A holistic approach to understanding and contextualizing the anti-work subreddit

Albert Lu, Aurimas Racas, Lawton Walker

Abstract

r/antiwork is a subreddit dedicated to exploring ideas about work-free life or work reform, seeking advice about poor working conditions or job hunting, and and venting about the status quo. r/antiwork exploded in popularity around mid-2021 due to pandemic-induced dramatic change in the traditional work dynamic. We use a variety of regression, classification, and other data science techniques to understand the nature of r/antiwork discussions, user interaction dynamics, and broader user participation on Reddit. Our primary goal is to shed light into the overall community of r/antiwork and understand whether it's a therapeutic hub for frustrated workers, the a beginning of new workers' rights movement, or something else entirely.

Introduction

The COVID-19 pandemic disrupted many social dynamics. One of the movements triggered by the pandemic was a case of extraordinarily high-volume voluntary separations dubbed "The Great Resignation". The voluntary rate of quits reached 2.9% in the US in August 2021, the highest since the data was first collected in 2020. (US Bureau Of Labor Statistics 2021).

Given the very recent nature of the phenomenon, research is ongoing to understand its root causes (e.g. (Parker and Clark 2022)). Several factors have been proposed, including worker burnout and re-assessment of importance of work / work-life balance in the light of the global pandemic. There are speculations that this dynamic may represent a beginning of a new workers' rights movement (Botz 2021).

In this project, we sought to better understand the dynamics behind the Great Resignation through one of its representations online, Reddit's /r/antiwork social community, a subreddit associated with contemporary labor movements and the anti-work movement. Established in 2013, it gained popularity in 2020 and 2021, doubling to 1.7 million subscribers in 2021 alone (Wikipedia 2020)

Our primary research goals were to understand the nature of discussions in r/antiwork, tying together information about prevalent discussion topics, post archetypes, subcommunities, user interaction dynamics, and finding other similar subreddits thorough common user iteractions.

Our work is based on data collected on r/antiwork spanning from its inception to February 2022, capturing over 200,000 posts. We also collected 6 million comments made by over 800,000 users, resulting in 4.8 billion interactions between users participating on the same discussion threads. All code and data is available at https://github.com/alu13/Antiwork_Analysis.

Related Work

Topic modeling is a common technique used to understand the nature of discussions in subreddits. Shen and Rudzicz (2017) use LDA to visualize anxiety topics in a variety of anxiety-related subreddits. Similarly, Pandrekar et al. (2018) analyze opiate-related subreddits using LDA topic modeling. Liu and Yin (2020) use LDA and semantic clustering to build rough topic models of the content of a weight loss subreddit. The performance of these three studies is variable, and the results are quite noisy, indicating the challenges with single subreddit topic modeling.

Much research has been done on classifying posts into their specific subreddit, but this research will be the first to classify posts of a single subreddit into overarching post archetypes. We believe the approach outlined in our work in a novel contribution to the field, albeit the task is challenging for a variety of reasons as we elaborate later.

We draw inspiration from several studies that explored user interactions on social communities. If r/antiwork represents a thriving community with long-term goals, we expect to find that some users develop outsize influence in the community, acting as opinion leaders. We build upon approaches used by Kilgo et al. (2016) and Huffaker (2010), in this regard. Hajian and White (2011) provide a survey of metrics used to identify opinion leaders in a social network, and develop an Influence Rank metric that, unlike most metrics that rely on network structure, is based on user-to-user interactions instead. While not directly applicable to our dataset due to a different nature of user interactions, we adapt their ideas to our work, too.

Beyond understanding if r/antiwork community has community leaders, we are also interested if we can observe other community dynamics. Most of the existing literature focuses on identifying user communities spanning multiple Reddit subreddits (e.g. Soliman, Hafer, and Lemmerich (2019), 2019). We apply similar methods but instead focus

on understanding if r/antiwork has sub-communities where users form cliques and tend to interact more frequently with each other and whether those interactions relate to specific topics.

In our work analyzing the the overlap between user posts in r/antiwork and other subreddits, we apply ideas similar to Andrei Kashcha (2022) Andrei Kashcha (2022), who implemented a graph visual that uses Jaccard similarity to show the similarities between different subreddits.

Topic Modeling

We first seek to analyze the topic space across all forms of content on r/antiwork. We chose to use LDA (Blei, Ng, and Jordan 2003) as it provides a powerful topic visualization of unlabeled data. To form a base hypothesis, some example topics we found from scrolling through r/antiwork are American oligarchs, minimum wage, teacher pay, 40 hour work week, work is hell, quitting, burnout/mental illness, bad interviews, general tips, and unionizing.

Method

We gathered all submissions on r/antiwork from January, 2014 to February, 2022. Many submissions take the form of images, which we convert into text by using Pytesseract OCR on the image and TextBlob to autocorrect spelling errors. In total, there are 224,009 submissions. 209,732 submissions were left after pulling from pushshift because of deleted image content and other API errors. 144,000 submissions were left after removing empty posts, deleted posts, and removed posts. It is unfortunate that we lose around 36% of potential data, but such is the nature of Reddit. The dataset statistics are shown in Table 1.

Num submissions	Max words	Min words	Avg words
144000	7074	0	150.6

Table 1: All submissions summary statistics

Raw text was heavily pre-processed. We first replaced all special characters with spaces, then lowercased all words. We used regular expressions to remove all URLs because they created unhelpful noise in testing. We considered lemmatizing and stemming all words, but we decided against it due to the unstable effects of lemmatizing on topic model performance (Schofield and Mimno 2016). The stopword list from the nltk package is very incomplete, so we added additional stopwords based on LDA testing. The definition of a stopword is quite loose, so define stopwords as words that appear in many different topics and don't provide unique insight into the topic. For example, we considered words like "work" or "money" to essentially function as stopwords, because they are extremely common in all posts within r/antiwork regardless of topic. We then added common bigrams.

We performed multiple LDA tests on the entire dataset. Evaluating topic models is a challenge itself, and we considered both coherence metrics and the simple eye test. Model coherence is defined as the average semantic similarity between words in each topic. Our model coherence scores hover at around .4 to .5 from a scale of 0 to 1 regardless of hyperparameter tuning. However, we did not find model coherence to be useful for this particular analysis because the semantic similarity of events or themes should not be high. For example, consider T13 on Figure 1. The words "Amazon", "warehouse", "weather", and "storm" clearly represent a coherent topic, but the intra-topic semantic similarity is not high. Optimizing for coherence scores rewarded preprocessing steps that created uninteresting and repetitive topics rather than a more representative topic space.

The eye test consists of looking at the topics and seeing if they are coherent. This method was much more effective at evaluating the coherence of our topic models. However, this method runs the risk of evaluator bias, because our interpretation of topic coherence may not match another evaluators interpretation. Using this approach, we found that our model produces the best topics at around 20-25 topics. An example topic space is shown in Table 2.

Results

The topics are evidently coherent. Some topics such as "T1: work from home", "T2: worker's rights", "T3: student debt", "T24: interviews", and "(T18, T21): social movements" are expected from a subreddit centered around work-related topics. Additionally, the topic model reveals specific events such as "T13: the Amazon warehouse tragedy".

The topic model also reveals some information about post archetypes, or the types of posts on r/antiwork such as memes or personal stories. The topic "T23: Twitter artifacts" reveals an archetype of social media screenshots through common words found on screenshots of social media posts. The topic "T20: r/antiwork Meta Posts" reveals a class of posts discussing r/antiwork through subreddit direction, moderator activity, and potential activism.

However, topic modelling on all subreddit content is not a sufficient comprehensive analysis of the content on r/antiwork. First, LDA is most effectively used in a clear and diverse topic space with relatively standard inputs, such as news or research papers. LDA or other topic modeling algorithms may not be effective on extremely noisy inputs with multiple modalities. We would ideally separate r/antiwork into standard post archetypes then run topic models on each of these analysis to obtain cleaner results.

Additionally, topics are just one dimension of subreddit analysis; post archetypes are another. In fact, some post archetypes are directly related to specific topics. All personal stories are about poor work environments and quitting. Topics and post archetypes are connected, and isolating one or the other will lead to an incomplete analysis. Because of the importance of this multi-dimensional analysis, we also analyze the types of posts on r/antiwork as well as the more nuanced topic models within a single post archetype.

Post Archetypes

What post archetypes are present in r/antiwork? Many different archetypes are visible on the front page of r/antiwork:

Topics	Key Words
T1: Work from home	Covid, store, home, come, manager
T2: Workers Rights	Minimum, wage, living, labor, movement
T3: Student Debt	School, college, student, debt, parents, rent
T13: Amazon Warehouse Tragedy	amazon, weather, warehouse, storm
T18: Social Movement	life, enough, system, help, everyone, better
T20: r/antiwork Meta Posts	f***, mods, posts, subreddit, antiwork
T21: Social Movement 2	Union, world, rights, society, human, free
T23: Twitter artifacts	Reply, share, comment, please, thank
T24: Interviews	interview, manager, asked, questions, start

Table 2: All data topic models

Image-based memes, meme tweets, tweet quips, long personal stories, angry non-personal posts, legal advice, news, and personal images (bad workplaces).

Dataset

We sampled and categorized around 1,000 posts on r/antiwork. 850 posts were randomly sampled from all posts and 150 posts came from a weighted sampling where each month (from 2015 to 2022) had an even chance of being sampled from. The objective of the weighted sampling was to sample post archetypes on old r/antiwork that may have disappeared when the subreddit became popular. The posts were classified in two different ways. The first was to classify by a downstream objective of LDA topic modeling and is shown in Table 3. The second was to classify by a more granular downstream objective of providing interesting and explainable visualizations and is shown in Table 4. There are miscellaneous image and text categories in both categorizations. Miscellaneous images include media and images that are difficult to categorize or appear fewer than 5 times such as work Christmas presents, tone-deaf signs at work, and personal ads. Miscellaneous text includes quotes, poems, advice, and witch-hunting among other categories. "r/antiwork Doubt" in Table 4 represents posts that question the direction of the subreddit or the activity of the moderators.

Archetype	Count
Link to News Article	112
Misc Text	74
Serious Question	111
Memes	89
Misc Images	150
Personal Story/Venting	286
Random Thoughts	74
Social Media Screenshots	123

Table 3: LDA Post Archetypes

However, there are no clear ground truth post archetypes and these annotations and categories were created by a single annotator. Thus, these categories and counts may represent a biased distribution of the underlying dataset. These categories will need to be validated in the future.

Archetype	Count
Link to Meme Article	6
Link to News Article	104
Misc Text	51
Serious Questions	111
Memes	89
Music	8
High-Quality Video	12
Misc Images	120
Meme Videos	8
Personal Story/Venting	122
r/antiwork Doubt	13
Random Thoughts	74
Images of Tweets	66
Images of News Articles	32
Call to Action	10
Images of Job Requirements	12
Images of LinkedIn Posts	9

Table 4: Content Post Archetypes

There were also many interesting insights in comparing the differences between weighted sampling and unweighted sampling. Weighted sampling reveals the topography of r/antiwork pre-pandemic, and unweighted sampling is saturated with pandemic content. Pre-pandemic content is centered around abolishment of work through links to news articles, music and video suggestions, and small discussions. Post-pandemic content is centered around work reform and poor working conditions through personal stories, discussion questions, memes, and miscellaneous image content. Links to news articles have turned into screenshots of headlines and tweets, and topics have shifted dramatically.

Method

The data in the content-centric post archetypes is very imbalanced and some categories have less than 10 samples. Thus, we primarily focus on the classifying LDA-centric post archetypes and will leave content categorizations to future work.

We first built a random-forest 8-class classifier with a variety of text features. We used 12 features: is image, the num-

ber of comments, the score of the post, the post link (typically reddit for text content, i.reddit for images, and random websites for links to news), post sentiment, post sentiment value, post emotion, title sentiment, title sentiment value, title emotion, post length, and # of unique words in the post. We also implemented support vector machines and Logistic Regression on the count-vectorized text representation of the post. Finally, we fine-tuned a BERT classifier on post text alone. We also implemented a BERT binary classifier to get higher single class accuracy. Due to time constraints, we were only able to train on personal stories. The ablation study is in table 5.

Evaluation

Our random forest outperformed logistic regression and SVM by a wide margin with an accuracy of .63 and a MacroF1 score of .55. Based on random forest feature importance metrics, sentiment, length, count, link, and is image features were the most important in classification.

BERT 8-class fine-tune outperformed the random forest, with accuracy of around .7 and a macro F1 of .64. While the accuracy is far above random, it is not good enough to confidently score all posts with a specific post category. Interestingly, BERT, SVM, and logistic regression achieved the highest test score accuracy and F1 when they had minimal regularization and were heavily overfitting to the train set. The training accuracy was around .9-1.0.

BERT binary fine-tune achieved .90 accuracy and a .88 Macro F1. Additionally, an error analysis showed that model errors were most likely to mix classes of personal story and not a personal story, such as a personal story that centered around a discussion question or random thoughts that involved some personal elements.

Classifier	Accuracy	Macro F1
Random Forest w/ non-text features	.63	.52
SVM body	.55	.45
SVM body + title	.51	.40
Logistic Regression body	.57	.48
Logistic Regression body + title	.51	.40
BERT 8-class fine-tune	.70	.64
BERT binary on personal stories	.90	.88

Table 5: Post Archetype Classifiers

Topic modelling of personal stories

The high accuracy and promising error analysis on the BERT binary classifier allowed us to classify each post in the larger dataset into personal and non-personal stories. Summary statistics are present in table 6.

Personal stories are relatively single topic, with all posts relating to work and money in some way. Thus, the number of topics is reduced to 10 from 25, and stop-word removal is a bit more aggressive (any word that appears in more than 5 topics is removed to improve topic diversity). Other than these two hyperparameter changes, the rest of the pipeline is identical to LDA on all data. The results are in table 7.

Num Submissions	Max words	Min words	Avg words
43170	5555	7	284.1

Table 6: Personal Story Summary Statistics

The results are very nuanced. Within 10 topics, we can see multiple different frames of poor working conditions alongside important topics such as COVID-19 and job seeking. This is a promising sign that finding nuance through archetype-specific topic models is an effective strategy.

Community analysis

Next, we turned to understanding the community structures in r/antiwork. Our research questions are as follows:

- **R1**: Are the community sub-structures and what do those substructures represent?
- **R2**: Who are the most connected community members and is connectedness related to member attributes, such as their activity track record and longevity?
- **R3**: Are member attributes predictive of their post popularity which may imply underlying power structures?

Data

To answer the research questions, we collected user interactions on all r/antiwork submissions. For each user, we captured number of posts, comments they have made in r/antiwork, the net votes they received on those comments and posts, and the number of interactions with other r/antiwork users. We capture two types of interactions. A direct interaction is recorded when user A makes a top-level comment under user B post or a child comment to user B parent comment (and vice-versa). Indirect interaction is recorded whenever user A and B participate (via post or comment) on the same discussion thread, independent on the number of discussion levels in between them.

In total, we capture 814K unique users participating in r/antiwork during the period of analysis who collectively made 155K posts and 5.98M comments. This resulted in 7.7M direct and 4,785M (4.8 billion) indirect interactions in total. We exclude posts that were made by users whose profiles were since deleted (61K posts) and comments made two active bots (AutoModerator and Magic_Eye_Bot, total of 53K comments). The data was collected using Reddit's API and its PRAW library. No significant data cleaning was required. Summary statistics of are shown in Table 8 ("All Users" column).

In line with prior research (Thukral et al. 2018), we find that user activity follows a power law distribution (Figure 1). While some of our analysis is performed with the entire dataset, we focus our work on the most active users as we believe they are the most representative of the community. As a threshold, we choose to include only the users that have interacted (posted or commented) on the subreddit at least 100 times. This subset comprises 7,807 users with their summary statistics presented side-by-side in Table 8.

Topics	Key Words
T1: Job seeking	position, experience, salary, jobs
T2: Poor Working Conditions	life, family, leave, part, f***
T3 Covid	home, covid, sick, health, quit, done
T4: Poor Working Conditions 2	s***, office, everyone, trying, put, leave
T5: Poor Working Conditions 3	asked, team, help, staff, call
T6: Thinking	think, find, way, lot, good, next
T7: Poor Working Conditions 4	come, office, store, call, christmas, last

Table 7: Personal Stories Topics

Statistic	All users	Selected users
Total number of users	814,350	7,807
Total posts	155,184	28,317
Total comments	5,978,678	1,744,481
Total post karma	76,130,301	16,876,913
Total comment karma	88,031,504	31,238,430
Avg. posts per user	0.19	3.63
Median posts per user	0.00	1.00
Avg. comments per user	7.34	223.45
Median comments per user	2.00	154.00
Median activity window (days)	0.00	125.00
Mean activity window (days)	45.65	233.95
Mean post karma	490.58	596.00
Mean comment karma	14.72	17.91

Table 8: Overall user statistics

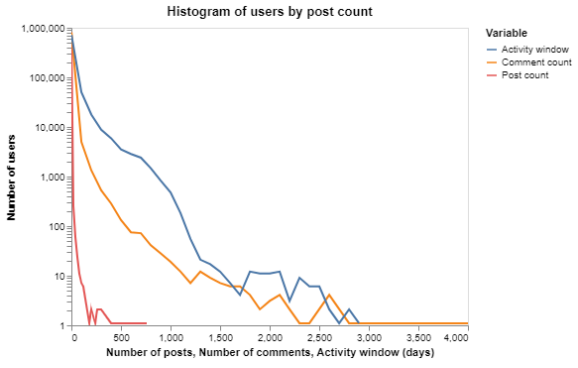

Figure 1: User activity distributions on r/antiwork

Sub-community identification

First, we explore we can discover any patterns on user interactions. This may uncover directly unobservable user preferences related to themes prevalent on r/antiwork or uncover users that are familiar with each other to the degree that they are more likely to participate in discussions together.

We use spectral clustering techniques (Von Luxburg 2007) with unnormalized graph Laplacian and mini-batch K-means as the clustering algorithm to identify user subnetworks in r/antiwork. We consider both direct and indirect interactions, and experiment with raw interaction counts (which are highly correlated with how long a user has been on r/antiwork) and user-normalized counts.

Overall, we do not find evidence of strong subcommunities. The are only few eigenvalues close to zero in the graph Laplacians, and clustering analysis reveals that these are typically small (less than 100) clusters that represent short-lived users who have have participated on a few threads only. The clusters formed by K-means step of the analysis are not well separated, as indicated by low silhouette scores (typically below 0.4).

The most meaningful results were found when analysing indirect user interactions based on raw interaction counts. We were able to retrieve a total of 12 clusters, of which 6 had at least 100 users. We performed U-Mann-Whitney tests to compare the clusters pair-wise based on total post karma, number of posts, number of comments, average post karma, and length of their longevity, and identified 2 clusters that were statistically different (p-value of 0.01) in 20 out of 25 comparisons (5 features compared vs. 5 other clusters). One of those clusters comprised 164 users with an median activity window of 266 days, the longest among all clusters. They also had a highest average post karma (median of 26.1), and highest number of posts (median 2) per user. Arguably, the users belonging to this cluster (”User cluster #10”) could be labelled as the ”clique of power users” of r/antiwork, and it would be interesting to further dive into understanding who these users are - something we leave for further research.

User connectedness

Having not found any strong evidence of user subcommunities, we next explored whether strongly connected users possess any specific characteristics. In line with other similar research (Hajian and White 2011), we used pagerank as the connectedness metric. Similarly to the community-based analysis, we analysed only users with over 100

posts/comments, and used both direct and indirect interaction definitions (normalized across the total user interactions).

We found that user connectedness varies significantly, with most connected users being 50x more connected than the least connected ones when direct interactions are used. We then explored whether user connectedness is associated with other user attributes. Specifically, we fit a log-log specification of a linear regression with pagerank as the response and user specific features: number of posts, comments, mean post karma, mean comment karma, longevity (time since first they joined r/antiwork) and total activity window (in days).

We found that all variables except longevity were significant, with overall adjusted R^2 of 35.4%, when direct interactions are used for pagerank calculations. Total comments made, average post karma and comment karma were all positively associated with connectedness, while activity window and total posts made had a negative association (see Table 9). After performing further exploratory data analysis, we concluded that, somewhat counter-intuitively, the most connected users are likely the ones that participated in the subreddit for a very short time, in a few discussions only ("one-hit wonders"). As a result, their interactions were very focused and led to high connectedness metrics.

Variable	Coefficient	95% conf. int
Intercept	-10.43	[-10.49; -10.37]
Number of posts	-0.02	[-0.02; -0.01]
Number of comments	0.27	[0.26; 0.28]
Average post karma	0.02	[0.02; 0.03]
Average comment karma	0.04	[0.03; 0.05]
Longevity	0.01	[0.00; 0.03]
Activity window length	-0.03	[-0.05; -0.02]

Table 9: User connectedness and other attributes

A similar analysis using indirect interactions to define connectedness yield a smaller range of connectedness values (14x between the most connected and least connected user), and a smaller relationship between connectedness and other user attributes (adjusted R^2 of 19.9%). The only significant variables related to indirect connectedness were number of comments made, longevity and length of activity window.

All in all, this analysis further supports the findings that there are no strong community sub-structures on r/antiwork. Long-standing users do not tend to be more strongly connected, and connectedness is largely an outcome of sporadic participation.

Topic distribution among users

Previously, we described the analysis to uncover latent topics discussed on r/antiwork. We attempted to combine topic analysis with analysis on user structures, too. Specifically, we compared topic distribution of posts made by all users, users with at least 100 posts/comments, most connected users, and users from our previously found "Cluster #10". We selected the topics where the distributions differed the most and visualize those differences below.

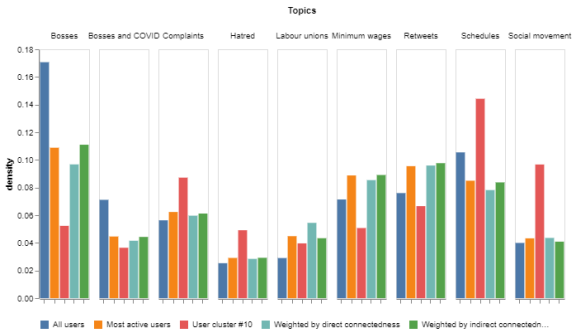

Figure 2: User activity across topics

There are a couple of interesting observations:

- Users from cluster #10 appear to be leaning towards more negative topics expressing general dissatisfaction, and stand out as the ones who focus on systemic discussions, and, surprisingly, work schedules.
- Top users overall, as well as most connected users post less about topics related to personal experiences (e.g. bosses and managers) and instead rely more on retweets and discussion of minimum wages.

User popularity

Finally, we explore whether past user track record is predictive of their future post popularity. We build on the approach used by Kilgo et al., who use total top-level comments as their dependent variable, with user karma and longevity as covariates. We extend this approach by:

1. Using post scores as the dependent variable; we believe that user votes is a better metric to measure popularity of a post given that a user can only contribute to it once.
2. We include total post karma, average post karma, longevity, total comments, total posts, and user connectedness for each of the user as covariates. This allows us to account for both the past quantity and quality of submissions, and, unlike Kilgo et al., we are able to use r/antiwork-specific metrics and not Reddit-wide ones. Longevity is calculated as the time period between the first activity of the user on r/antiwork and the date the post.
3. We expect that posts may attract varying levels of attention simply due to the topic covered, and thus we include topics as controlling variables.
4. User-specific features were estimated using a cut-off of 10 January, 2022, and posts used to assess predictive power were taken from the period January 10-16, 2022. As our overall dataset was collected in early February 2022, we believe the 3-week window is sufficient to limit any censorship effects on post popularity. After removing posts that were posted by deleted users, there are 5,527 posts in this time window.

Our results show that these factors have limited predictive power (adjusted R^2 of 0.19). Out of the factors considered,

statistically significant effects relate to total post karma and number of posts (p-value < 0.001) as well as some of the indicator variables for the topics. We find that number of posts a user has previously posted has a negative predictive effect on post score (0.54% less karma points expected for every 1% increase in total number of posts), whereas total post karma has a positive effect (0.36% more karma points expected for every 1% increase in total user karma). Given that average post karma is not significant, we can interpret the result as an indication that users posting a lot of content are rewarded negatively, unless their average post quality is high. At the same time, just having high average post quality (without the associated quantity) is not sufficient to predict future post popularity.

The results are thus similar to Kilgo et al., who also found that total user karma was the only significant variable, however in our case the impact itself is order of magnitudes smaller (using an identical specification to Kilgo et al., we find the impact to be 0.009 on a linear scale vs. 0.168 reported in their paper). It indicates that r/antiwork may not have strongly pronounced opinion leaders, and it tends to penalize users that create a lot of low-scoring content.

Subreddit Overlap

Activity in other subreddits

Further, we thought that it was important to analyze the other subreddits antiwork users participate in. It helped us characterize the users and gain a high level understanding of what drives them. As this information is not directly available, we used PRAW API to find subreddits where the most active r/antiwork users (defined as ones that made 100+ posts/comments) have recently posted or commented as the proxy, and developed r/antiwork user profiles in this way. Additionally, we found overlap in the themes and activity of anti-work and the other threads. These themes were characteristic of young progressive users who want to drastically reform the current economic system and functiomn of work as a whole.

Through analyzing basic summary statistics and visuals, we can see that the frequency of posts of users with over 100 activity is highly skewed to the left. Most users only have a few posts with outliers posting excessively. We can see this in Figure 6.

When analyzing the subreddits with the most post activity from our high activity antiworkers ordered by activity are r/antiwork, r/AskReddit, r/Showerthoughts, r/LateStageCapitalism, r/NoStupidQuestions r/politics, r/memes, r/PoliticalHumor lostgeneration, and r/unpopularopinion and comment activity . Just from the titles of these threads, we can see the theme of questioning the current economic system, status quo, and political actions that created the modern industrial complex. Additionally, we see undertones of humor, likely sardonic and used as a coping mechanisms for frustrations with the contemporary workforce.

When comparing these subreddits to the topic models we can see that the topics ”union”, ”rights”, ”covid”, and ”society” relate to the threads r/politics and r/LateStageCapitalism as these topics pertain to policy making that is done through government entities.

Using (Unknown 2022) we compute a similarity matrix where the score indicates the likelihood an antiwork user posts in that thread compared to an average reddit user. The numerical score is the likelihood an antiwork user posts or comments in that other subreddit compared to an average reddit user.

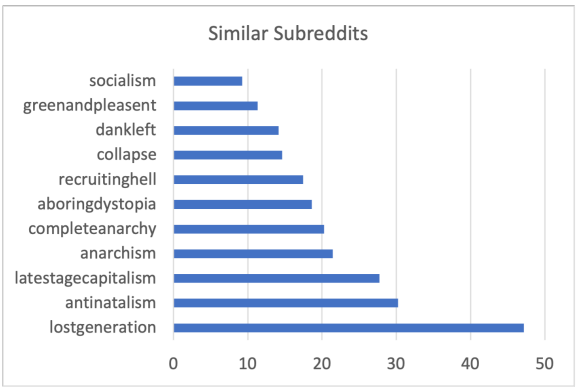

Figure 3: Similarity Scores for /r/antiwork

Next, we run the same analysis but on the subreddits with the most activity by top antiwork users by posts then comments. We include a the baseline r/AskReddit and an antithetical group r/conservatives. For posts we found the similar communities, we find that r/LateStageCapitalism and r/lostgeneration form communuties with significant overlap. These subreddits share similar ideas of going against current labor practices. These communties share very strong similarity scores.

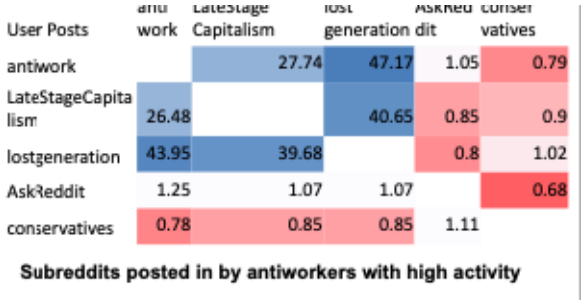

Figure 4: Similarity Scores for Top User Posts

For comments we see that r/news, r/WhitePeopleTwitter, and r/worldnews form similar communities. These subreddits have different themes of general news from posts. Additionally, these scores are significantly lower than posts. This indicates that top users will post about things that are more economically and politically motivated, but will comment more frequently on general news subreddits. General news can be less dense and theoretical which is likely why there were significantly more comments than posts in our data sample.

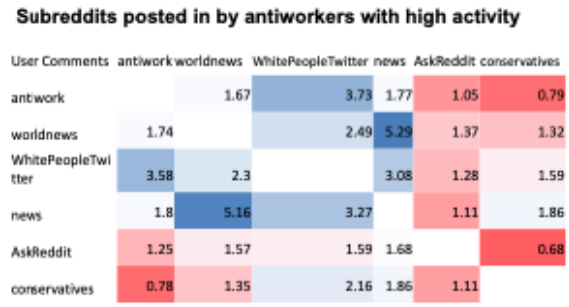

Figure 5: Similarity Scores for Top User Comments

Finally, we evaluated if this data could predict the user karma. We ran Regression Analysis on quantity of posts in the top 10 related subreddits as a predictor of total user post and comment karma in r/antiwork. Our User Post model explained 23.4% of the variation in the data with Antiwork, LateStageCapitalism, and lostgeneration having the largest coefficients. Our Comment model only explained 4% of the variation in the dataset.

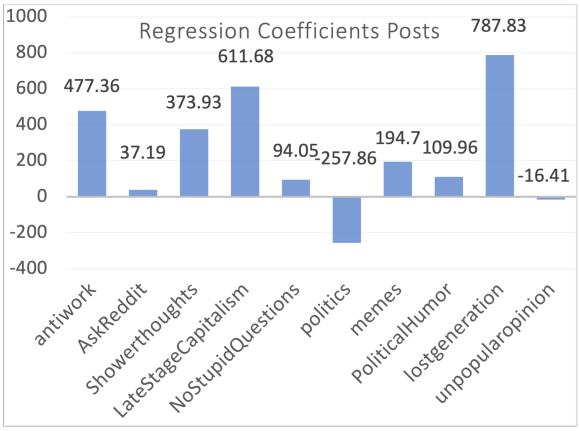

Figure 6: Post Users Regression Coefficients

Work Division

Albert:	Topic and Post Analysis
Aurimas:	User Analysis
Lawton:	Activity in other subreddits

Conclusion

Overall, we were able to gather a variety of insights into r/antiwork throughout our work. On a high-level, our findings indicate that while r/antiwork may have been a more ideological or philosophical subreddit in the past, it has become predominantly a venting space focused on personal stories related to work, with little indication of strong subcommunity structures and opinion leaders. This, perhaps, is not so surprising given its explosion in popularity which likely made its dynamics similar to ones observed in Reddit overall.

We showed that nuanced analysis of a subreddit is important to draw detailed insights. Our novel contribution of identifying and classifying post archetypes was critical in being able to uncover a more detailed discussion topic space on r/antiwork, something that would not have been possible using the approaches typically used in such research. Similarly, we showed that incorporation of subreddit specific user attributes (e.g. post karma on r/antiwork vs. total user karma) results in better explanatory performance of models identifying opinion leaders, and that different interaction type modes need to be considered when attempting to identify sub-community clusters. Similar to prior research Kilgo et al., we found that user longevity is not indicative of their ”power” on the subreddit, and that it is mostly earned via a track record of high quality submissions and discussions. Finally, we showed that r/antiwork users are also active in other subreddits related to similar social topics, and that participation in those subreddits is associated with higher overall post karma accumulated on r/antiwork. These findings are in line with other research that analysed political communities on Reddit and found evidence of cross-subreddit participation (Soliman, Hafer, and Lemmerich 2019).

Future Work

Despite all the analysis performed, we believe there is significantly more work that could be done to retrieve further insights.

We were only able to classify post archetypes related to personal stories. Extending this work to other post types and then building out topic models for each post type separately would likely yield significantly more insights. Second, LDA-centric post archetypes lacks nuance. Many unintuitive steps were taken to improve performance, such as considering job requirements social media and music a miscellaneous image category. The content-centric post archetypes would provide more nuanced and accurate information. Third, Reddit posts are multi-modal. This paper only covers text classification techniques. Using image classifiers would drastically improve the performance of image-based post archetypes which are particularly prevalent in the content-centric post archetypes.

Given that r/antiwork experienced a significant growth over the last few years, time-series analysis would likely yield further insights into its development, and could be applied to all aspects (topics, user interaction dynamics, broader Reddit participation) covered in our work.

Finally, it would be interesting to specifically focus on real-world influence of r/antiwork, as this should be the real outcome if it is a true reflection of a social movement. Understanding if the content and interactions of r/antiwork lead to its members changing jobs, taking to streets to protest or taking other real-world actions would be extremely insightful and definitely possible with a more tailored analysis.

All in all, r/antiwork data is extremely rich and complex, and there is much work to be done to properly use all of the information to reveal underlying subreddit traits. We've only touched the surface.

References

Andrei Kashcha, S. G. 2022. visualization of related subreddits. https://github.com/anvaka/sayit.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan): 993–1022.

Botz, L. 2021. The great resignation: A workers' movement in America. URL https://internationalviewpoint.org/spip.php?article7364.

Hajian, B.; and White, T. 2011. Modelling influence in a social network: Metrics and evaluation. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 497–500. IEEE.

Huffaker, D. 2010. Dimensions of leadership and social influence in online communities. *Human Communication Research* 36(4): 593–617.

Kilgo, D. K.; Yoo, J. J.; Sinta, V.; Geise, S.; Suran, M.; and Johnson, T. J. 2016. Led it on Reddit: An exploratory study examining opinion leadership on Reddit. *First Monday* 21(9). doi:10.5210/fm.v21i9.6429. URL https://journals.uic.edu/ojs/index.php/fm/article/view/6429.

Liu, Y.; and Yin, Z. 2020. Understanding Weight Loss via Online Discussions: Content Analysis of Reddit Posts Using Topic Modeling and Word Clustering Techniques. *J Med Internet Res* 22(6): e13745. ISSN 1438-8871. doi:10.2196/13745. URL https://www.jmir.org/2020/6/e13745.

Pandrekar, S.; Chen, X.; Gopalkrishna, G.; Srivastava, A.; Saltz, M.; Saltz, J.; and Wang, F. 2018. Social Media Based Analysis of Opioid Epidemic Using Reddit. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2018: 867–876. ISSN 1942-597X. URL https://pubmed.ncbi.nlm.nih.gov/30815129.

Parker, R.; and Clark, B. Y. 2022. Unraveling the great resignation: Impacts of the COVID-19 pandemic on Oregon workers. *SSRN Electron. J.* .

Schofield, A.; and Mimno, D. 2016. Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics* 4: 287–300. doi:10.1162/tacl_a_00099. URL https://aclanthology.org/Q16-1021.

Shen, J. H.; and Rudzicz, F. 2017. Detecting Anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 58–65. Vancouver, BC: Association for Computational Linguistics. doi:10.18653/v1/W17-3107. URL https://aclanthology.org/W17-3107.

Soliman, A.; Hafer, J.; and Lemmerich, F. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, 259–263.

Thukral, S.; Meisheri, H.; Kataria, T.; Agarwal, A.; Verma, I.; Chatterjee, A.; and Dey, L. 2018. Analyzing behavioral trends in community driven discussion platforms like reddit. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 662–669. IEEE.

Unknown. 2022. User/Commenter Overlap. https://subredditstats.com/subreddit-user-overlaps.

US Bureau Of Labor Statistics. 2021. Quits rate of 2.9 percent in August 2021 an all-time high. URL https://www.bls.gov/opub/ted/2021/quits-rate-of-2-9-percent-in-august-2021-an-all-time-high.htm. Bureau of Labor Statistics, U.S. Department of Labor, *The Economics Daily*.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing* 17(4): 395–416.

Wikipedia. 2020. Antiwork. URL https://en.wikipedia.org/wiki/R/antiwork.