Bias in Reddit Data Collection

Sunyam Bagga

ABSTRACT

The goal of this assignment is to estimate how popular subreddits are on average. The number of subsrcibers to a subreddit is used as a proxy for popularity here. For collecting a sample of subreddits, I use two different sampling techniques: (1) the first approach randomly samples a list of 500 subreddits from a static list of nearly all subreddits that exist on Reddit, and (2) the second approach is more dynamic in nature in that it samples subreddits by looking at the newest posts submitted to Reddit. In this study, I found that the mean and standard deviation of the number of subscribers varied drastically for the two sampling approaches. This shows how the data collection method used for a study can play a crucial role in the overall statistic we are measuring.

INTRODUCTION

In this growing field of computational social science, there has been an increasing number of studies aiming to predict or measure human behaviour (and "hidden" attributes) using data from the Internet, particularly social media sites like Twitter, Reddit, Facebook etc. In most studies, a lot of emphasis is given to the methods used and the resulting accuracy of those methods. Some of them fail to acknowledge the bias that might be present in their dataset, and seldom discuss the implications of the data collection method used.

The question of how the data is collected is important as it can lead to different outcomes of a study. Moreover, in order to make the findings useful for the research community, it is necessary for authors to address how the data collected might impact the findings of their study. There is a growing need to discuss the scope and the biases of their dataset, and the applicability of their methods on other kinds of datasets.

This paper shows how the data collected for a study may introduce biases which can have a significant impact on the overall findings of the study. Specifically, we calculate the average popularity of a set of subreddits using the number of subscribers to a subreddit as a proxy for popularity. Two different techniques are used for sampling a list of subreddits from all of Reddit. As discussed in detail in the Results section, we find that the average number of subscribers highly depends on how we sample the subreddits.

METHODOLOGY

The popularity is calculated using the mean of the number of subscribers of all sampled subreddits. In order to sample a list of subreddits, the following two techniques are used:

- *RandomApproach*: a list of 500 subreddits is randomly selected from a static list of nearly all subreddits that exist on Reddit.
- *StreamApproach*: a list of 500 subreddits is dynamically selected by looking at the newest posts submitted to Reddit.

Implementation Details

I used the Python Reddit API Wrapper (PRAW) to implement these two sampling approaches.

RandomApproach: I first shuffled the list of nearly all subreddits and then looped through them in order to extract the number of subscribers for each subreddit. I break out of the loop once I extract 500 subreddits. The reason I did not directly select 500 subreddits from the list is because this large (possibly outdated) list also included some banned subreddits, and it was not possible to extract the number of subscribers for those subreddits.

StreamApproach: For this, I used the *SubredditStream* class of the PRAW library, which provides submission and comment streams from Reddit. I loop through all the new posts submitted to Reddit using the SubredditStream class and store the corresponding names of the subreddit to a list k. I break out of the loop when I have 500 subreddits in k. Then, I extract the number of subscribers for these 500 subreddits using PRAW's Subreddit class.

After sampling a list of 500 subreddit along with their number of subscribers for both these approaches, I compute two statistics: mean and standard deviation of the number of subscribers.

RESULTS

The values of the statistics computed for both sampling approaches is shown in Table 1. Perhaps unsurprisingly, the mean and standard deviation are significantly different for the two approaches highlighting the importance of data collection methods on the outcomes of a study. As can be seen in the table, the average number of subscribers using the *StreamApproach* is almost a thousand times the average number of subscribers using *RandomApproach*. A similar trend can be seen for the standard deviation in the number of subscribers.

	Mean	**Std-Dev**
RandomApproach	785.6	8439.6
StreamApproach	756530.8	3039260.2

Table 1. The mean and standard deviation of the number of subscribers to the subreddits sampled using the two approaches.

DISCUSSION

As discussed in the previous section, the mean and standard-deviation obtained using both approaches are very different. When we sample using the *StreamApproach*, we naturally extract the more "active" subreddits since we are querying for the newest posts submitted to Reddit. On the other hand, *RandomApproach* gives us a list of completely random subreddits from all of Reddit. This explains the very high average number of subscribers for *StreamApproach* as compared to the *RandomApproach*. Another interesting thing to note is the very high standard deviation for both approaches. This shows that there is a lot of variation in the number of subscribers for different subreddits. Therefore, the sampled list of subreddits obtained using both approaches has some highly popular subreddits and some not-so-popular subreddits.

For the purpose of estimating the 'active' popularity of Reddit, I would argue that *StreamApproach* would be ideal. This is because it looks at the newest posts submitted to Reddit which means that we sample those subreddits that current users engage with. On the other hand, *RandomApproach* can return any random list of subreddits that could possibly be relatively 'inactive' and perhaps even banned. That being said, deciding which approach is better for sampling subreddits will ultimately depend on the end goal of our study.