ORIGINAL ARTICLE

Journal Section

How biased are social media searches? (working title)

Can Yang[1*] | Xinyuan Xu[2*] | Bernardo Pereira Nunes[3*] | Jônatas Castro dos Santos[4†] | Sean Wolfgand Matsui Siqueira[5†]

[1]Australian National University, Canberra, ACT, Australia

[2]UNIRIO, Rio de Janeiro, RJ, Brazil

Correspondence

Can Yang, Australian National University, Canberra, ACT, Australia

Email: can.yang@anu.edu.au

Present address

[†]Australian National University, Canberra, ACT 2601, Australia

TBC

KEYWORDS

Personalisation algorithms, Filter bubble, Social media, Search mechanism

1 | INTRODUCTION

Social media has become an integral part of everyday routines, enabling us to communicate ideas and engage in real-time interactions with others. With a growing number of individuals are devoting considerable time to social media platforms[1][2], there is also increasing concerns about the potential drawbacks of social media platforms. This is mainly caused by the personalisation algorithms.

Personalized algorithms derive information from a diversity of data features to provide a better experience for their users by improving the accuracy of search results. Unlike traditional search engines (e.g., Google and Bing), social media platforms incorporate social data, including users' previous activities as well as their social networks, into their personalisation algorithms.

An important feature of social media platforms is the spontaneous dissemination of the content. People tend to be more willing to accept content that agrees with their own views. Because the social media tend to retain users, the personalization algorithm is used to recommend content that is consistent with users' views which lead to the bias in social media. The bias is that users cling to their own opinions and believe that all opinions that contradict theirs are wrong. However, such personalized algorithms with bias would lose control over the content. It will negatively impact our society, limiting users' opinions and creating so-called filter bubbles, as described by Eli Pariser ([**?**]).

Bias on social media Filter bubbles are defined as a state of intellectual isolation. According to Eli Pariser, who coined the term, personalisation algorithms can create "a unique universe of information for each of us which fundamentally alters the way we encounter ideas and information" ([1]). This concept is often conflated with the term "echo chambers" - the closed systems in which people's beliefs are amplified or reinforced by constant communication and repetition, and thus are isolated from the opposition. People are able to find information in an echo chamber that reinforces their existing views without encountering opposing views, which can lead to the unexpected occurrence of confirmation bias ([2]). explaination?

bias effect Nowadays, online social media users are getting enormous and the influence of social media is even greater. The bias that emerges in social media can also have a greater impact on society (e.g. cause vaccine hesitancy/hinder the vaccination) as well as politics leading to increased polarization and extremism([3]).

bias?

[1] https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide

[2] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

To the best of our knowledge, this paper is the first of the kind to investigate:

- the bias on three social platforms: Twitter, Reddit and Tumblr;
- the comparisons of degree of bias in different social media platforms; and
- the factors that affect bias (i.e. the carry-over effect, the effect of cookies, and the effect of language on social media search results);

2 | RELATED WORKS

2.1 | Web Search

Prior study has focused on the personalisation algorithms employed in online search engines.

Hannak et al.[4] carried out pioneering work on Measuring Personalization of Web Search. They experiment on different topics to measure personalisation in the Google search engine. To reduce noise, all queries were performed at the same time and they used static DNS to direct all query traffic to a specific Google server. One of the most notable findings is the carry-over effect, Google will personalize search results based on the previous query in specific time duration. After the finding of carry-over effect, the following studies on Google have waited for 10min between each query in order to avoid its effects([5],[6],[7],[8]). The degree of personalization is quantified by the Jaccard index and Edit distance metrics. As the first investigation on the personalization bias, it led to a series of subsequent studies that also used these two metrics ([7],[8]) and two noise reduction methods([6],[7],[8]). The results show that measurable personalisation is determined by whether the user is logged in to his/her Google account and the user's geographic location.

Kliman-Silver et al.([5]) continued to investigate the impact of location in Google search personalization. It shows that Google search results will be personalized according to GPS coordinates and the location-related search terms show higher personalization

Le et al.([6]) posed a new metric as k-edit (edit distance/k) to investigate the personalisation in Google News. Finally, they conclude that the browsing history is the main feature of Google News personalisation.

Dillahunt et al.[9] is also inspired by [4].But they improve the illustration method (using Fruchterman-Reingold force-directed algorithm [10]) and introduce the metric of Kendall Tau ranking Distance (KTD) to better understand the filter bubble phenomenon in Google and Bing searches.

2.2 | Social Media Search

Only a few studies have looked at the applications of personalisation algorithms on social media platforms.

Kruikemeier [11] examined the potential impact of using Twitter in election campaigns. By tracking Twitter activity and the tweet characteristics of campaigning candidates, they revealed the role of online social platforms in campaigns. Because of the interactivity of Twitter, the use of Twitter can lead to more votes for candidates. This research brings the growing concern about social media bias.

Hargreaves et al. [12] shows the existence of the filter bubble phenomenon in the Facebook News Feed personalisation algorithm. It present three main findings: (i) Facebook filtering algorithms tend to recommend results that are consistent with the political leanings of its users; (ii) this tendency is concentrated at the top of the results; and, (iii) neutral users can also receive uneven results.

Kulshrestha et al., 2018 [13] proposed a framework for quantifying search bias that can not only capture the bias in the search results output by the search system but also identify the source of the bias (input data or ranking system). They experiment to compare the bias between Twitter and Google. The results show the bias appears in both Twitter and Google, but they are not consistent in them.

Santos et al. [14] investigated the impact of personalisation algorithms of social media search. They used three different agent namely anti, pro and neutral who were, respectively, opposed, supported or neutral to the topic. Besides using Jaccard index and Damerau–Levenshtein distance. they first use semantic similarity to quantify search personalisation [15]. They found strong evidence that there is no bias in Twitter on the selected topic. And the semantic similarity and edit distance are closely related. This paper is the base for this research project and here we presented a more in-depth analysis as well as consider a number of factors overlooked in their research (e.g., the carry-over effect, cookies, etc.). We also develop a more robust tool that is capable of verifying personalisation in multiple social media platforms.

2.3 | Classification of Search Bias

Previous studies also investigated the classification of personalization bias.

Pitoura et al. [16] focus on defining and measuring bias and consider two types of bias: (i) user bias and (ii) content bias. The first is the bias in the information received by the user according to users' attributes and the latter refers to the bias in the information delivered to the user.

Nikolov et al. [17] posed homogeneity bias and popularity bias. homogeneity bias is the tendency of platforms to generate traffic to a small number of websites disproportionately (higher number of clicks on certain specific websites), whereas popularity bias is the tendency of an application to generate traffic to an already popular website. To quantify bias, they compared five widely used web platforms and found that the higher Popularity bias will bring higher Homogeneity bias.

Kulshrestha et al., 2017 [18] divided the bias into categories: (i) input bias: whether the result filtered from the corpus is the real result of the query; (ii) ranking bias: the bias introduced by the ranking system, and (iii) the output bias: the bias in the ranking list output by the search system.

Robertson et al.[19] divided into average bias and weighted bias. The average bias does not consider the order of search results whereas the weighted bias considers the top results to be more relevant. After run some experiments, the authors reported significant differences in the biass for different search queries and components.

Our research is mainly inspired by Hannak et al. [4] and Santos et al. [14]. Hannak et al.[4] is outstanding study as they start to investigate the filter bubble effect on web search engine. It designed a series of experiment to investigate the factors of Google personalization Search. Our research adjust their experiment process to adapt for social media platforms. The carry-over effect and cookies experiment was guided from Hannak et al. [4]. Santos et al. [14] is a base of our research. They investigate the bias in Twitter search detailed through different metrics, different sessions and each result tab etc. It is worth mentioning that they introduced semantic similarity to measure bias, which is a meaningful approach and was also used in this research to quantify the bias in social platforms.While they only explored whether there was bias in Twitter, our study extends the bias research even further by drawing inspiration from other studies. Dillahunt et al.[9] introduce the Kendall Tau metric to quantify the bias which is also used in our research. The noise treatment method of Hannak et al.[4] and Kliman-Silver et al. [5] guided our research.

TABLE 1 Experiments Report - Reddit

Experiments	Queries list	Controlled noises	Number of runs	Evaluation results
Cookies	EN terms[3]	DNS, IP address, Query synchronisation	20	Almost no cookie used in Reddit search
Language	EN + CN terms	DNS, IP address, Query synchronisation	20	Reddit match search term to retrieve results
Polarization	EN terms	DNS, IP address, Query synchronisation A/B testing	20	No bias appears in Reddit
Carry-over effect	Carry-over terms	DNS, IP address, Query synchronisation	20	Carryover effect appears in Reddit since the re-rank of results

TABLE 2 Experiments Report - Tumblr

Experiments	Queries list	Controlled noises	Number of runs	Evaluation results
Cookies	EN terms	DNS, IP address, Query synchronisation	20	
Language	EN + CN terms	DNS, IP address, Query synchronisation	20	No results for Tumblr
Polarization	EN terms	DNS, IP address, Query synchronisation A/B testing	20	No bias appears in Tumblr
Carry-over effect	Carry-over terms	DNS, IP address, Query synchronisation	20	Carryover effect appears in Tumblr since the re-rank of results

TABLE 3 Experiments Report - Twitter

Experiments	Queries list	Controlled noises	Number of runs	Evaluation results
Cookies	EN terms	DNS, IP address, Query synchronisation	20	Cookie is used in Twitter personalizatio
Language	EN + CN terms	DNS, IP address, Query synchronisation	20	Twitter retrieve results based on langua
Polarization	EN terms	DNS, IP address, Query synchronisation A/B testing	20	There's bias in Twitter for anti agent
Carry-over effect	Carry-over terms	DNS, IP address, Query synchronisation	20	Carryover effect appears in Twitter, and it is cyclical

3 | EXPERIMENTS

Table Experiment - what you are controlling queries list - what keywords you have used Noise controlled - what noise can be control, what noise cannot be control? (explain why in text) Number of times of Running Results Similarities of evaluation results three tables for each Social media platform consistent of experiments setup!

3.1 | Workflow

The research process can be divide into four steps: Designing Experiments, data collection, Running Experiments, Evaluating Results. The data collection is responsible to collect the data used in the experiment, including the search terms, posts and social accounts. The experiment will query for these search terms to analyse the results. The posts and social accounts are used to set up the agent profiles. Experiment setup comes up with the detailed bias study on social media. There are mainly three experiments to check examine the bias in social media: carry-over effect, Cookies and polarisation problem. Besides the experiment setup, the noise treatment methods are especially important, because they represent the integrity and reliability of our experiment. We proposed some metrics to quantify the bias in social media. Previous work mainly uses the Jaccard index and edit distance as the metric, however, our research expend the evaluation method and introduced the Kendall Tau index and more robust semantic similarity method compared to Santos et al. [14].

3.1.1 | Data collection

Given a list of polarised topics, we need to collect search terms, posts and user accounts that are considered to be polarised. For the list of polarised topics, three main topic categories were considered in our experiments: Health, Politics, and General Purposes.

For the Health category, we chose COVID-19 as the main topic. The reason behind the COVID-19 topic choice is the polarisation between pro- and anti-vaccination movements. On social media, these groups have been growing in the last years reaching millions of people according to [20, 21].

For the Politics category, we chose the US Presidential Election, 2020, as the main topic. The polarisation between the two dominant US parties, democrats and republicans, has long been a case of study in political science, but recently it has been exacerbated by social media ([22]).

As for the General Purposes category, we chose multiple trending topics in 2020. Unlike political topics, these topics normally do not have evident opposing viewpoints. They cover a wide range of topics including celebrities, music festivals, games, films, etc. Generic topics, as a contrast to contrasting topics, are used to explore the impact of social media personalisation in non-contrasting topics.

After choosing the topics, we used the Google Trends (2020 trend topics) to manually select polarised search terms. The COVID-19 topic ended up with 80 polarised search terms, the US Presidential Election 2020 with 60 search terms and the General Purpose topic with 28 search terms. A few examples of the polarised search terms used during the US Presidential Election are: (pro-trump) "GodBlessPresidentTrump" and (anti-trump) "VoteBlue" (in reference to Trump's opponent Joe Biden).

Finally, the selected terms were used in the search mechanisms of Twitter, Reddit and Tumblr. For each polarised term, we collected the top 20 accounts or accounts returned in each social media platform. We manually reviewed each account and post to ensure that it was a representative post or account of a polarised view. This is important because the posts and accounts will be used by the BiasChecker tool and the agents (pro and anti) will create their polarised profiles by following or liking these accounts and posts.

3.1.2 | Noises Control

According to previous work ([4]), noise in this kind of experiments is unavoidable. Therefore, we will take into consideration the known noises below to reduce the noise in the results and our analysis.

Static DNS

For the timeliness of services in various regions, social media service providers usually set up servers for different regions. This means that the data held by different servers may not be synchronised due to network delays. This may cause each query to access and retrieve information from a different server, leading to different search results. To avoid this situation, a static host is configured to ensure that information is accessed and retrieved from a specific server.

IP Address and Geo-Location

Although a previous study reported the influence of location on Google's search personalisation ([5]), the role of location on social media search mechanisms has been little explored. In our experiments, we do not use mobile devices with built-in GPS and we disabled location sharing. In addition, we configured all computers used in our experiments to use the /24 subnet, which can give us a maximum range of two meters in locating the device used by a user, reducing location-related noise.

Query synchronisation

The rate at which online users post on social media platforms is very high and because of that it is required that the queries are issued at the same time. Therefore, BiasChecker was developed in such way that queries are issued at the same time to avoid discrepancies in the results returned.

A/B testing

Finally, we consider the noise that can be introduced by A/B testing implemented by social media platforms. A/B tests are often used by social media platforms to measure the engagement of their users, and to do this they randomly show a small percentage of users a different set of results for a given search query. To eliminate the noise introduced by A/B tests, we ran the same experiments (identical settings) on multiple agents and discarded noisy results.

3.1.3 | Experiments Setup

Our experimental design follows the route of first exploring possible factors that influence social media search personalization and then conducting a polarization experiment on social media by controlling for these factors. This study focus on the impact of the carry-over effect, cookies and languages on social media search. Since Hannak et al.[4] discovered the carryover effect in google searches, there have been many studies ([11],[12],[13],[14]) on social media bias, but no experiments have been conducted on the carry-over effect in social media. To my knowledge, this study is the first to examine the carry-over effect in social media. Also inspired by Hannak et al. [4], no one has investigated the use of cookies in social media. The language experiment was first proposed by me. This experiment was designed to check how different languages affect the search result and the semantic meaning used in social media search. This experiment follows an intuitive approach, setting up two agents with the same configuration to simultaneously execute queries in different languages with the same meaning. As for the polarization experiment, it simulates agents from three perspectives: Pro, Anti and Neutral. The evidence of the appearance of social media bias can be obtained by comparing the three agents results. A detailed description of the carryover effect, cookies and polarization experiments follows.

Experiment - Carry-over effect

Carry-over effect means that a query B may be affected by a given previous query A, if the time interval between queries A and B is lower than a given time t.

To verify whether a social media search mechanism uses previous searches to personalise the results of a subsequent search, we need to setup our experiments to identify the existence of the carry-over effect. The verification process is exhaustive as we need to set up a control group and also run queries A and B in different time intervals. The main idea is not only to identify the carry-over effect, but the time t that a query A will not influence the results obtained from a second query B.

The procedure runs as follows:

We initiate two agents with identical configuration. Agent A issues query q_1 while Agent B waits for a predetermined time interval t to issue the query q_2. After interval t has passed, both Agents will run query q_2. The search results are then compared to verify if the results returned are different. The carry-over experiment is executed 50 times for different query pairs and different t time intervals. The average results are reported in the next section.

Experiment - The role of cookies

Since cookies are used to store the user's browsing activity online, social media platforms can use them to personalise the search results performed by their users. To investigate whether cookies affect the search results in social media platforms, we set up multiple agents and checked the search results against control groups. The set up takes into consideration a few scenarios: (i) with users logged in and with cookies; (ii) with users logged in and without cookies (after every search, cookies are cleared from browsers); (iii) with users logged off and with cookies; and, (iv) with users logged off and without cookies. The control groups, that is, the agents that will be used to measure the differences in the search results will perform the queries and clear the cookies after each query is issued. The purpose of setting the control group is to explore the impact of logged in/off users and the removal of cookies in the search results. All other agents will perform the same queries but varying the configurations (i-iv). This experiment is performed by four agent instances. Two of the agents will log in to the social platform with some setup profiles. Also, two of the agents will clear their cookies after finishing the query.

This experiment can tell us a lot about the use of cookies by different social platforms. We expect that the search results will present differences before and after the cookie is cleared, and that these differences will be mainly related to the information stored in the cookies.

Experiment - Polarization

The polarisation experiment aims to investigate the presence of bias in personalised search results through polarised topics. There are three types of agents: anti-agent, pro-agent and neutral agent. The anti-agent is the agent that opposes a given topic. The pro-agent is the agent that supports a given topic. Finally, the neutral agent is a fresh profile without a view/opinion with regards to any topic.

To set up agents, we configure our tool to incrementally interact with polarised topics. The interactions are made in the form of likes, follows, retweet, etc.

It is important to note that we control the noise by running two or more identical agent instances to measure the noise introduced by, for instance, A/B testing. Moreover, the polarisation check is the last experiment to be executed as we will be able to know what are the factors and noises (e.g., carry-over effect, language, cookies, location, etc.) that influence the results and hence provide a more accurate analysis of the results obtained.

Experiment - Language

The experimental setup for language is relatively straightforward, it compares the results of queries with the same meaning in different languages. When exploring the influence of language on the personalisation of the results, we noticed that some platforms are not fully prepared to work with other languages, in this case, Chinese. Further experiments on other languages are required to generalise the results. For this experiment, we translated the English terms used into Chinese using Google translate. The translations provided by Google were carefully reviewed to ensure the terms were accurately translated.

3.1.4 | Evaluation Metrics

To evaluate the results returned by the social media search mechanisms, we used four metrics widely used in the literature [4, 5, 23, 7]: the Jaccard index and the Edit Distance.

Besides those evaluation metrics, we also use the Kendall Tau index to evaluate the results. Unlike Jaccard and Edit Distance, it considers the position of the results [9].

Last but not least, as in [14]),we used a semantic metric to measure the semantic similarity between the results.

Each of the metrics used in this report is explained in what follows.

Jaccard Index

The Jaccard index([24]) is used to compare the similarities and differences between a limited sample set. It is defined as the intersection size divided by union size of two sample set. The mathematical representation of the index is shown below:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

It is the total number of similar entities between sets divided by the total number of entities. However, the Jaccard index is largely affected by big data sets which decrease the accuracy of result. This is because a large data set may significantly increase the number of concurrent sets while keeping the intersection constant. Since we will not be using large datasets in our experiments, Jaccard i well suited as a metric

Edit Distance

Edit Distance, also known as Levenshtein distance([25]), refers to the minimum number of editing operations required to convert two strings from one to the other. There are only three editing operations, including replacing one character with another, inserting a character, and deleting a character. The smaller the edit distance, the greater the similarity between the two strings. Edit distance compares the similarity between them by the number of operations. However, it is necessary to ensure that the lengths of the search results are equal. Therefore, we collect 10 entries of search results in each query.

Kendall Tau Index

Kendall Tau([26]) is used to measure the ordinal correlation between two sets of data. When the observations have a similar order between the two variables, the Kendall correlation between the two variables will be very high. Conversely, when the observations are ranked differently between the two variables, the Kendall Tau index will be low. As the order of search results can affect how users interact online, in addition to Jaccard and Edit distance, we also use Kendall Tau.

Semantic Similarity

Much of the previous work has been on general search engines, and the content of web pages is not well organised, so sentence embedding is not able to generalise the web content in this situation. Posts in social media are usually short and therefore easily recognised and embedded by machine learning models.

The semantic similarity is measured by the cosine similarity([27]) of the results of two text embeddings - a vectorial representation of texts. In this work, we use the LaBSE model from [28] to generate the text embeddings with the goal of measuring the similarity between search results and tell whether deviations are found. Previous works, like [14], used the MUSE model [15] to measure search results similarity, however, given the performance of transformer-based machine learning techniques for natural language processing and that MUSE is used for multilingual word embedding, not sentence embedding, we opted to use LaBSE. LaBSE is a Language-agnostic BERT Sentence Embedding adapted from the multilingual BERT [29]. LaBSE combines the pre-training of the masked language model (MLM) and the translation language model (TLM) with the translation ranking task using a two-way dual encoder, which will produce a more accurate sentence representation.

The advantage of semantic similarity, as opposed to the metric listed above (e.g., Jaccard Index, Edit distance and Kendall Tau), is that it takes into account the extent to which the different search results are similar in terms of content. The metrics presented above simply treats the entries of a search result as elements of a collection or array, ignoring the textual content of the social media search results. We know that they are actually the same for two posts with identical content but different links, while, Jaccard and Kendall Tau index treats these two posts as two different items.

Sentence similarity provides further analysis of the similarity of search results. It is based on cosine similarity which is a commonly used similarity measure in information retrieval. For two lists of search results M and N, semantic similarity (inspired by [30]) will be formally defined as:

$$semantic\ similarity = \frac{1}{|M||N|}\sum_{i\in M}\sum_{j\in N}\frac{m_i \cdot n_j}{||m_i||||n_j||}$$

where the m_i and n_j is the element of search results M,N

3.1.5 | Running Experiment

Carry-over effect

All Agents must complete their queries before proceeding to the next query. It means that, in the carry-over effect experiment, while the first agent executes the first query, the second agent does nothing. For this, we added a special query ("$carry - over$") to the second agent that only acts as a synchronisation token and does not actually execute the first query.

Based on [5], geographic location in Google search engines significantly personalises search results. Therefore, the search terms for the carry-over experiment were created in such a way that the first search term was based on a different geographical location and the second search term was a search term that we had collected previously. Two widely debated topics are used, US election (within the US) and COVID-19 (worldwide), in the second query, thus the first term always the countries or states related to the second term.

The occurrence of the carry-over effect is confirmed, if, e.g., after the first search for "American", the results obtained for the search query "COVID-19 case update", returns results related to "American COVID-19 cases".The extent of carry-over effect will be finalised using a series of metrics.

The DNS, IP address and Query synchronisation was controlled in this experiment.

Cookies and (non-) Authenticated Users

In this experiment, we set up four agents corresponding to different cookies and whether authenticated. The cookies will be cleaned after each query. Our experimental environment is chromium + puppeteer and the puppeteer provides the control commands to clean the cookies of chromium. During the experiment, we control the noise from DNS, IP address and query synchronisation. We set up static DNS by changing the host file of the agent. The A/B testing is not considered in this experiment, because the A/B test can also be seen as the result of the use of cookies and authentication. We repeat this experiment 10 times to analyse the mean result.

Language

The language experiment comes from an intuitive idea that search results should be different using different languages. Our experimental setup is also simple, comparing search results in different languages with the same meaning. Furthermore, we can understand the utility of semantic meaning in social media searches. We translate the collected search terms. The translations are provided by google translate and are manually checked for consistency of meaning. This experiment also controls for noise except A/B testing.

Polarized Topic

This experiment controls all noise. There are three types of agents in the polarization experiment, each with two identical instances, being run on the same host, in order to control for A/B tests and other uncontrollable noise. We expect the same agents to receive identical results, so any differences obtained can be considered to be due to other noise. The experiment was conducted 5 times to obtain an average value.

4 | RESULTS

4.1 | Carry-over effect

The figure 1 shows the results using the Jaccard index and Kendall Tau on Twitter, respectively. The x-axis indicates the time interval between the first and second queries. The y-axis indicates the level of overlap between the results obtained for the second query and the control query.

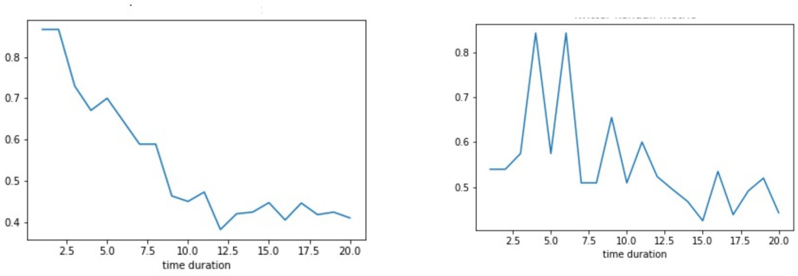
Note that in the first experiment, with almost no time interval between two queries, the results shows some different(approx 10% of the results are dissimilar) indicating that one query may affect the subsequent one. As the interval increases, the Jaccard index decreases, implying that its difference is gradually increasing.

After running the experiments for 30 query pairs, the results with greater time intervals showed on average a larger discrepancy between the results confirming the carry-over effect on Twitter. Curiously, when the time interval greater than t=10 min, no difference is found in Jaccard index metric and after that a constant average of supposedly personalisation in the results is found (approx. 50% of the results same). The Kendall Tau results seems irregular, but it also appears similar results after t = 10 min, fluctuating around 0.5 . We hypothesise that there is a cyclical carry-over effect. On the other hand, it is clear that the carry-over effect exists on Twitter search mechanism.

Figure 2 shows the results using the Jaccard index and Kendall Tau on Reddit, respectively. The Kendall Tau results present significant differences when compared to the Jaccard index. We note here that although the results presented using Jaccard indicate a small difference, Kendall Tau experiment shows Reddit re-rank the results more. Therefore, the carry-over effect is also found in Reddit, but instead of presenting different search results, it mostly re-rank the results to its users.

Finally, Figure 3 shows the results using the Jaccard index and Kendall Tau on Tumblr, respectively. Similarly, Tumblr also personalise the results based on previous queries issued by its users. The Jaccard doesn't have any difference, but Kendall tau have a difference approx 15% on average. Based on our results, we can affirm that there is carry-over effect on Tumblr.

Twitter

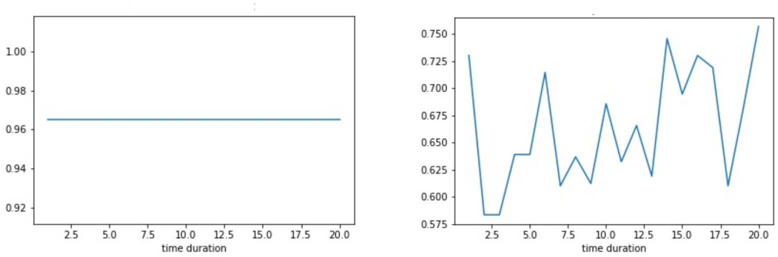


(a) The jacccard index result of the carryover effect in Twitter

(b) The Kendall Tau of the carryover effect in Twitter

FIGURE 1 The result of carryover effect experiment in Twitter

Reddit

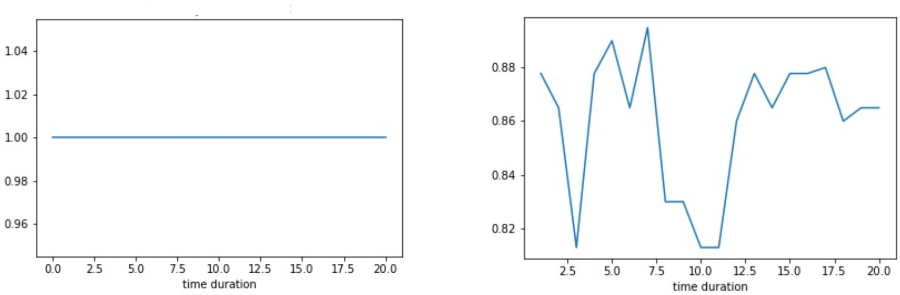


(a) The jacccard index result of the carryover effect in Reddit

(b) The Kendall Tau of the carryover effect in Reddit

FIGURE 2 The result of carryover effect experiment in Reddit

Tumblr



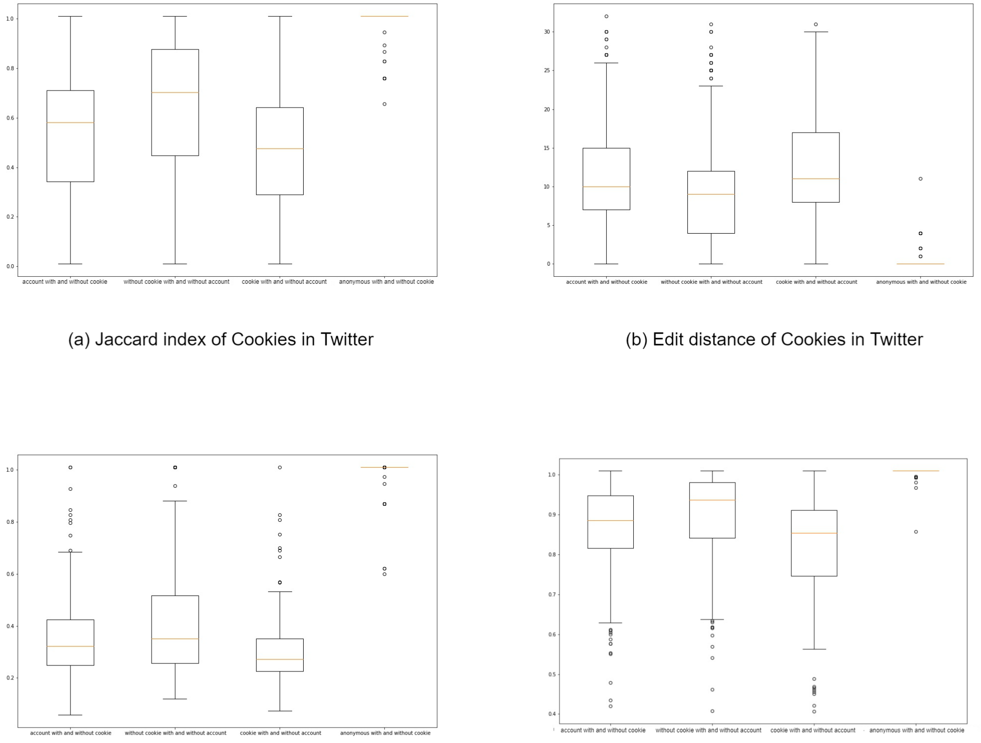
(a) The jacccard index result of the carryover effect in Tumblr

(b) The Kendall Tau of the carryover effect in Tumblr

FIGURE 3 The result of carryover effect experiment in Tumblr

4.2 | Cookies and (non-) Authenticated Users

4.2.1 | Twitter



(a) Jaccard index of Cookies in Twitter

(b) Edit distance of Cookies in Twitter

(c) Kendall Tau of Cookies in Twitter

(d) Semantic similarity of Cookies in Twitter

FIGURE 4 The result of cookies experiment in Twitter

Figure 4 shows a comparison between the search results for Twitter using different cookie settings. The first column of the figures shows the comparison between the two agents logged into the social platform. The Jaccard index ranges from 0.4 to 0.7 with an average of 0.54. This indices suggest that the Twitter strongly personalise the results based on cookies information. The edit distance and Kendall Tau also demonstrate evidence of this.

Figure 4(d) shows that the semantic similarity among the results is very high (μ=0.86) ranging from 0.8 to 1.0. The similarity indicates that although the results are different they are all content-wise related.

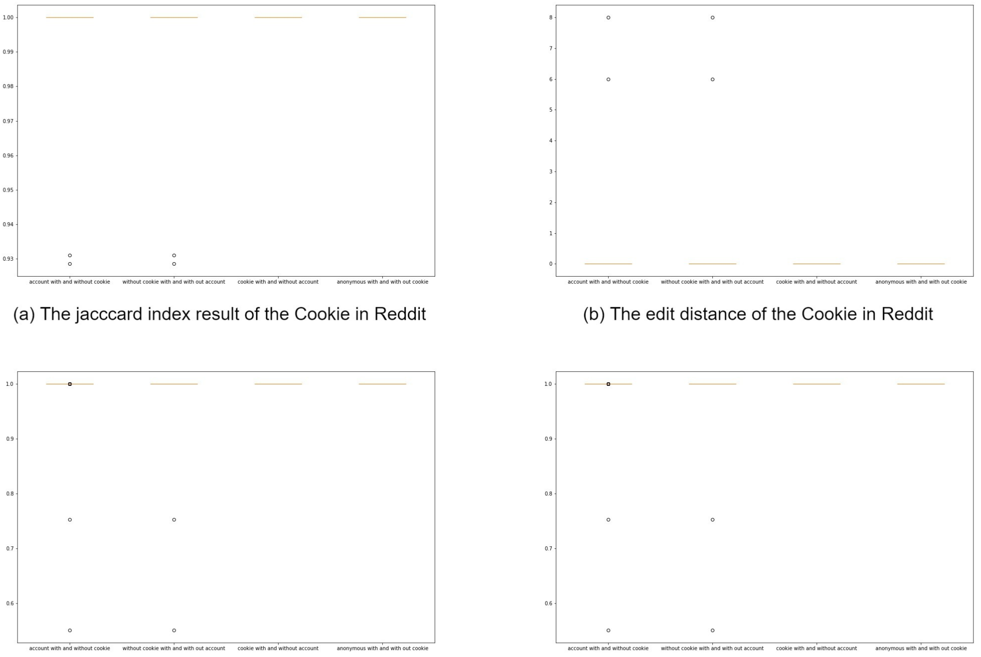
Cookies can be used with authenticated users or not and, therefore, we also carried out experiments to understand whether users that are logged into Twitter are subject to personalisation. The second column in Figure 4 shows the results for non- and authenticated users disregarding cookies information (cookies are delete after each query). The Jaccard index ranges from 0.4 to 0.9 (μ=0.64). This indicates that Twitter personalises based on the users' activity records, demographic information, etc. Again, the semantic similarity is high (μ=0.89), ranging from 0.8 to 1.0, which indicates that results, although different, are semantically consistent.

Now, considering that users are using cookies (i.e., users are being tracked),from the third column in Figure 4, the Jaccard index obtained in our experiments ranges from 0.3 to 0.7 (μ=0.47). The lower Jaccard index compared with the experiment carried out without cookies indicates greater personalisation in the results. The mean difference between the results with and without cookies is 0.17. Similarly as in previous experiments, the semantic similarity is high (approx. 0.8).

Finally, we carried out an experiment where the users use the anonymous/incognito mode. The Jaccard index obtained is in its majority 1.0 (μ=0.98). The semantic similarity is also close to 1.0.

In summary, Twitter uses information stored in cookies to offer a personalised experience to its users. Furthermore, if a user is authenticated, Twitter will take advantage of the users' profiles information and activities in its platform to offer an even more personalised experience. Finally, to avoid personalised results on Twitter, one should set the anonymous/incognito mode.

4.2.2 | Reddit



(a) The jacccard index result of the Cookie in Reddit

(b) The edit distance of the Cookie in Reddit

(c) The Kendall tau result of the Cookie in Reddit

(d) The Semantic simliarity of the Cookie in Reddit

FIGURE 5 The results of cookies experiment in Reddit

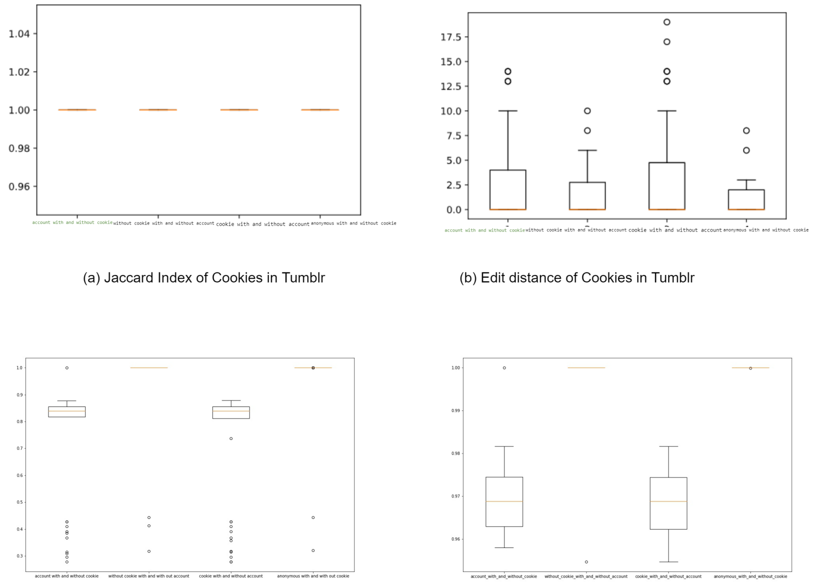
Figure 5 shows the results received in Reddit. Unlike Twitter, from the experiments carried out, Reddit rarely use information stored in cookies to offer a more personalised experience to its users. shows that whether cookies are retained or not leads to difference not exceed 0.05 and their semantics are almost identical. In subsequent parallel experiments across platforms to investigate the effect of other noise on the results, we discovered that such extent of difference is negligible compared with other noises. The evidence suggests that personalisation is rarely based on the storage of cookies in the Reddit platform.

4.2.3 | Tumblr

As in Twitter, Tumblr also uses the information stored in cookies to personalise its results. However, the effect of cookies on Tumblr is relatively small. We can see an average difference of 0.17 between whether cookies are kept on the login account and whether log into Tumblr while keeping cookies. These two comparisons also show a slight difference in semantic similarity. This was a surprising result, as we expected Tumblr to introduce the same level of personalisation based on cookies regardless of whether the account was logged in or not.

Recall the setups are: (i) Authenticated Users + cookies; (ii) Authenticated Users; (iii) cookies; and, (iv) Non-authenticated Users. The results show that there is no difference in the results of (ii), (iii) and (iv). And, (i) shows a slight difference from all three of them.

The only reasonable explanation for this result would be that Tumblr only uses cookies for personalised searches when the user is logged in. When the user is not logged in the search results will not rely on cookies information.



(a) Jaccard Index of Cookies in Tumblr

(b) Edit distance of Cookies in Tumblr

(c) Kendall Tau of Cookies in Tumblr

(d) Semantic similarity of Cookies in Tumblr

FIGURE 6 The result of cookies experiment in Tumblr

4.3 | Language

The experimental setup for language is relatively straightforward, it compares the results of queries with the same meaning in different languages. When exploring the influence of language on the personalisation of the results, we noticed that some platforms are not fully prepared to work with other languages, in this case, Chinese. Further experiments on other languages are required to generalise the results. For this experiment, we translated the English terms used into Chinese using Google translate. The translations provided by Google were carefully reviewed to ensure the terms were accurately translated. Below we present the results for Reddit, Twitter and Tumblr.

4.3.1 | Reddit

Reddit does not provide an advanced search with the option to set a specific language for the returned results.

Figure 7 shows the number of results obtained by different languages between different tabs on the Reddit platform, with blue representing English search results and orange representing Chinese search results.

The number of English search results on the Reddit platform is, intuitively, much higher than the number of Chinese search results. This reveals the unevenness of English and Chinese posts on the Reddit platform.

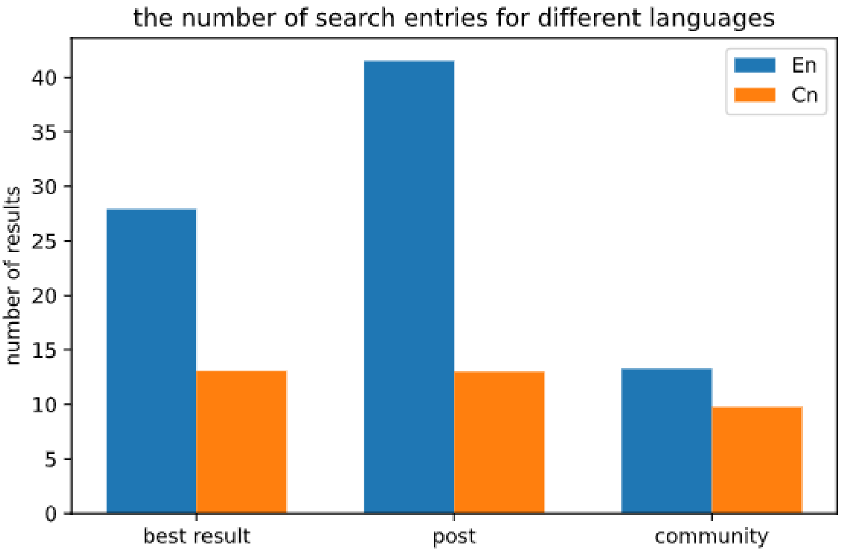
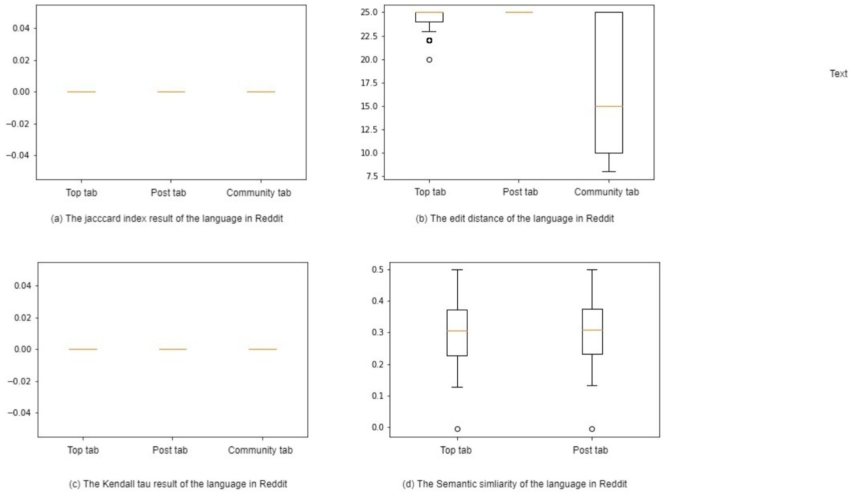


FIGURE 7 The Number of results retrieved for each tab in Reddit



(a) The jacccard index result of the language in Reddit

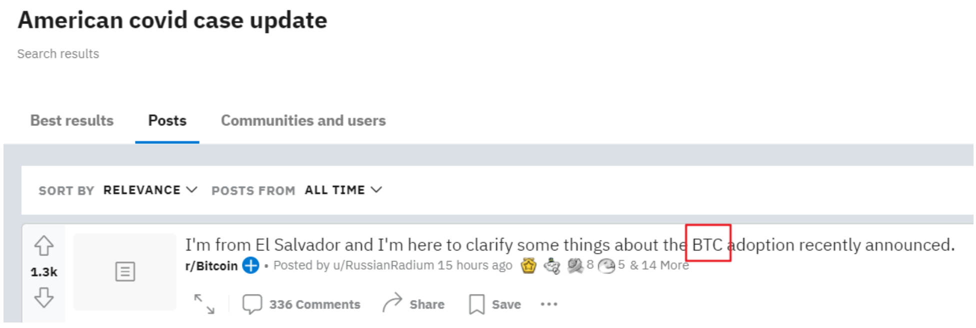
(b) The edit distance of the language in Reddit

(c) The Kendall tau result of the language in Reddit

(d) The Semantic simliarity of the language in Reddit

FIGURE 8 The result of Language experiment in Reddit

As Figure 8 shows, the Jaccard Index and Kendall Tau have the value of 0. In Reddit, even without the advanced search for language filtering, using two terms with different languages, different results are presented. The results show that no search results are the same. Recall that the community tab only contains the communities and users which does not have semantic information thus the results with semantic similarity only contain the top result and post tab. Focusing on the semantic similarity in the results, it was found that the semantic similarity in the top result and post tab was relatively low, primarily in the range of 0.25-0.4, implying that the search results are seldom relevant even in the semantic level.

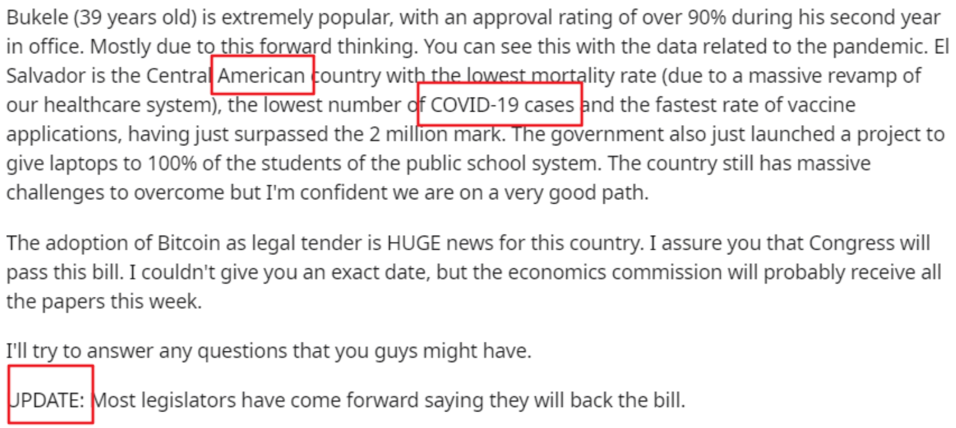

American covid case update

Search results

I'm from El Salvador and I'm here to clarify some things about the BTC adoption recently announced.

r/Bitcoin • Posted by u/RussianRadium 15 hours ago 8 5 & 14 More

FIGURE 9 Example of a search result in Reddit that does not match the semantics of the search term



Bukele (39 years old) is extremely popular, with an approval rating of over 90% during his second year in office. Mostly due to this forward thinking. You can see this with the data related to the pandemic. El Salvador is the Central American country with the lowest mortality rate (due to a massive revamp of our healthcare system), the lowest number of COVID-19 cases and the fastest rate of vaccine applications, having just surpassed the 2 million mark. The government also just launched a project to give laptops to 100% of the students of the public school system. The country still has massive challenges to overcome but I'm confident we are on a very good path.

The adoption of Bitcoin as legal tender is HUGE news for this country. I assure you that Congress will pass this bill. I couldn't give you an exact date, but the economics commission will probably receive all the papers this week.

I'll try to answer any questions that you guys might have.

UPDATE: Most legislators have come forward saying they will back the bill.

FIGURE 10 The content of the example shown in figure 9 in Reddit

Such unusual findings prompted us to investigate further. By examining the search results, we found that more than 1/3 of the Chinese searches did not receive any results. We then manually queried the search terms and discovered that some of the content was completely unrelated to the query.

Initially, I was attributed this to the lack of Chinese posts on Reddit and an improper understanding of the Chinese query. So Reddit will be more personalised based on the language - Chinese - rather than on the meaning of the search term.

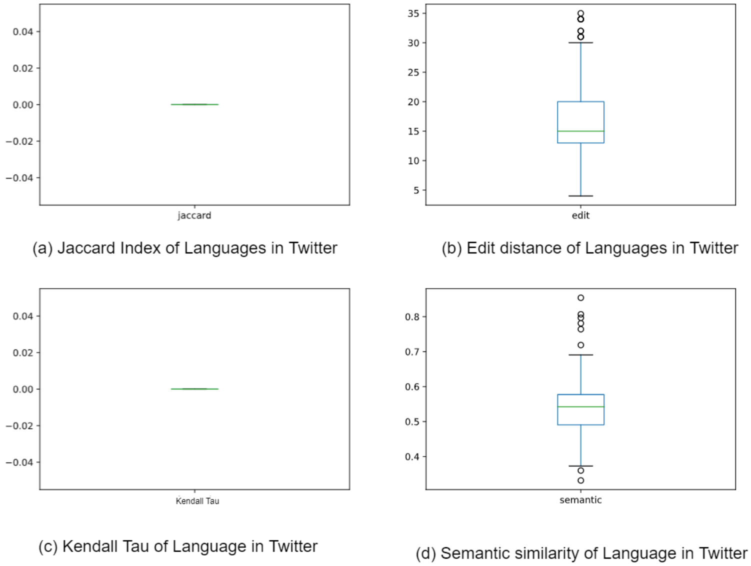
Subsequently, the contradiction in this assumption was cleared. The search for "American covid case update" in English (as shown in Figure 9) does not take into account the inherent semantics of the query as a whole, but rather on the search terms as separate words. The red boxes in Figure 10 shows these words and the contexts they appear.

We cannot affirm that Reddit does not consider the semantics of the search terms, relying on this example. However, we hypothesise that searches in Reddit will take little or no account of the semantic content of search terms. This hypothesis requires additional investigation.

4.3.2 | Twitter

The Twitter platform provides an advanced search function, where we can set the language of interest for the retrieved results. For this, the user has to add into his/her search the text ’lang:’ followed by the language code (e.g. ’en’ for English or ’zh-cn’ for Chinese).

The main search results on Twitter are shown in Figure 11.



(a) Jaccard Index of Languages in Twitter

(b) Edit distance of Languages in Twitter

(c) Kendall Tau of Language in Twitter

(d) Semantic similarity of Language in Twitter

FIGURE 11 The result of Language experiment in Twitter

Observe that the result of the Jaccard index and Kendall Tau is 0, which is to be expected. The semantic similarity between the results are relevant (ranging from 0.4 to 0.8) showing that the results obtained using different languages are semantic related. This outcome shows that the search results in Twitter will be personalised by language and to a considerable extent.

4.3.3 | Tumblr

The number of Chinese users in Tumblr is rare and therefore the number of posts in Chinese is very limited. For this reason, we were unable to conduct comparative experiments using Tumblr in both Chinese and English languages. As future work, we intend to experiment with other more popular languages used on Tumblr.

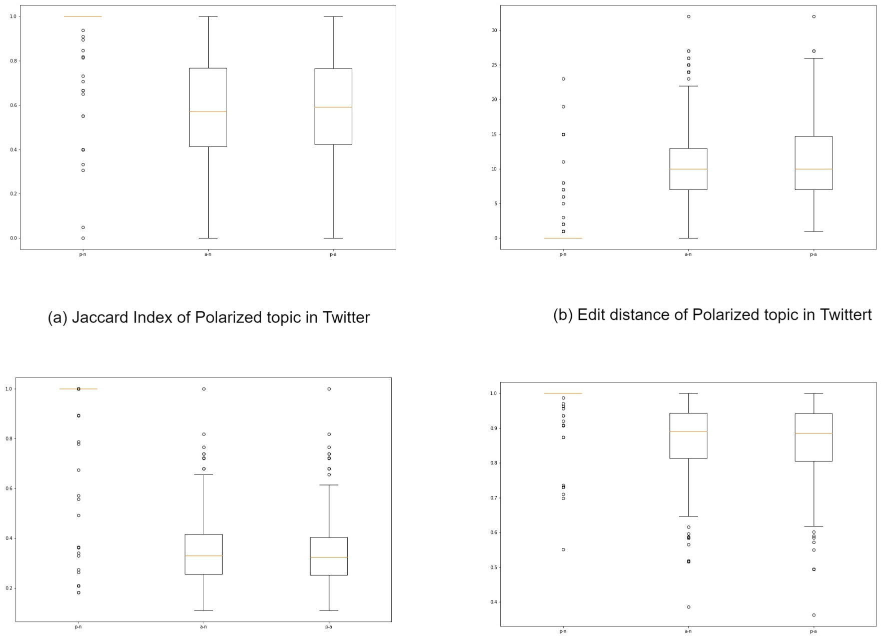
4.4 | Polarised topics

4.4.1 | Twitter

When running the polarisation experiments, we ran parallel experiments where, for each class of agents, we ran multiple instances at the same time, having exactly the same configuration. Here, we compared each group of agents (pro, anti and neutral). We can see from the Table 4 that agents with the same configuration get almost the same results. However, there are still some of them that are slightly different. For the pro group, on average, 0.06 of the result is different, while the anti and neutral results are about 0.03 different. The results for the pro group were on average 1.5 results different, while the anti and neutral were about 0.6 different. In addition, the semantic similarity in the results is highly overlapping, up to 0.99. The results here provide us with an indication of the degree of influence of other noises. It provides us with a benchmark for measuring the true deviation. When the error in the search results exceeds the error due to other noise, it can be considered personalised.

TABLE 4 The parallel experiment results in Twitter

	PRO	Neutral	Anti
Jaccard index	0.94106	0.987136	0.97148
Edit distance	1.50335	0.56451	0.67785
Kendall Tau	0.91941	0.96380	0.94415
Semantic similarity	0.97729	0.99999	0.98617



(a) Jaccard Index of Polarized topic in Twitter

(b) Edit distance of Polarized topic in Twittert

(c) Kendall Tau of Polarized topic in Twitter

(d) Semantic similarity of Polarized topic in Twitter

FIGURE 12 Polarized experiment results in Twitter

TABLE 5 The mean of the different metrics in Twitter

	P-N	A-N	P-A
Jaccard index	0.93553	0.57094	0.57537
Edit distance	2.44	11.24615	11.34693
Kendall Tau	0.91506	0.33940	0.33582
Semantic similarity	0.96149	0.84451	0.84008

First, we can see that for all indicators, the results for the pro agent are almost identical to those for neutral, while the results for anti agent and pro agent differ more from the others. It has an average Jaccard of 0.93 for pro-neutral and about 0.57 for anti-neutral and pro-anti which means that more than 1/3 of the results in anti and pro are different, regardless of their order. The Kendall Tau between pro and neutral is 0.92, while the difference between anti and neutral is almost 0.34. The Kendall Tau index takes into account the order of the search results, so we can see that while 2/3 of the search results are the same, half of them appear in a different order.

The difference between pro and neutral agents is relatively low approximate to noise level which means the configuration of PRO-agents does not play an important role. On the other hand, the differences between ANTI-agents and Pro-agents are significant, so we can assume that opposition to COVID vaccines is highly individualised and that the number of people opposed to vaccines is likely to be relatively small and therefore more personalisation will be applied.

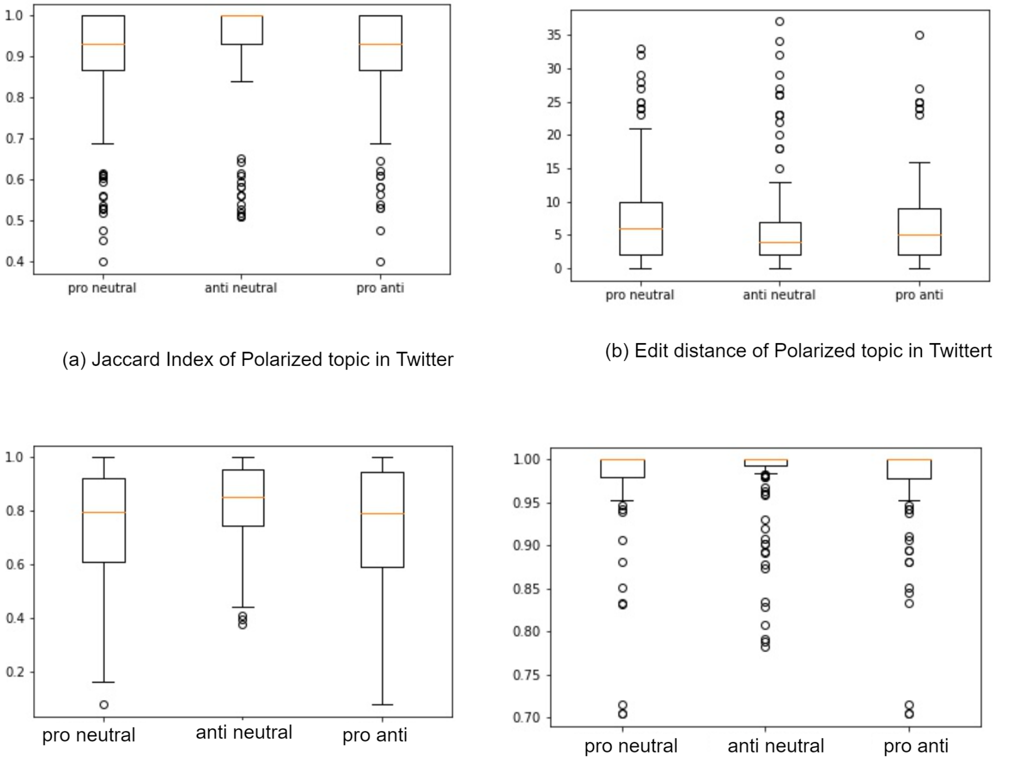
The large gap between the anti and other groups and the slight gap between pro and neutral suggest that the anti group is to some extent trapped into a filter bubble.

4.4.2 | Reddit

TABLE 6 The parallel experiment results in Reddit

	PRO	Neutral	Anti
Jaccard index	0.99001	1.0	0.99151
Edit distance	0.86956	0.26086	0.32456
Kendall Tau	0.99999	0.99363	0.98197
Semantic Similarity	0.99472	1.0	0.98793

Table 6 shows the results of running multiple identical agents. Each column represents an agent group. All results show very slight differences. For Jaccard index, Kendall Tau and Semantic similarity, all agents were close to 1, with almost no difference. The edit distance for the pro agent has an average of 0.8 differences in search results, neutral agent has 0.3 and anti agent has only 0.2. The level of variation of these agents will provide a baseline for our subsequent exploration of their personalisation magnitude.



(a) Jaccard Index of Polarized topic in Twitter

(b) Edit distance of Polarized topic in Twittert

(c) Kendall Tau of Polarized topic in Tumblr

(d) Semantic similarity of Polarized topic in Tumblr

FIGURE 13 Polarised experiment results in Reddit

Figure 13 shows the polarised results in Reddit. First, the metric using the Jaccard index shows an average result of 0.89 for pro-neutral, and 0.93 for anti-neutral, and 0.91 for pro-anti. While, edit distance shows a different on average of 8 search result entries between pro and neutral agent. Anti-neutral and anti-pro also differ by an average of almost 7 search result entries. The Kendall Tau metric indicates that 0.25 of the results for pro and neutral are inconsistent, and by comparing this with the Jaccard index we find that 15% of the results are the same but not in the same order. Pro and anti agents have the same results but over 16% of the results are inconsistent in order. Anti and neutral agents are less different, with 10% of the results being inconsistent. The semantic similarities are all very high, almost the same as the noise level shown above.

There is a significant difference in the results in the best tab, a difference of about 10% in the content of the results and an even greater difference in the ordering of the search results. Here, the neutral agent is set to the anonymous mode, so the results of the neutral agent are used as a baseline. Therefore, we can conclude that there is personalisation in the best tab.

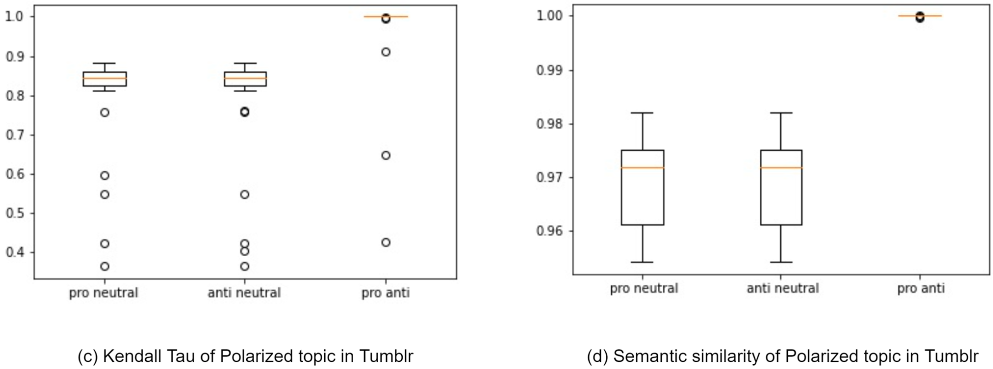
Note that the results among the three different agents presented similar personalisation levels, therefore, Reddit is relatively balanced among opposing viewpoints.

4.4.3 | Tumblr

TABLE 7 The parallel experiment results in Reddit

	PRO	Neutral	Anti
Jaccard index	0.99999	1.0	0.99999
Edit distance	0.16379	0	0.12746
Kendall Tau	0.98416	1.0	0.98416
Semantic Similarity	0.98986	1.0	0.98986

Table 7 shows the results of running multiple identical agents in Tumblr. Each column represents an agent group. The results in pro and anti agents are symmetric. This indicate that Tumblr search results will be adjusted based on the account profiles.



(c) Kendall Tau of Polarized topic in Tumblr

(d) Semantic similarity of Polarized topic in Tumblr

FIGURE 14 Polarised experiment results in Tumblr

Figure 14 shows the results for the Tumblr platform. Pro and anti agents return almost identical results. Some difference is found between pro-neutral and anti-neutral agents. The differences here are consistent with the differences found when exploring cookie information earlier.

The results obtained for Tumblr indicates that Tumblr does not offer biased results for polarised topics. Therefore, users are less likely to experience the filter bubble effect in Tumblr.

5 | DISCUSSION

graphs in the current results section if there is a bias? Why there is a bias? Uncontrolled noises? The effects of users? Limitations?

6 | CONCLUSION

Acknowledgements

Supporting Information

references

[1] Pariser E. The Filter Bubble: What the Internet Is Hiding from You. Penguin Group , The; 2011.

[2] Nickerson R. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Review of General Psychology 1998;2:175 – 220.

[3] Tucker J, Guess A, Barbera P, Vaccari C, Siegel A, Sanovich S, et al. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. SSRN Electronic Journal 2018 01;.

[4] Hannak A, Sapiezynski P, Molavi Kakhki A, Krishnamurthy B, Lazer D, Mislove A, et al. Measuring Personalization of Web Search. In: Proceedings of the 22nd International Conference on World Wide Web WWW '13, New York, NY, USA: Association for Computing Machinery; 2013. p. 527–538. https://doi.org/10.1145/2488388.2488435.

[5] Kliman-Silver C, Hannak A, Lazer D, Wilson C, Mislove A. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In: Proceedings of the 2015 Internet Measurement Conference IMC '15, New York, NY, USA: Association for Computing Machinery; 2015. p. 121–127. https://doi.org/10.1145/2815675.2815714.

[6] Cozza V, Hoang VT, Petrocchi M, Spognardi A. Experimental measures of news personalization in Google News. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2016;9881 LNCS:93–104. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84992694506{\&}doi=10.1007{\%}2F978-3-319-46963-8{_}8{\&}partnerID=40{\&}md5=338fff38e3d4ef3d46fa9cd594bcfac8.

[7] Salehi S, Du JT, Ashman H. Examining Personalization in Academic Web Search. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media HT '15, New York, NY, USA: Association for Computing Machinery; 2015. p. 103–111. https://doi.org/10.1145/2700171.2791039.

[8] Hoang VT, Spognardi A, Tiezzi F, Petrocchi M, De Nicola R. Domain-specific queries and Web search personalization: Some investigations. In: ter Beek M H LAL, editor. Electronic Proceedings in Theoretical Computer Science, EPTCS, vol. 188 Open Publishing Association; 2015. p. 51–58. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954499520{\&}doi=10.4204{\%}2FEPTCS.188.6{\&}partnerID=40{\&}md5=ecfe2373fe1487d9a85f1a100e1a687f.

[9] Dillahunt TR, Brooks CA, Gulati S. Detecting and visualizing filter bubbles in Google and Bing. In: Conference on Human Factors in Computing Systems - Proceedings, vol. 18 Association for Computing Machinery; 2015. p. 1851–1856. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954305091{\&}doi=10.1145{\%}2F2702613.2732850{\&}partnerID=40{\&}md5=aba847464dee0114f572b3136d62f8ae.

[10] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software: Practice and Experience 1991;21(11):1129–1164. https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380211102.

[11] Kruikemeier S. How political candidates use Twitter and the impact on votes. Computers in Human Behavior 2014;34:131–139. http://www.sciencedirect.com/science/article/pii/S0747563214000302.

[12] Hargreaves E, Agosti C, Menasché D, Neglia G, Reiffers-Masson A, Altman E. Biases in the Facebook News Feed: A Case Study on the Italian Elections. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2018. p. 806–812.

[13] Kulshrestha J, Eslami M, Messias J, Zafar MB, Ghosh S, Gummadi KP, et al. Search bias quantification: investigating political bias in social media and web search. INFORMATION RETRIEVAL JOURNAL 2019;22(1-2, SI):188–227.

[14] C dos Santos J, W M Siqueira S, Pereira Nunes B, P Balestrassi P, R S Pereira F. 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, New York, NY, USA: Association for Computing Machinery; 2020. p. 267–276. https://doi.org/10.1145/3394231.3397917.

[15] Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, et al., Multilingual Universal Sentence Encoder for Semantic Retrieval; 2019.

[16] Pitoura E, Tsaparas P, Flouris G, Fundulaki I, Papadakos P, Abiteboul S, et al. On Measuring Bias in Online Information. SIGMOD Rec 2018;46(4):16–21. https://doi.org/10.1145/3186549.3186553.

[17] Nikolov D, Lalmas M, Flammini A, Menczer F. Quantifying Biases in Online Information Exposure. Journal of the Association for Information Science and Technology 2019;70(3):218–229. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24121.

[18] Kulshrestha J, Eslami M, Messias J, Zafar MB, Ghosh S, Gummadi KP, et al. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing CSCW '17, New York, NY, USA: Association for Computing Machinery; 2017. p. 417–432. https://doi.org/10.1145/2998181.2998321.

[19] Robertson RE, Jiang S, Joseph K, Friedland L, Lazer D, Wilson C. Auditing Partisan Audience Bias within Google Search. Proc ACM Hum-Comput Interact 2018;2(CSCW). https://doi.org/10.1145/3274417.

[20] Burki T. The online anti-vaccine movement in the age of COVID-19. The Lancet Digital Health 2020 Oct;2(10):e504–e505. https://doi.org/10.1016/S2589-7500(20)30227-2.

[21] Johnson NF, Velásquez N, Restrepo NJ, Leahy R, Gabriel N, El Oud S, et al. The online competition between pro- and anti-vaccination views. Nature 2020 Jun;582(7811):230–233. https://doi.org/10.1038/s41586-020-2281-1.

[22] Olson J. Whiteness and the Polarization of American Politics. Political Research Quarterly 2008;61(4):704–718. http://www.jstor.org/stable/20299771.

[23] Le H, Maragh R, Ekdale B, High A, Havens T, Shafiq Z. Measuring Political Personalization of Google News Search. In: The World Wide Web Conference WWW '19, New York, NY, USA: Association for Computing Machinery; 2019. p. 2957–2963. https://doi.org/10.1145/3308558.3313682.

[24] Jaccard P. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. Bulletin de la Societe Vaudoise des Sciences Naturelles 1901 01;37:241–72.

[25] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 1966 Feb;10:707.

[26] KENDALL MG. A NEW MEASURE OF RANK CORRELATION. Biometrika 1938 06;30(1-2):81–93. https://doi.org/10.1093/biomet/30.1-2.81.

[27] Cosine similarity. Wikimedia Foundation; 2021. https://en.wikipedia.org/wiki/Cosine_similarity.

[28] Feng F, Yang Y, Cer D, Arivazhagan N, Wang W, Language-agnostic BERT Sentence Embedding; 2020.

[29] Devlin J, Chang MW, Lee K, Toutanova K, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; 2019.

[30] Rahutomo F, Kitasuka T, Aritsugi M. Semantic Cosine Similarity; 2012. .