Detecting why Google Flu Went Wrong by Bias Analysis

Group 15

Student Names:

Anisha Samant

Sheetal Rajgure

Issam (Sam) Tamer

Siwei (David) Ran

School: *UC Davis*

Course Number and Name*: BAX 401 Information, Insight and Impact*

Instructor*: Ashwin Aravindakshan*

Date: *2022-9-29*

Table of Contents

Executive Summary

Google Flu Trends (GFT), developed to detect influenza epidemics using search engine query data, overtly failed in 2013 when it drastically overestimated the flu levels. Even with continuous updates & tweaks in the algorithm since 2009, GFT could only respond to temporary glitches but the larger and actual problem at hand remained unsolved. In this report, we recognize the issues with GFT, analyze why GFT failed and provide mitigations and preventions to safeguard against failure.

Various data, algorithm and user biases contributed to the incorrect predictions. We identify the ways & types of biases that plagued GFT: measurement bias, representation bias, algorithm bias, content production bias, omitted variable bias, population bias and popularity bias. Furthermore, there are other ways GFT could have failed with user interaction bias, presentation bias, sampling bias and red team dynamics coming into the picture. To prevent many of these biases and issues, we suggest Google to increase independent variables in data, supplement the GFT algorithm with traditional research, regularly update relevant keywords and cross-validate on more and recent data points.

Introduction

Google created GFT with an objective of providing real time accurate data monitoring of influenza activity in the US (and later expanded globally) based on certain key searches. Being aware that not every search activity would correspond to a flu case, Google was certain that a pattern would emerge by aggregating the queries that would correctly evaluate the level of flu activity. Although GFT started with reliable predictions, it was not able to correctly estimate the non-seasonal H1N1 outbreak [1]. Soon, the discrepancies began increasing with GFT continuing to overestimate the number of flu cases as online search behaviors changed along with Google's own search algorithm.

Problem Formulation

Using millions of search query terms in their database and identifying the best performing ones, GFT had a high chance to identify keywords that structurally had no relation to the flu. For example, “high school basketball” being a keyword was a strong hint that GFT’s algorithm was overfitting the cases and it would likely not be able to explain future trends [2]. In the next section, we will outline certain biases that affected GFT leading the model to be more prone to failure.

Issues/Biases seen in GFT that led to its Failure

When analyzing the frequency of search terms used without understanding the context or ‘why’s’ behind those searches, then forming a causative relationship with actual cases via a linear algorithm, biases arise when predicting influenza trends. This method was not able to take into account any spikes in search queries related to high media coverage or online trends, nor correctly differentiate any non-flu related seasonal search terms, leading to inaccurate conclusions. The CDC has also reported that a very small percentage of doctor visits for “influenza-like illness” (ILI) actually test positive for influenza [6]. Therefore, Google searches cannot be a reliable identifier for influenza (Measurement Bias).

Google’s database was such that each search query was exact and sequential without combining and changes in linguistics, positions, synonyms or spelling. This made their algorithm prone to content production bias. Semantics, style of writing, spellings can vastly differentiate among different age groups, genders, ethnic backgrounds and geographical areas. This also goes hand in hand with representation bias as when the data set only consists of certain search queries, it lacks the diversity to accurately represent the population or recognize anomalies. It fails to consider factors such as income differences or access to education that directly limit the fairness in data collection. Especially during the time period this model was active, the internet, smartphones and Google were fairly new and not as accessible as they are today. Individuals belonging to lower socioeconomic backgrounds would have lower access to such resources and are less likely to turn to google for medical assistance. Older

populations who are more susceptible to flu might rely more on traditional routes rather than on Google for their diagnosis. If there had to be a situation wherein an outburst of a strain happened in a low-income area, GFT would have again missed the prediction by a great margin. These lead to the population bias in GFT: the user population of the platform (web searchers) is different from the original target population. The actual web searchers could be care-takers or younger digital savvy members of the family and not the people who fell sick or older people who are more likely to get sick.

GFT was directly dependent on Google's search algorithm that underwent many changes during this time period. Google kept modifying its searches to provide better results for the users, and ads became more popular. The fact that their searches started recommending more of what other people were searching for points to popularity bias. The more relevant searches that show up, the more they will be clicked on, making them more popular and highly ranked in their search algorithm. GFT, assuming that the search volume is based on how many people have the flu, is likely to overestimate the flu prediction during times when there is a panic like situation due to news coverage which is a core reason for GFT's overestimation of the prevalence of flu in 2011-12.

The true need for an algorithm like GFT instead of traditional methods used by the CDC was for its ability to make predictions with a shorter lag and to predict during times of uncertainty. GFT heavily relied on large historical search query data from the past five years and measured its performance to select the top performing queries. Ultimately, though, the model was created based on only the 42 top performing key phrases. This represents the sampling bias in GFT, shown when it was not able to estimate ILI during 2009's H1N1 outbreak [3]. The trends in historical data used to create the model could not be accurately generalized for future predictions, especially during unusuality.

Omitted variable bias also existed in GFT. Selected search queries were the only independent variable collected. However, as discussed previously, there are several other factors that should have been taken into consideration when building the model. Those other factors include lagged ILIs, seasonal effects,

population density, and various demographic data (income levels, access to education, race, age, gender, etc.).

Issues/Biases that Could have Occurred

As a generation that experienced COVID, it was interesting to see how the pandemic was politicized and political views were also used as a variable to explain rise or drop in cases. Political campaigns were using COVID to further their agenda and much talk about this was seen on social media. This sort of user interaction bias could have also affected GFT, throwing off the accuracy of its predictions. In the middle of say a presidential campaign during an outbreak, users could have triggered and affected the search queries based on their political bias, with GFT's algorithm having no measure to combat it.

We anticipate that GFT could have potentially faced an issue of red team dynamics in which web searchers intentionally attempt to manipulate the data-generating process to meet their own economic, political or business goals [1]. For instance, if an advertiser has potential business gain through certain types of search queries showing up, they could try to manipulate the process by skewing the search terms used.

GFT could have also faced presentation bias when only recommended search results are displayed on the user's screen. Additionally, if the first few links aren't what the user is looking for, the user might search again -- inflating the number of searches further.

The model was trained with the nine surveillance regions of the United States. However, the system was generalized and rolled out to 29 countries worldwide suggesting that the system might have faced sampling bias in those countries where linguistics, demographics, resources and culture might be completely different to that in the US.

Safeguarding Against Failure

As discussed, the existing GFT algorithm came with many flaws, but many of these flaws can be mitigated or prevented. One primary way is to increase the independent variables used in the data.

Three examples are: data from a more diverse set population (i.e., other social media platforms like Twitter, Facebook, and TikTok), the density of news articles published about flu symptoms, and lagged CDC ILI data. All of these would mitigate some of the biases introduced by the existing algorithm: a broader demographic would be more representative of the population, news articles about an unrelated symptom could control for keyword usage spikes, and lagged CDC ILI data would help mitigate some of the remaining lurking variables. However, including more big data should be done with caution to avoid missing key points of context.

An additional solution would be to include more traditional research to supplement GFT.

By supplementing the GFT algorithm with traditional research, key context can be added that would mitigate inaccurate predictions during unusual times. A key example would be the spread of COVID. During such an unprecedented event (due to the lack of extensive data available during previous global pandemics), a data-based algorithm would have no frame of reference to accurately predict spread. In this case, supplementing GFT with data from experiments, such as samples of population density in retail stores, would more accurately account for anomalies than the online data could alone.

Google is also recommended to regularly recalibrate and update their model with more relevant keywords to correct outdated misinformation. Regular cross-validation of their model on more and current data points will also help reduce overfitting.

Conclusion

GFT's algorithm accurately modeled the *correlation* between keyword searches and flu infection rate, but this quickly became unreliable when the true causes of flu infection did not follow the same trend. Limiting GFT's algorithm to a single, correlated variable introduced numerous different types of biases, which ultimately undermined the integrity of its predictions. While some of these shortcomings could be mitigated by introducing more internet-collected data, a truly accurate and longstanding model for flu cases cannot exist without the context of external, experiment-based data.

References & Notes

1. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (1970, January 1). *The parable of google flu: Traps in big data analysis*. Science. Retrieved September 28, 2022, from https://gking.harvard.edu/publications/parable-google-flu%C2%A0traps-big-data-analysis

2. Ginsberg, j., Mohebbi,M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009, Feb 19). *Detecting influenza epidemics using search engine query data*. Nature News. Retrieved September 28, 2022, from https://www.nature.com/articles/nature07634#online-methods

3. Butler, D.(2013, February 13). *When Google got flu wrong*. Nature news. Retrieved September 29, 2022, from https://www.nature.com/news/when-google-got-flu-wrong-1.12413

4. Walsh, B. (2014, March 13). *Google flu trends failure shows drawbacks of Big Data*. Time. Retrieved September 29, 2022, from https://time.com/23782/google-flu-trends-big-data-problems/

5. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021, July). A Survry on Bias and Fairness in Machine Learning. ACM Computing Surveys. Retrieved September 29, 2022, from https://dl.acm.org/doi/pdf/10.1145/3457607

6. Salzberg, S. (2014, March 23). *Why google flu is a failure*. Forbes. Retrieved September 29, 2022, from https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/?sh=427dba9e5535

Appendix

While Brownstein's approach of using volunteers to self-report symptoms instead of physician data is great, it is not representative of the population. Sampling Bias refer to when some individuals of a population are systematically more likely to be selected. In this approach, people who do not want to

report due to privacy concerns would not self-report to the organization, which means those people will not be considered as data. Additionally, people who do not have access to Boston Children's hospital will most likely not know and participate in this system. So, the sampling might only represent the population who are local citizens.