Natural Language Processing to Analyze Abuse and Domestic Violence Subreddits During COVID-19

Amie Kong
Khoury College of Computer Sciences
Northeastern University
kong.am@northeastern.edu

Abstract

With school closures and stay home orders being implemented as a result of COVID-19, it is unclear how victims of abuse and domestic violence were impacted. Analyzing the effect of COVID-19 on the community of domestic violence and abuse support groups will provide data to healthcare providers and social workers to address the concerns of those victims. With social media being the most accessible platform for global users, people have turned to Reddit as a form of "seeking solace." The aim of this study is to use natural language processing (NLP) to analyze subreddit submissions of domestic violence and abuse support groups (r/abuse, r/domesticViolence) before and during the initial stages of the COVID-19 pandemic. By extracting NLP features from posts in each subreddit, comparisons can be made on each subreddit before and during the pandemic. From the Linguistic Inquiry and Word Count analysis, the top psychological processes that make up abuse and domestic violence subreddit posts include social processes, cognitive processes, and relativity. The sentiment analysis demonstrated a high rate of negative sentiment posts in both the abuse and domestic violence subreddits. Based on the Latent Dirichlet Allocation (LDA) topic model results, the domestic violence subreddit during the pandemic had more posts pertaining to "children" and "abuse". As for the abuse subreddit, there were more posts related to specific topics such as "police" and "relationship" during the pandemic, which suggests that victims were opening up about issues regarding the police and relationships. The identification of sensitive topics over time is beneficial in providing social workers who help victims of abuse and domestic violence to help them understand their thought processes.

1 Approach

English stop words were removed from posts and preprocessing depended on the analysis. VADER sentiment analysis and Linguistic Inquiry and Word Count (LIWC) were features extracted from the body text of each post using an installed vaderSentiment package and a LIWC software [1]. For extracting the polarity scores for VADER sentiment analysis, the original text was used because uppercased text has significance[3]. From those feature extraction methods, detection of patterns or trends in the subreddit will be useful in highlighting any drastic changes during the pandemic. The subreddit datasets with posts in the prepandemic time frame act as the baseline for this project. Creating a topic model using Latent Dirichlet Allocation (LDA) will uncover any noticeable topics that arose from the mid-pandemic in the subreddit, whether it be a new topic or a decline or increase in the term frequency of a topic or word [6]. The code for the LDA model was modeled after kapadias' tutorial. [1]

2 Data Downloading and Preprocessing

Datasets were extracted using the Pushshift API (pushshift.io) from 4 different subreddits (r/abuse, r/domesticViolence, r/ptsd, r/CSEducation) during different time frames: pre-pandemic (November 1, 2018 to November 1, 2019), mid-pandemic (January 15, 2020 to May 1, 2020), and a control time frame to compare the quantity of submissions before and during the pandemic (January 15, 2019 to May 1, 2019). Another dataset was created as a separate experimental group: (r/covid) in a shortened time frame of January 15, 2020 to April 1, 2020 to account for the influx of post submissions and it made it easier to handle the data. Only posts including a title and a body were included to exclude

[1] https://github.com/kapadias/mediumposts/blob/master/nlp/published_notebooks

posts with just images. The datasets were then saved as CSV files and distinguished by the different time frames and subreddit. The number of submissions for each dataset range from 792 to 8682 due to the popularity of certain subreddit groups (e.g. r/ptsd and r/home) that possess more posts overall. Dropping and querying of the datasets was necessary to normalize the data into the following columns: Post Id, Title, Body, Author, Publish Date, and Total No. of Comments. Using the NLTK and redditclean package the "Body" column text of each submission post was further cleaned and preprocessed by removing English stopwords, punctuation, and url links.

3 Subreddit Analysis Using LIWC Feature Extraction

Linguistic Inquiry Word Count (LIWC) features from the body of each subreddit was extracted to perform an analysis on the dominant traits present in each subreddit [1]. 93 features were extracted using the LIWC software which includes the general LIWC features such as word count, analytic, cloud, tone, pronoun, as well as other features that detect other scenarios such as anger, sad, social (all LIWC features extracted in the following order):

'WC', 'Analytic', 'Clout', 'Authentic', 'Tone', 'WPS', 'Sixltr', 'Dic', 'function', 'pronoun', 'ppron', 'i', 'we', 'you', 'shehe', 'they', 'ipron', 'article', 'prep', 'auxverb', 'adverb', 'conj', 'negate', 'verb', 'adj', 'compare', 'interrog', 'number', 'quant', 'affect', 'posemo', 'negemo', 'anx', 'anger', 'sad', 'social', 'family', 'friend', 'female', 'male', 'cogproc', 'insight', 'cause', 'discrep', 'tentat', 'certain', 'differ', 'percept', 'see', 'hear', 'feel', 'bio', 'body', 'health', 'sexual', 'ingest', 'drives', 'affiliation', 'achieve', 'power', 'reward', 'risk', 'focuspast', 'focuspresent', 'focusfuture', 'relativ', 'motion', 'space', 'time', 'work', 'leisure', 'home', 'money', 'relig', 'death', 'informal', 'swear', 'netspeak', 'assent', 'nonflu', 'filler', 'AllPunc', 'Period', 'Comma', 'Colon', 'SemiC', 'QMark', 'Exclam', 'Dash', 'Quote', 'Apostro', 'Parenth', 'OtherP'

The psychological processes summed by the 9 subprocesses (affective, social, cognitive, perceptual, biological, informal language, and drives), were extracted from each subreddit to study the general psychological processes used in posts, as well as for comparison before and during the pandemic [1].

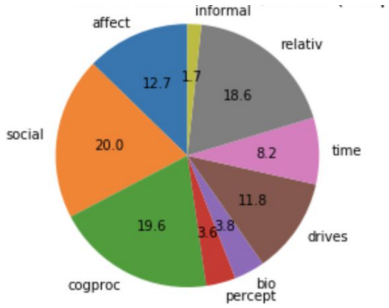

Figure 1: Pie graph of breakdown of the 9 subprocesses that make up the psychological processes in the Linguistic Inquiry and Word Count scores of the r/abuse subreddit in the mid-pandemic data set.

A change in the psychological processes categories of cognitive processes and relativity existed for the CS Education subreddit, which is the control group of this study. During the pandemic, the frequency of words that fall under the cognitive processes category increased by 4 percent and relativity increased by 2 percent. It also has a higher portion of language classified under "drives" (summed up by categories: achievement, power, and reward scores); therefore, we can deduce that this subreddit group primarily includes posts that relate to drive. As for the other control dataset of the PTSD subreddit, all areas of psychological processes remain consistent before and during the pandemic with the top two processes that generalize the PTSD subreddit being cognitive processes (e.g. insight, causation, discrepancy, tentative, certainty) and relativity (e.g. motion and space) [1].

For the abuse subreddit, the top processes that generalize this data set are social, cognitive, and relativity; however, no change in frequency distribution of the categories was detected in pre-pandemic vs the mid-pandemic dataset. Similarly, the top processes that generalize the domestic violence subreddits are social, cognitive, and relativity, with no change detected in the frequency distribution of the psychological processes in the pre-pandemic vs the mid-pandemic dataset.

4 Extracting VADER Features for Sentiment Analysis

Valence Aware Dictionary and Sentiment Reasoner (VADER) is a "a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media." It measures the strength and direction of sentiment across an entire text and the algorithm adjusts for negations and

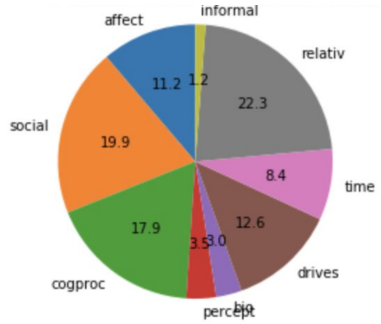

Figure 2: Pie graph of breakdown of the 9 subprocesses that make up the psychological processes in the Linguistic Inquiry and Word Count scores of the r/domesticViolence subreddit in the mid-pandemic data set.

booster words, returning the proportion of the text that is negative, positive, neutral, and a combined score [3]. VADER scores were extracted for each subreddit by importing SentimentAnalyzer from the NLTK package. The polarity scores were concatenated onto the data frame and an analysis on the compound scores were done to examine the positive, neutral, and negative posts of each subreddit.

Compound scores generated from the sentiment values for our control groups, CS Education and PTSD subreddits, showed a contrast in sentiment; with CS Education posts primarily falling under neutral to positive before and during the pandemic and PTSD posts falling under the negative range for both time periods.

Compound scores for the abuse subreddit group consist primarily of posts falling under negative sentiment; however, there was a 8 percent increase in the percentage of posts falling under neutral during the pandemic. A separate control data set, including posts from January 1, 2018 to May 15, 2018 of the r/abuse subreddit, was extracted a few months before the pre-pandemic period to minimize this increase as being due to chance.

For the domestic violence subreddit, a 12 percent drop in the number of neutral posts during the pandemic and a 9 percent increase in the number of negative posts were observed. When compared with the control dataset, including posts from January 1, 2018 to May 15, 2018 of the r/domesticViolence subreddit, the mid-pandemic dataset closely resembled the statistics of the control dataset; hence, suggesting that it was due to chance.

	Neutral	Negative
Control	15%	54%
Prepandemic	14%	55%
Midpandemic	22%	48%

Table 1: Table for VADER compound scores for the r/abuse subreddit shows that many submissions have more negative than positive sentiment.

	Neutral	Negative
Control	19%	45%
Prepandemic	33%	38%
Midpandemic	21%	47%

Table 2: Table for VADER compound scores for the r/domesticViolence subreddit shows that many submissions have more negative than positive sentiment.

Vader Feature Analysis On Abuse Subreddit: For Figure 1 and Figure 2, VADER compound scores were calculated for each subreddit post in r/abuse and r/domesticViolence subreddit. The tables show the percent breakdown of negative, neutral, and positive compound scores of each subreddit. The compound scores range from -1 to 1, with -1 being a strong negative and 1 being a strong positive. This was done for pre-pandemic and mid-pandemic posts.

5 Latent Dirichlet Allocation Topic Modeling on Subreddits

A probabilistic modeling approach of topic modeling, LDA (Latent Dirichlet Allocation), was performed on each data set to detect the top underlying topics and if there exists any noticeable changes for each subreddit prepandemic vs midpandemic [4]. In LDA, each word in a text document derives from a topic and the topic is selected from each document over topics. The probability of a word given document $P(w|d)$ is equal to:

$$\sum_{t \epsilon T} p(w|t,d)\, p(t|d)$$

$td = P(t|d)$ is the probability distribution of

topics in documents $wt = P(w|t)$ is the probability distribution of words in topics

With T being the total number of topics and W being the total number of words in our vocabulary for all documents. Assuming conditional independence, $P(w|t, d) = P(w|t)$, making $P(w|d)$ equal to:

$$\sum_{t=1}^{T} p(w|t)\, p(t|d)$$

In other words, the dot product of td and wt for each topic t.

LDA was used to detect any topic changes in the extracted pre-pandemic subreddit submissions vs the mid-pandemic subreddit submissions. A bag-of-words corpus was created, which was used to create an LDA model and 5 topics were created. Models were also generated multiple times with different numbers of words for each topic to address the consistency of topics. Using the scikit package, a manually chosen LDA model with 5 topics was then applied to the r/abuse subreddit to assess the distribution of topics, allowing for comparison between the distribution of subreddit submissions pre-pandemic vs mid-pandemic. Furthermore, the gensim package was installed to generate the interactive pyLDAvis visualization maps [2].

5.1 Jensen-Shannon Divergence

The Jensen-Shannon Divergence was used to generate the intertopic distance map to help show the similarities between topics provided by the LDA model outputted for each subreddit data set. Based on the Kullback-Leibler divergence, the Jensen-Shannon divergence is a popular method of measuring the similarity between two probability distributions [6]. The Jensen-Shannon divergence measures the similarity between two distributions; therefore, by applying Jensen-Shannon divergence to the topic assignment for the subreddit submission, it will allow us to measure the distance and similarity between subreddit submission in the particular subreddit data set . For probability distributions P and Q, Kullback-Leibler divergence of Q from P is defined to be:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Thus, the Jensen-Shannon Divergence of Q from P is defined as:

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

r/abuse Topics (Pre-pandemic)	Terms Extracted by LDA Topic Model
Topic 1	im dont mom want says f*****g gets people
Topic 2	time told feed said things want relationship friends
Topic 3	mom school room house time family money told
Topic 4	mom dad time told mother brother remember said
Topic 5	abuse want help feel people abusive life abused

r/abuse Topics (Mid-pandemic)	Terms Extracted by LDA Topic Model
Topic 1	time feel mom told things want said dad
Topic 2	people abuse want time life feel think mother
Topic 3	father people think abuse wife started pain said
Topic 4	time police told think life people started friend
Topic 5	deleted used nick female better girl person

Table 3: This table shows the topics extracted from the r/abuse subreddit using the LDA topic model.

where $M = \frac{1}{2}(P + Q)$

6 Results

Topics Found Using LDA Topic Modeling On Abuse Subreddit: Table 1 shows the topics extracted from the r/abuse subreddit. Figures 3 and 4 are intertopic distance maps for the r/abuse subreddit before and during the pandemic. The terms found in the topics included familial terms such as different variations of "mom" and "dad" before and during the pandemic. There is a lack of mention of "relationship", "school", "room," and "money" during the pandemic, which were in the top 30 salient terms before the pandemic.

Topics Found Using LDA Topic Modeling On Domestic Violence Subreddit: Table 2 shows the topics extracted from the r/domesticViolence subreddit. Figures 1 and 2 are inter-topic distance maps for the r/domesticViolence subreddit before and during the pandemic. The noticeable area of Topic 1 in Figure 2 suggests that Topic 1 of "children people time help abuse violence domestic going" is a popular topic during the pandemic. There are terms in Topic 1 of r/domesticViolence during the pandemic that does not appear in the topics before the pandemic such as "children," "domestic," and "violence." These terms were not in the top 30 salient terms in the pre-pandemic data set.

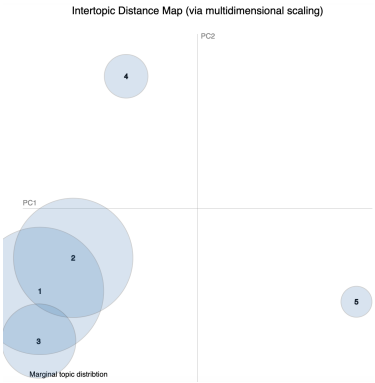

Figure 3: The intertopic distance map shown for r/abuse before the pandemic demonstrates Topic 1 and Topic 2 as being the most prevalent topic since the areas of those topics are the greatest. The distance between the epicenters of Topics 1, 2, and 3 are closer than the distance between Topic 1 and Topic 5, meaning that those topics are least similar.

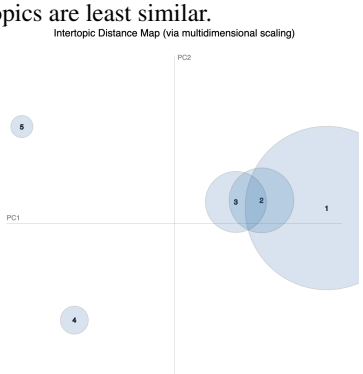

Figure 4: The intertopic distance map shown for r/abuse during the pandemic demonstrates Topic 1 as being the most prevalent since the area is tripled that of the other topics. Topics 2 and Topic 3 rank similarly in prevalence along the subreddit posts, since the area of the circles are equivalent, while Topic 4 and Topic 5 are the least prevalent. The distance between the epicenters of Topics 1, 2, and 3, are closer than the distance between Topic 1 and Topic 5 and Topic 1 and Topic 4, suggesting that those topics are not that similar.

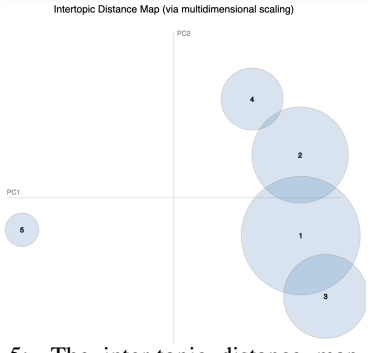

Figure 5: The inter-topic distance map shown for r/domesticViolence before the pandemic demonstrates Topic 1 as being the most prevalent topic since the area is the greatest, while Topics 2 and 3 share similar prevalence along the subreddit posts, since the areas of those circles are equivalent. The distance between the epicenters of the circles of Topic 1, 2, and 3 are closer than the distance between Topic 1 and Topic 5.

r/domesticViolence Topics (Pre-pandemic)	Terms Extracted by LDA Topic Model
Topic 1	said time things feel started told want leave
Topic 2	house told said want mom going day went
Topic 3	going time home order ex right told said
Topic 4	time removed told going feel years things life
Topic 5	Want help feel family time abuse abusive things

r/domesticViolence Topics (Mid-pandemic)	Terms Extracted by LDA Topic Model
Topic 1	children people time help abuse violence domestic going
Topic 2	police violence feel want said things going years
Topic 3	sister want abuse tell feel time started told
Topic 4	removed years want help abusive life away day
Topic 5	house family help want time parents leave feel

Table 4: This table shows the topics extracted from the r/domesticViolence subreddit using the LDA topic model.

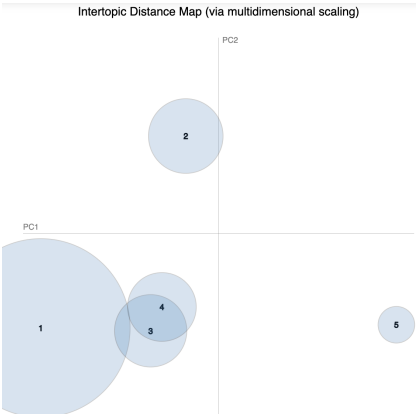

Figure 6: The inter-topic distance map shown for r/domesticViolence during the pandemic demonstrates Topic 1 as being the most prevalent since the area is tripled that of the other topics. Topics 2, 3, and 4 share rank similarly in prevalence along the subreddit posts, since the area of the circles are equivalent, while Topic 5 is the least prevalent. The distance between the epicenters of Topics 1, 3, and 4 are closer than the distance between Topic 1 and Topic 5.

7 Conclusion

From the LIWC analysis, the top psychological processes that comprise abuse and domestic violence reddit word text are similar, with social processes, cognitive processes, and relativity being the top 3 categories. No significant change existed in the frequency of words that fall under the 9 psychological processes in the abuse, domestic violence, and PTSD subreddit posts.

The sentiment analysis generalized a high rate of negative posts in abuse and domestic violence subreddits when compared to the CS Education subreddit, which demonstrated the majority of posts falling under neutral-positive.

Based on the topic model results, the domestic violence subreddit during the pandemic had more posts pertaining to "children" and "domestic violence". The rise in presence of the terms "children", "domestic," and "violence" during the pandemic is alarming since these terms were not in the top 30 frequent terms used in the discussions before the pandemic. It could suggest that there are more discussions regarding children and domestic violence.

As for the abuse subreddit, the mention of the terms "relationship", "school", "room," and "money" during the pandemic were not in the top 30 frequent terms, when it was used frequently before the pandemic. Besides that, there were no noticeable changes or trends in the intertopic distance maps for r/abuse before and during the pandemic. Perhaps, scraping more subreddit submissions would improve our model and detect any trends since our data set for the mid-pandmeic is half the size of the pre-pandemic data set.

Better understanding the thoughts of victims of abuse and domestic violence allows social workers to assist the victims by having knowledge of sensitive topics that the victims are concerned about.

As for future work, attempting other LDA topic models and using coherence scores as a metric to measure the similarities of the topic results outputted by each model would be beneficial in analyzing the performance of the topic model results, rather than completing an observational study for topic modeling.

References

[1] Pennebaker JW, Booth RJ, Boyd RL, Francis ME. LIWC 2015 Operator's Manual. Austin, TX: Pennebaker Conglomerates Inc; 2015.

[2] Tomar, A. Topic modeling using Latent Dirichlet Allocation(LDA) and Gibbs Sampling explained!. 2019, July 25

[3] Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., Talkar, T. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. Journal of medical Internet research. *Journal of medical Internet research, 22(10), e22635.* 2019, July 25

[4] Park, A., Conway, M., Chen, A. T. Examining Thematic Similarity, Difference, and Membership in Three Online Mental Health Communities from Reddit: A Text Mining and Visualization Approach. . *Computers in human behavior, 78, 98–112. https://doi.org/10.1016/j.chb.2017.09.001* 2018

[5] Minna Lyons, Nazli Deniz Aksayli, and Gayle Brewer. Mental distress and language use: Linguistic analysis of discussion forum posts.. *Computers in Human Behavior, 87:207–211.* 2018

[6] Blair, S. J., Bi, Y., Mulvenna, M. D. Aggregated topic models for increasing social media topic coherence.. *Applied Intelligence, 50(1), 138-156. doi:10.1007/s10489-019-01438-z* 2019