COVID, Conflict, and Reddit: An Application of Topic Modeling

Annabelle Strong ‘23 | Advisor: Smaranda Sandu

Background and Research Questions

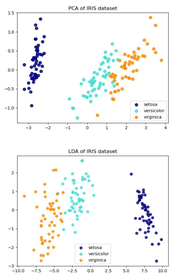
The boom of the social internet over the last two decades has ushered in a new era of complex public interactions. With a global network available to anyone with a computer, social dynamics are becoming ever more visible and quanti!able through social media sites like Reddit. This study aims to examine a subset of such data, Reddit’s r/AmITheAsshole (hereafter referred to as AITA). AITA is a subreddit in which individuals who may be in the wrong write about their situation and get an assessment of blame from other users. Given the acute popularity of AITA and the crowdsourced moral code it produces, the subreddit provides a uniquely accessible and authentic insight into the types of conflict experienced by users. This study aims to detect and quantify patterns in these categories over the last !ve years, if any exist, to better understand the impact of worldwide social shifts necessitated by the COVID-19 crisis on social discord.

To accomplish this goal, this study parses submission dumps for each month between and including June 2016 and June 2021 from the PushShift archive. After extracting the posts of interest—namely AITA posts with the original submission text intact—we will apply topic modeling techniques to identify common sets of language both across and within the months and years in question. These topic sets will then be compared with respect to time, allowing us to discover and examine any latent patterns.

Research Questions

1. **What were the most common topics in the dataset?**
2. **How similar are these topics across temporal groups?**
3. **What, if any, significant shifts in these topics occurred in early 2020?**

What is Topic Modeling?



Comparison of PCA (above) and LDA (below) projections of famous Iris dataset

Topic modeling is a technique in Natural Language Processing that identifies themes in a set of documents. It is a type of clustering algorithm, which means that it uses geometric proximities between vectorized texts to create groups of documents. As such, topic modeling can be unsupervised or supervised: it can make groups with no regard or need for manual categorization data, or it can use such data when available to optimize its clusters’ integrity with respect to them.

This study employs the former. More specifically, we will use Linear Discriminant Analysis (LDA) with no seeded topics. This model takes a set of vectors and a desired number of topics K and finds a plane onto which the projection of those vectors is most saliently separated into K distinct groups. It’s conceptually analogous to a Principal Component Analysis (PCA) model that is supervised with respect to optimizing separability between classes rather than variance within and that is optionally capable of categorical (label-conscious) supervision.

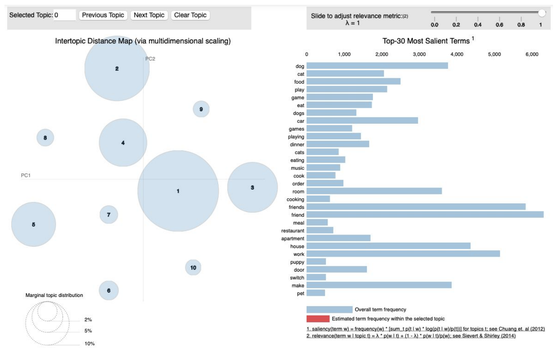
Data and Methods

The dataset extracted from the PushShift dumps contains 297,442 submissions posted by 243,728 different users on 1,794 individual days during the 5-year period of interest. The vast majority of these posts are from the latter half of this window, with 2019 marking the 25th percentile for year of submission.

We tokenize these documents using term frequency-inverse document frequency (TF-IDF) as the quantifier. This method assigns an importance score to each word in the corpus that is directly proportional to the frequency of that word within each document and inversely proportional to its frequency across the corpus.

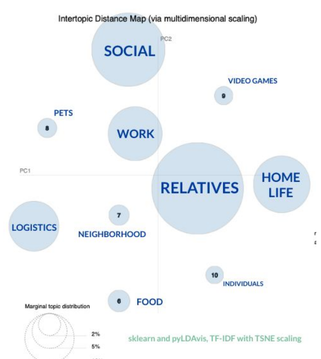
Using the resulting document-topic matrix (DTM), we train a sklearn LDA model to extract K=10 topics as defined by highly correlated probabilities between words. After plotting the result using pyLDAvis with TSNE scaling, we manually label the resulting topics based on observed commonalities in the words within each set.

Results



The figure above is the LDA result for K=10 topics extracted from the full dataset and visualized in pyLDAvis. The left section shows abstract representations of the extracted topic clusters, sized by the collective frequency of constituent tokens and situated on the projection plane selected by the model to maximize cluster separability. The right section shows term frequency for the top 30 tokens in the corpus.

Discussion



1. What were the most common topics in the dataset?

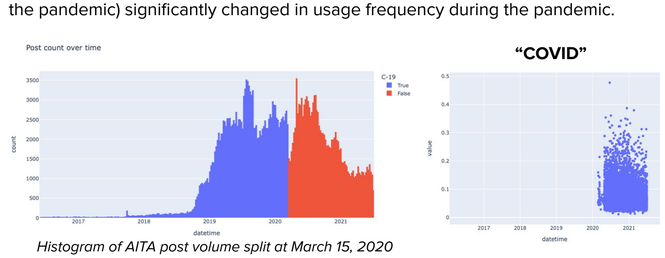
The figure on the left shows the topics from the previous figure manually labeled based on the most common words in each. Unsurprisingly, the top topics—*Relatives*, *Social*, *Home Life*, *Work*, *Logistics*, *Food*, *Neighborhood*, *Pets*, *Video Games*, and *Individuals*—are interpersonal in nature. *Relatives* contains words like "mom", "family", "dad", and "sister", while *Social’s* top words were "friend" and "friends". Many of the other topics are indirectly interpersonal as well—*Neighborhood*, for example, contains words related to noise and space complaints, while *Individuals* is simply a collection of names (primarily female). This observation is hardly unexpected; the objective of AITA is to determine fault in interpersonal struggles. That being said, these findings confirm such topic bias and give credence to the idea that any shift in the subreddit around the onset of COVID-19 is likely to be one of volume rather than of topic.

2. How similar are these topics across temporal groups?

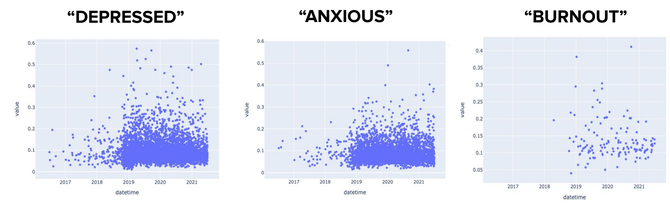
Applying LDA to each individual month represented in the dataset did not reveal any major inconsistencies. While the ranking of the ten topics discussed above would occasionally vary slightly and the bottom couple topics were sporadically replaced with other uncommon ones, the topics were generally the same across time. While not shown here, these observations are supported by statistical analyses discussed in the full paper: calculating model perplexity and Hellinger distances for each instance of a given topic over time shows low dissimilarity across the board with explainable outliers as the only exceptions..

3. What, if any, significant shifts in these topics occurred in early 2020?

The data do not suggest that any significant shift in topics occurred in response to the COVID-19 pandemic. As discussed above, the topics were fairly consistent across time with minor exceptions. Many of the trends observed seem instead to be a result of increased submission volume beginning in late 2018 (see the figures below), which in turn increased the number and frequency of new words compared to earlier months. With respect to the data at hand, however, none of the words chosen for manual examination (so selected for semantic relevance to the pandemic) significantly changed in usage frequency during the pandemic.



Histogram of AITA post volume split at March 15, 2020



References

Comparison of LDA and PCA 2D projection of Iris dataset. scikit. (n.d.). Retrieved from https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_vs_lda.html

Strong, A. (2022). *COVID, Conflict, and Reddit: A 5-year retrospective analysis of r/AmITheAsshole.* Unpublished manuscript. Retrieved from https://annabellestrong.com/files/COVID__Conflict__and_Reddit.pdf

Scan for tull paper