The University of Vermont

CSYS 300: Principles of Complex Systems

Analyzing Parent's Conversation Surrounding Vaccination on BabyCenter

Megan Ardren, Spencer Dooley, Alex Friedrichsen, Carter Ward

January 21, 2022

Contents

1 Introduction

Vaccine hesitancy is an important subject both politically and practically in the safety of our society. Our project seeks to better understand the language and vocabulary differences between groups who differ in their stances towards vaccines. We use the online website BabyCenter, a website for parents to discuss early childhood rearing, to collect three dictionary samples: a general-discussion sample, a general vaccine-centered discussion sample (with a pro-vaccine bias), and a vaccine hesitant (anti-vaccination bias) sample. We use these samples to analyze the sentiment of all posts within our samples and determine a sentiment score for each group as well as the words contributing most heavily to that score. Moreover, we create rank-divergence comparisons of these samples using allotaxonographs. Our general sample collects posts from 2018-2020 whereas our other two samples are up to present. This time differential allows more meaningful comparison of the changes in attitudes towards vaccines over time. These comparisons give insight into the different vocabulary and sentiment present in groups with different perspectives on vaccines the type of which we discuss in our results section. The importance of this research is clear: an understanding of the language differences between two polarized groups is a basis for changing one of the group's opinions. Opinion change is well researched to be more frequent when new information is delivered from a source that one can identify with. Language is a key barrier to identity that stands between two opinions. Our project can potentially jump start the path to influencing more stubborn anti-vaccination individuals.

2 Methods

2.1 Data Collection

We collect data from conversations surrounding vaccines in order to get at the issue we inspect: vaccine hesitancy. We chose to collect data from the site BabyCenter, an online forum angled towards providing information about early childhood development and parenting as well as open discussion. We collected text data from several different groups on the site with an aim to find general, hesitant, and non-hesitant samples. We first identified potential groups that might fit these requirements by using BabyCenter's built in search function and keywords such as *vaccine*, *vaccination*, *pro-vax*, *anti-vax*, and others.

One challenge we discovered was that several of the groups we wanted to sample from were private and wouldn't be accessible by our forthcoming data scraper program. In order to mitigate this issue and ensure a large enough sample, we accessed the data from BabyCenter in two main scripts. The first archives the "birth-club" discussion groups for months between 2018 and 2020 which comprises our general sample. This script accesses historical site records. The second script scrapes data from ongoing discussion group posts including the groups in our hesitant and non-hesitant samples. The hesitant sample includes the groups "alternative_parenting" and "crunchy_mamas" while the non-hesitant sample includes the groups "pro_vaccination_immunization" and "pro_vaccination_support_and_info". The text from each of the discussion threads from within the

communities is cleaned and then stored in a Firebase (no-SQL) database in a hierarchy as shown below.

Hierarchy of Firestore

$$Sample \rightarrow Community \rightarrow Thread \Rightarrow Comments \rightarrow (Author/Date/Text)$$

The Python libraries used in the data pipeline include Spacy library, a natural language processing library, and several usual Python libraries such as Numpy, Pandas, etcetera. In total our data collection process resulted in several thousands of words in our dictionaries for visualization.

2.2 Allotaxonographs

To visualize the difference in vocabulary used between two different systems, or in this case, different groups of parents, we utilized allotaxonographs. The text from all posts and comments on each post were collected and split into a list of all words used in that group's online discussions. From this list of words used, vocabulary (unique words) and their frequencies within each separate group. At this point, we can utilize the power of allotaxonographs to compare these group's varying vocabularies.

Allotaxonographs show the rank of a word in both systems, where each axis represents the rank of that word in each system. The closer to rank 1 the word is, the more talked about that word is in that system. Each cell in the allotaxonograph is colored by the number of words there are per that cell, this conveys lots of words are mentioned only a couple of times, where as only a couple of words are mentioned repeatedly. Allotaxonographs have the potential to highlight differences in vocabularies between two different groups as well as show which words are more or less popular within a certain population.

2.3 Sentiment Shifts

The sentiment shift graph is a type of word shift graph, a family of graphs where words are weighted according to a lexicon of scores and displayed in a horizontal bar chart format. When working with and analyzing the results of word shifts, it is important to keep in mind that they work best on larger dictionaries of text. Our data is sufficiently large to use word shifts without worrying about the methodology, herein weighted averages, behind them.

We construct our word shift graphs using the Shifterator Python package developed by Gallagher et al [1]. Shifterator greatly simplifies the creation of word shift graphs. After data collection, we load the hesitant posts, non-hesitant posts, and general posts into separate text dictionaries. We then load our lexicon; the sentiment shift graphs we implement use the English labMT lexicon developed by the Computational Story Lab at the University of Vermont [2]. This lexicon was constructed by finding the top 5000 most used words on several popular media sources, then feeding the resulting list of 10022 words to workers using Amazon Mechanical Turk (MT) where workers were asked to rate the happiness of words on a scale from 1 to 9. This lexicon comes preloaded inside the Shifterator package. We proceed to create a sentiment_shift object using the WeightedAverageShift function by passing in the combinations of text dictionaries we want to compare, the English labMT lexicon, and a stop_lens equal to (4,6). This stop_lens excludes words with neutral emotional valence

from our graphs, so that we may only look at the relative tails of our distribution. The sentiment shift graph is useful in its ability to provide clear and easily comprehensible comparison of emotional valences for words in two text dictionaries. One is able to pick out which words carry the most emotional weight and can make further inferences from there.

3 Results

3.1 Exploratory Analysis

To get a more in-depth look into each group corpus we graphed the distribution of valence scores for each of the threads (post + replying comments), as can be seen in Figure 1. We found that all of the distributions were normal and centered around 4.98 with a very small range of 4.94 to 5.05. One reason we could be seeing such neutral sentiment is that all word's that were able to be scored were included. If we were to only include words that had a high or low valence (ie. less than 4 or greater than 6) then we might be able to counteract the majority of words that are neutral from drowning out the polarizing sentiment words in the overall sentiment score for the thread.

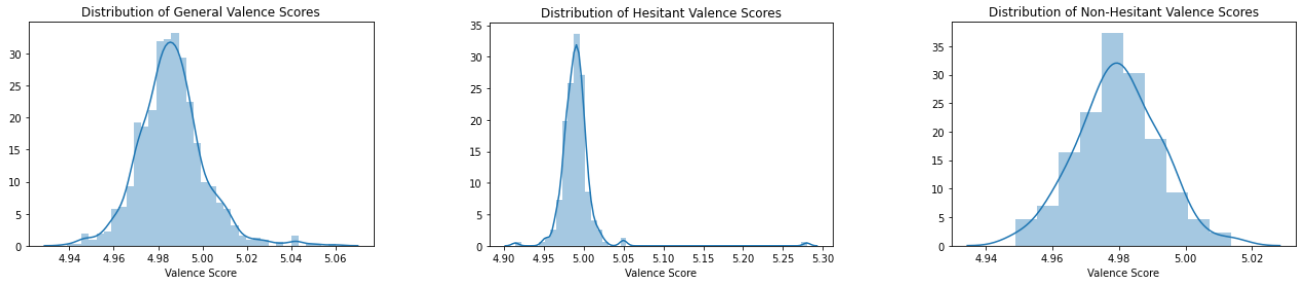

Figure 1: Distribution of Thread Valence for Each Group Corpus

3.2 Allotaxonographs

Below, figures containing allotaxonographs are our three different sample comparisons, comparing each of our three sample against each other. Figures 2-4 are comparison using all posts and comments within each group, whereas Figures 5-7 are comparisons using an anchor word of vaccine (so a more vaccine-specific sample). The first figure shows the non-hesitant sample vs. the general sample. We see these two systems are not completely symmetrical, and words such as *vaccine*, *vaccines*, and *vaccination* are some of the most talked about words for the non-hesitant sample where as the general sample does not possess these words. One thing to keep in mind is that, while our non-hesitant sample was taken directly from vaccine-centered discussion boards, our general sample was take from discussions that were not necessarily vaccine-centric. The second allotaxonograph shows the hesitant sample vs. the general sample. Similarly to the previous comparison, we see some more specific vocabulary being used in the hesitant sample both vaccine related or not such as, *vaccine*, *vaccines*, *organic*, and *gluten*. The hesitant sample was taken from discussions

which were not necessarily about vaccines, but ones that were apt to mention alternative medicines/treatments. For our final comparison, between the hesitant and non-hesitant groups we see interesting behaviour. One thing that stands out is that the words *vaccine* and *vaccines* are spoken about slightly more in the non-hesitant group, which can likely be attributed to these posts being specifically about vaccines whereas the hesitant group ones are not necessarily vaccine-specific in this comparison. While that is true, both of these words are spoken about more overall between the two systems then they were before. From these visualization, we see both groups talking heavily about the topic of vaccines, we will now investigate the sentiment shifts around these words.

For the second set of allotaxonographs (Figures 5-7), we see similar relationships overall. In Figure 5, comparing the non-hesitant and general samples again we see the systems are overall 'similar' as neither diverge from the center line by much. Moving to the vaccine-specific comparison of the hesitant sample and the general sample, we see somewhat similar systems once again. In the final vaccine-specific comparison, between the hesitant and non-hesitant sample, the systems are slightly dissimilar, interestingly the word *vaccine* was more talked about in the non-hesitant group.

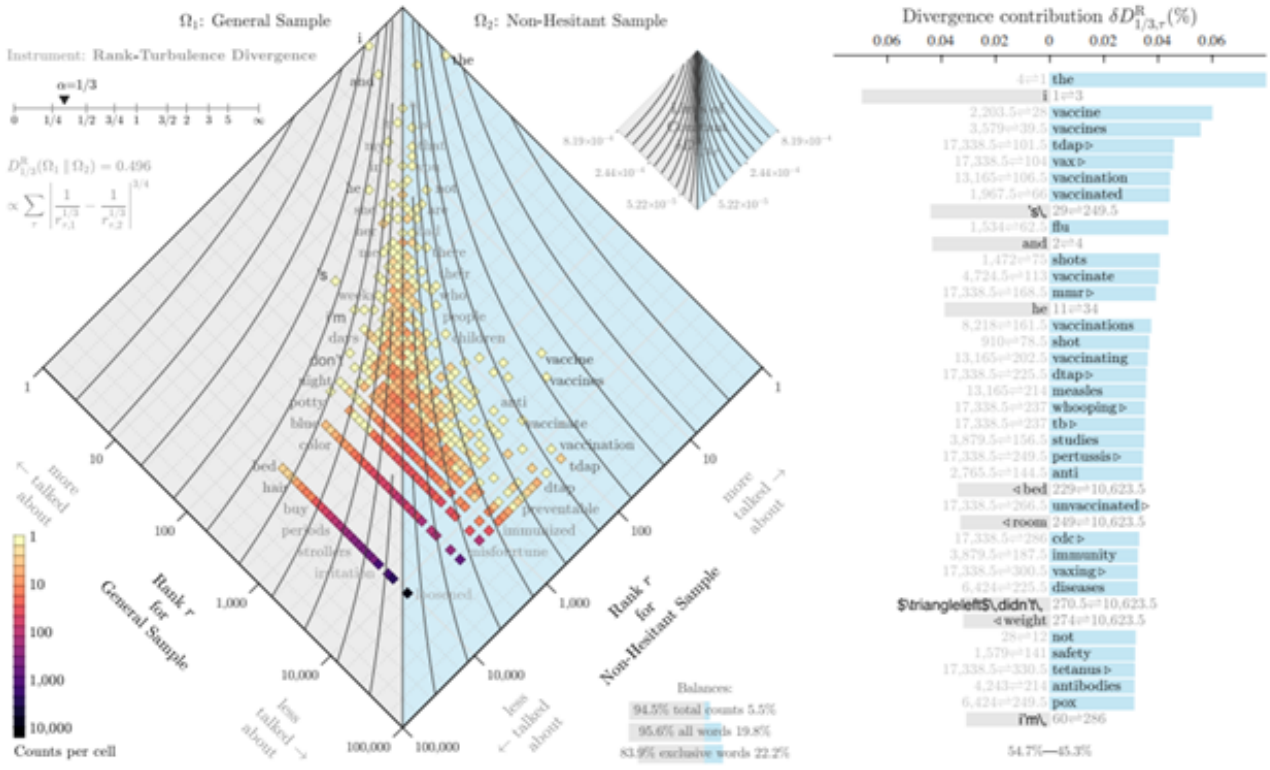

Figure 2: Non-Hesitant Sample vs. General Sample

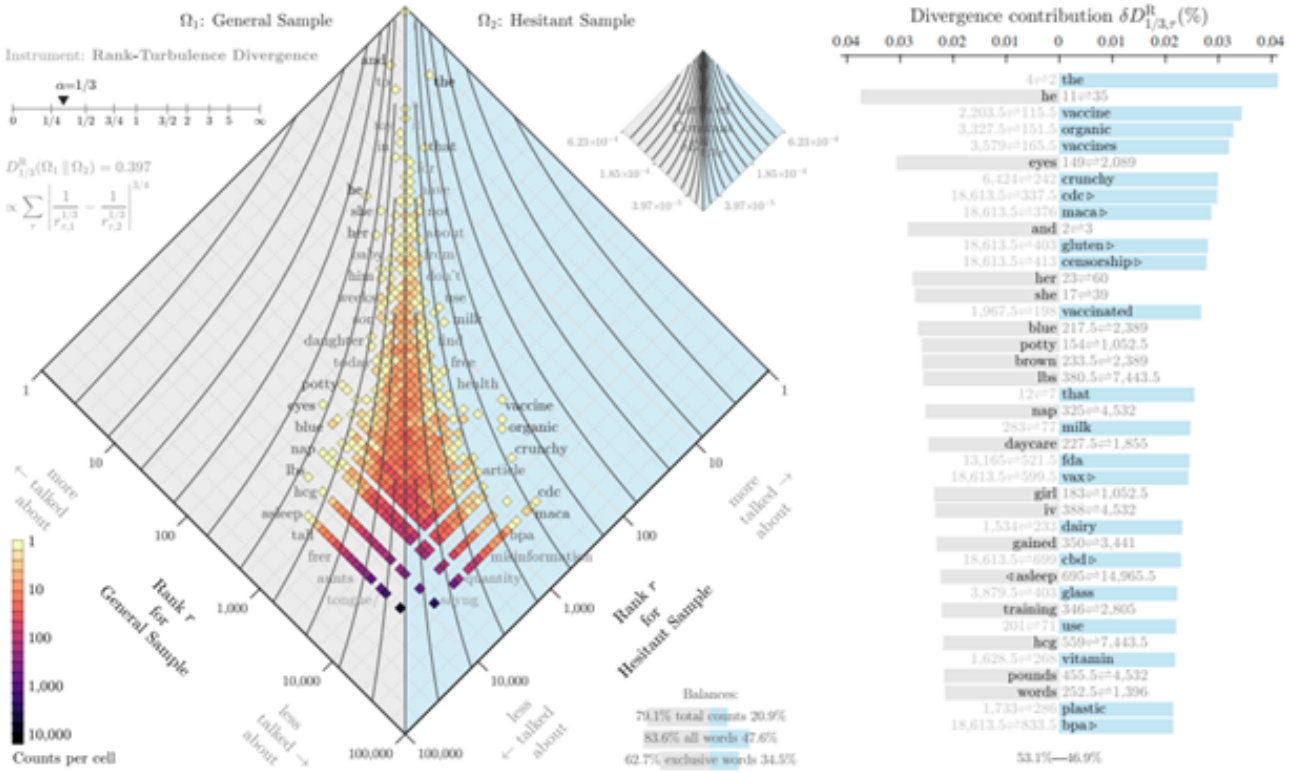

Figure 3: Hesitant Sample vs. General Sample

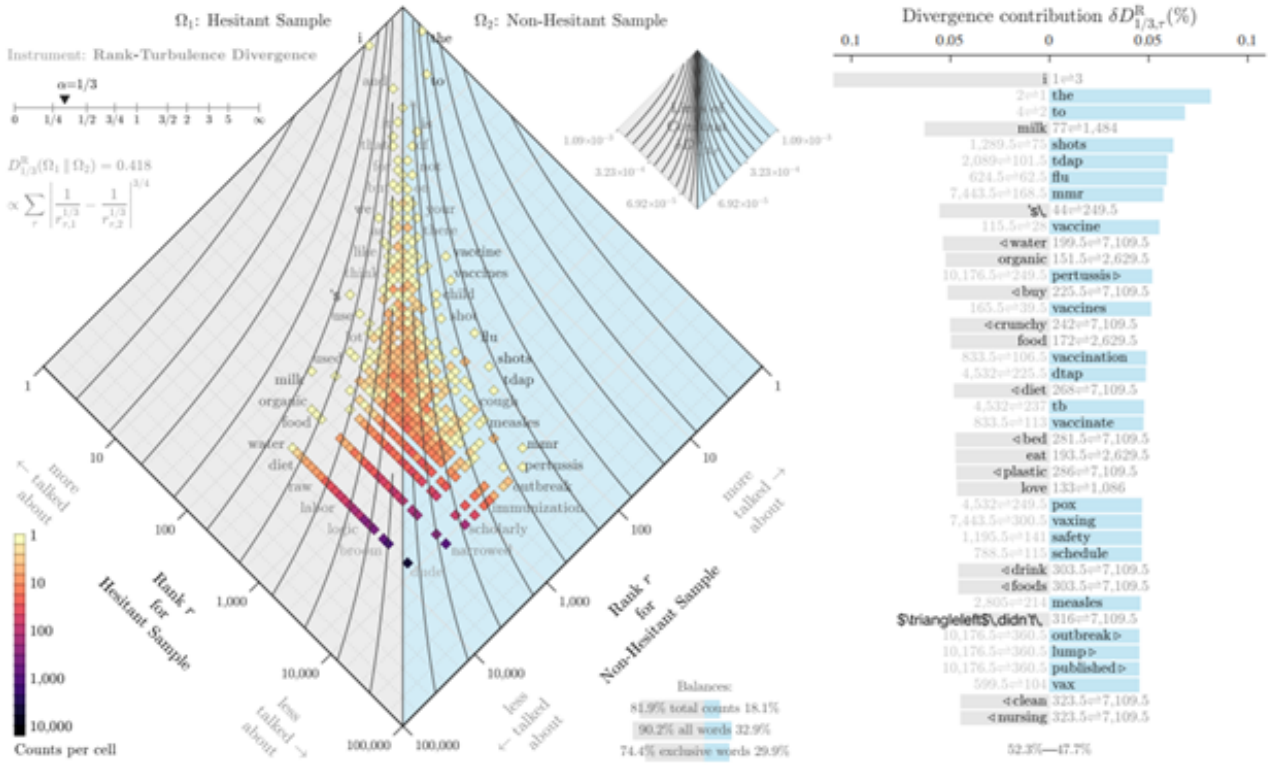

Figure 4: Non-Hesitant Sample vs. Hesitant Sample

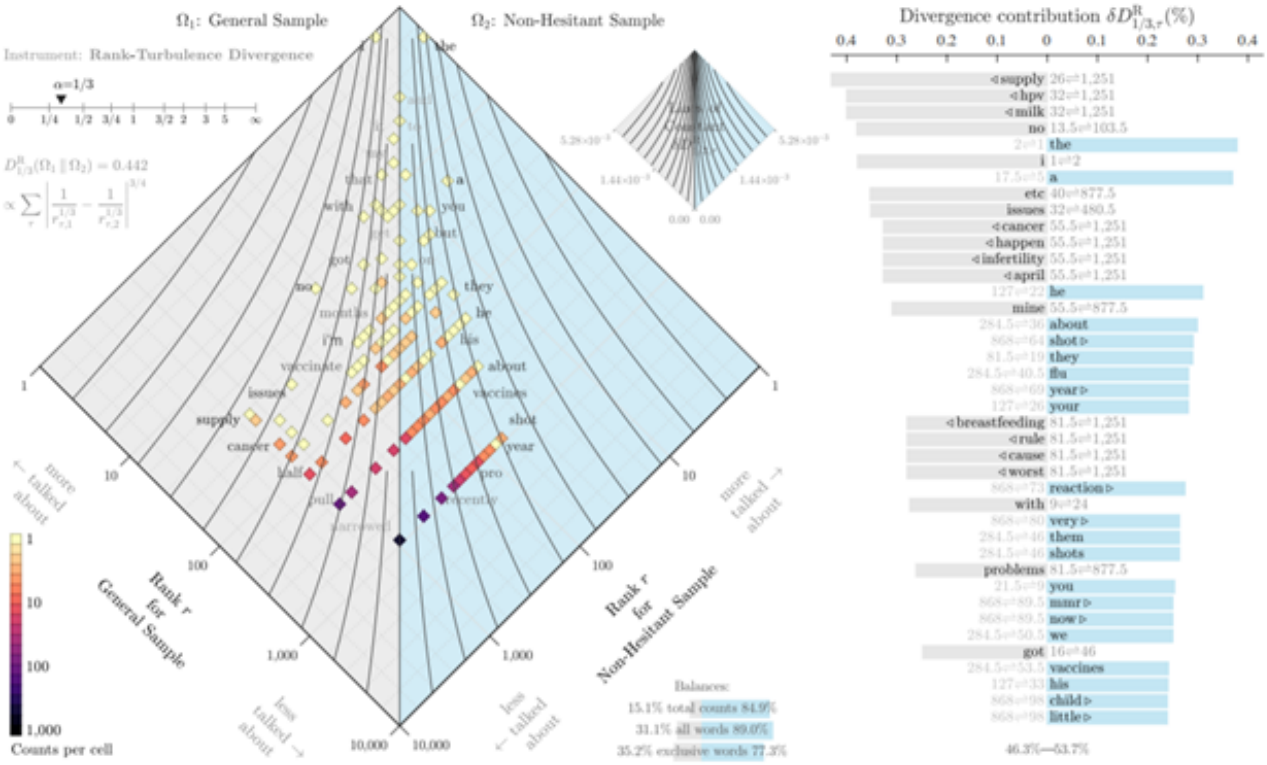

Figure 5: Non-Hesitant Sample vs. General Sample (Vaccine-Specific posts)

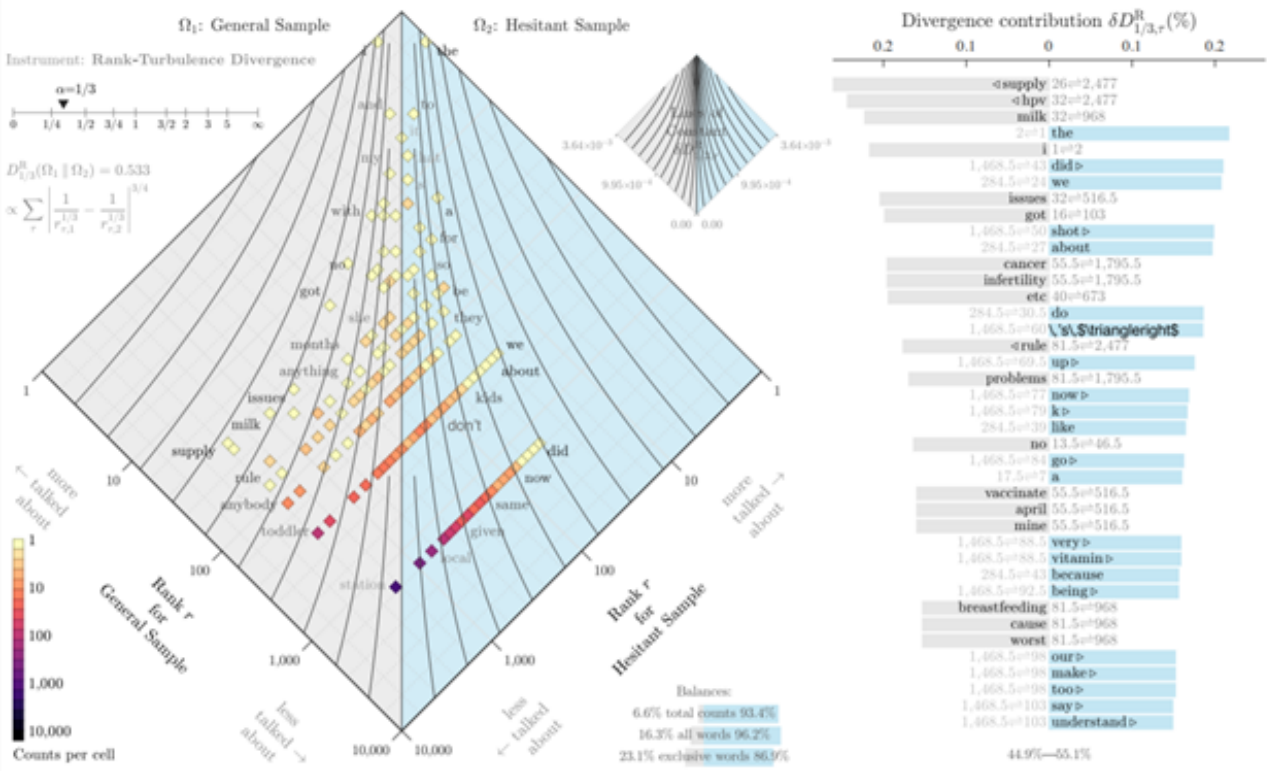

Figure 6: Hesitant Sample vs. General Sample (Vaccine-Specific posts)

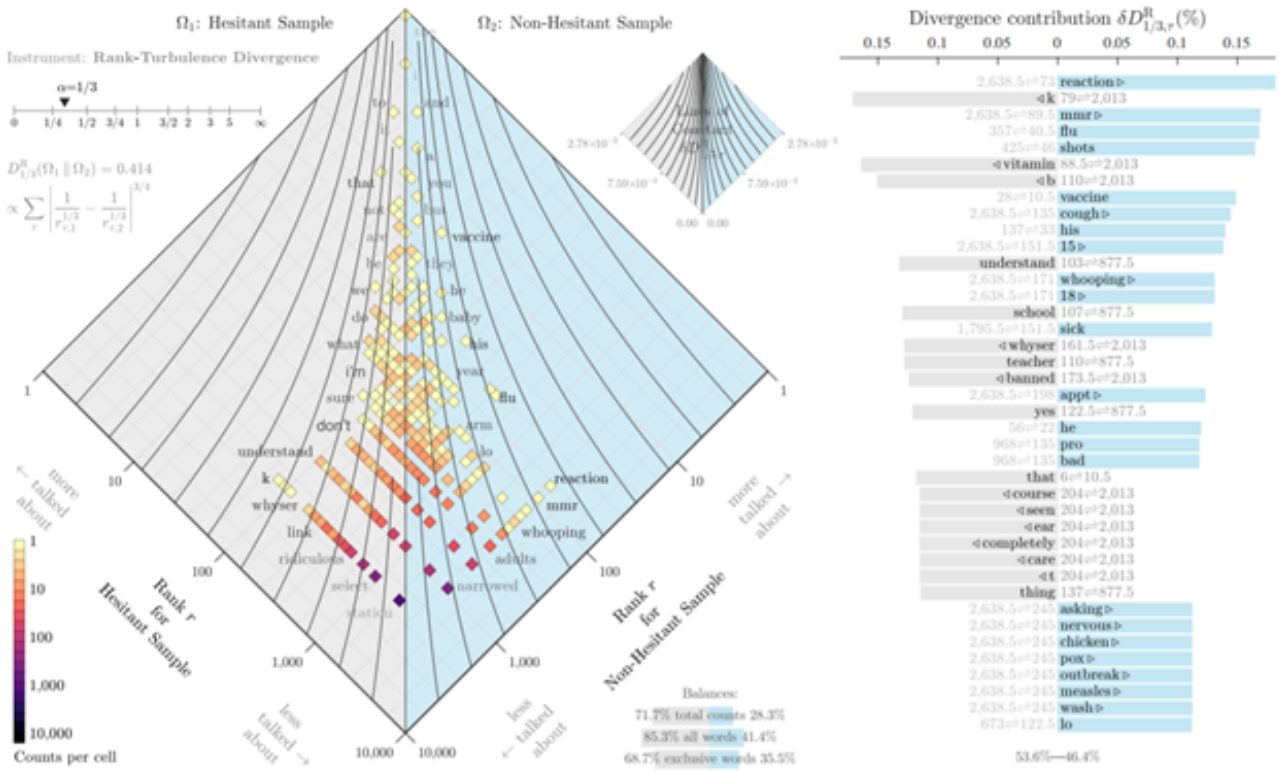

Figure 7: Non-Hesitant Sample vs. Hesitant Sample (Vaccine-Specific posts)

3.3 Sentiment Shifts

Using the Shifterator python package [2] we created figures that display the differences in frequency of words that contributed most to the difference in sentiment between two corpuses. This allows us to explore words and topics that are more present and significantly different in terms of sentiment in conversation within groups that are labeled vaccine hesitant, vaccine non-hesitant, and a general sample from the website.

From the sentiment shift between vaccine hesitant groups and vaccine non-hesitant groups (Figure 5) we observe that the sentiment of words used in the threads on non-hesitant boards has a lower sentiment (5.69) than the hesitant boards (5.99). By looking at the words in the figure it is clear that words associated with vaccination and disease are more frequent in the non-hesitant group. These words have negative sentiment which is what is driving the difference in sentiment between the groups.

From the sentiment shift between vaccine hesitant groups and general conversation groups (Figure 6) we observe that they have almost identical average sentiment. We see that there are some words that are more frequent in the hesitant group that have to do with vaccination such as shot and disease, but there are also words that appear in other topics discussed in the hesitant group such as organic and natural.

In the sentiment shift between vaccine non-hesitant groups and general conversation groups

(Figure 7) we observe very similar results to the comparison of non-hesitant groups and hesitant groups (Figure 2). By looking at the words in the figure it is clear that words associated with vaccination and disease are more frequent in the non-hesitant group. These words have negative sentiment which is what is driving the difference in sentiment between the groups.

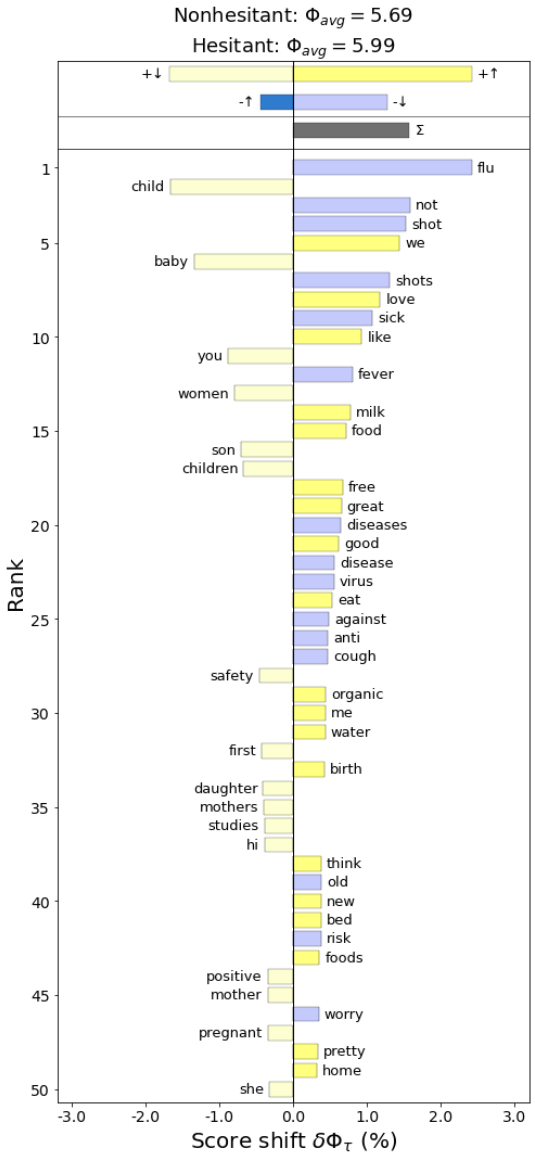

Figure 8: Sentiment Shift Between Vaccine Hesitant Groups and Vaccine Non-Hesitant Groups (The +/- shows the word is more/less happy than the average sentiment of the nonhesitant group. The upwards arrow indicates that the word is more prevalent and the downwards arrow indicates the word is less prevalent in the non-hesitant group than the hesitant group.)

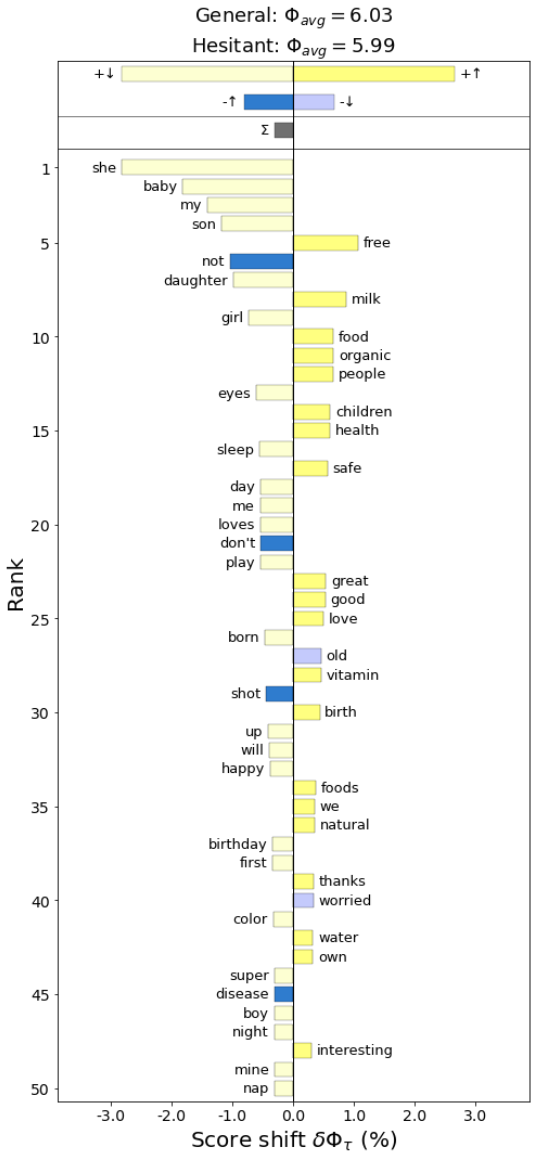

Figure 9: Sentiment Shift Between General Groups and Vaccine Hesitant Groups (The +/- shows the word is more/less happy than the average sentiment of the general group. The upwards arrow indicates that the word is more prevalent and the downwards arrow indicates the word is less prevalent in the hesitant group than the general group.)

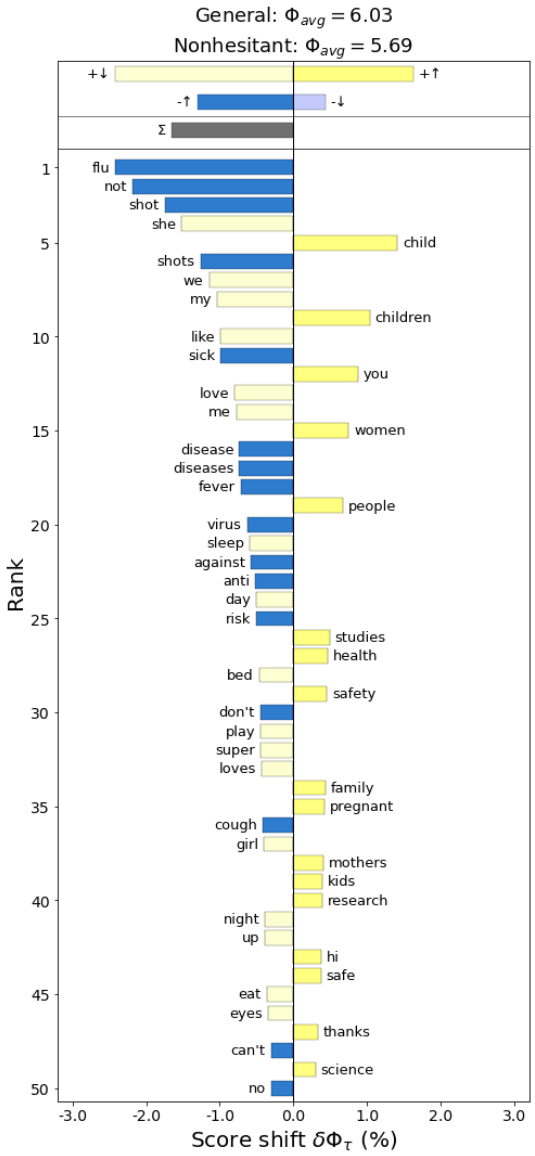

Figure 10: Sentiment Shift Between General Groups and Vaccine Non-Hesitant Groups (The +/- shows the word is more/less happy than the average sentiment of the general group The upwards arrow indicates that the word is more prevalent and the downwards arrow indicates the word is less prevalent in the non-hesitant group than the general group.)

To get a more focused view, we created subsets of the hesitant and non-hesitant groups that only included threads where the original post contains the word "vaccine".

The word shift graph in Figure 8 shows which words contributed most to the shift in sentiment between the vaccine specific subset vaccine hesitant groups and vaccine non-hesitant groups. The total average sentiment of the nonhesitant text is 5.61 while the total average sentiment of the hesitant text is 5.82. The word "flu" contributes most to the change in sentiment. It appears that people in the nonhesitant groups are talking more about health and the process of vaccination like

"flu", "shots", "sick", and "virus", which are all negative words. Conversation about vaccines in the hesitant group appears to be more broad, including words like "understand", "love", "teacher", and "school".

The word shift graph in Figure 9 shows which words contributed most to the shift in sentiment between our general sample and vaccine specific subset of the hesitant groups. The total average sentiment of the general text is 6.03 while the total average sentiment of the hesitant text is 5.82. The increase in frequency of negative words "shot" and "not", and the decrease in the positive word "she" contribute most to the change in sentiment.

The word shift graph in Figure 10 shows which words contributed most to the shift in sentiment between our general sample and vaccine specific subset of the non-hesitant groups. The total average sentiment of the general text is 6.03 while the total average sentiment of the non-hesitant text is 5.61. This change in sentiment is mostly influenced by the increase in frequency of vaccine related words which are scored negatively, such as "flu", "shot", and "virus".

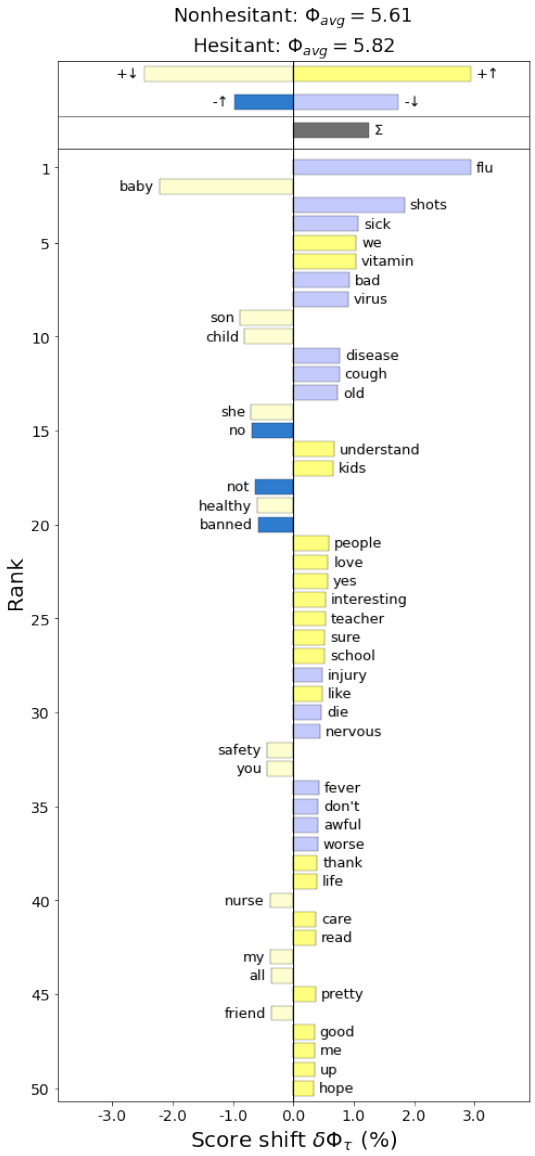

Figure 11: Sentiment Shift Between Vaccine Hesitant Groups and Vaccine Non-Hesitant Groups Only Including Posts with the Word "vaccine" (The +/- shows the word is more/less happy than the average sentiment of the nonhesitant group The upwards arrow indicates that the word is more prevalent and the downwards arrow indicates the word is less prevalent in the hesitant group than the non-hesitant group.)

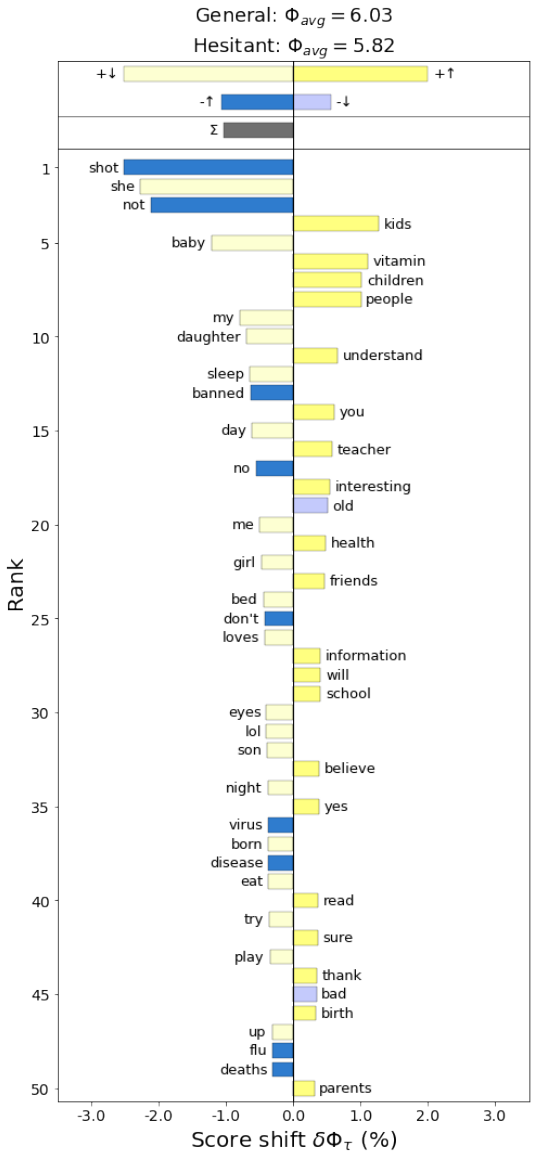

Figure 12: Sentiment Shift Between General Groups and Vaccine Hesitant Groups Only Including Posts with the word "vaccine" (The +/- shows the word is more/less happy than the average sentiment of the general group. The upwards arrow indicates that the word is more prevalent and the downwards arrow indicates the word is less prevalent in the hesitant group than the general group.)

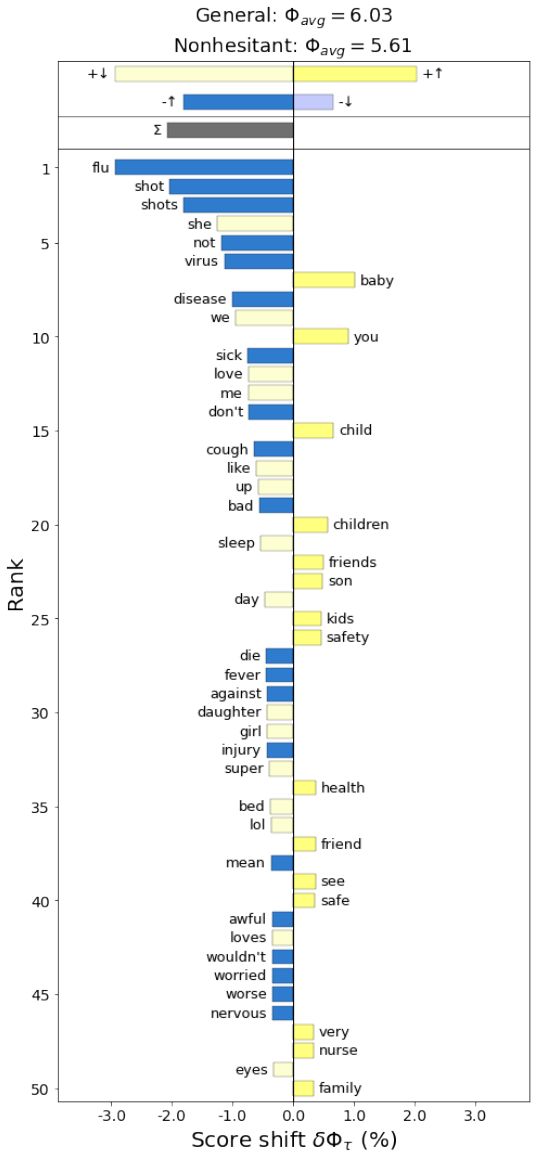

Figure 13: Sentiment Shift Between General Groups and Vaccine Non-Hesitant Groups Only Including Posts with the word "vaccine" (The +/- shows the word is more/less happy than the average sentiment of the general group The upwards arrow indicates that the word is more prevalent and the downwards arrow indicates the word is less prevalent in the non-hesitant group than the general group.)

4 Discussion

The results of our project gave us insight into language and sentiment of online discussion groups with differing opinions surrounding vaccination and has given us many ideas and methods to explore further.

We ran into the challenge of comparing groups of different nature. There are groups on BabyCenter whose only purpose is to be a place of discussion about vaccination that is specifically non-hesitant (ex. Pro-Vaccination Immunization). Unfortunately there are no groups of this type for vaccine hesitant conversation. We have found groups where there is vaccine hesitant discussion, but it is not the only/main topic of the group. This was apparent by looking at the words in the figure it is clear that words associated with vaccination and disease are more frequent in the non-hesitant group. These words have negative sentiment which is what is driving the difference in sentiment between the groups.

Based on these results we created subsets of the hesitant and non-hesitant groups that only included threads where the original post contains the word "vaccine". We found that both groups had lower average sentiment, and that the hesitant sample had a greater drop. In the future we want to create the subset using more anchor words associated with vaccination in addition to 'vaccine' to make sure we are encompassing as many posts related to vaccination as possible.

One trend that we see in the sentiment shifts is a higher frequency of words that directly relate to vaccinations in the nonhesitant conversations, it is possible that conversation in the nonhesitant groups is centered more about the vaccine process while conversation in the hesitant groups is broader.

Based on some of the higher frequency words in the hesitant conversation such as "teacher", "school", "understand", they might be talking more about the social repercussions of not having their children vaccinated.

In the allotaxonographs we see varying vocabulary frequencies between different groups with different opinions surrounding vaccines. Since allotaxonographs are a tool we can use to visualize the words used the most within certain groups, the vocabulary choice is very important. While this allowed us to see which groups used vaccine-specific words more when compared to each other, We also saw many common English words appear in our allotaxonographs such as (the, did, so) which clouded our analysis. We originally did not remove these common words because the initial analysis was meant to look at the frequency of *all* words pertaining to a certain group, however, removing these stop-words could provide us a more meaningful analysis in the future.

References

[1] Frank M. R. Mitchell L. Schwartz A. J. Reagan A. J. Danforth C. M. Dodds P. S. Gallagher, R. J. Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts.

[2] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE 6, no. 12*, 2011.