, atestando que a defesa ocorreu com a participação dos componentes abaixo listados.

Aprovada por:



Sean Wolfgand Matsui Siqueira, D.Sc. – UNIRIO

Fabrício R. S. Pereira, D.Sc. – UNIRIO

Tadeu Moreira de Classe, D.Sc. – UNIRIO

Jonice de Oliveira Sampaio, D.Sc. – UFRJ

RIO DE JANEIRO, RJ - BRASIL

OUTUBRO DE 2020

Catalogação informatizada pelo(a) autor(a)

S237

Santos, Jônatas Castro dos
Investigating personalization in Twitter Search through an experiment on polarized accounts and opinions on the Brazilian Social Welfare Reform / Jônatas Castro dos Santos. -- Rio de Janeiro, 2020.
72

Orientador: Sean Wolfgand Matsui Siqueira.
Coorientador: Bernardo Pereira Nunes.
Dissertação (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2020.

1. Personalização. 2. Busca em Mídias Sociais. 3. Polarização. 4. Filter bubble. I. Siqueira, Sean Wolfgand Matsui, orient. II. Nunes, Bernardo Pereira, coorient. III. Título.

"Oh, the depth of the riches both of the wisdom and knowledge of God!

How unsearchable are His judgments and unfathomable His ways!

For WHO HAS KNOWN THE MIND OF THE LORD, OR WHO BECAME HIS COUNSELOR?

OR WHO HAS FIRST GIVEN TO HIM THAT IT MIGHT BE PAID BACK TO HIM AGAIN?

For from Him and through Him and to Him are all things. To Him be the glory forever.

Amen."

**Romans 11.33-36**

I dedicate this dissertation to my wife that most believed I could continue to the end of it.

And for my parents that prayed tirelessly for my success.

Santos, Jônatas Castro. **INVESTIGATING PERSONALIZATION IN TWITTER SEARCH THROUGH AN EXPERIMENT ON POLARIZED ACCOUNTS AND OPINIONS ON THE BRAZILIAN SOCIAL WELFARE REFORM**. UNIRIO, 2020. 71 pages. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

## RESUMO

Os algoritmos de personalização desempenham um importante papel na maneira que as plataformas de busca entregam os resultados para os usuários. Há muitos estudos empíricos sobre os efeitos desses algoritmos em mecanismos de busca como o Google e Bing, porém, há poucos relatos a respeito de personalização em busca de mídias sociais. Este estudo exploratório visa entender e quantificar os limites da personalização nos resultados de busca do Twitter. Nós desenvolvemos uma metodologia de medição e agentes automatizados para treinar um par de contas polarizadas no Twitter e coletar resultados de busca dessas contas de forma simultânea. Os agentes foram executados no contexto político da Reforma da Previdência do Brasil. Nossos resultados mostraram um significativo quantitativo de diferenças de personalização quando comparamos resultados de buscas de contas novas com contas "usadas". Encontramos poucas evidências de diferenças de personalização entre os resultados de dois perfis que seguiram contas com pontos de vista polarizados acerca de um mesmo assunto, entretanto, não podemos anular a hipótese sobre *filter bubbles*.

**Palavras-chave: Personalização, Busca em Mídias Sociais, Polarização, Filter bubble**

## ABSTRACT

Personalization algorithms play an essential role in the way search platforms fetch results to users. While there are many empirical studies about the effects of these algorithms on Web searches like Google and Bing, reports about personalization on social media searches are rare. This exploratory study aims to understand and quantify the limits of personalization in Twitter search results. We developed a measurement methodology and agents to train a pair of polarized Twitter accounts and simultaneously collected search results from these accounts. The agents were run in a political context, the Brazilian Welfare Reform. Our findings show a significant amount of personalization differences when we compare search results from a new fresh profile to non-fresh ones. Little evidence for differences between results of two profiles that followed different accounts with polarized viewpoints about the same topic was found -- the filter bubble hypothesis cannot be null.



**Keywords: Personalization, Social Media Search, Polarization, Filter bubble**

# Contents

# List of Figures

# List of Tables

# 1. Introduction

An estimated 3.6 billion people were using social media in 2020[1]. The advent of social media has caused a tremendous impact on the way people deal with information on the Internet. While they act in helping people to share information with others, they also impact on the way people shape their opinions.

For instance, Twitter, one of the most popular social media microblogging services, reached over 186 million daily active monetizable users worldwide in the second quarter of 2020[2]. The most-followed Twitter accounts include celebrities and politically exposed people. It is estimated that over 500 million tweets are delivered per day[3], and the company's annual revenue amounted to almost U$ 3.46 billion in 2019[4].

In order to sustain that big amount of revenue, social media platforms usually depend on advertizing incomes and thence need to retain users. The personalization algorithms play an important role in this goal as it helps on feeding users with information of their interest.

Usually, personalization algorithms capture information about user's interests, preferences,

---

[1] According to https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (Accessed on August 9th, 2020)
[2] According to https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/ (Accessed on August 6th, 2020)
[3] According to https://www.internetlivestats.com/twitter-statistics/ (Accessed on August 6th, 2020)
[4] According to https://www.statista.com/statistics/204211/worldwide-twitter-revenue/ (Accessed on August 6th, 2020)

web history and contextual information (e.g., time, location) to derive optimally customized information based on various approaches that should vary depending on the system goals and requirements (BOZDAG, 2013; HAIM *et al.*, 2018). These algorithms have done a tremendous contribution to society as they are the core engine of most well-known search and social media platforms.

On the other side, personalization algorithms end up creating an invisible barrier that blocks users from confronting topics – the well-known filter bubble phenomenon. According to PARISER (2011), search engines and social media provide users with non-confronting information to increase their on-time within their platforms.

Besides the social media companies' goals on retaining users, the filter bubble has been claimed to cause polarization on social networks (FLAXMAN *et al.*, 2016) - when there are two conflicting groups with different opinions on a topic. Polarization is a concerning issue to democracy systems as it can lead to users receiving biased information, which can foster intolerance to opposing viewpoints, ideological segregation and antagonism in mainstream political and societal issues (GARIMELLA *et al.*, 2018). Many studies reported polarization on social media (GARIMELLA *et al.*, 2016; GUERRA *et al.*, 2013; MORALES *et al.*, 2015).

As an example of polarization in social media, we highlight a long-lasting trending topic that emerged on Twitter on early 2019: the *Brazilian Social Welfare Reform*. On March 22-23, 2019, the hashtags **#FightForYourRetirement** (*#LutePelaSuaAposentadoria*) and **#ISupportTheNewWelfare** (*#EuApoioNovaPrevidencia*) about the Brazilian Social Welfare Reform became evident on the Twitter trending topics (ESTADÃO, 2019). During this period, this political concern was the central topic of many media streams, social networks and street protests. We use this topic as a case of study for this work.

Given the polarization issue, it is important to understand the dynamics of personalization algorithms on social media, such as Twitter. Twitter delivers a personalized social media search experience to the users: the Twitter Search. From a single query term input, a user can search for tweets and profiles over the platform. Once a user submits a query, the Twitter Search presents the results in five filters (displayed in tabs): top, most recent, people, photo and video. Each of these filters provides a different set of results. The Twitter Search engine makes use of personalization algorithms to select and rank lists of tweets to users.

On Twitter, users usually follow others for sympathizing with their opinion. If Twitter takes it into account when deciding the order of search results, it may result in increasing polarization. Alongside the polarization issue, most consumers do not know that search results are personalized, yet users tend to place blind faith in the quality of search results (PAN et al., 2007). Hence, the Twitter Search interface provides different filters that may behave differently with each user. Thus, it is critical to understand the extent of these differences.

## 1.1 Goals

This master thesis aims to investigate the extent of personalization in the specific context of *social media search*. It is essential to understand how social media results are personalized based on user profiles to understand the polarization on the Web.

It is important to state that opening the black-box of personalization algorithms is not the goal of this study. Hence, we want to measure the effects of the personalization algorithms among polarized users on social media search.

## 1.2 Research Questions

This study focuses on the following research questions attached to some hypothesis:

**RQ1:** Do the personalization amounts change between Twitter Search tabs, advanced date filters or query terms?

- **H1:** Except for the recent tab, all the Twitter Search tabs present significant personalization.
- **H2:** There are statistically significant personalization differences as we change the date filters.
- **H3:** There are statistically significant personalization differences when we change the query terms.

**RQ2**: How much does the act of following accounts due to sympathizing with an opinion about a political topic may cause the Twitter Search personalization to provide different results for polarized users?

- **H1**: Personalization differences increase as we increase the polarization between profiles.

## 1.3 Research Methodology

For answering these questions, we make a quantitative research by experiment. We run an empirical experiment that consists of training fresh Twitter profiles with different viewpoints. Then, we developed software agents to follow social media profiles with different viewpoints related to a specific topic and execute automated search sessions. Similarity metrics are used to evaluate the hypothesis and answer the research questions.

The agents were divided into *PRO* and *ANTI*, representing a reform supporter and non-

supporter, respectively, and they strictly followed profiles with the same political inclination. As a baseline and to determine whether the results deviate from one to another profile, we created a *NEUTRAL* agent. The neutral agent does not follow any Twitter profile; therefore, Twitter personalization algorithms should not be able to perform any inference based on this agent.

We measure the amount of personalization of these sessions using existing similarity metrics for information retrieval (Jaccard Index and Edit distance), as well as a novel metric (Semantic Similarity), and evaluate the collected data from our experiment by doing quantitative analysis. We verify our hypothesis through non-parametrical statistical tests.

## 1.4 Main Contributions

The main scientific contributions of our work are threefold:

- A semantic similarity metric to enhance prior methodologies on measuring personalization in Web search (HANNÁK *et al.*, 2013, 2017; LE *et al.*, 2019);
- An empirical study to understand at what extent the number of followers affects the search results in social media;
- An empirical study to understand to what extent social media search results are affected by the search filters (videos, images, profiles, and top searches).

We also highlight some artifacts as technical contributions:

- The *Multiquerier tool*[5] – a system that instantiates bot agents to make search queries and collect the results on social media searches.
- The *2019 Brazillian Pension Reform Tweets (2019-BPRT)* dataset of **4,527 row**s

[5] Available at https://github.com/jonatascastro12/twitter-search-personalization-research

containing search results from simulated polarized users, as well as **67,240 tweets** (12,529 *ANTI* + 54,711 *PRO*) that were used to train our polarized accounts[6].

This master thesis was summarized in the paper "Is There Personalization in Twitter Search? A Study on polarized opinions about the Brazilian Welfare Reform" that was published on the 12th International ACM Conference on Web Science in 2020 (SANTOS *et al.*, 2020).

## 1.5 Work Outline

This master thesis is organized as follows:

- On chapter 2 we cover important topics about Web personalization and filter bubble;
- On chapter 3 we explain our experiment on measuring personalization for polarized users on Twitter Search, as well as, specify the Multiquerier tool;
- On chapter 4 we evalute the data collected through the experiment;
- On chapter 5 we discuss our results;
- We conclude our work on chapter 6.

[6] Available at https://github.com/jonatascastro12/twitter-search-personalization-research

# 2. Background

## 2.1 Web personalization

*Web personalization* is a subtopic of the *personalization* field, which can be defined as the "process whereby products and services are tailored to match individual preferences utilizing consumer data" (ALEXANDER, 2007; MONTGOMERY *et al.*, 2009). Although we can establish a parent-child relation between *personalization* and *web personalization*, personalization commonly refers to web personalization – there is often interchangeable use of the terms (SALONEN *et al.*, 2016). **Thus, for the purpose of this dissertation, we stick the personalization concept as: "the process of individualized matching to users' preferences through automated processes in the web environment".**

Many fields of research, such as marketing and information systems, have the web personalization as a focus area. Web personalization addresses human-computer interactions, but there is a prevalence of the technological focuses, while topics like consumer research and psychology emerge within models that supplement the technological spotlights. Therefore, many papers focus on topics similar to recommender systems, data collection and processes, or user profiling (SALONEN *et al.*, 2016).

The personalization process should help users to deal with the tremendous amount of

available information, and the personalization algorithms play an important role in automating this process. Personalization algorithms capture information about user's interests, preferences, web history and contextual information (e.g., time, location) to derive optimally customized results based on various approaches that should vary depending on the system goals and requirements (BOZDAG, 2013; HAIM *et al.*, 2018).

A relevant field in which personalization has been crucial is Information Retrieval (IR).

## 2.2 Information retrieval

In this section, we want to cover some critical concepts of Information Retrieval as summarized in Figure 1 that shows a standard process of IR (BAEZA-YATES *et al.*, 2011). We can divide the IR process into three subprocesses: the *retrieval process*, the *ranking process,* and the *indexing process*.

The **retrieval process** aims to provide *documents* from within a collection that is relevant to an arbitrary user *information need*. The *information need* is communicated to the system by a *query*. While the information need is the topic about which the user desire to know more, the query is the "code" the user sends to the system to communicate the information need. A document is *relevant* if it is one that the user perceives as containing information of value regarding their personal *information need* (BAEZA-YATES *et al.*, 2011; MANNING *et al.*, 2008). The most primitive kind of retrieval task, *Boolean Queries*, simply output if a document is relevant or not (MANNING *et al.*, 2008).

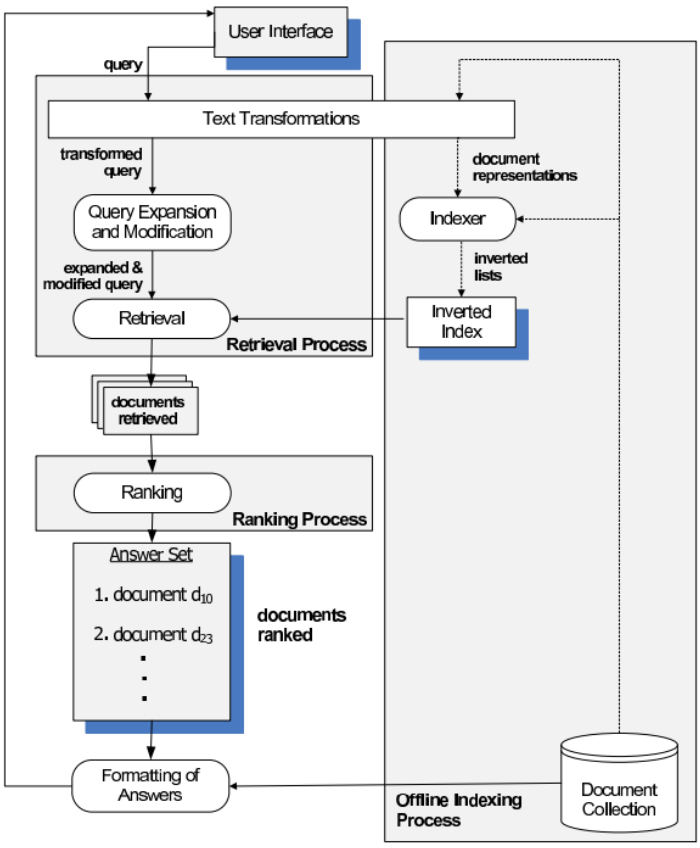



Figure 1: IR process – indexing, retrieval, and ranking documents (BAEZA-YATES *et al.*, 2011)

Regarding the query, it should be enhanced to improve the retrieval process. First, the query goes through text transformations where it is parsed and cleaned – e.g., fixing spelling mistakes and removing stopwords. Then, the query can be expanded or modified – this might be, for instance, concatenating suggestions or replacing certain terms with synonyms from a thesaurus[7] (BAEZA-YATES *et al.*, 2011).

The output of the retrieval task is a collection of documents. However, when we deal with big collections, the resulting number of matching documents can far exceed the number a human user could possibly sift through. For this reason, we need the **ranking process**.

---

[7] "Thesaurus" is a dictionary of synonyms.

In this process, we compute scores for each document regarding the query. This score tells about the *likelihood* of relevance of a document to the user about the query. So, the collection can be ordered by the scores descending, so that the first document of the list is considered the most relevant one. *Ranking* is critical on the IR process as it is directly linked to the quality of the results that are perceived to the users (BAEZA-YATES *et al.*, 2011).

In order to return a relevant document efficiently, we go through the **indexing process**. In this process, we build indexes upon the content of each document. The most popular index strategy is called *inverted indexes,* where all the distinct words from the documents collection are linked to a list of documents that contain it. This strategy is crucial on large and distributed IR systems like web searches. In this kind of systems, the system must provide search over billions of documents stored on millions of computers (BAEZA-YATES *et al.*, 2011; MANNING *et al.*, 2008).

Classical IR systems are based only on a query search. Given a collection of documents and a user information need, they aim to provide relevant resources to the user (PALTOGLOU *et al.*, 2010). However, different users may have distinct desires for a same keyword, which causes a query to be ambiguous (SANDERSON, 2008).

## 2.3 Personalized information retrieval

The traditional “one-size-fits-all” search strategy that returns the same search results to the same query without considering who submits it or under what circumstances limits search engine performance in providing relevant search results.

In this sense, a **personalized information retrieval** (PIR) can help to solve this issue by taking into account other information that is connected (or not) to the user, beyond the query itself (BOUHINI *et al.*, 2016; DAOUD *et al.*, 2011). Personalization is considered the most effective

technique that integrate the user in the retrieval process basing on the explicit and implicit elicitation of its preferences (FAKHFAKH *et al.*, 2016).

PIR enhance the IR process by complementing explicit user requests with implicit user preferences to meet individual user needs better. The main task of PIR is to exploit the user's profile and integrate it during the IR process (BOUHINI *et al.*, 2016). In the past two decades or so, PIR has received extensive attention in both academia and industry (ALEXANDER, 2007; BAEZA-YATES *et al.*, 2011; LIU *et al.*, 2020; SALONEN *et al.*, 2016).

The **user profile** is an especially important aspect of PIR. It can be used both on expanding the query or changing the indexation and ranking of the documents process (BOUHINI *et al.*, 2016). The PIR process constructs the user profile, usually based on the user interests that are captured according to the user's interactions inside an application (e.g., visited pages, *liked* items, people *followed*, previous searches). Beyond the user's direct interactions, PIR would also consider other contextual information like geo-localization.

On past approaches, PIR consider just implicit social annotations that correspond to the user's tags at system resources. This "social tagging" process is commonly known as a *folksonomy* (XU *et al.*, 2008; YEUNG *et al.*, 2008).

Social tagging is quite common on online social networks (DING *et al.*, 2009). The advantage of social tagging is that a single tag can be shared among the users. For instance, there is a high probability that a user's interest is shared with his neighbor inside a social network. Consequently, this aspect can be availed on both user profiles.

Many past works have proposed modeling the user profile in a Vector Space Model (CAI *et al.*, 2010; VALLET *et al.*, 2010). Usually, the social tags, query terms and documents are represented as vectors on the same space, so that it is possible to match documents with queries

and user's profile computing the cosine similarity (BOUHINI *et al.*, 2016).

A more recent facet for PIR is the introduction of semantics in the user profile to improve the relevance of search results as well as better understanding the user's intentions (MOHAMED *et al.*, 2017; RAFA *et al.*, 2018). The main idea over a semantic oriented PIR is considering not only a set of related terms but contemplating the meaning of these terms and the user's cognitive needs. The approaches over semantics link user information, social annotations, and contextual features like geolocation onto knowledge graphs, such as ontologies (RAFA *et al.*, 2018). Once all this information is linked to a single knowledge space, it can be used to expand the query to better express user needs.

## 2.4 Web search personalization

Although the PIR idea attracted researches for decades (BELKIN, 1993; TAYLOR, 1968), the web search industry started implementing personalized web searches only in the early 2000's. According to the Google Press Blog, *Google Web Search* introduced the "Personalized Search" idea in 2004 (GOOGLE, 2004) and released the feature in 2005 (GOOGLE, 2005). Only in late 2009, Google would deliver personalized search results for all users, even those without Google accounts (GOOGLE, 2009). Google would later use Google Plus social network to improve the personalization experience (GOOGLE, 2012). For example, web pages that were shared by a user from the same circle would be included in the search results. On April 2, 2019, Google plus was discontinued for personal use though[8], disabling the personalization effects on the search results.

Beyond Google, Microsoft introduced "Localized Results" and "Adaptive Search" on *Bing*

[8] https://support.google.com/plus/answer/9217723?hl=en

*Search* in 2011[9], customizing search results using the user's location and previous search history, respectively. In 2013, Bing added links shared by Facebook's users' friends alongside normal search results.

These previous personalization initiatives are just examples of explicitly mentioned personalization features on popular web searches engines. The personalization algorithms behind the scenes are proprietary, and their real impact is unknown for the users. For this reason, methods for measuring the impact of personalization algorithms are necessary. Before describing these methods, we run over a known issue about the personalization algorithms.

## 2.5 Filter bubble and echo chambers

The personalization mechanism on search results leads straight to a known issue called the filter bubble effect (PARISER, 2011). When two different users search for the same query at the same time, they have a chance of receiving different results. It happens because users are only given results that the personalization algorithm thinks they want. Consequently, the algorithm hides other results that may be important to the user. In other words, metaphorically, the algorithms put users into isolated universes (bubbles) of information to which users would have an affinity (according to the algorithmic filter).

PARISER (2011) outlines three dynamics of the filter bubble. First, *"you are alone in it"*, which means that it can make the user apart of other information, opposing to common media streams like a TV channel where other users would share the same interest about a single channel. Second, *"the filter bubble is invisible"* – the filtering occurs without the user perception. This lack

[9] https://searchengineland.com/bing-results-get-localized-personalized-64284

of perception is particularly dangerous for political viewpoints as the personalization algorithm assumes the user interest in a topic. This assumption can be biased, and the user may not have the chance to review it[10]. Furthermore, third, *"you don't choose to enter the bubble"* – comparing to a traditional media, the user can choose a newspaper or TV channel knowing the kind of filter is being chosen. This kind of choice does not happen with personalized filters. The author complements that the problem on those dynamics is that the personalization mechanisms "drive up profits for the web sites that use them" (PARISER, 2011).

An additional term that is very related to *filter bubbles* is the *echo chambers*. SUNSTEIN (2009) claims that social media companies put the users into polarized groups that amplify their viewpoints. He argues that individuals are largely exposed to conforming opinions only. Personalization algorithms may cause either filter bubbles or echo chambers.

Away from the filter bubbles and echo chambers, there are other theories not necessarily related to personalization that could explain polarization. We cite the *selective exposure* effect – "the idea that people purposefully select information matching their viewpoints" (STROUD, 2010). While the filter bubble is a direct implication from machines, the selective exposure effect would be a natural human desire. LE *et al.* (2019) state that algorithmic personalization can intensify the selective exposure beyond the user's choice. It results in a vicious cycle that can contribute to polarizing the society.

We summarize the three cited concepts that are used to explain polarization in Table 1.

---

[10] Some platforms are providing some more control about this aspect. See, for instance: https://twitter.com/settings/account/personalization

Table 1: Concepts that are used to explain polarization

| Concept | Reference | Description |
|---|---|---|
| **Caused by personalization algorithms** | | |
| *Filter Bubbles* | (PARISER, 2011) | Personalization algorithms put users into invisible information bubbles that reinforce their point of view. |
| *Echo chambers* | (SUNSTEIN, 2009) | Social media companies put users into polarized groups that amplifies their previous beliefs. |
| **Caused by natural human desire** | | |
| *Selective exposure* | (STROUD, 2010) | People purposefully select information matching their viewpoints. |

There is a long-running debate about whether filter bubbles and echo chambers produced by personalization algorithms are really a concern for the society (FLAXMAN *et al.*, 2016). On one side, there is the potential information segregation, on which people may choose to consume only contents that agree with their previously held beliefs. It leads to a serious concern to the democracy systems, where it is fundamental that all individuals have access to a variety of opinions (LASSEN, 2004).

On the other side of the debate, some researches argue that social media, in general, increases the exposure to a diversity of ideas (BENKLER, 2006). MESSING *et al.* (2012) show that there is a significant amount of links between polarized users on online social networks, which increases the possibility of diverse content discovery. Also, HOSANAGAR *et al.* (2014) explain that personalized recommendation systems help users expand their interests in media consumption.

While the real consequences of the filter bubble effect are still in debate, we cannot negate the existence of the algorithmic personalization and how they could hide certain contents from users. We reinforce that most consumers do not know that search results are personalized, yet users tend to place blind faith in the quality of search results (PAN *et al.*, 2007).

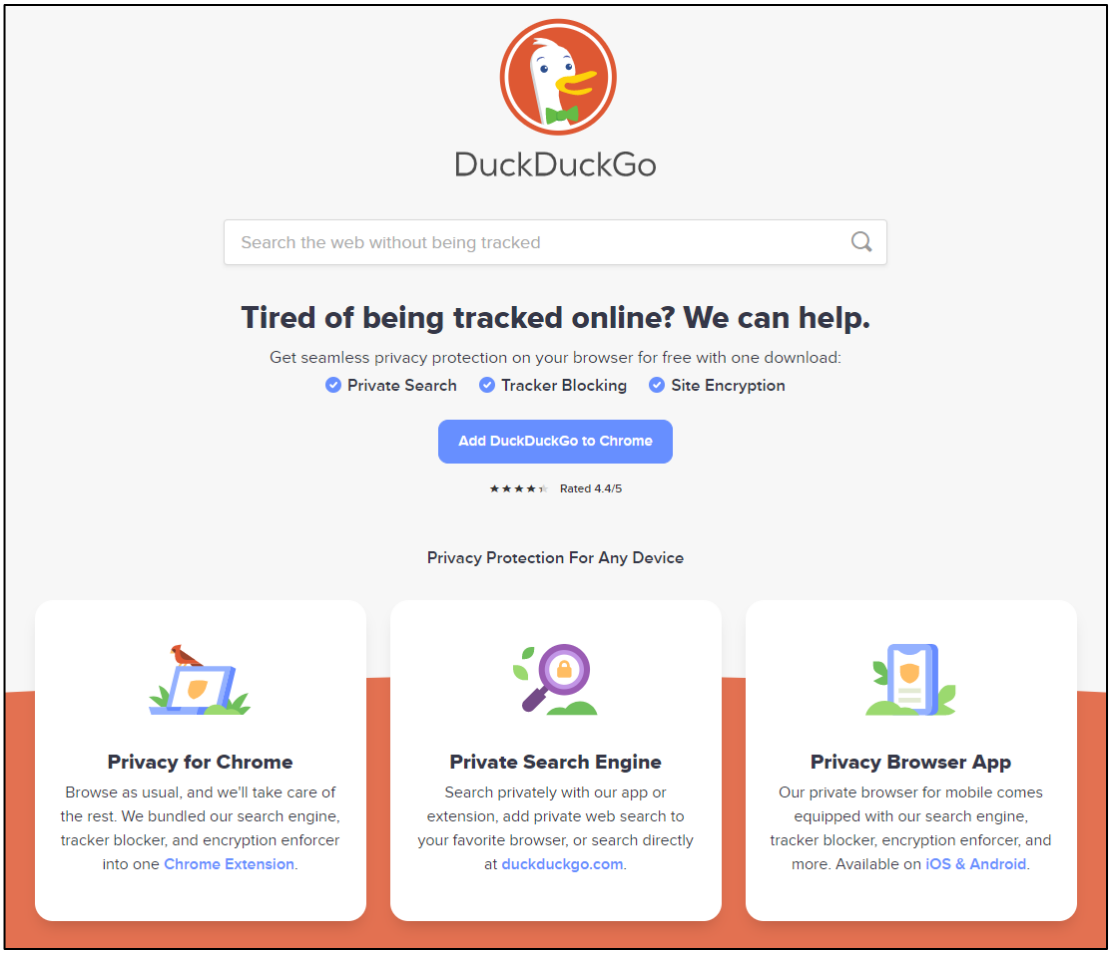



Figure 2: DuckDuckGo homepage screenshot from July 2020

As a result of these potential issues, the academia and the market have worked on solutions to handle the filter bubble effect. One of the solutions on the web search market is the rise of non-personalized search platforms such a DuckDuckGo[11]. Such platform promises that privacy is guaranteed, and user data is never shared, while the search results remain the same for any user.

On the academia, studies have proposed manners to detect (DILLAHUNT *et al.*, 2015; TRAN *et al.*, 2015), avoid (BOZDAG *et al.*, 2015) and proof the existence of filter bubbles on web search, news and social media (COURTOIS *et al.*, 2018; PUSCHMANN, 2019).

[11] https://duckduckgo.com/

As the filter bubble is a direct consequence of personalization and given the potential concerns to the society, it is important to provide mechanisms to measure the impact of personalization on platforms that users touch. In the following section, we cover related work for measuring personalization on a web search.

## 2.6 Measuring personalization on web search

HANNÁK *et al.* (2013) started a research line for measuring personalization on Web search motivated by the Filter Bubble phenomenon. They introduced a methodology to quantify personalization in Web search results using demographic and tracking data (such as user agent, navigation, browsing history and IP address) as features.

The methodology is conceptually simple: "running multiple searches from different instances for the same queries and compare the results". Nevertheless, in order to account the differences in the returned results to personalization requires considering multiple factors, such as temporal changes in the search index, consistency issues in distributed indices, A/B tests being run by the search provider (HANNÁK *et al.*, 2017).

The results are compared by measuring the differences between them. Along with HANNÁK's research, at least three metrics are used for quantifying these differences: the Jaccard Index, edit distance, and Hamming distance.

The ***Jaccard Index*** (JACCARD, 1901) views the results as sets. So it is stated as the size of the intersection over the size of the union, where $0$ represents no overlap between the lists; and $1$ indicates equal sets (Eq. 1). This metric looks for the presence or absence of the elements but does not account for the documents ranking.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

The ***Hamming distance*** (HAMMING, 1986) is the number of positions differences (replaces) between to lists. This metric is limited by the fact that the lists must have equal sizes. Note that when the hamming distance is 0, the sequences are identical, but when it is 10, for two 10-length sequences, they are totally different. We can express the metric with the mathematical expression for the *Manhattan distance* (Eq. 2) considering the position $i$ for two vectors $x$ and $y$:

$$D(A,B) = \sum_{i=1}^{k} |x_i - y_i| \tag{2}$$

The ***edit distance***, or the *Levenshtein distance* (DAMERAU, 1964), is a sum of the number of insertions, deletions, substitutions or swaps to make different lists equal. Therefore, it can look into the differences in the ranking of results. Note that when the edit distance is 0, the sequences are identical, but when it is 10, for two 10-length sequences, they are totally different. Mathematically, for two sequences $A$ and $B$ with lengths $i$ and $j$ respectively, the *Levenshtein distance* is defined as in Eq. 3:

$$E_{A,B}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ min \begin{cases} E_{A,B}(i-1,j) + 1 \\ E_{A,B}(i,j-1) + 1 \\ E_{A,B}(i-1,j-1) + 1_{(A_i \neq B_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{3}$$

where $1_{(A_i \neq B_j)}$ is the indicator function equal to 0 when $A_i = B_j$ and equal to 1 otherwise.

## 2.7 Related Work

We summarize a list of the main related work in Table 2 that can be directly compared to our work. HANNÁK *et al.* (2013) reported that Google and Bing search engines approximately personalize 11.7% and 15.8% of their results, respectively (HANNÁK *et al.*, 2017). KLIMAN-

SILVER *et al.* (2015) showed that Google personalizes the search based on the user location, especially for queries related to local businesses. Likewise, SALEHI *et al.* (2015) proposed a methodology to quantify personalization in the academic context. They observed slight personalization differences between personalized (Google Search) and non-personalized (DuckDuckGo) search engines for academic topics. HAIM *et al.* (2017) investigated the influence of search results on suicidal topics.

Table 2: Comparison between studies on Measuring Personalization in Web Search

| **Reference** | **Search platform** | **Metrics for measuring personalization** | **Semantic analysis from content** | **Context** |
|---|---|---|---|---|
| (HANNÁK *et al.*, 2013, 2017) | Google Search, Bing, DuckDuckGo | Jaccard Index, Edit distance | No | General |
| (KLIMAN-SILVER *et al.*, 2015) | Google Search | Jaccard Index, Edit distance | No | General |
| (SALEHI *et al.*, 2015) | Google Search, StartPage | Jaccard Index, Hamming distance | No | Academic |
| (HAIM *et al.*, 2017) | Google Search | Jaccard Index, Edit distance | No | Suicidal |
| (COURTOIS *et al.*, 2018) | Google Search | Jaccard Index, Kendall's Tau[12] | No | Social-political |
| (PUSCHMANN, 2019) | Google Search and Google News | Jaccard Index, Edit distance | No | Social-political |
| (LE *et al.*, 2019) | Google News | Jaccard Index, Edit distance | No | Social-political |
| **Our Work** | **Twitter Search** | **Semantic Similarity,** Jaccard Index, Edit distance | **Yes** | Social-political |

[12] A form of edit distance

Strictly for the social-political context, COURTOIS *et al.* (2018) found no evidence for personalization under their collected data, while PUSCHMANN (2019) found that the similarity of the results (Jaccard Index) on Google Search does not drop below 70% for parties and 80% for candidates indicating a few evidence for personalization. On the other hand, LE *et al.* (2019) observed significant personalization based solely on browsing history - they trained and compared browser profiles pro and anti-immigration at Google News.

Besides the main list of related work, a newer set of studies audits other aspects of the search result pages. For instance, ROBERTSON *et al.* (2018) provide an audit for Google Search that considers and identifies various components of the result page (e.g., video card, news card, embedded Twitter), rather than just the ordinary result items. Finally, HU *et al.* (2019) analyze search snippets from Google Search on the political context. They found that 54%-58% of snippets amplify partisanship. However, they express the need for applying semantic metrics in their approach as they only consider the presence of lexicons to account for differences in the search snippets.

Our work mainly differs from previous ones as (1) it semantically analyses the content of the tweets to explain personalization differences of Twitter search results, while we conduct an in-depth experiment and analysis to (2) understand to what extent the number of followers affects the search results in Twitter as well as (3) to understand how search filters in Twitter personalize search results.

# 3. Measuring personalization for polarized users on Twitter search

In this chapter, we present our methodology for measuring search personalization for polarized users on Twitter search. Our methodology leads us to measure the personalization caused by the user adherence of an opinion about a general topic.

We assume the premise that users express their opinion about a general topic by interacting with social media (e.g., following some Twitter account). These interactions serve as input to the personalization algorithms so that they can deliver compatible content to the users when they execute searches.

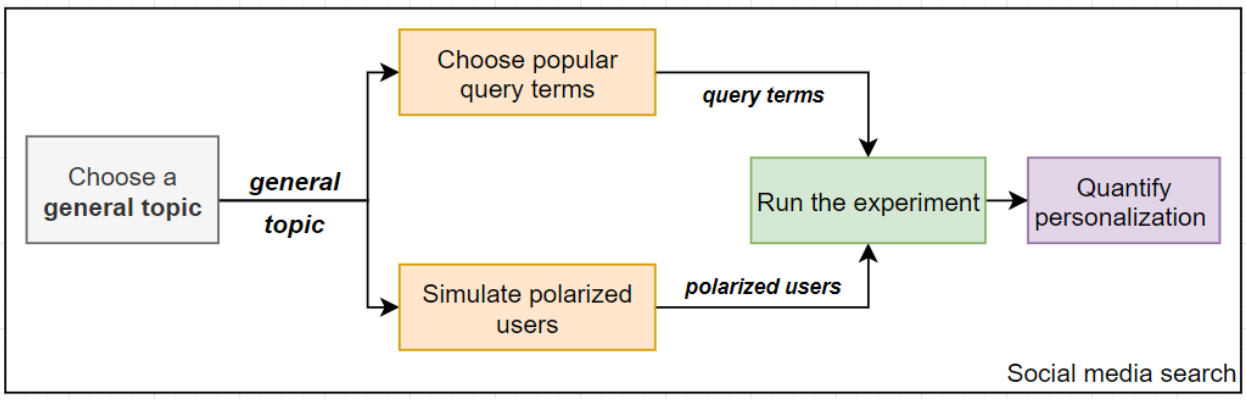



Figure 3: Methodology overview for measuring personalization for polarized users on Twitter Search

In our methodology, we want to simulate these interactions to "force" the development of polarized profiles at the social platform.

Therefore, we need three assets to execute our experiment:

- **the general topic** – we choose the Brazilian Social Welfare Reform
- **polarized user simulation mechanism (agents)** – the agents represent users on Twitter that are against/in favor the Brazilian Social Welfare Reform
- **popular query terms** – we came with 28 political query terms

In further sections, we explain the process we took to output each of these assets.

## 3.1 Choosing a general topic

As the general topic of our experiment, we piggyback on an important Brazilian debate in the early 2019's: *the Brazilian Social Welfare Reform* or the *Brazilian Pension Reform*. From this debate, we want to simulate the polarized users against and in favor of the Reform.

On March 22-23, 2019, the hashtags **#FightForYourRetirement** (*#LutePelaSuaAposentadoria*) and **#ISupportTheNewWelfare** (*#EuApoioNovaPrevidencia*) about the Brazilian Social Welfare Reform became evident on the Twitter trending topics (ESTADÃO, 2019). During this period, this political concern was the central topic of many media streams, social networks and street protests. Therefore, we found an opportunity to simulate our polarized users.

## 3.2 Simulating polarized users

For simulating polarized users executing queries over Twitter Search, we developed the

**Multiquerier tool**[13] based on the JavaScript Puppetter library[14] to train Twitter accounts and capture Twitter search results on an automated browser. We call the child instances of this tool: agents. Each agent simulates a kind of user that is capable of log in a Twitter account, follow a set of profiles, and execute a sequence of queries on Twitter Search. We instantiated three parallel agents:

- one that represents a user against the Reform - we named it “ANTI";
- one that represents a user that supports the Reform - we named it “PRO";
- one that represents a neutral user - we named it “NEUTRAL".

The latter is a baseline user intended to measure the differences from the previous ones.

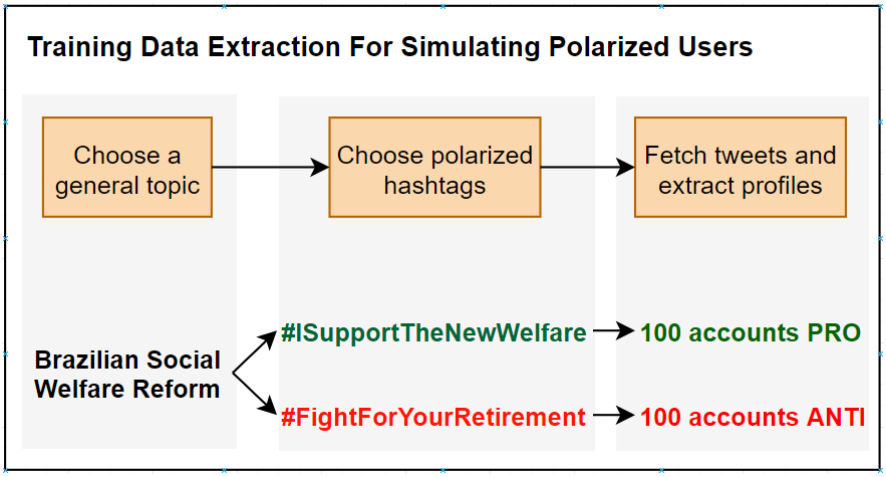



Figure 4: Training Data Extraction for Simulating Polarized Users

For the sake of training our agents, we needed to fetch some accounts that they would follow. These accounts should represent users that issue opinions against or in favor of the Brazilian Welfare Reform. To fetch these accounts, we first scrapped search results for the

[13] Source code publicly available at https://github.com/jonatascastro12/twitter-search-personalization-research

[14] https://developers.google.com/web/tools/puppeteer

polarized hashtags from our context in which **#FightForYourRetirement** represents *ANTI* tweets and **#ISupportTheNewWelfare** represents *PRO* tweets.

We assume that the profiles that tweeted these hashtags are representative of the polarized users. So, we captured tweets from March 9th, 2019 to November 6th, 2019. We fetched a set of 12,529 tweets for *ANTI* and 54,711 tweets for *PRO*. From these tweets, we extracted 3,952 unique accounts for *ANTI*, and 13,317 for PRO. Then, we balanced these numbers by ordering each set of accounts by the number of followers. Finally, we retrieved the top-100 profiles for each group. The overall process over these steps is summarized in Figure 4, and we summarize this data in Table 3. These data is part of the public available **2019-BPRT dataset**[15].

Table 3: Training data extraction summary

| | **ANTI** | **PRO** |
|---|---|---|
| **Original Hashtag** | `#LutePelaSuaAposentadoria` | `#EuApoioNovaPrevidencia` |
| **Translated Hashtag** | `#FightForYouRetirement` | `#ISupportTheNewWalfare` |
| **Number of captured tweets** | 12,529 | 54,711 |
| **Number of extracted unique profiles** | 3,952 | 13,317 |
| **Sample of the first 5 profiles (number of followers)** | `@teleSURtv` 1,807,063 | `@MomentsBrasil` 670,980 |
| | `@LulaOficial` 1,410,886 | `@kimpkat` 535,933 |
| | `@MarceloFreixo` 1,191,742 | `@MBLivre` 478,712 |
| | `@ptbrasil` 894,335 | `@Desesquerdizada` 318,426 |
| | `@GuilhermeBoulos` 701,716 | `@Biakicis` 301,587 |

*Note: Overview of the data used to train ANTI and PRO agents*

## 3.3 Choosing popular query terms

Our goal in planning the queries is to understand whether certain types of terms could influence users when querying. We summarize out query planning process in Figure 5. So, we

[15] 2019-BPRT dataset is available at https://github.com/jonatascastro12/twitter-search-personalization-research

harvested daily Brazilian trending topics from *a day before* (March 21th) to *a day after* (March 24th) the apex of the debate.

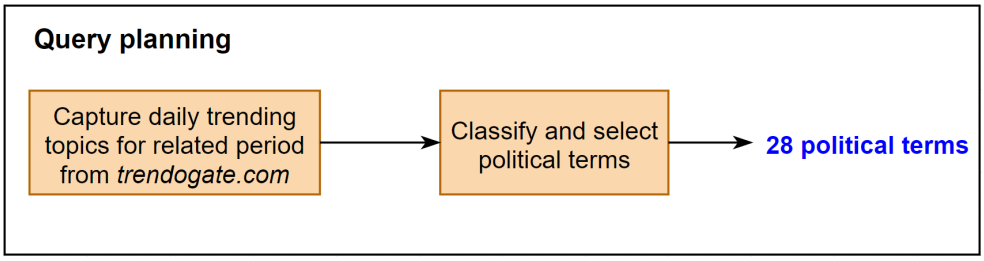



Figure 5: Choosing popular query terms: query planning

As Twitter does not provide free historical data, we scrapped the trending topics from *trendogate.com*. Therefore, we manually classified 200 trending topics into general categories from IAB Categories[16], including politics. Considering only the political context, we found out 28 topics and then performed two classifications onto these topics.

First, into *politicians*, *political issues*, and *humor and satire* (Class 1) (Table 4). To obtain more quality on these classifications, we randomized the set of trending topics and asked two research colleagues also to perform the classification, executing an agreement analysis.

Afterward, we considered a second classification (Class 2) that says whether the term expresses an *opinion* or is *informative* (Table 4). Thus, we were able to verify personalization based on these categories.

[16] Content taxonomy from IAB Tech Lab (https://www.iab.com/guidelines/taxonomy/); these categories are used on Twitter API for advertisement purposes (https://developer.twitter.com/en/docs/ads/campaign-management/api-reference/iab-categories)

Table 4: Political-related query terms

| Original Pt. term | Translated En. term | Class1 | Class2 |
|---|---|---|---|
| **Articulacao** | **Articulation** | Political Issues | Informative |
| **#OuReformaOuQuebra** | **#OrReformOrBreak** | Political Issues | Opinion |
| **NovaPrevidencia** | **NewWelfare** | Political Issues | Informative |
| **ProSul** | **ProSouth** | Political Issues | Informative |
| **#LutePelaSuaAposentadoria** | **#FigthForYourRetirement** | Political Issues | Opinion |
| **#LavaJato** | **#CarWash** | Political Issues | Informative |
| **#PergunteSobrePrevidencia** | **#AskAboutWelfare** | Political Issues | Informative |
| **#EuApoioNovaPrevidencia** | **#ISupportTheNewWelfare** | Political Issues | Opinion |
| **ARGEPLAN** | **ARGEPLAN** | Political Issues | Informative |
| **Lava-Jato** | **Carwash** | Political Issues | Informative |
| **PMDB** | **PMDB** | Politician | Informative |
| **Marun** | **Marun** | Politician | Informative |
| **MoreiraFranco** | **MoreiraFranco** | Politician | Informative |
| **Pezao** | **Pezao** | Politician | Informative |
| **Bretas** | **Bretas** | Politician | Informative |
| **CoronelLima** | **ColonelLima** | Politician | Informative |
| **Aecio** | **Aecio** | Politician | Informative |
| **EduardoCunha** | **EduardoCunha** | Politician | Informative |
| **FreixoePauloTeixeira** | **FreixoandPauloTeixeira** | Politician | Informative |
| **Sarney** | **Sarney** | Politician | Informative |
| **Temer** | **Temer** | Politician | Informative |
| **Pinochet** | **Pinochet** | Politician | Informative |
| **DilmaRousseff** | **DilmaRousseff** | Politician | Informative |
| **#LulaLivreDomingoSDV** | **#FreeLulaOnSunday** | Politician | Opinion |
| **Michelzinho** | **Michelzinho** | Humor and Satire | Opinion |
| **FaltaaDilma** | **MissingDilma** | Humor and Satire | Opinion |
| **Vampirao** | **BigVampire** | Humor and Satire | Opinion |
| **AteaDamares** | **EvenDamares** | Humor and Satire | Opinion |

*Note: The first column refers to the original terms in Brazilian Portuguese that were used in the experiment, and the second is an English version for a better context. Class 1 and Class 2 refer to our manual classification.*

In order to analyze the results of the Twitter Search according to the topics, it was important to consider the same timespan. Then, we decided to explore Twitter Advanced Search[17] that provides tags for time filtering (since and until[18]) to fetch results from *before the apex* (until: 2019-03-22), *during the apex* (since:2019-03-22 until:2019-03-24) and *after the apex* (since:2019-03-24). For each term from our list, we run a query with the term alone (no filter) and queries with the time filtering tags. This way, we can check for differences in personalization within these time constraints. We summarize these filters in Table 5.

Table 5: Time filtering from Twitter Advanced Search

| Filter | Description |
|---|---|
| `"until:2019-03-22"` | Before the apex of the discussion |
| `"since:2019-03-22 until:2019-03-24"` | During the apex of the discussion |
| `"since:2019-03-24"` | After the apex of the discussion |
| `""` | No filter |

## 3.4 Running the experiment

In this section, we explain the details of our experiment runs. We ran ten sessions of queries, and we incremented ten followings in each session. The steps of a session are described in Figure 6.

So, in the first session, we had each account following ten profiles, while in the last session, we had each account following 100 profiles. It gave us 1,680 sets of results per session, and a total of 5,600 triples (*ANTI/PRO/NEUTRAL*) of sets of comparable results, including all queries (28), filters (4) and tabs (5). Each set of results contains at least ten tweets per tab, resulting in the total

[17] https://twitter.com/search-advanced

[18] The until filter tag is not inclusive, so the *end* goes until 11:59:59PM of the previous date.

amount of 168,000 tweets. Each session took from 2-3 hours to complete.

```
input  : Twitter credentials, session id, profiles, terms,
         filters, and tabs
output : Search results for the session
profiles ← set of profiles (100);
terms ← list of political terms (28);
filters ← the list of filters to be concatenated to the term (4);
tabs ← list of tabs from Twitter Search (5);
Login(credentials)
if agent is not NEUTRAL then
    accounts_to_follow ← Pop 10 first accounts from
      profiles
    foreach account in accounts_to_follow do
        Follow(p)
    end
end
foreach term in terms do
    foreach filter in filters do
        RunQuery (term + filter)
        foreach tab in tabs do
            ClickOnTab(tab)
            CaptureAndSave(session, term + filter, tab,
              10)
        end
        Wait (60)
    end
end
```

Figure 6: Algorithm for an agent section

After running our queries, we saved each set of term results in a file with a unique filename. Then, we merged all the files into a single data file and removed all errors and inconsistencies (unbalanced results, *null* results) regarding the data collection process. We ended with a dataset of 4,527 rows. The full dataset is publicly available at a GitHub repository[19].

[19] Dataset publicly available at https://github.com/jonatascastro12/twitter-search-personalization-research

### 3.4.1 Noise treatment

We treated two main sources of noise in our agent running environment. First, we handled the timing noise, similar to (LE *et al.*, 2019). The manager server triggers the agent's actions simultaneously. It is very important for our experiment since we can control the timing factor. This way, we decreased the probability of differences between the results in the function of running the queries at different times. It is one of the advantages of running an automated execution rather than manually running the experiment from real-user profiles.

Also, we created fresh profiles for *ANTI/PRO/NEUTRAL* agents within an interval not higher than five minutes between each account creation. Additionally, our profiles were created with male names and mobile numbers from the same network carrier. We have not followed any account on the sign-up form, and we have not enabled the option that allows Twitter to track user usage on websites outside Twitter[20].

Another possible source of noise could be the location. A previous study on other search platforms reported high personalization in the function of location (KLIMAN-SILVER *et al.*, 2015). Twitter gives clues that it personalizes its content based on geolocation[21]. However, it is not clear whether Twitter applies this personalization to the search results. Thus, our agents run on the same machine, so that possible geolocation differences did not influence the search results. The machine was located in the city of Rio de Janeiro, Brazil.

---

[20] https://help.twitter.com/en/using-twitter/tailored-suggestions

[21] https://twitter.com/settings/account/personalization

### 3.4.2 Implementation of the Multiquerier Tool

The non-functional requirements for the Multiquerier tool are (1) unlimited on *scalability* – it should be able to run an unlimited number of agents; (2) *portability* – it should be able to run on different platforms (e.g., Windows, Linux); (3) *distributed* – it should be able to run over different hosts, so that, we can test different IPs; (4) search platform *generic* – it should be able to be used on different search platforms; (5) queries should run *simultaneously* and in *real-time*.

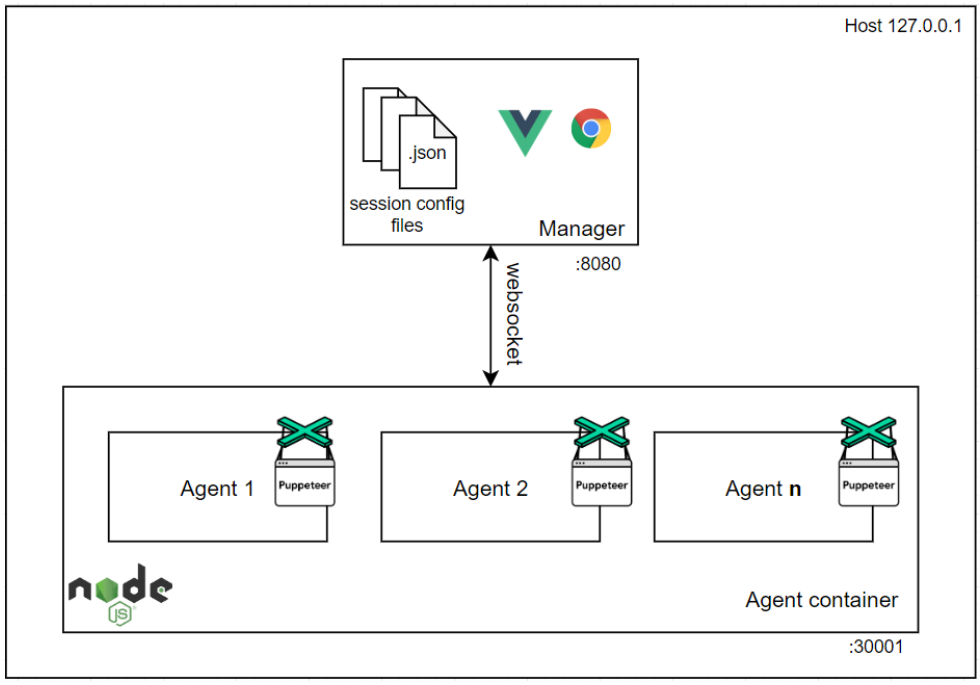



Figure 7: The Multiquerier architecture

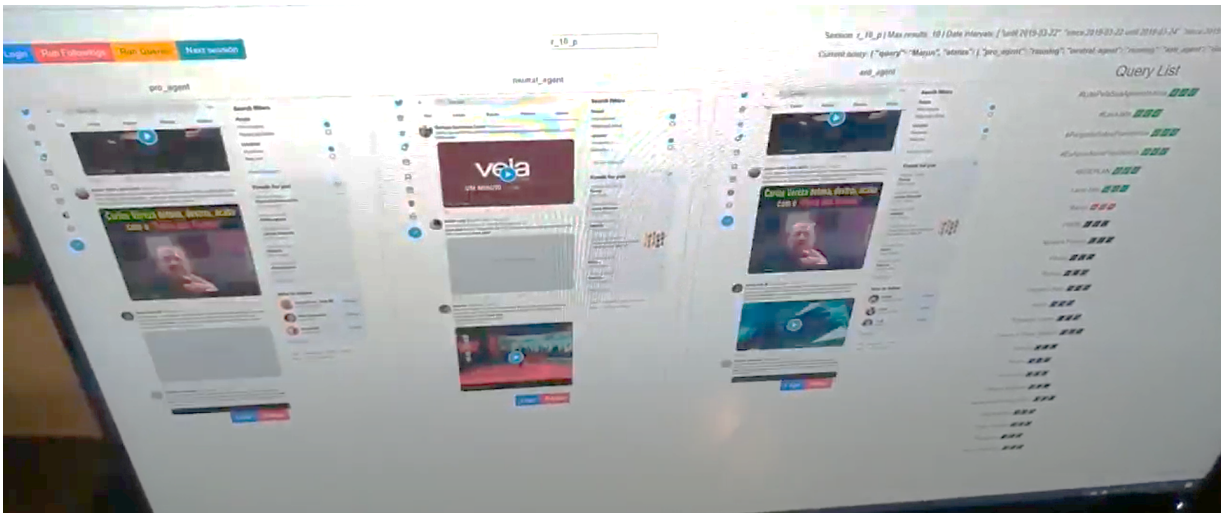



Figure 8: The Multiquerier user interface: it shows the sessions running in real-time. On the left side, we see the PRO agent snapshot. In the middle, we see the NEUTRAL agent snapshot, while, on the right side, we see the ANTI agent. At the right sidebar, we see the list of query terms. This screenshot was taken during the experiment running.

Keeping these requirements in mind, we designed our tool on a distributed architecture, as we show in Figure 7. Although we ran our experiment on a single machine for convenience, the containers can run on different hosts.

The manager is a web application running directly on the browser. It was implemented using Javascript and Vue.js framework for enhancing the tool user interface (Figure 8). It contains the logic to trigger the actions to all agents simultaneously in real-time. The manager should be initialized with a set of session config files (Figure 9). A single session should log in into a Twitter account, follow some profiles and them execute some queries on Twitter. The session config file contains:

- a **`session_id`**, that identifies each session;
- **`max_results`**, that is the maximum number of results that should be captured on each Twitter tab;
- the list of **`queries`**, that contains the query terms that will be searched in series;

- the **`agents`**' configurations, that contains an **`id`** for identification (e.g., pro, anti and neutral), credentials for login (**`username`** and **`password`**) and the **`platform`** identification (the tool is designed to be generic – we want to support other search platforms in the future - only Twitter is implemented for now);
- the list of **`profiles`**, for training the agent – the agent will follow each of them;
- the list of **`date_intervals`** that are used on the Twitter advanced filter, they are concatenated on the queries.


Figure 9: The Multiquerier session config file example

Once the session config files are placed, all session information can be sent from the

manager to the agent container, which is a Node.js application that can instantiate multiple Puppeteer instances. Puppeteer is a Javascript open-source library that allows to instantiate and control headless browsers (e.g., Chrome instances). The “headless” term means that the browser can be initialized as a background task without necessarily show a user interface. The agent container initializes as many Puppeteer instances as specified on the session file. Furthermore, the manager starts as many sessions as the number of session config files are placed.

The agent container communicates with the manager through a websocket connection. The websocket allows all communication to happen in real-time. Since it opens one connection, it allows the interchangeable exchange of asynchronous messages (Figure 10).

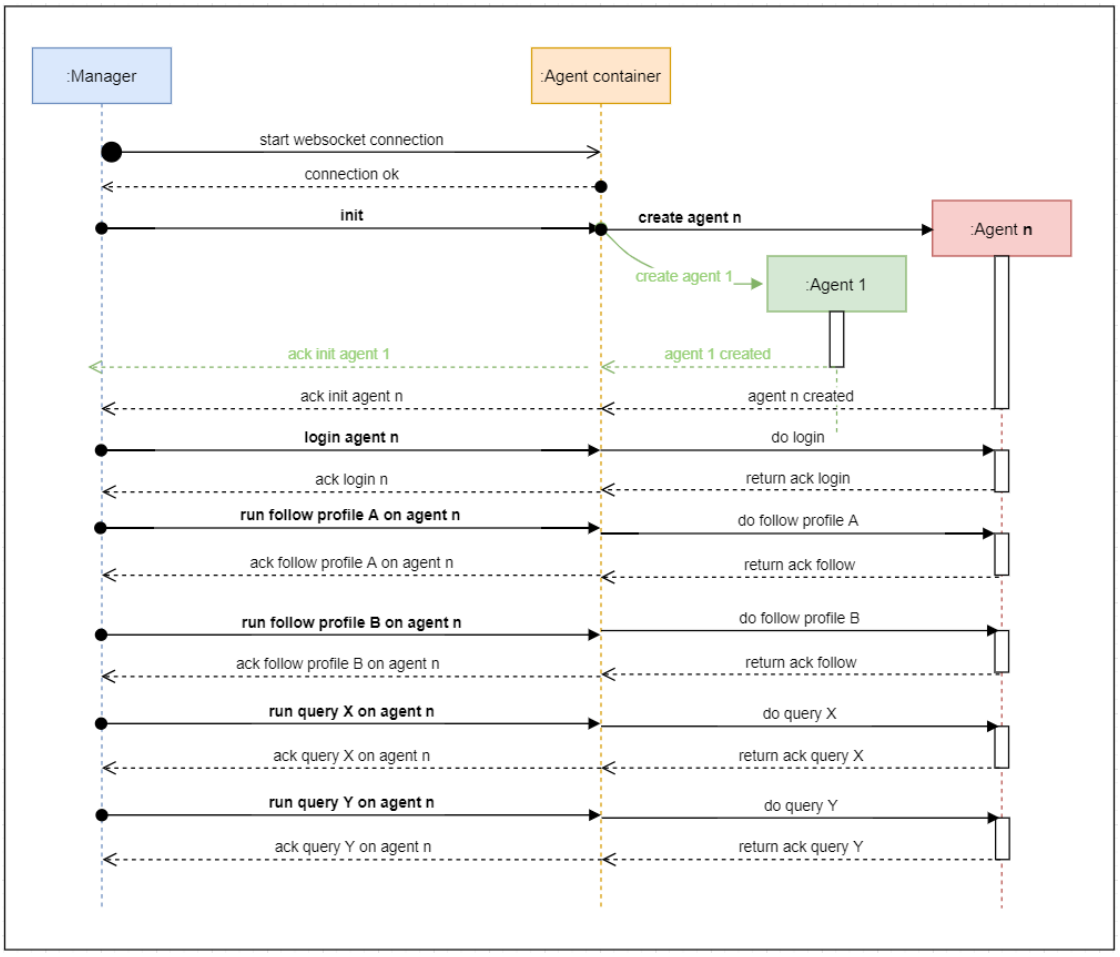



Figure 10: Sequence diagram that shows the exchange of messages between the manager and agent container

The session flow starts with the manager opening a connection with the agent container when the manager is started. Once the connection is opened, the manager initializes the process by sending the **`init`** message to the agent container. The **`init`** message triggers the simultaneous creation of all the agents that are specified at the config file. The agent container confirms that each agent was created by sending an **`ack init agent`** message from each initialized agent.

Once all the agent's creation is confirmed, the manager sends the login message simultaneously to all the agents. It makes each agent open the Twitter login page, type username and password, as specified in the **`agents[agent_name]`** config file and hits the login button. The agent container confirms the login is done for each agent, and the manager starts the followings actions by sending the **`run follow profile "A" on agent "n"`** to the respective agent as specified on the **`profiles[agent_name]`** list from the config file – where **`"A"`** is a name from a Twitter account, and **`"n"`** is the identifier for an agent. When each following action is finished, the agents send an **`ack follow profile "A"`** to signal that.

When all the followings are finished, the manager synchronously starts to run the queries among the agents. The manager sends a **`run query "X" on agent "n"`** to execute a search on Twitter for the term **`"X"`** from the agent **`"n"`**. When the search is finished, the agent sends an **`ack run query "X" on agent "n"`**, so that the manager can continue to the next query. The manager iterates through the list of queries as specified at the **`queries`** key from the config file.

Once all the queries from the queries key are done on all the agents, the manager looks for the next config file to repeat a new session flow.

### 3.4.3 Challenges encountered during the experiment

We faced some challenges during the experiment. The first challenge is regarding the creation of fresh Twitter accounts. To be successful in our experiment, we need to create testing

Twitter accounts to simulate our users. In the beginning, Twitter requires only an email to identify a real user. It gives the possibility to place a mobile number, although it is not required. So, we created a fake email account from a provider.

However, when we started running some sessions, Twitter suddenly requires placing a real mobile number to make an SMS confirmation. Our first cheap solution was using a paid SMS service that provides a temporary mobile-phone number. The issue is that the mobile number is temporary and allows us to receive only one SMS message. Unfortunately, Twitter would ask for a new SMS validation, which makes us lose our account.

Our new solution was buying some real mobile phones and some SIM cards to make some real mobile-phone numbers available. So, as we follow the sessions running, Twitter may suspect that we are running automated tasks and prompt a new phone validation – we manually place the SMS confirmation at the prompt.

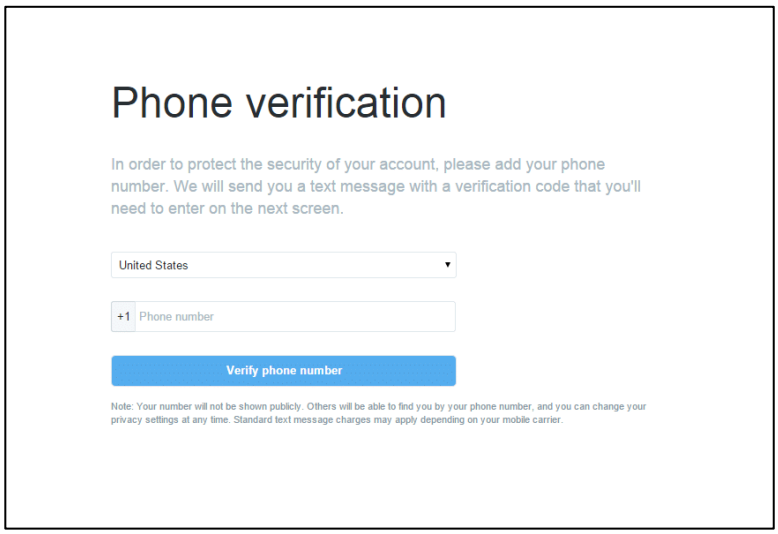



Figure 11: Twitter phone verification prompt screen

Another challenge was regarding the computer memory management during the execution of the experiment. When we first tried executing our experiment, it stopped on the second session

- eight sessions were missing – due to a memory leak. We realized that running the queries and scrolling the Twitter page would require a big amount of memory due to the download of several page resources (such as images and videos). Therefore, we decided to clear the browser's memory after each query (through a Puppeteer mechanism) and restart the agents after the end of the sessions.

## 3.5 Quantifying Search Personalization

We quantify personalization by calculating the difference between the results from the different types of agents (*ANTI/PRO/NEUTRAL*). First, we use two known metrics based on prior work (HANNÁK *et al.*, 2017; KLIMAN-SILVER *et al.*, 2015; LE *et al.*, 2019; PUSCHMANN, 2019; SALEHI *et al.*, 2015), the *Jaccard Index* (JACCARD, 1901), and the *Damerau–Levenshtein distance* (DAMERAU, 1964) or simply *edit distance*. Then, we introduce the use of a new metric that is capable of quantifying semantic differences.

While the metrics *Jaccard Index* and *Edit distance* are great to compute differences by checking the presence or absence of document identifications or changes in the ranking, they do not take into account the content itself. Search results can rather contain different identifications (*e.g., URL*), with different orders, but continue to have semantically similar contents.

Therefore, we introduce the ***semantic similarity*** metric based on sentence embedding. For calculating this metric, we need to convert our textual tweets into numbers. So we use a *state-of-art* sentence encoder model called *Multilingual Universal Sentence Encoder for Semantic Retrieval* (MUSE) (YANG *et al.*, 2019). This machine learning model converts our sentences into semantic rich vectors called *sentence embeddings*. These are 512-dimensional vectors that can extract semantic characteristics of a sentence. It means that if we input two different sentences to

MUSE, it will output two different vectors. Moreover, it allows us to input content from 16 languages, including English and Portuguese, using a unique semantic space. Thus, if we compare two sentences in different languages, but with the same meaning, it will output very similar vectors.

We use the outputs of the MUSE model as an input for our semantic similarity function. For a pair of vectors (sentence embeddings) $u$ and $v$, we do as Eq. 4. This similarity metric converts the traditional cosine similarity scores into angular distances that obey the triangle inequality, as suggests YANG *et al.* (2018).

$$s(u, v) = -arccos\left(\frac{uv}{\|u\|\|v\|}\right) \quad (4)$$

We first calculate the semantic similarity per pair of tweets within the two sets of results that have the same length. Then, we calculate the average similarity, characterizing the differences between the two sets (Eq. 5). Let $A$ and $B$ be two sets (of tweets) with the same length $n$, where $A_i$ and $B_i$ correspond to elements (tweets) of the set:

$$S(A, B) = \frac{\sum_{i=1}^{n} s(A_i, B_i)}{n} \quad (5)$$

Note that, when $S(A, B) = 0$, the set of sentences are completely different semantically, whether $S(A, B) = 1$, the set of sentences are very similar semantically.

# 4. Evaluation

We use $J(A,B)$ for the Jaccard index, $E(A,B)$ for the edit distance and $S(A,B)$ for the semantic similarity, where $(A,B)$ represents a pair of search results. We calculate these metrics over three pairs: *ANTI* and *NEUTRAL*, *PRO* and *NEUTRAL*, and *PRO* and *ANTI*. We summarize our set of calculated metrics in Table 6 and summarize the headers of our dataset in Table 7 with sample data.

Table 6: Metrics

| Metric Name | A | B | Pair Result Metric |
|---|---|---|---|
| Jaccard index | ANTI | NEUTRAL | $J(A,N)$ |
| Jaccard index | PRO | NEUTRAL | $J(P,N)$ |
| Jaccard index | PRO | ANTI | $J(P,A)$ |
| Edit distance | ANTI | NEUTRAL | $E(A,N)$ |
| Edit distance | PRO | NEUTRAL | $E(P,N)$ |
| Edit distance | PRO | ANTI | $E(P,A)$ |
| Semantic Similarity | ANTI | NEUTRAL | $S(A,N)$ |
| Semantic Similarity | PRO | NEUTRAL | $S(P,N)$ |
| Semantic Similarity | PRO | ANTI | $S(P,A)$ |

*Note: Result metrics for comparing the pairs of search result sets (A,B).*

We want to highlight the distinction of important concepts regarding our results: **"*personalization*"** vs. **"*personalization differences*"**. Having ***personalization*** means that a fresh user (that is using the search platform for the first time) and an old user (that took various actions on the platform along a period, e.g., clicking on various elements of results) used the same query, but encountered different results. We measure these differences when we compare *ANTI x NEUTRAL* or *PRO x NEUTRAL*.

Table 7: Sample of the dataset of Twitter Search

| session | term | class1 | class2 | filter | tab | E(A,N) | E(P,N) | E(P,A) | J(A,N) | J(P,N) | J(P,A) | S(A,N) | S(P,N) | S(P,A) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r_010_p | Articulation | Political Issues | Informative | until_2019-03-22 | top_tab | 10 | 10 | 0 | 0.11111 | 0.11111 | 1.00000 | 0.134976 | 0.134976 | 0.999751 |
| r_010_p | Articulation | Political Issues | Informative | until_2019-03-22 | latest | 0 | 0 | 0 | 1.00000 | 1.00000 | 1.00000 | 0.999837 | 0.999837 | 0.999833 |
| r_010_p | Articulation | Political Issues | Informative | until_2019-03-22 | people_tab | 0 | 0 | 0 | 1.00000 | 1.00000 | 1.00000 | 0.999638 | 0.999638 | 0.999630 |
| r_010_p | Articulation | Political Issues | Informative | until_2019-03-22 | photos_tab | 10 | 10 | 0 | 0.05263 | 0.05263 | 1.00000 | 0.084183 | 0.084183 | 0.999800 |
| r_010_p | Articulation | Political Issues | Informative | until_2019-03-22 | videos_tab | 10 | 10 | 0 | 0.25000 | 0.25000 | 1.00000 | 0.115051 | 0.115051 | 0.999677 |
| | | | | | ... | | | | | | | | | |
| r_100_p | Even Damares | Humor and Satire | Opinion | since_2019-03-22... | videos_tab | 8 | 6 | 2 | 0.666667 | 0.818182 | 0.818182 | 0.220291 | 0.501320 | 0.234108 |
| r_100_p | Even Damares | Humor and Satire | Opinion | since_2019-03-22... | videos_tab | 8 | 6 | 2 | 0.66667 | 0.81818 | 0.81818 | 0.220291 | 0.501320 | 0.234108 |
| r_100_p | Even Damares | Humor and Satire | Opinion | since_2019-03-24 | top_tab | 10 | 10 | 4 | 0.05263 | 0.05263 | 1.00000 | 0.124994 | 0.100881 | 0.563498 |
| r_100_p | Even Damares | Humor and Satire | Opinion | since_2019-03-24 | latest | 0 | 0 | 0 | 1.00000 | 1.00000 | 1.00000 | 0.999842 | 0.999842 | 0.999839 |
| r_100_p | Even Damares | Humor and Satire | Opinion | since_2019-03-24 | photos_tab | 10 | 10 | 5 | 0.25000 | 0.25000 | 1.00000 | 0.104858 | 0.089437 | 0.412785 |

*Note: The sample lists the top-5 and bottom-5 rows from the dataset that calculates the different metrics for the pairs of search results. $N = 4527$. Check the full dataset at: https://github.com/jonatascastro12/twitter-search-personalization-research/blob/master/datasets/twitter-search-results/exported_dataset_metrics.csv*

Having ***"personalization differences"*** means that two old users (that took different and various actions along a period, e.g., following polarized profiles) run the same query but encountered different results. We measure these differences when we compare the *ANTI x PRO* results.

## 4.1 Comparing the metrics

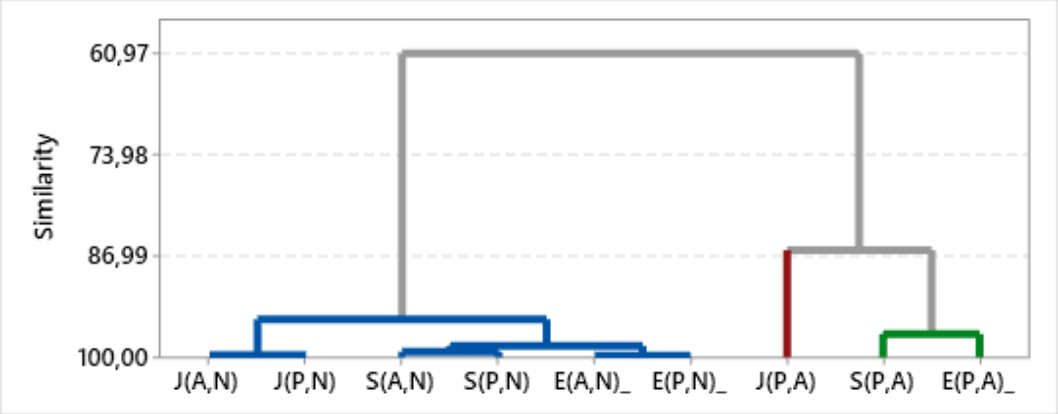



Figure 12: Dendrogram for all the metrics.
Complete linkage; Correlation Coefficient Distance

We first evaluate the correlation between our three metrics. For this evaluation, we standardized the edit distance to be compatible with the other metrics. Figure 12 shows a tree diagram that displays the groups formed by the clustering of variables at each step and their similarity levels (i.e., *dendrogram*). This graph gives us some clues about the correlation of our metrics.

First, looking at the bottom of the graph, we note that the pairs of metrics for PRO and ANTI personalization ($(A, N)$ and $(P, N)$) are strongly correlated ($\approx 99.75\%$). It is also a strong evidence that the *ANTI* and *PRO* agents received the same amount of personalization. We will verify that further in the text.

Second, concerning the $(P, A)$, the *semantic similarity* is strongly correlated to the *edit distance* ($\approx 97.03\%$). This finding makes sense because it shows how changing the order of results can highly impact on the semantic differences. However, $S(P, A)$ is a little less correlated to the *Jaccard index* ($\approx 86.25\%$).

Thus, for the next analysis, we will avoid repeating the metrics for $(P, N)$ as it follows almost the same distribution as $(A, N)$.

### 4.1.1 The semantic similarity

The Semantic Similarity ($S$) metric is complementary to the other metrics. The key point is: when comparing the differences between the lists of ranked documents, the semantic similarity metric takes into account the content of the document (e.g., tweet text), rather than simply comparing the document identifier (e.g., tweet ID/URL). The other metrics (Jaccard and Edit distance) would only take into account the document identifier.

For instance, we can typify the situation of two tweets that contain similar content that could be an important opinion about a topic. While the Jaccard and Edit distance metrics would consider two completely different documents, the Semantic Similarity metric would consider the similarity between the two documents because they present similar content.

We can detach these differences when we look at the scatter plot between $S$ and $E$ (Figure 13). When the edit distance is 10, it means that the results are completely different, and there is no intersection between the two results. However, we cannot affirm that the results content is not semantically similar. For instance, when $E(A, N) = 10$, we see that the $0 < S(A, N) \lessapprox 0.4$, but even when $E(P, A) = 2$ (e.g., *just two swaps probably*), the $S(P, A)$ can still reach very low values.

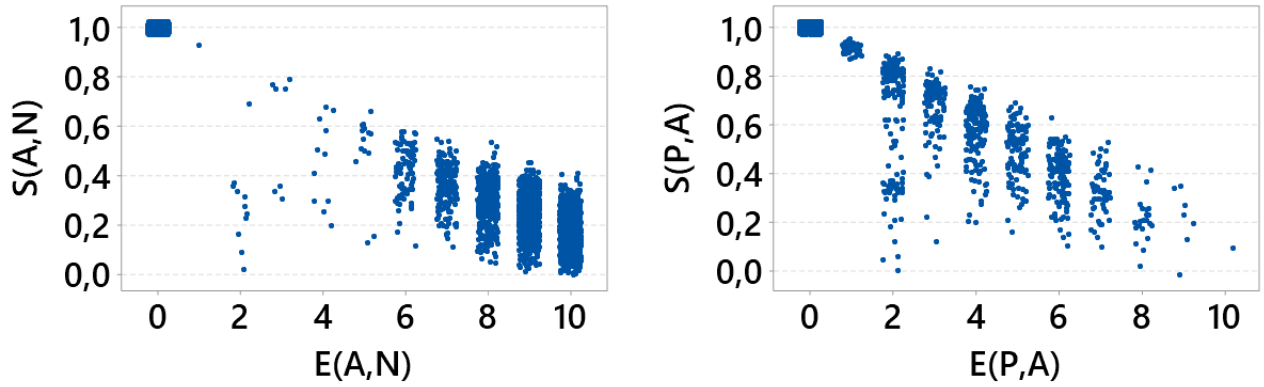



Figure 13: Scatter plot between S(A,N)×E(A,N) and S(P,A)×E(P,A)

## 4.2 Comparing personalization per tabs

Twitter Search interface presents five tabs with a set of results that we capture (Figure 14): *top*, *latest*, *people*, *photo* and *videos*. According to the Twitter Search FAQ page[22], the *top tab* shows “Tweets you are likely to care about most first", and it says an algorithm selects the content. However, it does not say much about the other tabs.

[22] Twitter Search result FAQs - https://help.twitter.com/en/using-twitter/top-search-results-faqs (accessed on January 31st, 2020)


Figure 14: Example of a query instance at the Twitter Search interface.
We mark the feature that we capture or label on our dataset. In this example, we query for "Brazilian welfare reform", and we label the results as from the "top" tab.

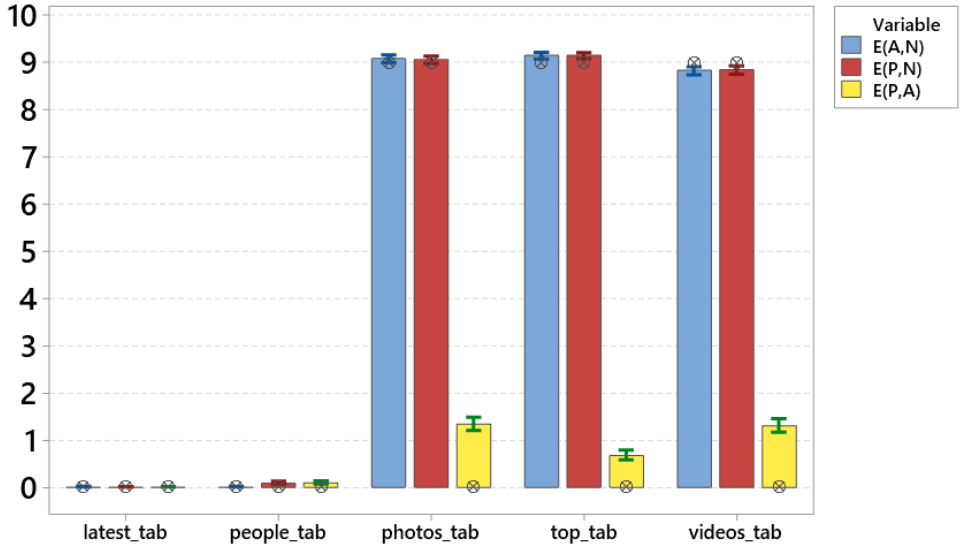



Figure 15: Edit distance per tabs

Thus, we start our analysis checking if there are differences in personalization between the Twitter Search tabs. Figure 15 shows the bar plot of means of the edit distance for each tab, where the $\otimes$ symbol indicates the median. We want to verify two hypotheses over this plot. First, we suspect that the *latest* and *people* tabs are never personalized (i.e., $E = 0$) **(RQ1.H1)**. Second,

we question if the *ANTI* and *PRO* agents present the same amount of personalization (i.e., $E(A,N) - E(P,N) = 0$ $J(A,N) - J(P,N) = 0$ $S(A,N) - S(P,N) = 0$) **(RQ2.H1)**.

For the first hypothesis, we use the Wilcoxon signed-rank test to check for the medians (Table 8). We cannot reject the null hypothesis that indicates non-personalization in all metrics on the *latest tab*. However, we cannot say the same for the *people* tab.

Table 8: Wilcoxon signed-rank test for “Latest" and “People" tabs personalization

| | **Latest tab** | | | | **People tab** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Statistics** | **P-Value** | **Median ($H_0$)** | $H_A$ | **Statistics** | **P-Value** | **Median ($H_0$)** | $H_A$ |
| **E(A,N)** | 3 | 0.186 | 0 | > | 1 | 0.5 | 0 | > |
| **E(P.N)** | 1 | 0.5 | 0 | > | 378 | 0 | 0 | > |
| **E(P.A)** | 1 | 0.5 | 0 | > | 406 | 0 | 0 | > |
| **J(A.N)** | 0 | 0.186 | 1 | < | 0 | 0.5 | 1 | > |
| **J(P.N)** | 0 | 0.5 | 1 | < | 0 | 0 | 1 | > |
| **J(P.A)** | 0 | 0.5 | 1 | < | 0 | 0 | 1 | > |
| **S(A,N)** | 471,279 | 1 | 0.9997 | < | 148,372 | 1 | 0.9998 | > |
| **S(P,N)** | 472,260 | 1 | 0.9997 | < | 134,995 | 1 | 0.9998 | > |
| **S(P,A)** | 457,637 | 1 | 0.9997 | < | 116,466 | 1 | 0.9998 | > |

For the second hypothesis, we want to verify whether the pair $(A,N)$ and $(P,N)$ for each metric is equal by applying the Mann-Whitney test to the distributions (Table 9). With a 95% of confidence level, we cannot reject the null hypothesis that the differences between the distributions are equal. From these results, we can conclude that both *PRO* and *ANTI* results are receiving the same amount of personalization. It does not mean, however, that they are receiving the same results, although the magnitude of these differences is very low. Therefore, for this observation, we need to perform a more in-depth analysis of $(P,A)$ metrics, as in Figure 5, the mean and median are distant from each other.

Table 9: Mann-Whitney between (A,N) and (P,N)

| $H_0$ | W-Value | P-Value* |
|---|---|---|
| $E(A,N) - E(P,N) = 0$ | 8,559,419.50 | 0.899 |
| $J(A,N) - J(P,N) = 0$ | 8,547,907.00 | 0.959 |
| $S(A,N) - S(P,N) = 0$ | 8,538,291.5 | 0.841 |

*Confidence level = 95%

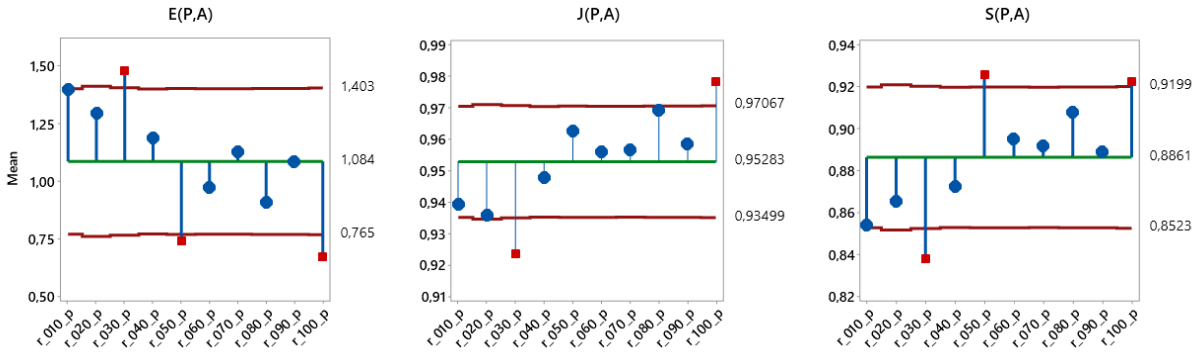
For the next analysis, we filter out the *latest* tab, as it does not manifest personalization, and the *people* tab, as is presents a slightly different kind of content. We will also stand only with $(P, A)$ metrics, as we have shown that both *PRO* and *ANTI* results are receiving the same amount of personalization.

## 4.3 Comparing personalization per session

In each session of execution in our experiment, we make our agents follow ten more accounts. We name each session with the identifier `r_<n>_p`, where *n* corresponds to the number of profiles that are followed in each session.

Thus, we start with ten followings at `r_010_p` but end with 100 followings at `r_100_p`. We would expect that the personalization differences increases as the number of followings increase. However, when we plot an analysis for the means with a $\alpha = 0.05$ of significance level (Figure 4.3), we see very low differences between the sessions, and we see more a trend for a decrease in the differences as the number of followings increases **(RQ2.H1)**.

The red dots speak for sessions that are outside the significance level. It means that for some reason, they presented atypical values of difference. In our case, the *30 followings* session (`r_030_p`) results are more personalized than the other groups, while the *50* and *100 followings* session are less personalized.




Results differences for $E(P,A)$, $J(P,A)$, and $S(P,A)$ per session

## 4.4 Comparing personalization by date filter

For each term that was queried by the agents, we concatenated some advanced filters from Twitter Search that delimit the time-period of the search results (Table 5).

We want to verify the personalization level when the user queries before, during or after the apex of the discussion on Twitter. Figure 16 shows the analysis for the means between these filters.

None of the means fitted at the interval of significance, but the *before apex* and *apex* period would present fewer differences than the *after apex* period. Moreover, if we do not apply any date filter (*no filter*), our results would be more different than placing any of the filters.

As we do not have statistical significance on these numbers, we cannot conclude **RQ1.H2.**

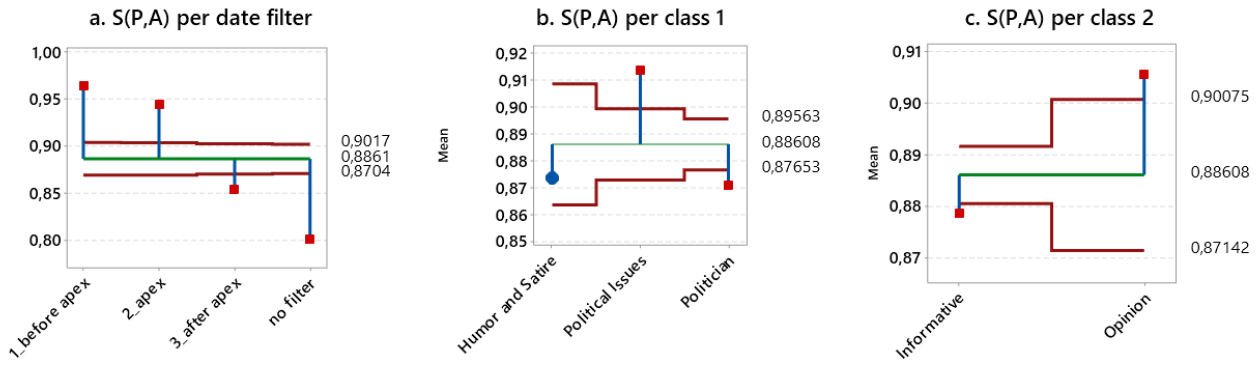



Figure 16: Results for S(P,A) per date filter, Class 1 and Class 2

## 4.5 Comparing personalization by terms

Before analyzing the individual terms, we study the terms classifications. As explained in previous sections (Section 3), we classified our query terms based on IAB categories. We want to examine whether the differences between *PRO* and *ANTI* results vary in the function of these classes. We show the analysis of means in Figure 16.b and Figure 16.c.

The graphics show that for class 1, queries for *Political Issues* would cause fewer differences than *Politician*, while *Humor and Satire* remain next to the mean. However, informative queries would provoke more differences than opinion. Although the magnitude of these differences is very low, this result is counter-intuitive because *opinion* terms are more likely to bring polarized results.

We finally investigate the results by the query terms. We want to verify when the *ANTI* and *PRO* agents would have more probability to receive different results. Looking into the mean analysis for $S(P,A)$ (Figure 17), we see some terms that are beyond the significance level ($\alpha = 0.05$). They represent atypical results that run out the centrality of the mean.

First, we list the red dots at the top: **`#FightForYourRetirement`**, **`#ReformOrBreak`**,

**ARGEPLAN**, and **Sarney**. For these terms, the differences would be the minimum as they cross the significance level at the top.

Second, we list the red dots at the bottom: **Aecio**, **Marun**, and **Pezao**. For these terms, the differences are more evident as they cross the significance level at the bottom.

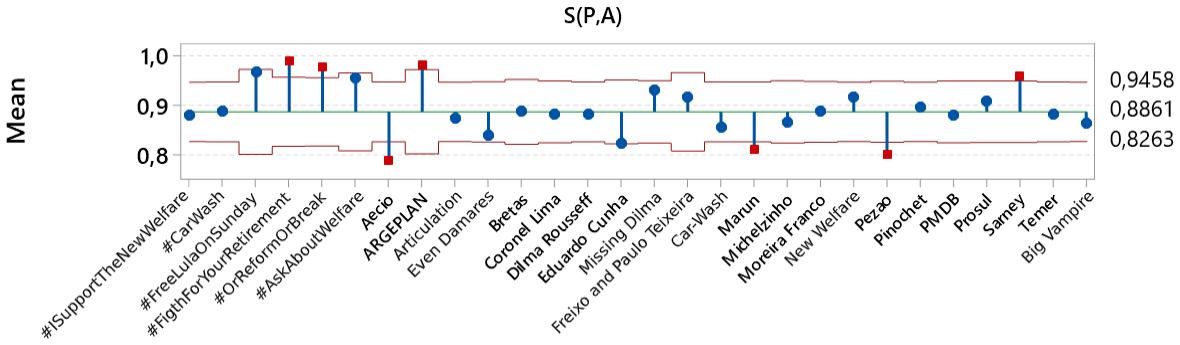



Figure 17: Results for S(P,A) per query term

We cannot make further assumptions for **RQ1.H3** *(There are statistically significant personalization differences as we change the query terms)* as the results are inconsistent, and the magnitude of these results is very low.

# 5. Discussion

In this section, we discuss some important topics that we cover in this work.

## 5.1 Twitter Search personalization

When one dive into Twitter help topics, one can find an effort from the Twitter team for showing transparency on the personalization mechanisms[23]. After a series of issues on political intervention (KRUIKEMEIER, 2014), Twitter has taken lots of actions to fight against these bad effects. Indeed, they have created a variety of gears that gives more control to the user on how Twitter content is personalized.

However, although there is more control for the user, the real impact of its personalization algorithms, at least on the search features, is still obscure. There is very scarce documentation on Twitter about the behavior of its search personalization algorithm.

The goal of our empirical study is to quantify the limits of Twitter Search personalization. Thus, we discuss our results according to the research questions.

**RQ1 -** ***Do the personalization amounts change between Twitter Search tabs, advanced date filters or type of query terms?*** For Twitter Search tabs **(RQ1.H1)**, yes. We statically

[23] https://help.twitter.com/en/search?q=personalization

concluded that the *latest* tab is not personalized at all. Although this might be obvious for someone, there is still scope for thinking in a kind of personalization for recent items.

We found very low evidence of personalization on the people tab, but we could not assert it statistically. Although our data do not confirm the absence of personalization, we observed an ambiguity on the concept of this tab. Rather than fetching only people references, it actually brings any kind of Twitter account, whether it represents a people, a place, institution, issue or any other entity.

About the other tabs (*top*, *photo*, and *video*), we found significant amounts of personalization for all the agents. There are very few differences between the polarized agents, which we will discuss at RQ2.

For the date filters **(RQ1.H2)**, we could not make any statistical conclusion as we do not found any significance in our data. For the query terms **(RQ1.H3)**, we verified the differences between the 28 terms alone and by placing the queries into two kinds of classification. For both cases, we found very few evidences of differences. We cannot make major statistical conclusions among these evidences.

**RQ2 - *How much does the act of following accounts due to sympathizing with an opinion about a political topic may cause the Twitter Search personalization to provide different results for polarized users?*** By our empirical results, we may risk saying *"very little"*. The main reason is that the magnitude of the metrics that compare PRO and ANTI results $(P, A)$ were relatively low when compared with the amount of personalization that we found for $(A, N)$ and $(P, N)$.

Even though those differences are very low, we should be aware that a simple swap on the ranking for the first items may have a large impact on the final meaning of the results. We raise two reasons for that. *First*, we showed in Section 4.1.1 that a low value for the edit distance might

trigger a low value for semantic similarity, i.e., the semantic meaning may severely be changed by changing a simple item in the search results. *Second*, if we focus on the first items of the search results from the *photo* and *video* tabs, we notice that a single tweet can fill the whole screen. Generally, an ordinary tweet fits all the screen on both the desktop and mobile versions of Twitter. It means that whether the algorithm changes just the first result, this change may cause a high impact for the user. It is something that we want to investigate in future work.

On the other hand, we want to highlight an important finding. Twitter drastically personalized the results yet on the first session of the experiment when the agents have followed only ten accounts. We have substantial data to say that the *ANTI* and *PRO* results are very different from the NEUTRAL results ($E(A,N)$ and $E(P,N)$ are next to 10, while $J(A,N)$, $J(P,N)$, $S(A,N)$, and $S(P,N)$ are next to 0).

We suspect that Twitter activated a kind of personalization profile that was very similar between the *ANTI* and *PRO*, but this profile was not very influenced as our agents follow new accounts.

## 5.2 Semantic similarity metric

One of the main contributions of our work is the introduction of the semantic similarity metric. Past studies on measuring personalization (HANNÁK *et al.*, 2013; KLIMAN-SILVER *et al.*, 2015; LE *et al.*, 2019; SALEHI *et al.*, 2015) were not able to compare differences based on the content. Generally, they rely on the URL (*or part of it*) as keys of the elements to compare for the differences between the search results.

Apropos, we note that the granularity of the other metrics would increase as the size of the result set increases. So if we have a limited size of the results, the other metrics may not be enough

to measure differences reliably. For instance, the metric space of the Edit distance metric for comparing 10-documents results would be $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The semantic similarity, however, would have an infinite metric space within the $[0 - 1]$ interval, even the number of results is limited in 10. Thus, besides the capability of compare sentences semantically, the semantic similarity metric allows making more detailed and granular comparisons.

## 5.3 Polarized hashtags

Although it was not the main focus of our work, we uncover some interesting results on the data that we collected for training our agents. We rather looked for two polarized hashtags about a topic, in our case, *the Brazilian Welfare Reform*. However, the terms apparently characterized very well a left-right spectrum of political polarization when looking into the fetched accounts (Table 3). For the ANTI-Reform agent, we found accounts that are more aligned with the left-leaning, while for the PRO-Reform agent, we found accounts that are more aligned with the right-leaning.

Although we would need a deeper social-political study to confirm those assumptions about the data, it sheds a light that we could use these accounts as a ground truth for classifying political leaning on future work.

# 6. Conclusion

## 6.1 Contributions

We made a controlled experiment to quantify personalization on Twitter Search. For executing the experiment, we built a set of tools to train agents that simulate polarized users and execute queries on Twitter Search.

Our results revealed significant **personalization** on Twitter Search when a user follows just a few accounts. Moreover, our results showed no personalization on the *latest* tab and very few on the *people* tab, but the *top*, *photo* and *video* tabs are very personalized.

When it comes to the political opinion preference, indicated by following other accounts for supporting an opinion, our results counterintuitively showed very little **personalization differences**. However, we cannot negate the Filter Bubble hypothesis for these cases because, as discussed previously, a few differences in the top-ranking results may cause a huge impact for the user on the meaning of the results (HAIM *et al.*, 2018).

Besides that, we contributed with a new metric to measure personalization on a web search. Through the semantic similarity metric, we can compare not only the document identifiers, or part of it (e.g., URL, domain name), but the content itself. We argue that other metrics cannot consider when two documents have different IDs but similar content. Another interesting aspect of the

semantic similarity is the capability of reading sentences from 16 languages. Most of the empirical studies are based on English content and a few in German. Our study is the first one of this kind to analyze Portuguese content.

## 6.2 Limitations

We recognize some limitations in our empirical work. **First**, our experiment was applied in a limited context regarding the topic of polarization - *the Brazilian Welfare Reform*. We may test our experiment on other topics, political and non-political, for future work.

**Second**, we limited our set of results in ten due to the convenience of capturing the data. However, it limited the variability of the *Jaccard index* and *Edit distance* metrics. On the other hand, it showed that the semantic similarity metric variability was not affected. One could say that this size of the result set is not sufficient to characterize the personalization differences. We argue that a minimum swap on the first items of the result set could severely impact the meaning of the search results to the user.

**Third**, regarding the noise treatment of our measurement methodology, we did not account for A/B test possibility and neither for a “carry-over" effect as previous work did (HANNÁK *et al.*, 2013; KLIMAN-SILVER *et al.*, 2015; LE *et al.*, 2019). Although none of these works studied Twitter Search, we cannot ensure the occurrence of these events that do not account for personalization.

**Forth**, we hold only five minutes between following the profiles (agents training) and executing the queries. It may have caused the “littleness" of differences for ANTI/PRO agents. Instead of immediately do the queries after following new accounts, we may make our agents hold some time before querying (*hours, days, or weeks?*) so that Twitter properly triggers a more

unbalanced personalization between the polarized agents.

## 6.3 Future Work

Finally, we want to brainstorm several ideas for future research.

**First**, we want to test more factors that can trigger personalization on Twitter Search, like the query history, browser history, location, and tweeting. Twitter's documentation and account settings gives clues that these factors are used to personalize user's contents. There is also a more recent feature flag to disable all the personalization features. We could use it to potentially improve the accuracy of our neutral agents.

**Second,** we want to use our methodology to investigate other social media search platforms, e.g., YouTube, LinkedIn and Instagram. We would need to adapt our tool regarding the singularity of such platforms.

**Third,** as a benefit of using the MUSE to calculate our semantic similarity metric, there is a space to apply our methodology in other languages or countries. We can make new experiments on context from other countries to check if the personalization algorithms behave differently from the Brazilian context.

**Forth,** also as a benefit of the MUSE, we can use such a model to identify political bias in the search results. As mentioned in the session 5.3, the accounts that were fetched for training our agents could be useful as a ground-truth for political-leaning classification.

# Bibliography


ALEXANDER, T., "Personalization: The state of the art and future directions". **Business Computing**, Handbooks in Information Systems. Emerald, 2007. v. 3. p. 3–40.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval: the concepts and technology behind search**. Second edition ed. New York, Addison Wesley, 2011.

BELKIN, N. J. "Interaction with Texts: Information Retrieval as Information-Seeking Behavior". 1993. 1993.

BENKLER, Y. **The wealth of networks: How social production transforms markets and freedom**. Yale University Press, 2006.

BOUHINI, C.; GÉRY, M.; LARGERON, C. "Personalized information retrieval models integrating the user's profile". 2016. 2016. p. 1–9.

BOZDAG, E. "Bias in algorithmic filtering and personalization", **Ethics and Information Technology**, v. 15, n. 3, p. 209–227, set. 2013. DOI: 10.1007/s10676-013-9321-6.

BOZDAG, E.; HOVEN, J. "Breaking the Filter Bubble: Democracy and Design", **Ethics and Inf. Technol.**, v. 17, n. 4, p. 249–265, dez. 2015. DOI: 10.1007/s10676-015-9380-y.

CAI, Y.; LI, Q. "Personalized Search by Tag-Based User Profile and Resource Profile in Collaborative Tagging Systems". 2010. event-place: Toronto, ON, Canada. New York, NY, USA, Association for Computing Machinery, 2010. p. 969–978. DOI: 10.1145/1871437.1871561.

COURTOIS, C.; SLECHTEN, L.; COENEN, L. "Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study", **Telematics and Informatics**, v. 35, n. 7, p. 2006–2015, 2018. DOI: https://doi.org/10.1016/j.tele.2018.07.004.

DAMERAU, F. J. "A technique for computer detection and correction of spelling errors", **Communications of the ACM**, v. 7, n. 3, p. 171–176, 1 mar. 1964. DOI: 10.1145/363958.363994.

DAOUD, M.; TAMINE, L.; BOUGHANEM, M. "A personalized search using a semantic distance measure in a graph-based ranking model", **Journal of Information Science**, v. 37, n. 6, p. 614–636, dez. 2011. DOI: 10.1177/0165551511420220.

DILLAHUNT, T. R.; BROOKS, C. A.; GULATI, S. "Detecting and visualizing filter bubbles in Google and Bing". 18, 2015. 2015. p. 1851–1856. DOI: 10.1145/2702613.2732850.

DING, Y.; JACOB, E. K.; CAVERLEE, J.; *et al.* "Profiling Social Networks: A Social Tagging Perspective", **D-Lib Magazine**, v. Volume 15, n. 3/4, abr. 2009.

ESTADÃO, C. **Reforma da Previdência cria "guerra" de hashtags no Twitter**. **Época Negócios**. 2019. Disponível em: https://epocanegocios.globo.com/Brasil/noticia/2019/03/reforma-da-previdencia-cria-guerra-de-hashtags-no-twitter.html. Acesso em: 23 maio 2020.

FAKHFAKH, R.; FEKI, G.; BEN AMMAR, A.; *et al.* "Personalizing information retrieval: A new model for user preferences elicitation". In: **2016 IEEE International Conference on Systems, Man and Cybernetics (SMC)**, out. 2016. Budapest, IEEE, out. 2016. p. 002091–002096. DOI: 10.1109/SMC.2016.7844548.

FLAXMAN, S.; GOEL, S.; RAO, J. M. "Filter bubbles, echo chambers, and online news consumption", **Public Opinion Quarterly**, v. 80, n. Specialissue1, p. 298–320, 2016. DOI: 10.1093/poq/nfw006.

GARIMELLA, K.; DE FRANCISCI MORALES, G.; GIONIS, A.; *et al.* "Quantifying Controversy in Social Media". 2016. event-place: San Francisco, California, USA. New York, NY, USA, ACM, 2016. p. 33–42. DOI: 10.1145/2835776.2835792.

GARIMELLA, K.; MORALES, G. D. F.; GIONIS, A.; *et al.* "Quantifying Controversy on Social Media", **Trans. Soc. Comput.**, v. 1, n. 1, p. 3:1–3:27, jan. 2018. DOI: 10.1145/3140565.

GOOGLE. **Google Introduces Personalized Search Services; Site Enhancements Emphasize Efficiency**. **Google News from Google**. 2004. Disponível em: http://googlepress.blogspot.com/2004/03/google-introduces-personalized-search.html. Acesso em: 27 jun. 2020.

GOOGLE. **Personalized Search for everyone**. **Google News from Google**. 2009. Disponível em: https://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html. Acesso em: 27 jun. 2020.

GOOGLE. **Personalized Search Graduates from Google Labs**. **Google News from Google**. 2005. Disponível em: http://googlepress.blogspot.com/2005/11/personalized-search-graduates-from_10.html. Acesso em: 27 jun. 2020.

GOOGLE. **Search, plus Your World**. **Google News from Google**. 2012. Disponível em: https://googleblog.blogspot.com/2012/01/search-plus-your-world.html. Acesso em: 27 jun. 2020.

GUERRA, P. H.; MEIRA JR, W.; CARDIE, C.; *et al.* "A measure of polarization on social media networks based on community boundaries", **Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013**, p. 215–224, 2013.

HAIM, M.; ARENDT, F.; SCHERR, S. "Abyss or Shelter? On the Relevance of Web Search

Engines' Search Results When People Google for Suicide", **Health Communication**, v. 32, n. 2, p. 253–258, 2017. DOI: 10.1080/10410236.2015.1113484.

HAIM, M.; GRAEFE, A.; BROSIUS, H.-B. "Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News", **Digital Journalism**, v. 6, n. 3, p. 330–343, 2018. DOI: 10.1080/21670811.2017.1338145.

HAMMING, R. W. **Coding and information theory. 1986**. Englewood Cliffs, NJ, Prentice Hall, 1986.

HANNÁK, A.; SAPIEŻYŃSKI, P.; KHAKI, A. M.; *et al.* "Measuring Personalization of Web Search", **arXiv:1706.05011 [cs]**, arXiv: 1706.05011, 15 jun. 2017.

HANNÁK, A.; SAPIEZYNSKI, P.; MOLAVI KAKHKI, A.; *et al.* "Measuring Personalization of Web Search". In: **Proceedings of the 22Nd International Conference on World Wide Web**, 2013. ACM, 2013. p. 527–538. DOI: 10.1145/2488388.2488435.

HOSANAGAR, K.; FLEDER, D.; LEE, D.; *et al.* "Will the global village fracture into tribes recommender systems and their effects on consumer fragmentation", **Management Science**, v. 60, n. 4, p. 805–823, 2014. DOI: 10.1287/mnsc.2013.1808.

HU, D.; JIANG, S.; E. ROBERTSON, R.; *et al.* "Auditing the Partisanship of Google Search Snippets". In: **The World Wide Web Conference**, 2019. San Francisco, CA, USA, ACM Press, 2019. p. 693–704. DOI: 10.1145/3308558.3313654.

JACCARD, P. "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines", **Bulletin de la Société Vaudoise des Sciences Naturelles**, v. 37, p. 241 – 272, 1901.

KLIMAN-SILVER, C.; HANNAK, A.; LAZER, D.; *et al.* "Location, Location, Location: The Impact of Geolocation on Web Search Personalization". In: **the 2015 ACM Conference**, 2015. Tokyo, Japan, ACM Press, 2015. p. 121–127. DOI: 10.1145/2815675.2815714.

KRUIKEMEIER, S. "How political candidates use Twitter and the impact on votes", **Computers in Human Behavior**, v. 34, p. 131–139, maio 2014. DOI: 10.1016/j.chb.2014.01.025.

LASSEN, D. "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment", **American Journal of Political Science**, v. 49, 1 fev. 2004. DOI: 10.2139/ssrn.475821.

LE, H.; MARAGH, R.; EKDALE, B.; *et al.* "Measuring Political Personalization of Google News Search". In: **The World Wide Web Conference**, 2019. ACM, 2019. p. 2957–2963. DOI: 10.1145/3308558.3313682.

LIU, J.; LIU, C.; BELKIN, N. J. "Personalization in text information retrieval: A survey", **Journal of the Association for Information Science and Technology**, v. 71, n. 3, p. 349–369, mar. 2020. DOI: 10.1002/asi.24234.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. New York, Cambridge University Press, 2008.

MESSING, S.; WESTWOOD, S. "Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online", **Communication Research**, 18 nov. 2012. DOI: 10.1177/0093650212466406.

MOHAMED, S.; ABDELMOTY, A. I. "Spatio-semantic user profiles in location-based social networks", **International Journal of Data Science and Analytics**, v. 4, n. 2, p. 127–142, set. 2017. DOI: 10.1007/s41060-017-0059-9.

MONTGOMERY, A. L.; SMITH, M. D. "Prospects for Personalization on the Internet", **Journal of Interactive Marketing**, v. 23, n. 2, p. 130–137, maio 2009. DOI: 10.1016/j.intmar.2009.02.001.

MORALES, A. J.; BORONDO, J.; LOSADA, J. C.; *et al.* "Measuring political polarization: Twitter shows the two sides of Venezuela", **Chaos: An Interdisciplinary Journal of Nonlinear Science**, v. 25, n. 3, p. 033114, mar. 2015. DOI: 10.1063/1.4913758.

PALTOGLOU, G.; THELWALL, M. "A Study of Information Retrieval Weighting Schemes for Sentiment Analysis". 2010. event-place: Uppsala, Sweden. USA, Association for Computational Linguistics, 2010. p. 1386–1395.

PAN, B.; HEMBROOKE, H.; JOACHIMS, T.; *et al.* "In Google We Trust: Users' Decisions on Rank, Position, and Relevance", **Journal of Computer-Mediated Communication**, v. 12, n. 3, p. 801–823, abr. 2007. DOI: 10.1111/j.1083-6101.2007.00351.x.

PARISER, E. **The filter bubble: what the Internet is hiding from you**. London, Viking, 2011.

PUSCHMANN, C. "Beyond the Bubble: Assessing the Diversity of Political Search Results", **Digital Journalism**, v. 7, n. 6, p. 824–843, 2019. DOI: 10.1080/21670811.2018.1539626.

RAFA, T.; KECHID, S., "A Semantic-Based Personalized Information Retrieval Approach Using a Geo-Social User Profile". In: ELLOUMI, M., GRANITZER, M., HAMEURLAIN, A., *et al.* (Org.), **Database and Expert Systems Applications**, Communications in Computer and Information Science. Cham, Springer International Publishing, 2018. v. 903. p. 301–313. DOI: 10.1007/978-3-319-99133-7_25. Disponível em: http://link.springer.com/10.1007/978-3-319-99133-7_25. Acesso em: 26 jun. 2020.

ROBERTSON, R. E.; LAZER, D.; WILSON, C. "Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages". In: **the 2018 World Wide Web Conference**, 2018. Lyon, France, ACM Press, 2018. p. 955–965. DOI: 10.1145/3178876.3186143.

SALEHI, S.; DU, J. T.; ASHMAN, H. "Examining personalization in academic web search". 2015. 2015. p. 103–111. DOI: 10.1145/2700171.2791039.

SALONEN, V.; KARJALUOTO, H. "Web personalization: The state of the art and future avenues for research and practice", **Telematics and Informatics**, v. 33, n. 4, p. 1088–1104, 2016. DOI: https://doi.org/10.1016/j.tele.2016.03.004.

SANDERSON, M. "Ambiguous Queries: Test Collections Need More Sense". 2008. New York,

NY, USA, Association for Computing Machinery, 2008. p. 499–506. DOI: 10.1145/1390334.1390420.

SANTOS, J. C.; SIQUEIRA, S. W. M.; NUNES, B. P.; *et al.* "Is There Personalization in Twitter Search? A Study on polarized opinions about the Brazilian Welfare Reform". In: **WebSci '20: 12th ACM Conference on Web Science**, 6 jul. 2020. Southampton United Kingdom, ACM, 6 jul. 2020. p. 267–276. DOI: 10.1145/3394231.3397917.

STROUD, N. J. "Polarization and Partisan Selective Exposure", **Journal of Communication**, v. 60, n. 3, p. 556–576, 19 ago. 2010. DOI: 10.1111/j.1460-2466.2010.01497.x.

SUNSTEIN, C. R. **Republic.com 2.0**. Princeton, NJ, Princeton University Press, 2009.

TAYLOR, R. S. "Question negotiation and information seeking in libraries", **College and Research Libraries**, v. 29, n. 3, p. 178–194, 1968.

TRAN, G. B.; HERDER, E. "Detecting Filter Bubbles in Ongoing News Stories". In: **23rd Conference on User Modelling, Adaption and Personalization**, 2015. 2015.

VALLET, D.; CANTADOR, I.; JOSE, J. M. "Personalizing Web Search with Folksonomy-Based User and Document Profiles". 2010. Berlin, Heidelberg, Springer Berlin Heidelberg, 2010. p. 420–431.

XU, S.; BAO, S.; FEI, B.; *et al.* "Exploring Folksonomy for Personalized Search". 2008. event-place: Singapore, Singapore. New York, NY, USA, Association for Computing Machinery, 2008. p. 155–162. DOI: 10.1145/1390334.1390363.

YANG, Y.; CER, D.; AHMAD, A.; *et al.* **Multilingual Universal Sentence Encoder for Semantic Retrieval**. 2019. Disponível em: https://arxiv.org/abs/1907.04307.

YANG, Y.; YUAN, S.; CER, D.; *et al.* "Learning Semantic Textual Similarity from Conversations". In: **Proceedings of The Third Workshop on Representation Learning for NLP**, 2018. Melbourne, Australia, Association for Computational Linguistics, 2018. p. 164–174. DOI: 10.18653/v1/W18-3022.

YEUNG, C. M. A.; GIBBINS, N.; SHADBOLT, N. "A Study of User Profile Generation from Folksonomies". mar. 2008. mar. 2008.