Search Auto-Completes Related to COVID-19 Yield Different Results in English and Spanish

Vivek Singh, PhD, Ishaan Singh, and Pamela Valera, PhD, MSW

Rutgers University

The coronavirus (COVID-19) has become a global epidemic, surpassing over 5.9 million cases and 300,000 deaths worldwide as of May 31, 2020, as per the World Health Organization. The global coronavirus pandemic has resulted in sharp increases in online search activity about the disease, its spread, and remedial actions. Hence, search engines can significantly influence public perception of the disease and the efforts undertaken by the public.

We are examining “autocomplete” results, a search engine feature that allows users to complete searches faster by populating information in the search engine’s text box as they type. When a user types “coronavirus is,” Google suggests autocompletes such as “coronavirus is it contagious”, “coronavirus is man-made,” etc. The function, while convenient, may contribute to bias that, if left unchecked, has the potential to lead to health inequality experienced by marginalized and racial minority groups by providing different results for similar inquiries.

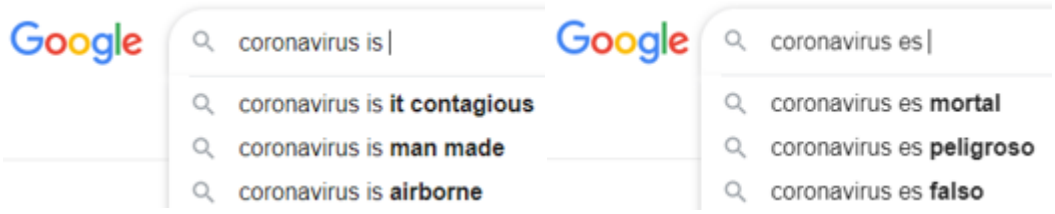

Figure 1: Differences across languages in search engine auto-completes pertaining to COVID-19

Preliminary investigation, which led to this National Science Foundation funded project, showed that the auto-completes for terms related to COVID-19 were quite different in English and Spanish (See Figure 1). For instance, in the first three auto-complete results for "coronavirus is ...," the English version had one question and two neutral-sounding statements. However, for the Spanish version "coronavirus es ..." all of the top auto-completes were unfavorable. We found similar differences in multiple COVID-19 related terms. Hence, we decided to study these differences systematically over time.

We have collected the autocomplete results for multiple COVID-19 related query terms since Mar 12, 2020, daily using Google’s search API, which allows one to specify the language for search (e.g., English, Spanish). In this article, we share the results based on 60 days of data (Mar 12, 2020- May 10, 2020) collected on a computer in the New York metropolitan area. The query terms included those related to the overall situation (“Coronavirus is ...”, “Flu is …”, and “Pandemic is …”) and those pertaining to precautions

("Hand sanitizer is …", "Hand washing is …", and "Face Mask is …"). There were between 0 and 20 autocompletes found for each of the terms on each of the days.

We placed a sentiment score (e.g., positive, negative, or neutral) for each of the autocompleted queries. The queries which were likely to evoke positive emotion in the reader (e.g., "coronavirus is curable") were assigned a positive sentiment label and those likely to evoke negative emotion in the reader (e.g., "coronavirus is deadly") were assigned a negative sentiment. This score was computed automatically using Amazon's sentiment recognition software, which returns the most likely sentiment for the text. The scores were averaged for each term using +1 for positive, -1 for negative, and 0 for neutral autocompletes, and these averages were tracked over time. Each of the query terms was also translated into Spanish, double-checked for accuracy, and passed on to the Spanish language version of the Google search API. The resulting autocompletes were similarly labeled for the sentiment.

We have compared the sentiments of the autocomplete queries in English and Spanish. See Figure 2(A) for results for terms "Coronavirus is …" and "Coronavirus es …" and Figure 2(B) for sample results for terms "Hand sanitizer is …" and "la desinfectante de manos es …". The blue line represents queries in English and orange lines are for queries in Spanish.

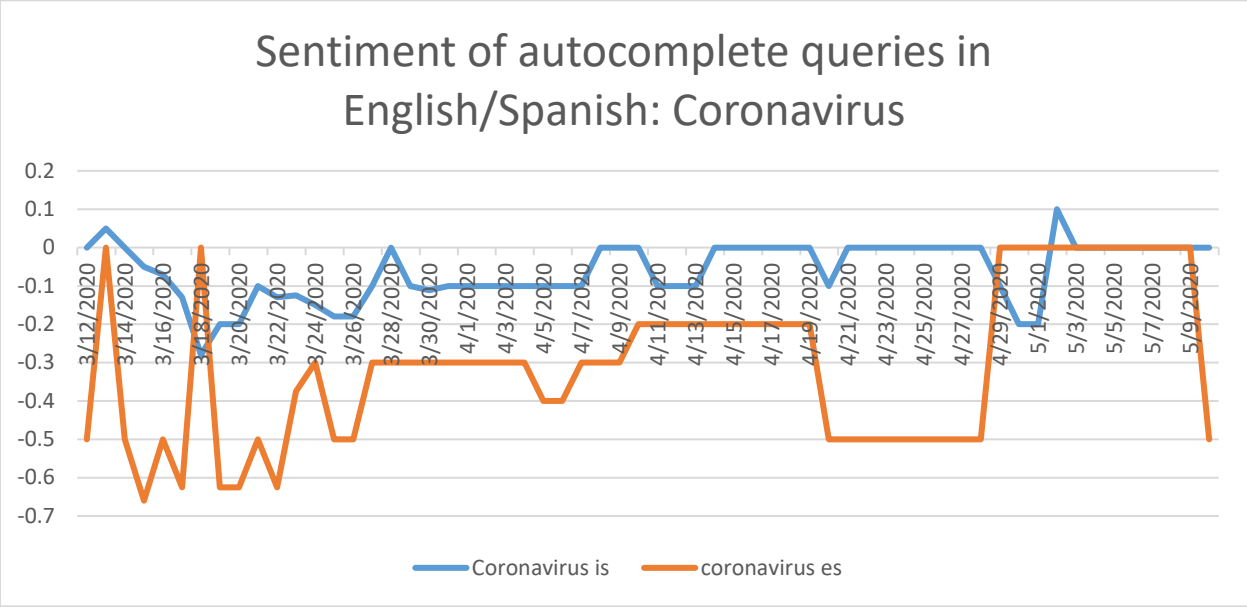

Figure 2(A): Sentiment of autocomplete queries in English and Spanish for "Coronavirus is …"

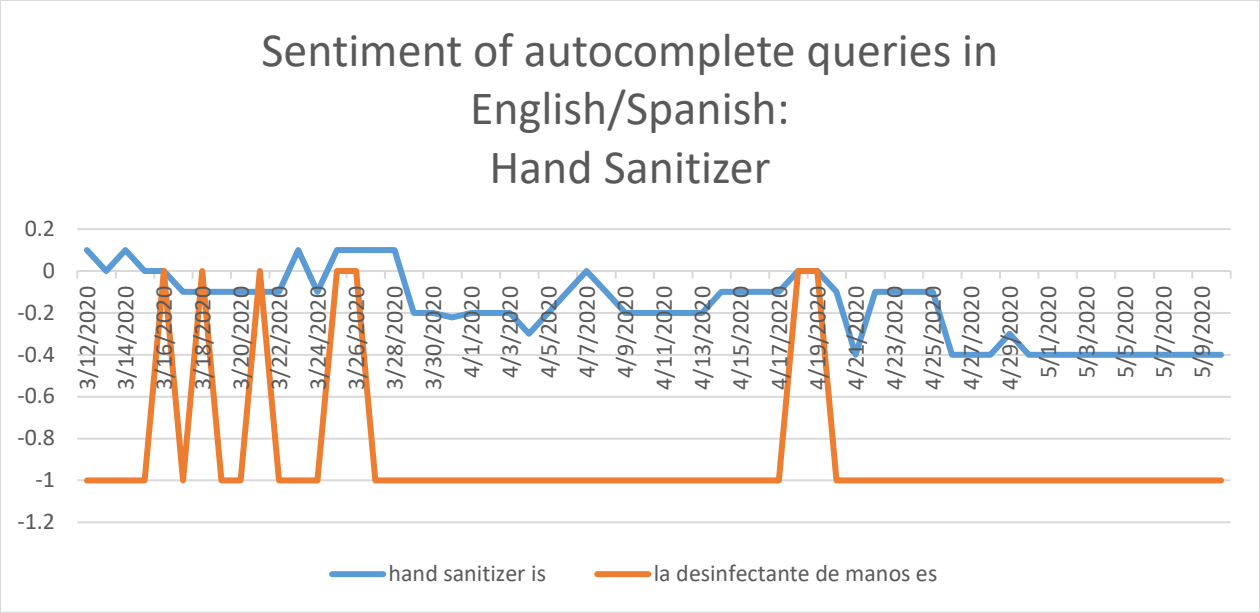

Figure 2B: Comparison of sentiment scores for autocomplete queries in English and Spanish: for "Hand sanitizer is …"

The major findings from the analysis were as follows. Most autocompletes yielded neutral to negative autocompletes. The average sentiment score for the abovementioned six terms in the two languages was -0.29. This is not surprising given the negative impact of the pandemic on all walks of human life. In addition, the *Spanish autocompletes were generally more negative than English autocompletes for the same terms*. The average sentiment score for English was -0.11 for English queries and -0.47 for Spanish queries. Lastly, *the sentiment fluctuated a lot more in Spanish autocompletes than in English autocompletes*. The median, standard deviation of sentiment scores in English, was 0.11, and in Spanish, it was 0.27.

*As such, Spanish speaking users may be getting different health information despite facing the same underlying health risks***.** For instance, a Spanish-speaking user who sees negative autocompletes for *hand sanitizer* may adopt a very different approach than an English-speaking user who sees more neutral autocompletes. Similarly, a Spanish-speaking user who sees large fluctuations in positive and negative sentiments for different health-related terms is more likely to be confused about taking action (e.g., using hand sanitizers) than someone who gets a more consistent message.

Varying autocomplete queries may influence users to think contrarily about the pandemic and take different precautions. Hence, these results suggest that urgent action is needed to counter information bias. These biases can provide unequal access to health care services to various segments of society. Little attention has been paid so far to unequal access to health information mediated by computer algorithms. We urge public health stakeholders and the search engine organizations to continuously audit autocompletes for potential bias and remove disparities in health information dissemination.

About the Authors:

Vivek Singh, PhD, is a faculty member in the School of Communication and Information at Rutgers University. Email: v.singh@rutgers.edu

Ishaan Singh is a graduate student in the Computer Science Department at Rutgers University.

Pamela Valera, PhD, MSW, is a faculty member in the School of Public Health at Rutgers University and a member of Cancer Health Justice Lab.