Is There Personalization in Twitter Search? A Study on polarized opinions about the Brazilian Welfare Reform

Jônatas C. dos Santos
jonatas.santos@uniriotec.br
UNIRIO
Rio de Janeiro, RJ, Brazil

Sean W. M. Siqueira
sean@uniriotec.br
UNIRIO
Rio de Janeiro, RJ, Brazil

Bernardo Pereira Nunes
bernardo.nunes@anu.edu.au
Australian National University
Canberra, ACT, Australia

Fabrício R. S. Pereira
fabriciorsf@uniriotec.br
UNIRIO
Rio de Janeiro, RJ, Brazil

Pedro P. Balestrassi
pedro@unifei.edu.br
Federal University of Itajubá
Itajubá, MG, Brazil

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. Peculiarly, little evidence for differences between two profiles that followed different accounts with polarized viewpoints about the same topic was found – the filter bubble hypothesis cannot be null.

CCS CONCEPTS

• **Information systems → Personalization**; *Web search engines*; *Social networks*; *Social networking sites*; Similarity measures.

KEYWORDS

Personalization, Twitter Search, Social Media Search

ACM Reference Format:
Jônatas C. dos Santos, Sean W. M. Siqueira, Bernardo Pereira Nunes, Fabrício R. S. Pereira, and Pedro P. Balestrassi. 2020. Is There Personalization in Twitter Search? A Study on polarized opinions about the Brazilian Welfare Reform. In *12th ACM Conference on Web Science (WebSci '20), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394231.3397917

WebSci '20, July 6–10, 2020, Southampton, United Kingdom
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7989-2/20/07…$15.00
https://doi.org/10.1145/3394231.3397917

1 INTRODUCTION

In order to improve user experiences, personalization algorithms end up creating an invisible barrier that blocks users from confronting topics – the well-known filter bubble phenomenon. According to Pariser [13], search engines and social media provide users with non-confronting information to increase their on-time within their platforms. The objective is very clear: the longer users remain on their platforms, the more value it has for an advertiser and, therefore, the more revenue it is likely to generate.

This paper aims to investigate the filter bubble phenomenon in the specific context of *social media search*, a still under-investigated research topic. It is important to understand how social media results are personalized based on user profiles to understand the polarization on the Web. Then, this study focuses on the following research question: "To what extent following polarized profiles affect users' search results on social media?". To answer this question, we automatically trained fresh social media profiles with different viewpoints. Briefly, we developed software agents to follow social media profiles with different viewpoints related to a specific topic.

The main contributions of our work are fourfold:

- A semantic similarity metric to enhance prior methodologies on measuring personalization in Web search [5, 11];
- An empirical study to understand at what extent the number of followers affect the search results in social media;
- An empirical study to understand at what extent the social media search results are affected by the search filters (videos, images, profiles, and top searches); and,
- According to our case study, we argue that polarized profiles do not retrieve different search results.

For the experiments, we instantiated our agents to work with Twitter, as it is a well-known social media platform with more than 330 million monthly active users[1]. Furthermore, we decided to use a world-wide confronting topic to verify our research question: politics – more specifically – the *Brazilian Welfare Reform*[2], a long-lasting trending topic in Twitter [1, 12]. The agents were divided into *PRO* and *ANTI*, representing a reform supporter and non-supporter, respectively, and strictly followed profiles with the

[1] According to https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[2] Also known as the *Pension Reform in Brazil* - https://en.wikipedia.org/wiki/Pension_reform_in_Brazil

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.

Our paper is organized as follows: Section 2 reviews related work. Section 3 presents the methodology that we used for preparing and executing our empirical experiment. Section 4 and Section 5 presents and discusses our results. Finally, Section 6 concludes the paper and presents future works.

2 RELATED WORK

Hannak et al. [5] 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. They reported that Google and Bing search engines approximately personalize 11.7% and 15.8% of their results, respectively (also reported in [6]).

Further work from Kliman-Silver et al. [9] showed that Google personalizes the search based on the user location, especially for queries related to local businesses. Likewise, Salehi et al. [16] proposed a methodology to quantify personalization in the academic context. They observed slightly personalization differences between personalized (Google Search) and non-personalized (DuckDuckGo) search engines for academic topics. Other works investigated whether queries for suicidal topics could be influenced by the search results or not [3], and the political personalization in the Google News search [11].

A newer set of studies audits other aspects of the search result pages, for instance, Robertson et al. [15] 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. [7] analyses 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; and, (2) all queries are concurrently issued to avoid temporal sensitive issues in the results retrieved. Besides that, we conduct an in-depth experiment and analysis to (3) understand to what extent the number of followers affects the search results in Twitter as well as (4) to understand how search filters in Twitter personalize search results.

3 METHODOLOGY

The methodology for measuring search personalization is inspired by the work from Hannak et al. [5], Le et al. [11].

3.1 Data Harvesting

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 [1]. During this period, this political concern was the central topic of many media streams, social networks and street protests. Therefore, we harvested daily Brazilian trending topics from *a day before* (March 21th) to *a day after* (March 24th) the apex of the discussion. As Twitter does not provide free historical data, we scrapped the trending topics from *trendogate.com*.

3.2 Query planning

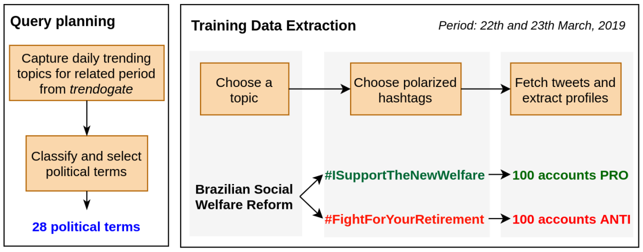

Query planning
Capture daily trending topics for related period from *trendogate*
Classify and select political terms
28 political terms
Training Data Extraction
Period: 22th and 23th March, 2019
Choose a topic
Choose polarized hashtags
Fetch tweets and extract profiles
Brazilian Social Welfare Reform
#ISupportTheNewWelfare
100 accounts PRO
#FightForYourRetirement
100 accounts ANTI

Figure 1: Training data extraction and query planning

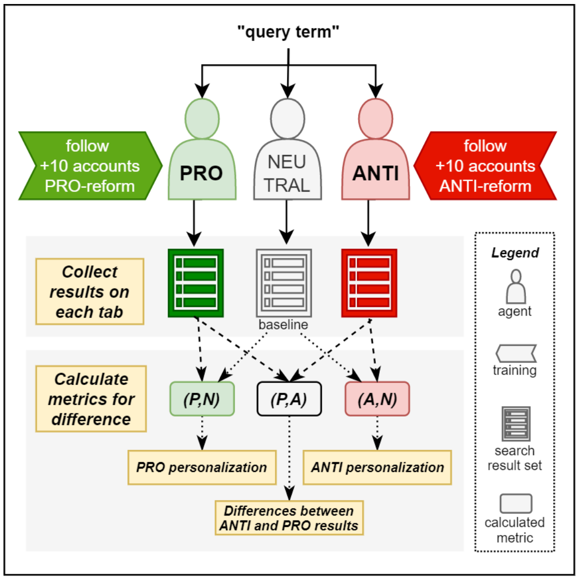

"query term"
follow +10 accounts PRO-reform
PRO
NEU TRAL
ANTI
follow +10 accounts ANTI-reform
Collect results on each tab
baseline
Calculate metrics for difference
(P,N)
(P,A)
(A,N)
PRO personalization
ANTI personalization
Differences between ANTI and PRO results
Legend
agent
training
search result set
calculated metric

Figure 2: Query execution instance

Our goal in planning the queries is to understand whether certain types of terms could influence users when querying. Therefore, we manually classified 200 trending topics into general categories from IAB Categories[3], including politics. Considering only the political

[3] 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 1: Political-related query terms

Original Pt. term	Translated En. term	Class 1	Class 2
Articulacao	Articulation	Political Issues	Informative
#OuReformaOuQuebra	#OrReformOrBreak	Political Issues	Opinion
Nova Previdencia	New Welfare	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	Car-Wash	Political Issues	Informative
PMDB	PMDB	Politician	Informative
Marun	Marun	Politician	Informative
Moreira Franco	Moreira Franco	Politician	Informative
Pezao	Pezao	Politician	Informative
Bretas	Bretas	Politician	Informative
Coronel Lima	Colonel Lima	Politician	Informative
Aecio	Aecio	Politician	Informative
Eduardo Cunha	Eduardo Cunha	Politician	Informative
Freixo e Paulo Teixeira	Freixo and Paulo Teixeira	Politician	Informative
Sarney	Sarney	Politician	Informative
Temer	Temer	Politician	Informative
Pinochet	Pinochet	Politician	Informative
Dilma Rousseff	Dilma Rousseff	Politician	Informative
#LulaLivreDomingoSDV	#FreeLulaOnSunday	Politician	Opinion
Michelzinho	Michelzinho	Humor and Satire	Opinion
Falta a Dilma	Missing Dilma	Humor and Satire	Opinion
Vampirao	Big Vampire	Humor and Satire	Opinion
Ate a Damares	Even Damares	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.

Table 2: 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

context, we found out 28 topics and then performed a classification into *politician*, *political issues*, and *humor and satire* (Class 1) (Table 1).

To obtain more quality on these classifications, we randomized the set of trending topics and asked two research colleagues to also 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 1). Thus, we were able to verify personalization based on these categories.

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[4] that provides tags for time filtering (`since` and `until`[5]) 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 2.

3.3 Training data extraction

In contemplation of extracting data from queries executed over Twitter Search, we developed a tool[6] based on JavaScript Puppetter library[7] to train our Twitter accounts on an automated browser. We call the instances of this tool: *agents*. Each agent logs in a Twitter account, follow a set of profiles, and execute a sequence of queries on Twitter Search. For each session, 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 control user intended to measure the differences from the previous ones.

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 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. We summarize this data in Table 3.

3.4 Agents setup and noise treatment

We treated two main sources of noise in our agent running environment. First, we handled the timing noise. The agents are Chrome-based browsers that connect to a manager server that triggers the agent’s actions simultaneously. This is very important for our study 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[8].

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

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

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

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

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

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

```
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
```

Algorithm 1: An agent session

Another possible source of noise could be the location. A previous study on other search platforms reported high personalization in the function of location [9]. Yet, Twitter gives clues that it personalizes its content based on geolocation[9]. However, it is not clear whether Twitter applies this personalization to the search results. Thus, our agents run on the same machine, so that it was not influenced by possible geolocation differences. The machine was located in the city of Rio de Janeiro, Brazil.

We also wanted to see if the amount of personalization would vary in function of the number of followings. Therefore, we run 10 sessions of queries, and we incremented 10 following on each session. So, in the first session, we had each account following 10 profiles, while in the last session, we had each account following 100 profiles. This 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 10 tweets per tab, resulting in the total amount of 168,000 tweets. Each session took from 2-3 hours to complete. We summarize the agent actions at the **Algorithm 1** and a query execution instance in **Figure 2**.

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[10].

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 [6, 9, 11, 14, 16], the *Jaccard Index* [8], and the *Damerau–Levenshtein distance* [2] or simply *edit distance*. Then, we introduce the use of a new metric that is capable of quantifying semantic differences.

The ***Jaccard Index*** could be defined 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. This metric looks for the presence or absence of the elements, but does not account for their order.

The ***edit distance*** computes the number of insertions, deletions, substitutions or swaps to make different lists equal. Therefore, it can look into the differences at 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.

While these metrics 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

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

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

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* model called *Multilingual Universal Sentence Encoder for Semantic Retrieval* [17] (MUSE). 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 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. 1. This similarity metric converts the traditional cosine similarity scores into angular distances that obey the triangle inequality as suggests Yang [18].

$$s(u, v) = 1 - arccos\left(\frac{uv}{\|u\| \, \|v\|}\right) \tag{1}$$

We first calculate the semantic similarity per pair of tweets within the two set of results that have the same length. Then, we calculate the average similarity, so we can characterize the differences between the two sets. (Eq. 2). Let A and B be two sets of results with the same length n, where A_i and B_i correspond to elements of the set:

$$S(A, B) = \frac{\sum_{j=1}^{n} s(A_i, B_i)}{n} \tag{2}$$

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

Before starting with our evaluation, let us denote our metrics from the previous section. 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 4, and summarize headers of our dataset in Table 5 with sample data.

4.1 Comparing the metrics

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 3 shows a tree diagram that displays the groups formed by 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 (≈ 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.

Table 4: 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)*.

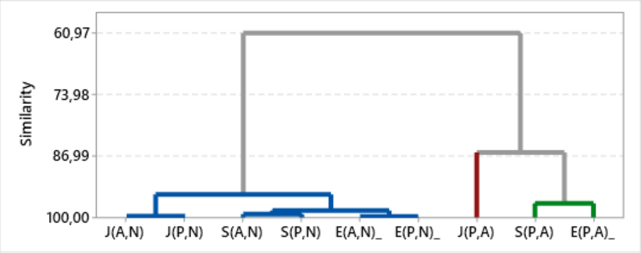


Figure 3: Dendrogram for all the metrics. Complete linkage; Correlation Coefficient Distance

Second, concerning the (P, A), the *semantic similarity* is strongly correlated to the *edit distance* (≈ 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* (≈ 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).

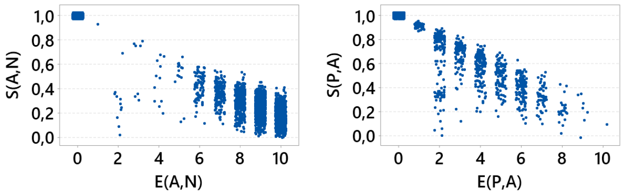


Figure 4: Scatter plot between $S(A, N) \times E(A, N)$ and $S(P, A) \times E(P, A)$

4.1.1 The semantic similarity . The Semantic Similarity (S) metric can point to differences that the other metrics cannot. The key point is: the S does not present differences from inside the content rather than simply compare the references (e.g. tweet *URL*). We can detach these differences when we look to the scatter plot between S and E (Figure 4).

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

Table 5: 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. $N = 4527$

semantically similar. For instance, when $E(A, N) = 10$, we see that the $0 < S(A, N) \lessgtr 0.4$, but even when $E(P, A) = 2$ (e.g., *just two swaps probably*), the $S(P, A)$ can still reach very low values.

4.2 Comparing personalization per tabs

Twitter Search interface presents five tabs with a set of results that we capture (Figure 6). According to Twitter Search FAQ page[11], the *top tab* shows "Tweets you are likely to care about most first", and it says the content is selected by an algorithm. However, it does not say much about the other tabs.

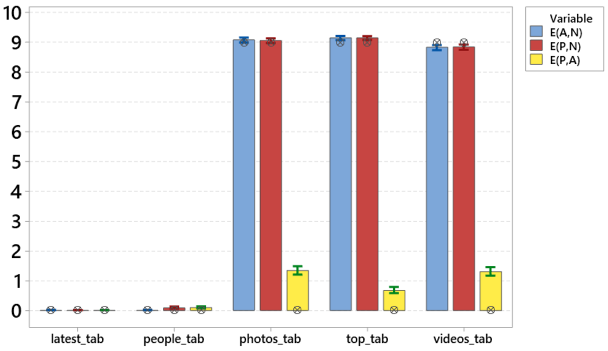


Figure 5: Edit distance per tabs

Thus, we start our analysis checking if there are differences in personalization between the Twitter Search tabs. Figure 5 shows the bar plot of means of the edit distance for each tab, where the $\bigotimes$ symbol indicates the median. We want to verify two hypotheses over this plot. First, we suspect that *latest* and *people* tabs are never personalized (i.e., $E = 0$). Second, we question if the *ANTI* and *PRO* agents present the same amount of personalization (i.e., $E(A, N) - E(P, N) = 0$ $J(A, N) - J(P, N) = 0$ $S(A, N) - S(P, N) = 0$).

For the first hypothesis, we use the Wilcoxon signed rank test to check for the medians (Table 6). We cannot reject the null hypothesis that indicates for non-personalization in all metrics on the *latest tab*. However, we cannot say the same for the *people* tab.

Table 6: 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	>

Table 7: 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%

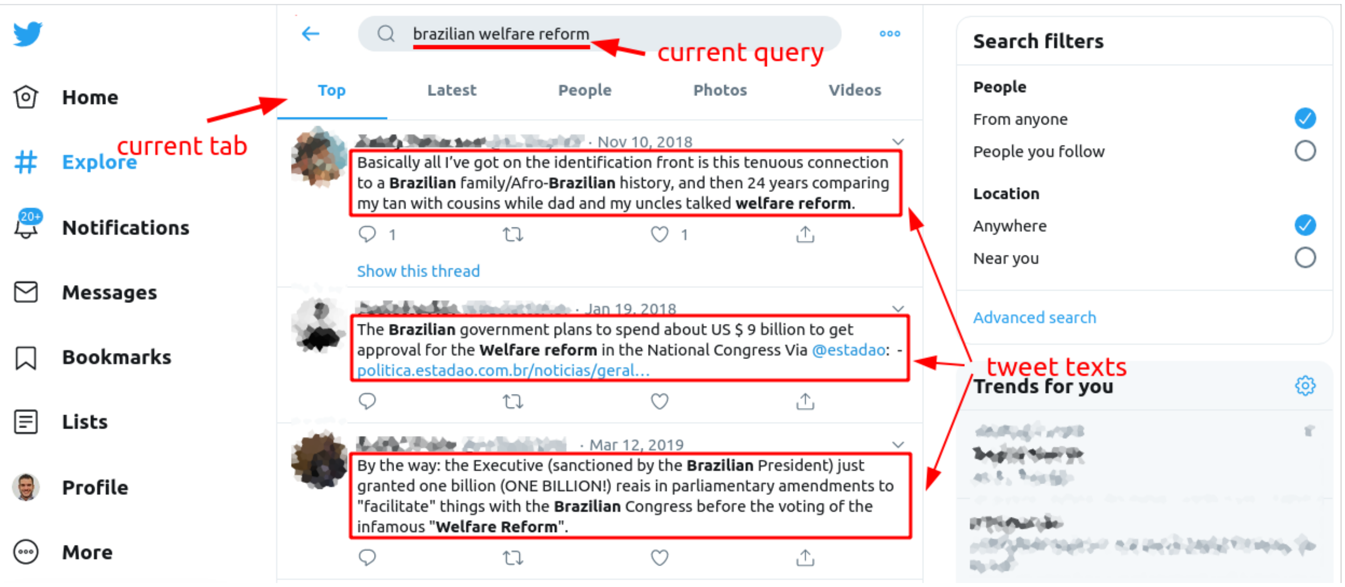
For the second hypothesis, we want to verify whether the pair (A, N) and (P, N) for each metric is equal by applying the Mann-Whitney test to the distributions (Table 7). With 95% of confidence level, we cannot reject the null hypothesis that the differences between the distributions are equal. From 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.

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

On each session of execution in our experiment, we make our agents follow 10 more accounts. Thus, we start with 10 followings at `r_10_p`, but end with 100 followings at `r_100_p`. We would

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



brazilian welfare reform

[redacted] · Nov 10, 2018

Basically all I've got on the identification front is this tenuous connection to a **Brazilian** family/Afro-**Brazilian** history, and then 24 years comparing my tan with cousins while dad and my uncles talked **welfare reform**.

1 1

Show this thread

[redacted] · Jan 19, 2018

The **Brazilian** government plans to spend about US $ 9 billion to get approval for the **Welfare reform** in the National Congress Via @estadao: - politica.estadao.com.br/noticias/geral…

[redacted] · Mar 12, 2019

By the way: the Executive (sanctioned by the **Brazilian** President) just granted one billion (ONE BILLION!) reais in parliamentary amendments to "facilitate" things with the **Brazilian** Congress before the voting of the infamous "**Welfare Reform**".

Search filters

People
From anyone ●
People you follow ○

Location
Anywhere ●
Near you ○

Advanced search

Trends for you

[redacted]

Figure 6: 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.

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 7), we see very low differences between the sessions, and we see more a trend for decreasing in the differences as the number of followings increases.

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, *30 followings* session (`r_30_p`) results are more personalized than the other groups, while *50* and *100 followings* session are less personalized.

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 2).

We wanted to verify whether the level of personalization would increase when the user queried before, during or after the apex of the discussion on Twitter. Figure 8 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.

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 function of these classes. We show the analysis of means in Figure 8.b and Figure 8.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 9), 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.

5 DISCUSSION

In this section, we want to run over some important topics that we cover in this paper.

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[12]. After a series of issues on affecting politics [10], 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.

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

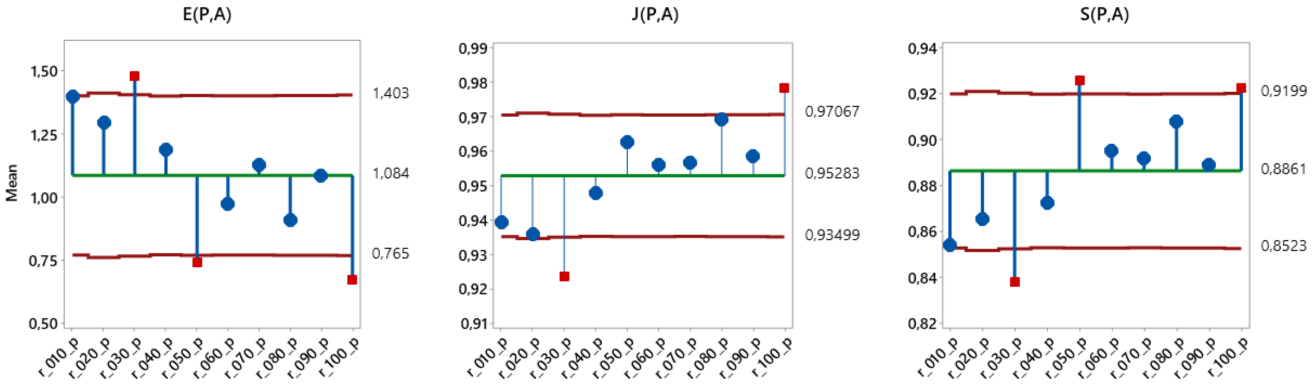


Figure 7: Results differences for $E(P, A)$, $J(P, A)$, and $S(P, A)$ per session

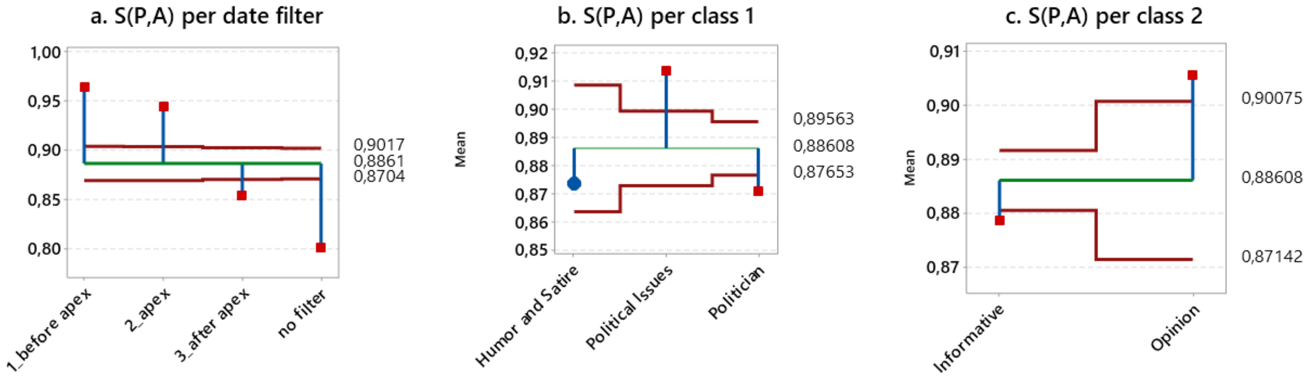


Figure 8: Results for $S(P, A)$ per date filter, Class 1 and Class 2

- translated.png

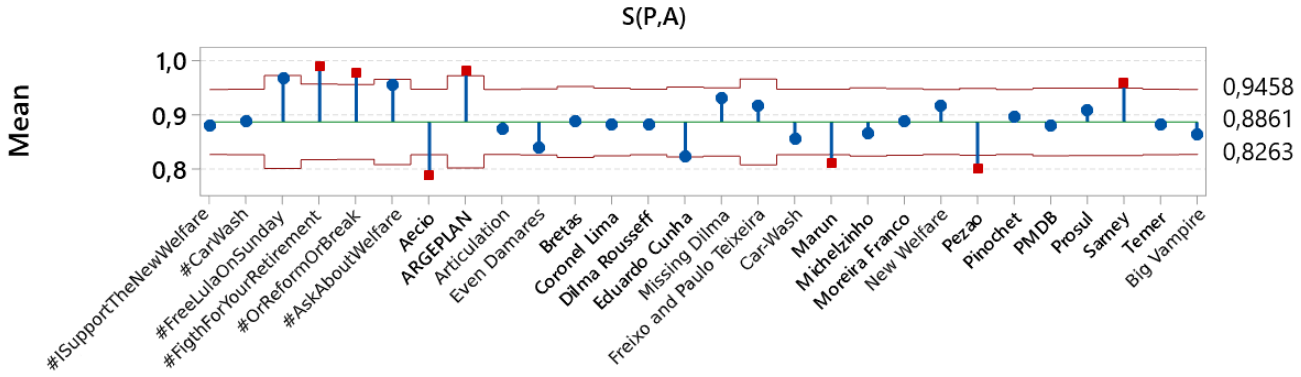


Figure 9: Results for $S(P, A)$ per query term

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 want to discuss if we can answer our main question: *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 compares 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).

We ask if this "littleness" of differences is due to the way we executed the data collection. 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 trigger a more unbalanced personalization between the polarized agents. We want to investigate these possibilities in future research.

Even 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 the just first result, this change may cause a high impact for the user. It is something that we want to investigate on 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 10 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.

Another interesting finding is the confirmation that the *latest* tab is not personalized. Even this might be obvious for someone, there is still scope for thinking in a kind of personalization for recent items.

The last observation is about the *people* tab. Although our data do not let to conclude for 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.

5.2 Semantic similarity metric

One of the main contributions of our paper is the introduction of the semantic similarity metric. Past studies on measuring personalization [5, 9, 11, 16] were not able to compare differences based on the content. Generally, they rely on the primary key (*or part of the PK*) of the elements to compare for the differences between the search results.

By the way, 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.

Besides the capability of compare sentences semantically, the semantic similarity metric allows making more detailed and granular comparisons.

Another interesting aspect for 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.

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. It sheds a light that we could use these accounts as a ground truth for classifying political leaning on future work.

6 CONCLUSION

Our experiment showed significant personalization on Twitter Search when a user follows just a few accounts. Besides, our results printed 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 showed very little personalization differences. Hence, we cannot negate the Filter Bubble hypothesis, because a few differences on the top-ranking results may cause a huge impact for the user on the meaning of the results [4].

We recognize several limitations in our empirical work. **First**, our experiment was applied in a limited context - *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 10 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 [5, 9, 11]. Although none of these works studied Twitter Search, we cannot ensure the occurrence of these events that do not account for personalization.

Finally, we summarize additional ideas for future research: (i) testing other factors that would trigger personalization on Twitter Search; (ii) using our methodology to investigate others social media search platforms, e.g., YouTube, LinkedIn, and Instagram; (iii) applying our methodology in other languages or countries; and, (iv) using the Multilingual Universal Sentence Encoder to identify political bias in the search results.

ACKNOWLEDGMENTS

This study was financed in part by the National Council for Scientific and Technical Development (CNPq) - Brazil, Grant 315374/2018-7.

REFERENCES

[1] Estadão Conteúdo. 2019. The Welfare Reform creates hashtags war at Twitter. https://epocanegocios.globo.com/Brasil/noticia/2019/03/reforma-da-previdencia-cria-guerra-de-hashtags-no-twitter.html

[2] Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* 7, 3 (March 1964), 171–176. https://doi.org/10.1145/363958.363994

[3] M. Haim, F. Arendt, and S. Scherr. 2017. Abyss or Shelter? On the Relevance of Web Search Engines' Search Results When People Google for Suicide. *Health Communication* 32, 2 (2017), 253–258. https://doi.org/10.1080/10410236.2015.1113484 cited By 14.

[4] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the Filter Bubble? *Digital Journalism* 6, 3 (2018), 330–343. https://doi.org/10.1080/21670811.2017.1338145 arXiv:https://doi.org/10.1080/21670811.2017.1338145

[5] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 527–538. https://doi.org/10.1145/2488388.2488435

[6] Anikó Hannák, Piotr Sapieżyński, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. 2017. Measuring Personalization of Web Search. arXiv:cs.CY/1706.05011

[7] Desheng Hu, Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Auditing the Partisanship of Google Search Snippets. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 693–704. https://doi.org/10.1145/3308558.3313654

[8] Paul Jaccard. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (1901), 241 – 272.

[9] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization, In Proceedings of the 2015 Internet Measurement Conference (Tokyo, Japan). *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC* 2015-October, 121–127. https://doi.org/10.1145/2815675.2815714 cited By 23.

[10] Sanne Kruikemeier. 2014. How political candidates use Twitter and the impact on votes. *Computers in Human Behavior* 34 (05 2014), 131–139. https://doi.org/10.1016/j.chb.2014.01.025

[11] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. 2019. Measuring Political Personalization of Google News Search. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2957–2963. https://doi.org/10.1145/3308558.3313682

[12] Sharon Meraz. 2017. Hashtag Wars and Networked Framing. In *Between the Public and Private in Mobile Communication*, Ana Serrano Tellería (Ed.). Routledge, New York, Chapter 16, 303–323. https://doi.org/10.4324/9781315399300-17

[13] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Group , The, UK.

[14] C. Puschmann. 2018. Beyond the Bubble: Assessing the Diversity of Political Search Results. *Digital Journalism* 7, 6 (nov 2018), 824–843. https://doi.org/10.1080/21670811.2018.1539626 cited By 2.

[15] Ronald E. Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 955–965. https://doi.org/10.1145/3178876.3186143

[16] Sara Salehi, Jia Tina Du, and Helen Ashman. 2015. Examining Personalization in Academic Web Search. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (Guzelyurt, Northern Cyprus) *(HT '15)*. Association for Computing Machinery, New York, NY, USA, 103–111. https://doi.org/10.1145/2700171.2791039

[17] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual Universal Sentence Encoder for Semantic Retrieval. arXiv:cs.CL/1907.04307

[18] Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning Semantic Textual Similarity from Conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Melbourne, Australia, 164–174. https://doi.org/10.18653/v1/W18-3022