COVID, Conflict, and Reddit

A 5-year retrospective analysis of r/AmITheAsshole

Annabelle Strong
as8@wellesley.edu
Wellesley College
Wellesley, Massachusetts, USA

ABSTRACT

Reddit's r/AmITheAsshole community is a popular and divisive forum in which individuals post descriptions of personal moral dilemmas or accusations for other users to judge. While the judicial dynamics of subreddit are well-documented, few studies have examined the implicit sociological data embedded in its posts. In response to this blind spot, this paper aims to track the common topics behind those submissions over a five-year period (between June 2016 and June 2021, inclusive) to see if and how the COVID-19 pandemic and subsequent social shift impacted the subjects, prevalence, and nature of interpersonal conflict. Through topic modeling analysis, we find that no such shift is evident in the data.

KEYWORDS

Reddit, topic modeling, LDA, conflict

ACM Reference Format:

Annabelle Strong. 2022. COVID, Conflict, and Reddit: A 5-year retrospective analysis of r/AmITheAsshole. In *Proceedings of Data and Text Mining (CS 315)*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Background and Motivation

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 quantifiable 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 (see [5],[2],[4],[9] for deeper analysis of the patterns and dynamics of such judgements), the subreddit provides a uniquely accessible and authentic insight into the types of conflict experienced by users. This paper aims to detect and quantify patterns in these categories over the last five 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 archives. 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.

1.2 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?

RQ1 and **RQ2** aim to establish the variety and consistency of topics present in AITA posts irrespective of larger social and temporal context. **RQ3** examines this data longitudinally and holistically to detect and understand any notable shifts that may have occurred.

2 LITERATURE REVIEW

2.1 Subreddit Demographics

2.1.1 "Don't Downvote A$$$$$$s!!": An Exploration of Reddit's Advice Communities. Cannon et al. (2021) analyzed posters' demographic information with respect to their post's linguistic features in AITA and the related r/relationships subreddit. Among other findings, they determined that the majority of AITA posters who disclosed their gender identity were women, and that the most heavily represented age group is 18-24, followed by 25-34. They also found that self-disclosed male posters were more likely to elicit a negative community ruling than their female counterparts, although they acknowledged that the rarity of strongly negative rulings limit the statistical power of this finding.[3]

2.2 Social Effects of COVID-19

2.2.1 Day-to-day relational life during the COVID-19 pandemic: Linking mental health, daily relational experiences, and end-of-day outlook. Merolla et al. (2021) found that participants on generally reported minimal daily interpersonal conflict during the pandemic—on a scale from 1("None") to 7("A great deal"), the mean was 1.46 with a standard deviation of .68. Other measures of social difficulty were also found to be relatively low. When asked to rate their agreement with the statement "Maintaining relationships was difficult and frustrating today" on an increasing scale from 1 to 7, participants reported an average of 2.98 (SD = 1.62), which

CS 315, Spring 2022, Wellesley College

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

corresponds to slight disagreement. Surprisingly, when asked to do the same to indicate their degree of satisfaction with their social interactions, the participants reported an average of 3.56 with a standard deviation of 1.02—just slightly on the negative side of neutral. This study was conducted on 120 students at a large American university, with an average age of 21. As such, it captures the social experiences during the pandemic of individuals in the age group most heavily represented in AITA . That being said, the small size of the survey limits the conclusions that can be drawn from the data.[6]

2.2.2 A Study on the Psychological Wound of COVID-19 in University Students. A similar study done by Padrón et al. in 2021 examined the sources of stress in 932 Spanish university students (most aged 18-26) and their coping strategies. The authors found that changes in interpersonal conflict was ranked lowest of the five stressors they investigated—below academic, social distancing, pandemic, and general overload respectively. These factors were all moderately and significantly correlated, with correlation coefficients ranging from .29 to .53. This general increase in each source of stress corresponds to an increase in reported mental health issues. Most participants felt that their mental health changed during the COVID-19 crisis, 57.5% labeling it as "worse" and 14.7% as "much worse". This decrease was more pronounced in women than in men, though it should be noted that women were more heavily represented in the participant group (75.5%). [7]

2.2.3 COVID-19 and Mental Health/Substance Use Disorders on Reddit: A Longitudinal Study. Alambo et al. established correlations between BERT-generated topics for r/Coronavirus and the subreddits r/anxiety, r/depression, r/Suicidewatch, r/opiates, r/OpiatesRecovery, and r/addiction for each month between January and October (inclusive) of 2020. They found that the correlation for r/anxiety spiked early in the crisis, with a peak correlate value of .8 in May of that year, up from around .4 in February. r/depression was more variably correlated, with peaks and lows cycling every three months or so. r/Suicidewatch, however, steadily increased in correlation, with its peak (roughly .6) in August. Together, these three trends suggest that Reddit users generally experienced increased anxiety at the beginning of the pandemic, followed by periodic depression and increasing suicidal thoughts in the subsequent months. A similar story can be seen in the three substance-addition subreddits. r/opiates roughly increased in topic correlation with r/Coronavirus before peaking in July and beginning to decrease, while r/OpiatesRecovery spiked in July and continued to have relatively higher correlation coefficients (over .3 in September compared to just over .1 in June) for the rest of the months studied. Correspondingly, r/addiction increased steadily throughout the year, though it slightly decreased after August. Taken together, the data implies that some users began or increased their opiate usage as the pandemic ramped up before turning to substance management and recovery communities in the fall. Combined with the findings in Merolla et al. and Padrón et al., it seems that the effects of the pandemic were largely internal (relating to mental health, substances, academics, etc.) rather than interpersonal in nature. If this were the case, we would expect to see little change in AITA topics during the onset of the pandemic, as submissions typically pertain to social issues. [1]

2.3 Longitudinal Topic Modeling

2.3.1 A Geometry-Driven Longitudinal Topic Model. A 2021 paper by Wang et al. developed and analyzed a methodology for calculating and tracking LDA-generated topics over time in social media data. Using tweets collected between February 15 and May 15th, 2020, the authors first temporally smoothed the corpus by combining temporally-close documents according to an exponential formula. More specifically, for each set of documents d_{t_1} associated with time t, a proportion of each other set of documents d_{t_0} was calculated using the formula

$$w(t_0, t_1) = \gamma^{|t_1 - t_0|} \tag{1}$$

with coefficient $\gamma = .75$. Applied to every pairwise combination between one time slice (e.g. one particular day) x and its n peers (self-inclusive), its set of weights W_x is calculated to be

$$W_x = \{w(t, x), 0 \leq t \leq n\} \tag{2}$$

$$= \{.75^{|t-x|}, 0 \leq t \leq n\} \tag{3}$$

For $t = 2$, for example, the corpus would contain 100% of d_2, 75% of d_1 and d_3, and so on:

$$W_2 = \{.5625, .75, 1, .75, .5625, .4219, ..., .75^{|n-2|}\} \tag{4}$$

This method ensured that each set of documents was representative of a soft range of time rather than a single point t without biasing them with text too far removed in time. The authors then applied Latent Dirichlet Allocation (LDA) to each corpus to extract a set of topics for each. These topics consist of a set of words that are likely to occur together in the given corpus, and therefore capture cohesive patterns in its texts. These topics were then compared pairwise using the Hellinger distance formula (Equation 5) to establish geometric similarity, which the authors noted is more effective for probability distributions than Euclidean distance. The resulting

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=0}^{n} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{5}$$

Figure 1: The Hellinger Distance formula

$K \times T$ matrix represents quantified similarity between ever topic K associated with each time T, with higher values indicating higher dissimilarity. The authors used these values to find topic trends using shortest-path techniques (such as Dijkstra's algorithm) on a 10-nearest neighbor graph derived from the Hellinger matrix. They separately graphed the distances represented by the Hellinger matrix using a variety of multidimensional scaling methods including t-distributed stochastic neighbor embedding (t-SNE), which is our chosen method for scaling the dimensionality of our own data. This paper relies heavily on these techniques proposed by Wang et. al, and will employ highly similar methodology to find topic trends in the posts we collect.[10]

3 DATA AND METHODS

3.1 Collection Strategy

The data required to build on these findings will be collected from PushShift API, a public database of Reddit posts and comments

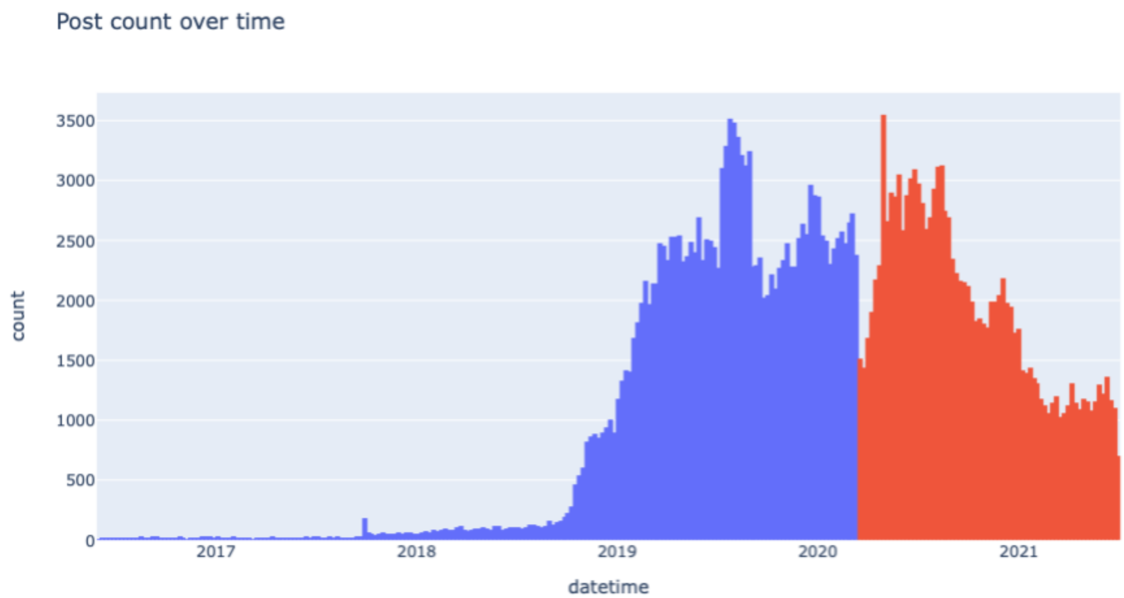


Figure 2: Histogram of collected submissions over time (divided at March 15th, 2020, roughly when US lockdowns started)

created by Jason Michael Baumgartner, Eddy Lazzarin, and Alexander Seiler. This method was chosen over the alternatives (such as the Reddit API or web crawling)[8] because of its cohesive and complete collection of submissions by month, which simplified the collection and organization process.

From its archives we downloaded compressed files of Reddit submissions, each containing every post for a particular month. As we were only concerned with the topics brought up by each post but not in their responses, we did not download the associated comment data. In total, we collected 61 such files for June 2016 through June 2021 (inclusive). From these we extracted all posts from AITA, filtering out those written by moderators and those without the post body intact. All in all, the raw dataset contained close to 300,000 unique posts (see Figure 3), each accompanied with the post's body text and details about its author's username, posting date and time, popularity (via comment and crosspost counts), etc., although for our purposes only the body text is of interest. The only modification made to this dataset was the addition of year, month, and day columns using data extracted from each submission's UTC string indicating its time of posting.

	num_comments	subreddit_subscribers	year	month	day
count	297442.000000	2.949030e+05	297442.000000	297442.000000	297442.000000
mean	72.988862	1.894884e+06	2019.570407	6.401288	15.850099
std	295.574866	8.095511e+05	0.770387	3.369811	8.835055
min	0.000000	5.045700e+04	2016.000000	1.000000	1.000000
25%	12.000000	1.202962e+06	2019.000000	4.000000	8.000000
50%	21.000000	1.874987e+06	2020.000000	6.000000	16.000000
75%	41.000000	2.744655e+06	2020.000000	9.000000	24.000000
max	99993.000000	2.891768e+06	2021.000000	12.000000	31.000000

Figure 3: Ranges and averages of quantitative columns

3.2 Data Summary

The final dataset 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. This corresponds with subscriber count growth over the years, which steadily increased from about 50,000 to almost 3 million. Comment counts also increased, but far less dramatically—most posts had fewer than 100 comments regardless of year posted. The biggest outlier is an open forum from June 1st, 2020 that garnered nearly 100,000 comments. As such, that submission will be excluded from analysis. Post length, on the other hand was fairly consistent: as Figure 4 shows, most posts had around 10^5 to 10^6 words in their body text, with a negative skew on the distribution and a maximum of roughly 10^8.

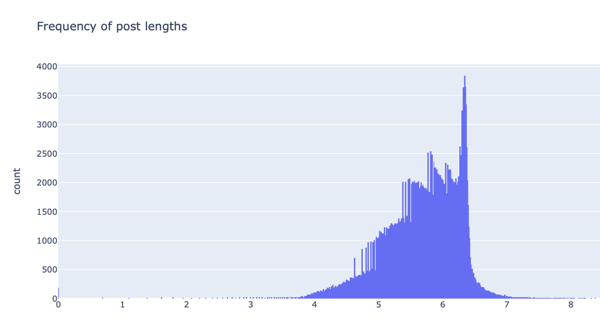


Figure 4: Histogram of logarithmically-scaled post word count

3.3 Methodology

To extract and interpret topics embedded in this corpus, we will employ the following strategy motivated by the methods proposed in Wang et al. [10]:

(1) Extract vocabulary from dataset and generate topics ($K = 10$) with LDA (**RQ1**)
(2) Group and temporally smooth the collected documents by month
(3) Apply LDA with the extracted collective vocabulary to each resulting corpus independently to extract a total of 610 topics ($K = 10$ for each $0 \leq t \leq 60$)
(4) Compute the perplexity values of each pairwise combination of trained LDA model and vectorized document
(5) Generate the similarity matrix $K \times T$ by calculating the Hellinger distance metric for each pair of topics (**RQ2**)
(6) Plot individual token frequencies over time and compare with COVID-specific language (**RQ3**)

Note that the choice to extract $K = 10$ topics from the corpus and its subsets is not necessarily optimal. Most previous work in topic modeling calculates an optimal value for K by comparing performance scores (such as perplexity and coherence) between models trained to extract different numbers of topics from the same data. While we encourage future work to do the same with respect to the data at hand, time and computing constraints restricted our K value to an integer large enough to capture topic variety, but small enough that pairwise calculations (steps 4 and 5 above) were able to be completed in a reasonable amount of time (though the operations required for step 5, described in 4.3.2, still took roughly nine hours total). We thus arbitrarily chose 10 for this value.

4 RESULTS

4.1 Vocabulary Extraction and Collective Topics

Before splitting the corpus into bins by month, we analyze the textual features of the entirety of the dataset to get a sense of overarching patterns. First, we vectorize the 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 an LDA model to extract $K = 10$ topics as defined by 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 (see Figure 5). The result answers our first research question **RQ1**, revealing that the top topics in AITA are *Relatives*, *Social*, *Home Life*, *Work*, *Logistics*, *Food*, *Neighborhood*, *Pets*, *Video Games*, and *Individuals* respectively.

4.2 Temporal Modeling

To get a more granular view, we group the data by month and apply the temporal smoothing technique described in Section 2.3.1. To summarize, we modify each group to include a exponentially decaying portion of each other group's documents based on temporal distance between the two sets. Figure 6 shows the number

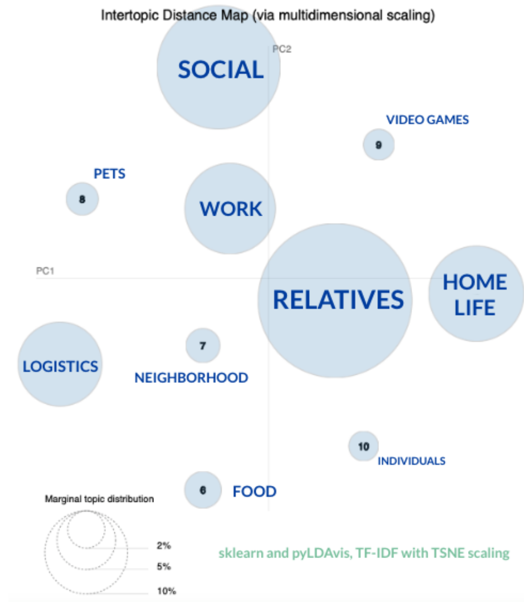


Figure 5: Topics for the full corpus

of documents from each month (y-axis) in the resulting 61 group compilations (x-axis). Having compiled these groups, we vectorize each corpus as we did the collective in Section 4.1, using the vocabulary generated in that section as a base to ensure consistent vectors. The resulting DTMs are then used to train individual LDA models.

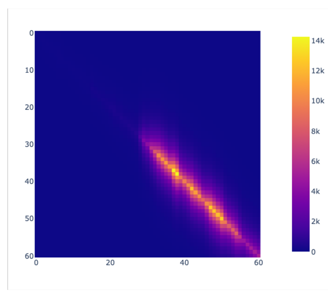
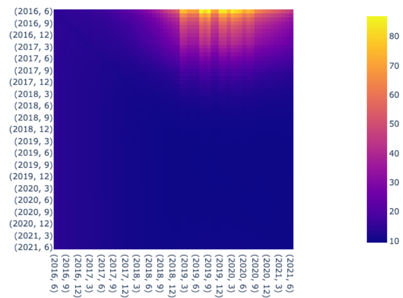


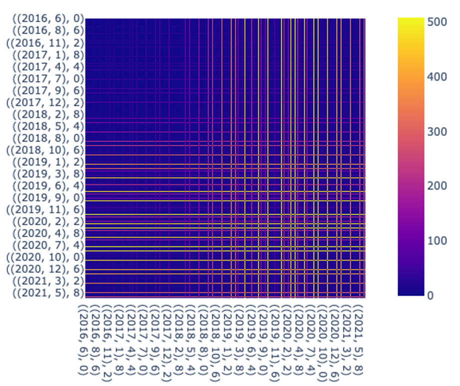
Figure 6: Number of documents in each group within the temporally-smoothed corpus

4.3 Similarity Analysis

4.3.1 Perplexity. Perplexity is a measure of how well a model predicts a sample (see [?] for more detail)(*error with auto-citation; intended reference is [4]*). To apply it here, we feed models trained on each month's data with the document-topic matrices of each other month's data to quantify the similarity between models. The result can be seen in Figure 7a. Bright colors indicate that the the model trained on the month indicated on the y-axis is not very predictive of data from the month on the x-axis. More specifically, the bright patch in the upper-right corner shows that models trained on 2016-2017 data were "surprised" by submissions from later years.



(a) Perplexity between each pair of months' models



(b) Hellinger distances between topic vectors

Figure 7: Topic similarity metrics

This implies that some kind of shift occurred around late 2018, causing the later datasets to contain words that earlier models didn't encounter in their training.

4.3.2 Hellinger Distance. To more closely examine this shift, we compare the topics generated by each of these models. As discussed in Wang et al. [10], this is done by calculating the Hellinger distance between the topics' probability distributions. We first extract the topic vectors and compile them into a 610×610 matrix, each row and column representing one of ten topics for each of the 61 months in the dataset. We then apply the Hellinger equation (see Figure 5) pairwise to this topic matrix to obtain an array of distance coefficients (as visualized in Figure 7b). The resulting striped heatmap shows highly consistent (though not constant) Hellinger distances between each topic and the others. Topics from the latter half of the dataset are more frequently dissimilar, though those that have high distance values to the majority of the nearly always have low distance values to each other. This suggests a split in topics, where the topics that appear in blue on the heatmap represent one set of cohesive topics and those that appear in orange or yellow represent another. It's noteworthy that topics in the second category span several years, including but not limited to the months since COVID-19 first became a global issue. This implies that such topics are not specific to the pandemic, instead representing a different subject of discussion that first becomes relevant around late 2018. That being said, the number of posts in AITA dramatically increased around the same time, so it's possible (if not probable) that document count is responsible for some degree of the apparent split in topics.

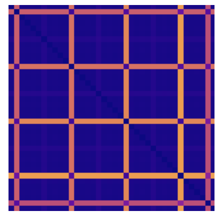
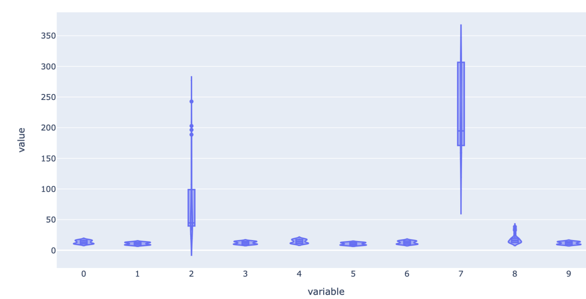


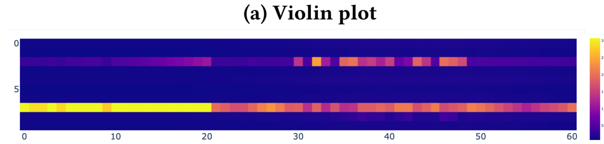
Figure 8: Figure 7b magnified between the coordinates ((2018, 10), 6) and ((2019, 2), 3)

From these results we also noticed consistently low Hellinger distances for a given topic number over time. As shown in Figure 8, the intersections of two bright bands representing the same topic number (in this case, most representing Topic 7) between months were noticeably darker than their surroundings. This implies that some degree of consistency might exist for each numbered topic.

To further investigate this possible trend across time, we filter each row of the Hellinger matrix to only include values where the constituent topics were of the same number (e.g. the values comparing Topic 3 from month 0 with Topic 3 for every other month). We flatten these lists by taking the median of each, which results in a 61×10 matrix where each cell contains the median dissimilarity between topic y for a given month x and the topic y values of the other months. As seen in Figure 9a, most topics have



(a) Violin plot



(b) Heatmap

Figure 9: Hellinger distances for each topic number over time

fairly constant Hellinger distances between their members. The exceptions are topics 2 and 7. Upon further investigation, the two share sets of words: "just", "like", "don"('t), "time", "friend", "told", "said", "didn"('t), and "really", among a few others. These words are among the top 10 listed for Topic 2 in the months prior to April 2018, while they are among the the top 10 for Topic 7 instead after that point. This switch seems to be at least partly responsible for the inconsistencies within the two topics—looking at Figure 9b, the month of the switch is clearly marked with a distinct change of color in both rows.

4.4 Token Trends

Considering this apparent lack of COVID-related topic changes, we look closer at the frequencies of certain words over time to see if any individual trends emerge. Figure 10 shows topic frequencies throughout the corpus for the words "anxious", "depressed", and "covid" respectively. These specific tokens were of interest due to the findings of Padrón et al. [7] and Alambo et al. [1], both of which indicated increased anxiety in depression in the groups studied, with the latter specifically investigating Reddit posts. The consonance between these results suggested that if there were to be any COVID-related shift on AITA, it would likely entail increases in the frequencies of anxiety- and depression-related submissions. The graphs do not support such an increase—while the term "covid" predictably experienced a sharp increase in usage in early 2020, the other two tokens don't appear to do the same. The only notable pattern—a dramatic increase in late 2018—is the same one we have previously discussed as a consequence of increased user activity. While we encourage future work to examine normalized figures that quantify the relative popularity of each term with respect to the subreddit's popularity, the data shown clearly indicates that no COVID-specific shift in these terms occurred.

That being said, there are a number of other terms that we did not manually track—such as "domestic violence" or similar—that may have been more volatile. Thus, our findings in this section should not be taken as a generalized conclusion about all potential COVID-related terms being uncorrelated with the onset of the pandemic; rather, just those explicitly discussed.

5 DISCUSSION

5.1 Research Question Results

5.1.1 RQ1: What were the most common topics in the dataset? Unsurprisingly, the top topics—*Relatives*, *Social*, *Home Life*, *Work*, *Logistics*, *Food*, *Neighborhood*, *Pets*, *Video Games*, and *Individuals* (see Section 4.1 and Figure 5)—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 topics.

5.1.2 RQ2: How similar are these topics across temporal groups? As discussed in Section 4.3.2, the top 10 topics appear to be fairly consistent between months. While these topics don't neatly match those extracted from the full data—the former consists of probability distributions extracted from each month's raw model, while the latter is a set of manual labels based on a TSNE-scaled pyLDAvis visualization (Figure 5)—their quantitative similarities over time do show that submissions on AITA have neither significantly shifted over time, nor been influenced by changing sociological factors. The only notable inconsistencies seem to reflect one or more shifts

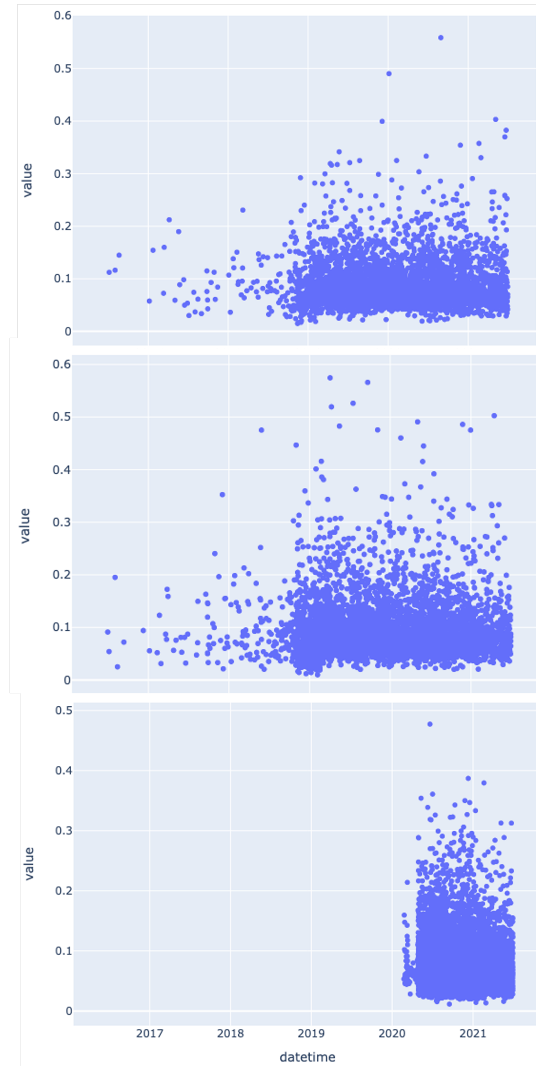


Figure 10: Token frequency over time: "anxious" (top), "depressed" (middle), and "covid" (bottom)

in the assigned numeric labels of various topics rather than their qualitative contents.

5.1.3 RQ3: What, if any, significant shifts in these topics occurred in early 2020? The data does not suggest that any shift in topics occurred in response to the COVID-19 pandemic. As shown by the analyses done in Section 4.3, the topics were fairly consistent across time with minor exceptions. Many of the trends observed (see Figure 7) seem instead to be a result of increased submission volume beginning in late 2018, which in turn increased the number and frequency of new words compared to earlier months. Future work might consider investigating the cause of this dramatic growth in the subreddit; though the topics being discussed didn't significantly change, the spike in user activity still might be in response to societal factors (such as the discord surrounding the 2020 US election, for example). With respect to the data at hand, however, none of the words examined correlated with with pandemic (Figure 10).

5.2 Validity

There are a few features of our study that could limit the validity of our findings. First of all, as noted in Section 3.3, our choice of $K = 10$ may not produce topics that accurately reflect subject trends within AITA. We also direct our analysis in Section 4.3.2 based on the noted observation of low distance values between topics with the same number. While this was an interesting point of investigation, it assumes that there is a consistent assignment of topic numbers between two similar models. While this assumption does not appear to be out of line, a more nuanced analysis, as in Wang et. al, would use shortest-distance formulas to find common topics across time groups instead of relying on potentially arbitrary topic numbers. We encourage future work to take such measures to generate a more accurate representation of topic trends in the dataset.

5.3 Ethical Considerations

Proferes et al., among others, has previously raised important considerations with respect to Reddit data analysis. They note that institutional review boards (IRB) tend to classify scraped Reddit data as "public" and therefore exempt from consent requirements [8]. It's unsuprising, then, that fewer that 15% of papers the authors reviewed in their meta-analysis even mentioned ethics reviews at all.That being said, data like quotes and media that would require attribution in other contexts are subject to competing needs to both protect the anonymity of the author (a key feature of Reddit) and give proper credit. Researchers most commonly resolve this issue by attributing such data to the authors' Reddit usernames, although identifying information is still a concern given the frequently highly personal nature of users' activity and contributions. Our data is no exception, but given the abstract nature of our analysis and the purpose of AITA—that is, publicly getting strangers' opinions on personal conflict—and the abstract nature of our analyses, we believe that our work and results pose few, if any, ethical issues.

6 CONCLUSION

Despite its well-documented impact on societal systems and norms over the past two year, COVID-19's impact on individuals is still a subject of much analysis. While previous literature has established negative mental health effects of the pandemic—namely, an increase in anxiety, depression, and addiction, among other struggles—we wondered about it interpersonal consequences. Did working from home reduce office conflict? Did quarantine increase familial tensions? As a popular forum that features discussion about highly personal conflicts, r/AmITheAsshole poses a uniquely raw insight into everyday disagreements. In this paper, we take advantage of this resource to search for trends in conflict over the course of the last five years: specifically, which topics were commonly discussed; how consistent these topics were over time; and if there was any shift in these topics around the time that COVID-19 became a worldwide crisis. We find that the most common topics, which were generally interpersonal in nature, were quantitatively consistent over the time period studied and did not seem to undergo any significant shifts in response to the pandemic. While we are confident in these findings, we note that they are limited by the analyses performed. More precisely, our analysis does not seek to label or understand the topics as calculated, and we do not attempt to connect these topics over time to get a quantitative sense of their changes (Wang et al. [10] accomplishes this with a shortest-paths algorithm, but we do not employ such methods in this study). We suggest that future work take these additional steps to better understand the nuances of AITA's topics beyond their quantitative similarities, and that it take changes in subreddit popularity into account to rule out activity-related confounds.

REFERENCES

[1] Amanuel Alambo, Swati Padhee, Tanvi Banerjee, and Krishnaprasad Thirunarayan. 2021. COVID-19 and Mental Health/Substance Use Disorders on Reddit: A Longitudinal Study. In *Pattern Recognition. ICPR International Workshops and Challenges (Lecture Notes in Computer Science)*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 20–27. https://doi.org/10.1007/978-3-030-68790-8_2

[2] Nicholas Botzer, Shawn Gu, and Tim Weninger. 2021. Analysis of Moral Judgement on Reddit. (Jan. 2021). http://arxiv.org/abs/2101.07664

[3] Emily Cannon, Bianca Crouse, Souvick Ghosh, Nicholas Rihn, and Kristen Chua. 2021. 'Don't Downvote A*******s!!': An Exploration of Reddit's Advice Communities. (2021). http://arxiv.org/abs/2109.09044

[4] Ion Stagkos Efstathiadis, Guilherme Paulino-Passos, and Francesca Toni. 2022. Explainable Patterns for Distinction and Prediction of Moral Judgement on Reddit. *J. ACM* (Jan. 2022). http://arxiv.org/abs/2201.11155

[5] Ethan Haworth, Ted Grover, Justin Langston, Ankush Patel, Joseph West, and Alex C Williams. 2021. Classifying Reasonability in Retellings of Personal Events Shared on Social Media: A Preliminary Case Study with r/AmITheAsshole. (2021).

[6] Andy J. Merolla, Christopher Otmar, and Carlos Ruvalcaba Hernandez. 2021. Day-to-day relational life during the COVID-19 pandemic: Linking mental health, daily relational experiences, and end-of-day outlook. *Journal of Social and Personal Relationships* 38, 8 (Aug. 2021), 2350–2375. https://doi.org/10.1177/02654075211020137 Publisher: SAGE Publications Ltd.

[7] Isabel Padrón, Isabel Fraga, Lucía Vieitez, Carlos Montes, and Estrella Romero. 2021. A Study on the Psychological Wound of COVID-19 in University Students. *Frontiers in Psychology* 12 (2021). https://www.frontiersin.org/article/10.3389/fpsyg.2021.589927

[8] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (June 2021). https://doi.org/10.1177/20563051211019004

[9] Parvathy Sarat, Rohit Mujumdar, Prathik Kaundinya, and Sahith Dambekodi. 2020. Can Machines Detect if you're a Jerk? (2020).

[10] Yu Wang, Conrad Hougen, Brandon Oselio, Walter Dempsey, and Alfred Hero. 2021. A Geometry-Driven Longitudinal Topic Model. *Harvard Data Science Review* 3, 2 (June 2021). https://doi.org/10.1162/99608f92.b447c07e