AN INTERACTIVE TOPIC MODEL OF SOCIAL MEDIA DATA TO UNCOVER HOW COVID-19 AFFECTS MENTAL HEALTH

Muthukumar Rajendran
mt911642@dal.ca
Dalhousie University
Halifax, Nova Scotia, Canada

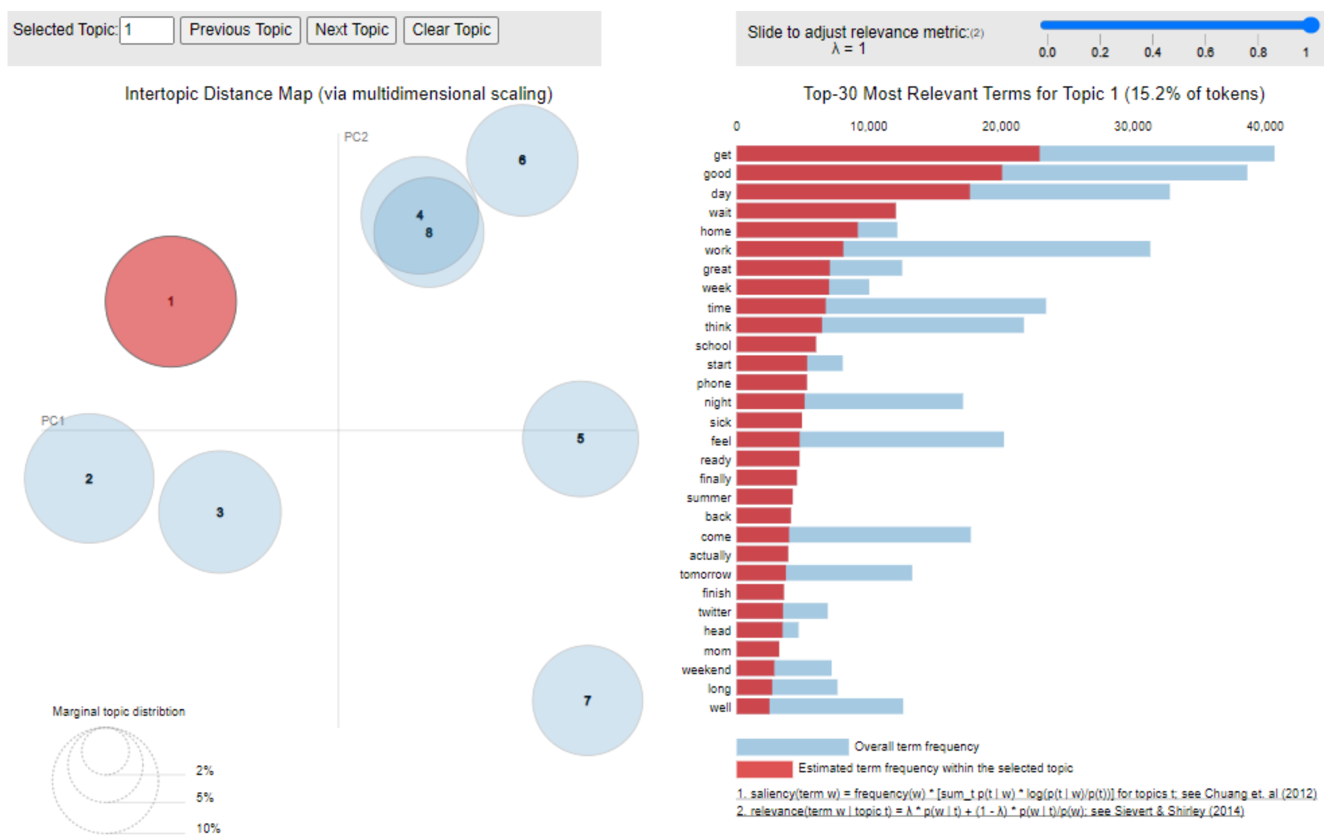

Figure 1: Topics using pyLDAvis visualization, with the global topic view on the left and the term bar-charts (with Topic-1 selected) on the right. The linked selections allow users to reveal aspects of the topic-term relationships compactly. The user has chosen one among the 9 topics to populate the barchart with the most relevant terms for that topic. The area of circle represents the importance of each topic over the entire corpus, the distance between the center of circles indicate the similarity between topics.Each bubble represents a topic. The larger the bubble, the higher percentage of the number of tweets in the corpus is about that topic

Fall '20, Dec 04, 2020, Halifax, NS

ACM ISBN 978-1-4503-2611-30/20/11…$15.00
https://doi.org/10.1145/1122445.1122456

ABSTRACT

Social media is the main online means of interaction among individuals during this COVID-19 pandemic situation. People are increasingly using social media, especially online communities, to discuss health concerns and seek support. Understanding topics, sentiment, and structures of these communities informs important aspects of health-related conditions. There has been growing research interest in analysing online mental health communities; however, analysis of these communities with health concerns has been limited. This paper investigates and identifies latent meta-groups of online communities with mental health-related conditions including depression and anxiety. Large datasets from online communities were crawled. We analyse topic-based features from posts made by members of these online communities. The work focuses on using LDA as the topic modelling approach to infer latent topics automatically from the corpus. The visualization of the discovered communities or topics is visualized by a visual library called pyLDAvis. This presents evidence of various topics in online mental health-related communities.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

datasets, Topic Modelling, neural networks, Interactive visualization

ACM Reference Format:
Muthukumar Rajendran. 2018. AN INTERACTIVE TOPIC MODEL OF SOCIAL MEDIA DATA TO UNCOVER HOW COVID-19 AFFECTS MENTAL HEALTH. In *Directed Studies course project report, Dec 04, 2020 - Halifax, NS.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The advantage of social media has been improving the quality of healthcare. Social media sites such as Facebook, Twitter, Reddit and Tumblr have become increasingly recognized as a promising platform for healthcare support and intervention. People have been moving away from their traditional communication, now meeting up online and using social media tools to make a different communication model. Online communities have been built up as forums for individuals to share information and advice in a variety of their daily life, especially their health beings. With the popularity of social networks, social media offers a low-cost sensing channel to analyse health behaviours of individuals and communities through their postings.

Mental healthcare is another area that gradually benefits from social media. By accessing social media sites (e.g. Reddit and Twitter), individuals with depressive symptoms can make connection with others for sharing their experiences, finding answers to health questions, and expressing themselves with many kinds of data.

Mental healthcare is another area that gradually benefits from social media. By accessing social media sites, individuals with depressive symptoms can make connection with others for sharing their experiences, finding answers to health questions, and expressing themselves with many kinds of data. In a social study, up to 41 percent of participants with depression answered that online communities help them to reconnect individuals and overcome their depressive states. Several latent patterns and factors within online depression-related communities dictate this process, including topics of discussion, posting behaviour, demographic information, relationship, interaction, and emotions. Furthermore, the core symptoms of depression, such as social withdrawal and sleep disturbance, are also characteristics of autism spectrum disorders (ASD) which are associated with the most psychiatric disorder as depression. In these conditions, social media can be seen a life saver.

COVID-19 is having a negative impact on peoples mental health, with many seeing their stress levels double since the onset of the pandemic. People are struggling with fear and uncertainty about their own health and their loved ones health, concerns about employment and finances, online education and the social isolation that comes from public health measures such as quarantining and physical distancing. A recent poll found that 50 percent of Canadians reported worsening mental health since the pandemic began with many feeling worried (44 percent) and anxious (41 percent). One in 10 Canadians polled said that their mental health had worsened 'a lot' as a result of COVID-19. Similar results were found in a survey of Canadian workers, where 81 percent reported that the pandemic is negatively impacting their mental health, indicating a significant drop in overall worker mental health since the beginning of COVID-19

The content of blog posts in social media sites is more subjective than others. It is suitable for a topic modelling on both individuals and community contexts. The features are either manually textual or predefined terms where words or phrases are mapped to sentiment-bearing scores. In addition, other works examined in features representing user's mobile behaviours and environment to perform the demographic and location prediction. However, profiling at multiple scales, from an individual to their joined community, is crucial within and across online communities. It needs the correct interpretation of features in context. Social media research has investigated global patterns of behaviours from large-scale data instead of seeking user-centric patterns. It requires to build tools for inferring and utilizing both the complex and dynamic relationships between online users from individuals to their community perspectives. While questionnaire-based methods continue to dictate the research by social scientists, data-driven methods for online communities with mental health-related concerns are still in early stages. Understanding risk factors and latent patterns in these online mental health-related communities is an important step in many aspects of coping with mental health issues, ranging from singling out individual aspect for support (e.g. relationship between long working hour and depression) to informing preventative healthcare policy (e.g. by looking at clustering of communities spatially).

This study aims to examine patterns and formations of online mental health-related communities including depression, anxiety, and general online communities. Using data crawled from Twitter and Reddit, we apply LDA topic modelling to automatically infer latent topics of interest in using generic words and the set of effective information among the communities. In addition to exploratory analysis, we demonstrate the usefulness of latent topics by further clustering these communities into meta-groups. Visualization of our clustering using pyLDAvis results indicate that these meta-groups of communities can be well separated when projecting on 2D spaces. This demonstrates the evidence of people getting affected mentally through online mental health-related communities, suggesting a possible angle for support and intervention.

Our contributions of this work are: (1) a novel problem on analysing both generic-based and sentiment-based topics in online mental health-related communities including those related to depression and anxiety symptom; (2) a visualization tool that helps to view the topic clusters interactively.

2 RELATED WORK

In this related work we will be explaining about the social media, the topic modelling approach that we use, about the topic interpretation and coherence and visualization of topics in an interactive way.

2.1 Twitter and Reddit

Twitter is one of the most popular social media in the world since it was first published in 2006. Various information can be obtained from social media Twitter which is now increasingly being used. Users use it for various needs, for example for public, government and business needs. Tweets shared by Twitter users cover a variety of specific topics. Users can share opinions about the shared tweet. Topics represent the contents of many tweets that discuss the same context. From these tweets can be found the main topics that are being discussed by many users at that time by conducting topic analysis. Twitter is present as a means of communication to exchange information about various events in the real world, short messages on Twitter generally reflect various events experienced by users in real-time.

Reddit is basically a large group of forums in which registered users can talk about almost anything you can imagine, from news, to pop culture, to technology, to comics, to film, to literature, to the weirdest things in the world, including Not Safe For Work stuff. Those specialized forums are called "sub-reddits," which are referred to as "topic" (example: Depression). There are more than 138,000 active sub-reddits. You can read and participate in all of them freely except private sub-reddits, which require an admission process. You can also subscribe to the sub-reddits, so their most popular posts appear on your personalized Reddit front page.

2.2 Social media and health

An increasing body of work has been interested in exploiting how social media can be used to infer the people health behaviours. In psychology, health behaviours shape the health and well-being of individuals, communities, and populations. Moreover, showed that social media can reshape healthcare in several ways (i.e. the way doctors and patients interact). By detecting changes of user-centric behaviours in social media, mental or behavioural health concerns are indicated. It is also revealed that social media data can be analysed to assess the role of sentiment, emotions, or mental status of a community as well as to identify most disease-related conditions and symptoms or mental health issues (e.g. depression or suicide). In addition, whenever posts have been made by someone, these posts can be immediately analysed to identify whether that person is an “at-risk” one in mental health or not. Then mental health support and interventions are considered to deliver self-help or proactive interventions for reducing the risk. Furthermore, some work found apparent evidences that people are progressively spending amounts of their online time to post messages about their health beings (e.g. depression) with treatment on virtual social networks such as Twitter, Facebook and Reddit[7].

2.3 Applied machine learning for community discovery

For learning latent topics from the content of posted messages, probabilistic topic modelling approaches, e.g. probabilistic latent semantic indexing (PLSI), latent Dirichlet allocation (LDA), or hierarchical Dirichlet processes (HDP) have shown to be effective in discovering latent topics from the corpus of blog posts[8]. Several studies used the standard parametric model LDA to learn latent topics from the content of blogs and tweets in the blogosphere for their research on mental health signals in social media. Using LDA to gain latent topics, found significant differences among study cohorts which are characterized by the latent topics of discussion, psycho-linguistic features, and tagged moods. Many studies investigated the impact of topics and language styles among users in different cohorts defined by mood tags, social connectivity, and age from the online depression community. However, the topic modelling approaches face a critical issue in determining the key parameters (e.g. the number of topics in LDA). These parameters are not always available and quite difficult to specify in advance. We address the above parametric limitation by employing the Bayesian non-parametric topic modelling in discovering latent topics from the content of blog posts made by users in online communities with and without mental health-related states

2.4 Topic Model Visualization Systems

A number of visualization systems for topic models have been developed in recent years. Several of them focus on allowing users to browse documents, topics, and terms to learn about the relationships between these three canonical topic model units (Gardner et al., 2010; Chaney and Blei, 2012; Snyder et al., 2013) [5]. These browsers typically use lists of the most probable terms within topics to summarize the topics, and the visualization elements are limited to bar-charts or word clouds of term probabilities for each topic, pie charts of topic probabilities for each document, and/or various bar-charts or scatter-plots related to document metadata. Although these tools can be useful for browsing a corpus, we seek a more compact visualization, with the more narrow focus of quickly and easily understanding the individual topics themselves (without necessarily visualizing documents).

3 METHODOLOGY

In this section, we briefly the background of Latent Dirichlet Allocation (LDA) [3], for inferring latent patterns/topics from data corpus. We then introduce different community representations using a variety of feature sets extracted from the content of blog posts made within the community. Finally, we visualize the topic clusters formed in a interactive way using the pyLDAvis library.

3.1 Topic Modelling

Topic modelling can be defined as a branch in unsupervised natural language processing which is used to represent the text documents or posts with the help of several topics, which can better explain the underlying information in that particular document or the post. It is similar to clustering. Here instead of clustering labels based on numerical data we will be working on text data.

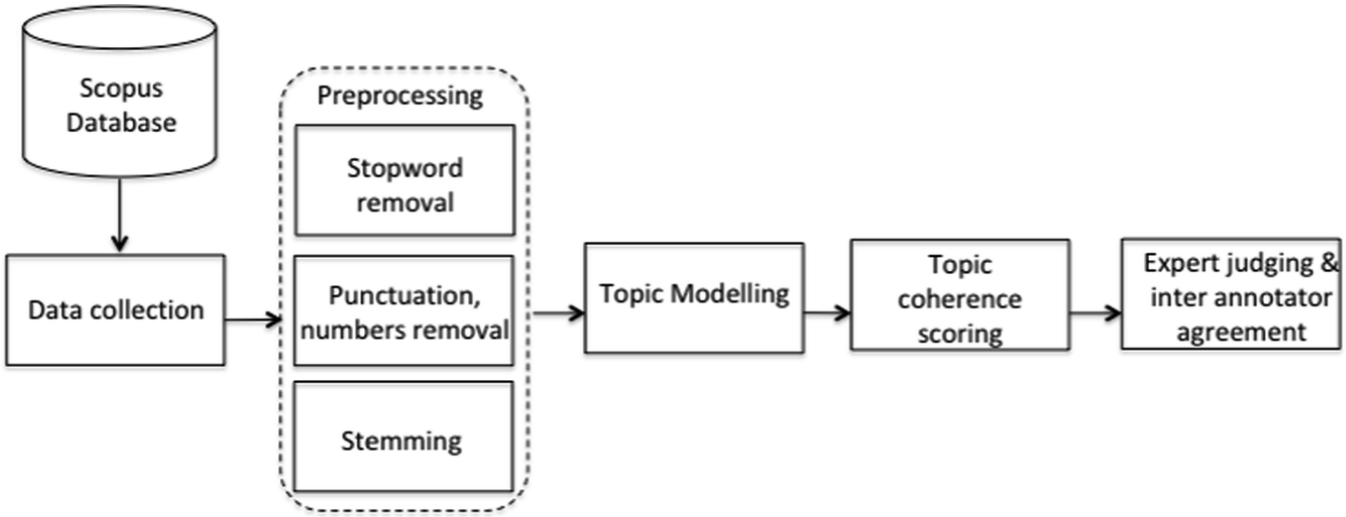

Figure 2: The general topic modelling workflow, where the user gives the corpus as input then pre-process the text. Then given to the topic modelling algorithm to get the topics and the topic coherence is calculated

In the text domain, document clustering aims to organize similar documents into groups, which is crucial for document organization, browsing, summarization, classification. Topic modeling develops probabilistic generative models to discover the latent semantics embedded in document collection and has demonstrated vast success in modeling and analyzing texts.

Document clustering and topic modeling are highly correlated and can mutually benefit each other. On one hand, topic models can discover the latent semantics embedded in document corpus and the semantic information can be much more useful to identify document groups than raw term features. In classic document clustering approaches, documents are usually represented with a bag-of-words (BOW) model which is purely based on raw terms and is insufficient to capture all semantics. Topic models are able to put words with similar semantics into the same group called topic where synonymous words are treated as the same. Under topic models, document corpus is projected into a topic space which reduces the noise of similarity measure and the grouping structure of the corpus can be identified more effectively.

3.2 Topic Interpretation and Coherence

It is well-known that the topics inferred by LDA are not always easily interpret-able by humans. A large user study that standard quantitative measures of fit, such as those summarized by Wallach et al. (2009)[12], do not necessarily agree with measures of topic interpret-ability by humans.

The Turbo Topics developed by Blei and Lafferty (2009) [2], a method of identifying n-grams within LDA inferred topics that, when listed in decreasing order of probability, provide users with extra information about the usage of terms within topics. This two-stage process yields good results on experimental data, although the resulting output is still simply a ranked list containing a mixture of terms and n-grams, and the usefulness of the method for topic interpretation was not tested in a user study. A method for ranking terms within topics to aid interpret-ability called Pointwise Mutual Information (PMI) ranking described by Newman et al. (2010) [6]. Under PMI ranking of terms, each of the ten most probable terms within a topic are ranked in decreasing order of approximately how often they occur in close proximity to the nine other most probable terms from that topic in some large, external "reference" corpus, such as Wikipedia or Google n-grams. Although this method correlated highly with human judgments of term importance within topics, it does not easily generalize to topic models fit to corpora that don't have a readily available external source of word co-occurrences.

Finally, a new statistical topic model that infers both a term's frequency as well as its exclusivity – the degree to which its occurrences are limited to only a few topics by Bischof and Airoldi (2012) [1]. They introduce a uni-variate measure called a FREX score ("FRequency and EXclusivity") which is a weighted harmonic mean of a term's rank within a given topic with respect to frequency and exclusivity, and they recommend it as a way to rank terms to aid topic interpretation. We propose a similar method that is a weighted average of the logarithms of a term's probability and its lift, and we justify it with a user study and incorporate it into our interactive visualization.

3.3 Latent Dirichlet Allocation (LDA)

There many topic modelling approaches available, but for this project we are going to use Latent Dirichlet Allocation also called as LDA. LDA is nothing but it is a distribution of distributions. More specifically, it is the distribution of topics in documents and distribution of words in the topic. LDA says is that each word in each document comes from a topic and the topic is selected from a per-document distribution over topics. Due to this we have two matrices, they are, the probability distribution of topics in documents and the probability distribution of words in topics.

$$\theta td = P(t|d) \tag{1}$$

$$\theta wt = P(w|t) \tag{2}$$

From equations 1 and 2 we can say that the probability of a word given document is equal to,

$$\sum p(w|t,d)p(t|d) \tag{3}$$

where T is the total number of topics. Also, let's assume that there is W number of words in our vocabulary for all the documents.

Latent Dirichlet Allocation (LDA) is a topic modelling algorithm used for extracting topics from a given collection of documents. It builds models in unsupervised mode, i.e., does not need labelled training data. Based on the assumption that a document contains a mixture of N underlying different topics, and the document is generated by these topics with different proportions or probabilities, LDA is able to find out the topics and their relative proportion, which are distributed as a Latent Dirichlet random variable. The algorithm's performance can be managed though assumptions on the word and topic distributions.

3.4 Interactive visualization

Recently much attention has been paid to visualizing the output of topic models fit using Latent Dirichlet Allocation (LDA). Such visualizations are challenging to create because of the high dimensionality of the fitted model. LDA is typically applied to many thousands of documents, which are modeled as mixtures of dozens (or hundreds) of topics, which themselves are modeled as distributions over thousands of terms. The most promising basic technique for creating LDA visualizations that are both compact and thorough is interactivity.

We use an interactive visualization system called as pyLDAvis [10] that attempts to answer a few basic questions about a fitted topic model: (1) What is the meaning of each topic?, (2) How prevalent is each topic?, and (3) How do the topics relate to each other? Different visual components answer each of these questions, some of which are original, and some of which are borrowed from existing tools.

This visualization (illustrated in Figure 1) has two basic pieces. First, the left panel of our visualization presents a global view of the topic model, and answers questions 2 and 3. In this view, we plot the topics as circles in the two-dimensional plane whose centers are determined by computing the distance between topics, and then by using multidimensional scaling to project the inter-topic distances onto two dimensions, as is done in We encode each topic's overall prevalence using the areas of the circles, where we sort the topics in decreasing order of prevalence.

Second, the right panel of the visualization depicts a horizontal bar-chart whose bars represent the individual terms that are the most useful for interpreting the currently selected topic on the left, and allows users to answer question 1, "What is the meaning of each topic?". A pair of overlaid bars represent both the corpus-wide frequency of a given term as well as the topic-specific frequency of the term.

The left and right panels of the visualization are linked such that selecting a topic (on the left) reveals the most useful terms (on the right) for interpreting the selected topic. In addition, selecting a term (on the right) reveals the conditional distribution over topics (on the left) for the selected term. This kind of linked selection allows users to examine a large number of topic-term relationships in a compact manner.

A topic in LDA is a multinomial distribution over the (typically thousands of) terms in the vocabulary of the corpus. To interpret a topic, one typically examines a ranked list of the most probable terms in that topic, using anywhere from three to thirty terms in the list. The problem with interpreting topics this way is that common terms in the corpus often appear near the top of such lists for multiple topics, making it hard to differentiate the meanings of these topics.

4 EXPERIMENTAL SETUP

LDA's approach to topic modeling is, it considers each document as a collection of topics in a certain proportion and each topic as a collection of keywords, again, in a certain proportion. When we provide the algorithm with the number of topics, all it does it to rearrange the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of topic-keywords distribution. Here, a topic is nothing but a collection of dominant keywords that are typical representatives. Just by looking at the keywords, we can identify what the topic is all about. The following are key factors to obtaining good segregation topics the quality of text processing, the variety of topics the text talks about, the choice of topic modeling algorithm, the number of topics fed to the algorithm, the algorithms tuning parameters.

4.1 Exploratory Data Analysis

We have two datasets, one from Reddit and one from Twitter. The Reddit data is collected using an API called PushShiftAPI and the twitter data is collected using the twitterAPI. To make a comparative analysis of mental health before the lockdown and after the lockdown, we need data before COVID and after the COVID. So, we collected the data from Reddit during 2018 and 2019 before the COVID and named it as the Reddit Pre-COVID dataset. The data that we collected during the pandemic situation, that is during the year 2020 is called the Reddit Post-COVID dataset. Since, we can't able to travel over the past in twitter, we collected the data from twitter during the period from April, 2020 to October, 2020. Also, we have one more dataset, that we collected from the kaggle. The time frame of the dataset that we collected from Kaggle is between mid of June to mid of August 2020.

4.2 Reddit Pre-COVID Data Analysis

The reddit pre COVID data statistics consists of 4 columns. The name of the columns are, subreddit, author, date and post. Here the subreddit is same as the hashtag. Each reddit post is associated with the subreddit like depression, anxiety and the other hashtags that are related to the mental health. This pre covid reddit dataset consists of reddit data collected for 2018 and 2019.

4.3 Reddit Post covid Dataset statistics

The Reddit post covid dataset consists of data that is collected during the pandemic year that is 2020. This dataset consists of same number of columns as like the reddit pre covid dataset. It has subreddit, date, author and post. We used the same hashtags for all the data collection.

Number of metadata attributes	4
Number of Reddit Posts	707,692
Total size in memory	668 MB
Average Reddit size in memory	1KB
Unique users	522136
Time Frame	January, 2018 to December, 2019

Table 1: Reddit Pre-COVID data profile. The Metadata includes subreddit, author, date and post

Number of metadata attributes	4
Number of Reddit Posts	320,364
Total size in memory	294 MB
Average Reddit size in memory	1KB
Unique users	291844
Time Frame	January, 2020 to September, 2020

Table 2: Reddit Post-COVID data profile. The Metadata includes subreddit, author, date and post

4.4 Twitter Dataset statistics

The information about the twitter data statistics is shown in the below table 3. This data has 7 columns namely the ID, Username, location, URL, hashtag, time and month. These columns help us in manipulating the data in a timely basis. We could see what tweet is posted and in which month it was posted and the hashtags that are associated with that tweet.

Number of metadata attributes	7
Number of tweets	165,325
Total size in memory	35.7 MB
Average tweet size in memory	1KB
Unique Users	85574
Time Frame	April,2020 - October, 2020

Table 3: Twitter data profile. Metadata includes ID, Username, location, URL, Hashtag, time and month

4.5 Twitter Kaggle Dataset statistics

This Twitter kaggle dataset is collected from the kaggle website. It has 13 columns namely the username, user location, user description, user created, user followers, user friends, user favorites, user verified, date, text, hashtags and source. Here the user created column tells us about, when the user account was created. The user follower column tells us about the total number of followers for that particular user. The user friends column tells us about the total number of friends that the user have. Lastly, the source column tell us from what device the tweet was created, say "browser for iPhone".

4.6 Creating Bigram and trigram models

Bigrams are pair of consecutive written units such as letters, syllables, or words. Trigrams are a special case of the n-gram, where n is

Number of metadata attributes	13
Number of tweets	179,108
Total size in memory	63.4 MB
Average tweet size in memory	1KB
Unique users	92272
Time Frame	July, 2020 to August, 2020

Table 4: Twitter Kaggle dataset profile. The Metadata includes username, user location, user description, user created, user followers, user friends, user favorites, user verified, date, text, hashtags and source

3. They are often used in natural language processing for performing statistical analysis of texts and in cryptography for control and use of ciphers and codes. Gensim's Phrases model can build and implement the bi-grams, tri-grams, quad-grams and more. The two important arguments to Phrases are min-count and threshold. The higher the values of these parameters, the harder it is for words to be combined to bi-grams.

4.7 Creating dictionary

The two main inputs to the LDA topic model are the dictionary(id2word) and the corpus. The dictionary is created using the gensims predefined method, for which the lemmatized bi-grams are passed as inputs that we got during the pre-processing. Gensim creates a unique id for each word in the document. Here the word could be both bigram or a unigram. The produced corpus is a mapping of word-id and word-frequency. For example, (0, 1) implies, word-id 0 occurs once in the first document. Likewise, word-id 1 occurs twice and so on. This is used as the input by the LDA model. If we want to see what word a given id corresponds to, then, pass the id as a key to the dictionary.

4.8 Building the topic model

Now, We have everything required to train the LDA model. In addition to the corpus and dictionary of bigrams and unigrams, we need to provide the number of topics as well. Apart from that, alpha and eta are hyper-parameters that affect sparsity of the topics. According to the Gensim documentation, both defaults to 1, the number of topics prior. chunk size is the number of documents to be used in each training chunk. update every determines how often the model parameters should be updated and passes is the total number of training passes.

5 EXPERIMENTAL RESULTS

In this section, we will be seeing about the results that we got after applying topic modelling on the pre-processed data. The topics that we got are visualized using the pyLDAvis and the pre-processed words are used to form the wordclouds to make a better comparison how good the topics are related with the wordclouds. Also, we made the t-SNE visualization with the text of two texts that we used for topic modelling.

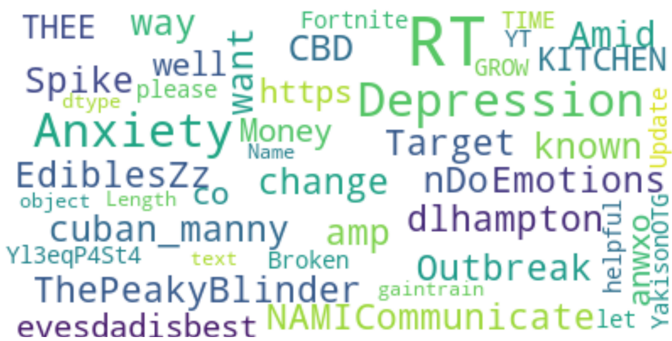

Figure 3: Wordcloud for Twitter dataset. This wordcloud is formed for the whole tweets in the twitter dataset. It is noted that anxiety and depression are the most discussed words.

Figure 4: Wordcloud for Reddit dataset.This wordcloud is formed for the whole Reddit posts in the Reddit dataset. It is noted that addiction and depression are the most discussed words.

5.1 pyLDAvis Visualization

So, using the pyLDAvis library we visualized the topic clusters in twitter and the reddit dataset. The figure 1 shows us the topic clusters formed in the twitter dataset. Eight topic clusters are formed on each dataset and top 30 most relevant tokens for the topic 1 is visualized. The user can able to choose between cluster and view the most frequently used tokens in that particular cluster.

5.2 Wordcloud visualization

We also formed the word cloud for both the datasets to get insights about the dataset before performing the topic modelling. We formed wordclouds for each month in the datasets that we collected to get a knowledge about what topic is being spoken. The figure 3 shows the wordcloud for the Twitter data and the figure 4 shows the wordcloud for the Reddit data.

5.3 t-SNE visualization

We also formed visualization of topic clusters using t-SNE [11] for both the Twiiter and Reddit datasets. Since the number of posts in the Redddit dataset is high than the number of posts in the twitter data, the t-SNE graph for the Reddit dataset as shown in figure 6

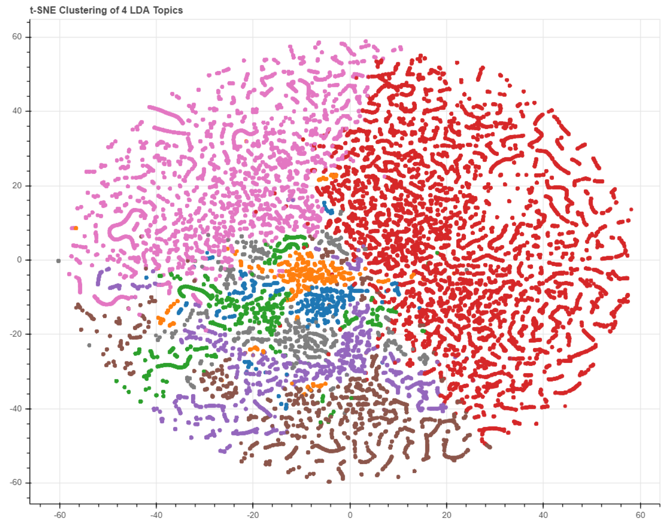

Figure 5: 2D visualization of tweets using t-SNE for Twitter dataset. Each point is a tweet. The tweets are plotted in the 2D plane based upon the similarity between each tweet. The points are colored by topic probability

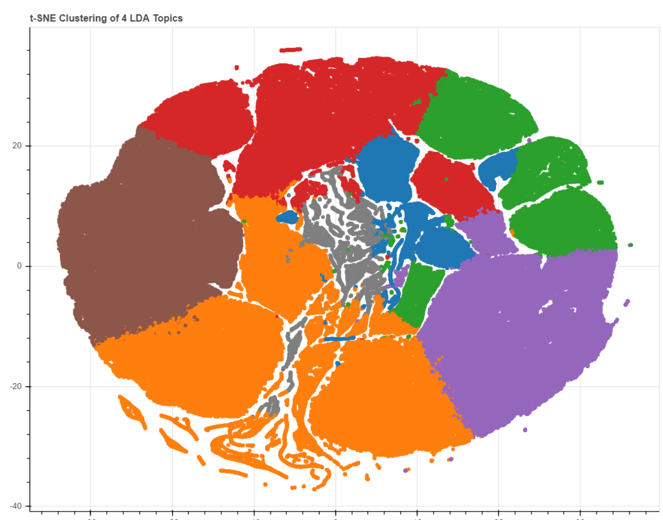

Figure 6: 2D visualization of reddit posts using t-SNE for Reddit dataset. Each point is a reddit post. The number of reddit posts are plotted in the 2D plane based upon the similarity between each posts in the topics group. The points are colored by topic probability. Since, the number of records are high they are closely packed.

more densely packed when compared to the twitter dataset which is shown in 5.

5.4 The LDA topics

We also created a dataframe and it is shown in figure 7 and 8 with the topics formed using the LDA model. We did for both the Twitter and the Reddit datasets. The most widely discussed topic in the Twitter dataset is about the mental health and care. This contributes about 91.36 percent and it is shown in figure 7. Whereas in the Reddit dataset, the most highly contributing topics are loneliness and parental care which is about 95 percent and the second most discussed topics are teaching, online and students which is about 90 percent and it is shown in figure 8

6 DISCUSSIONS

In this paper from the experiments that we did using LDA and the pyLDAvis it is known that there are many topic communities that are distinct from one another and the topics are are shown in table 8 and 7. Also, using the pyLDAvis the user can hover over the topics shown as bubbles and the most frequently used terms are shown once if the user clicks the bubble. User can increase the number of topics, but in this experiment we made the number of topics to be 8. We cannot obtain any other information using the pyLDAvis library. Only the frequently used terms can be shown and the distance between the topics can be shown. This is could be improved by the points that are mentioned in the future implementations section.

7 CONCLUSION

This paper analysed on online communities with mental health disorders using a variety of features from blog post corpus from Twitter and Reddit to discover meta-communities. We used the LDA algorithm to infer latent topics from the corpus which was built using hashtags, tweets, subreddits, Reddit posts and generic words in the posts made by users in online communities. Furthermore, the visualization library named pyLDAvis visualizes the online meta-communities which is discovered using the LDA algorithm. This is the evidence of sentiment-bearing differentiation in online mental health-related communities, suggesting a possible angle for building interventions that can bring help and support in mental healthcare. In addition to the advantages of pyLDAvis, there are several limitations to this tool, such as we could view the topics with a suitable title, we can able to view the relation between the clusters and we could not view what is the actual discussion or the summary that is present in that particular topic group. To overcome these limitations, we are currently building an interactive visualization tool, that overcomes all these limitations of this library and the user can find better insights through this tool.

8 FUTURE IMPLEMENTATIONS

As of now, we have seen the visualization of topics using pyLDAvis, an interactive library for visualizing the topics. We used LDA for generating these topics. There are many other topic modelling algorithms and visualization methods. Using multiple topic modelling algorithms and visualization libraries would help us form a study about which library could provide more information and which topic modelling algorithm could provide us more concrete topics.

Also, in this paper we did not perform any user study. So, the future implementation of this project will be allowing the user to choose between the topic modelling algorithm that they want to use. Once the topics are generated, the user could also choose between multiple visualization libraries and multiple visualizations of these topics could be shown to the user which makes them feel more better and get good insights about the data.

In pyLDAvis, the user can able to see only the frequently used tokens on selecting the particular topic group. This doesn't provide us more useful information and the user have to work further to get the actual discussion that is going in the data. To avoid this, once the user clicks the topic group, if we can able to show the actual message/s or post/s that has been discussed in that group, then, it would be great and the user can find all the information in a single place.

Again, there would be a problem if there are more number of tweets or posts. So, we could display both the posts and the summary of those posts using the available summarization techniques like Bert-Extractive summarizer. This helps us to get the central theme of the collection posts in that particular topic cluster. Thus, the system could be complete and the user can perform Exploratory data analysis, topic modelling and the data summarization in the same place.

Many other interactive visualizations that explains and helps us to choose an optimal model as described in [4] and techniques that helps us to get more accurate topics when working in short texts as described in [9]

REFERENCES

[1] Jonathan M. Bischof and Edoardo M. Airoldi. 2012. Summarizing Topical Content with Word Frequency and Exclusivity. In *Proceedings of the 29th International Coference on International Conference on Machine Learning* (Edinburgh, Scotland) *(ICML'12)*. Omnipress, Madison, WI, USA, 9–16.

[2] David Blei and John Lafferty. 2009. Visualizing Topics with Multi-Word Expressions. (07 2009).

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, null (March 2003), 993–1022.

[4] M. El-Assady, F. Sperrle, O. Deussen, D. Keim, and C. Collins. 2019. Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 374–384. https://doi.org/10.1109/TVCG.2018.2864769

[5] Matthew Gardner, Joshua Lutes, Jeffrey Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2020. The Topic Browser An Interactive Tool for Browsing Topic Models. (11 2020).

[6] David Newman, Jey Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 100–108.

[7] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and Content Analysis of Online Depression Communities. *IEEE Transactions on Affective Computing* 5 (07 2014), 217–226. https://doi.org/10.1109/TAFFC.2014.2315623

[8] Minsu Park, David W. McDonald, and Meeyoung Cha. 2013. Perception differences between the depressed and non-depressed users in Twitter. 476–485. 7th International AAAI Conference on Weblogs and Social Media, ICWSM 2013 ; Conference date: 08-07-2013 Through 11-07-2013.

[9] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy. 2018. Short-Text Topic Modeling via Non-Negative Matrix Factorization Enriched with Local Word-Context Correlations. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1105–1114. https://doi.org/10.1145/3178876.3186009

[10] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. https://doi.org/10.13140/2.1.1394.3043

[11] Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research* 9 (11 2008), 2579–2605.

[12] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009* 382, 139. https://doi.org/10.1145/1553374.1553515

	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.8904	check, follow, open, join, kill, wide, retweet, spread, put, conversation	[hate, publisher, overwhelmingly, nytime, progressiveleft, feminism, marxist]
1	1.0	0.7839	alone, hard, never, live, mast, struggle, ask, stream, man, plea	[vaccine, pass, final, stage, first]
2	2.0	0.8708	fight, awareness, today, follower, mentalhealth, fan, person, sometimes, pushup, great	[count, necessarily, size, dog, fight, size]
3	3.0	0.9128	suicide, https, co, support, listen, prevention, day, amp, sadly, national	[famous, salmon, amp, shrimp, justmunchie, philly, food, kabob, co, sbseyxrpli]
4	4.0	0.8501	friend, help, love, death, true, lose, picture, call, form, begin	[anxious, anxious, anxious, anxious, anxious]
5	5.0	0.8264	life, week, save, available, male, understand, serious, show, community, black	[drive, crazy, behappyquote, positivity, positivevibe, smile]
6	6.0	0.9136	rt, copy, depression, mental, health, feel, much, let, family, talk	[care, mental, health, talk, mental, health, whenever, normalize, talk, mental, health, endthest...
7	7.0	0.8221	always, month, demonstrate, repost, post, late, depth, reach, family_member, give	[crown, tee, online, wor]

Figure 7: Topics in Twitter dataset. Here the Topic Num is the index of a topic. The keywords column has the topic keywords. The topic Perc Contrib explains the total contribution of that particular topic throughout the corpus. The represntative text shows the most relavent tweet for that topic. We found topics that are distinct from on another like friendship, suicide, anxeity and depression

	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.9025	teacher, school, much, class, help, teaching, experience, high, education, problem	[lamictal, lamotrigine, experience, drug, suffer, major, depression, psychotic, simptom, much]
1	1.0	0.9833	feel, people, teach, die, thing, life, kid, always, way, bad	[people, ask, feel, usually, fine, feel, real, anymore, wish, feel, well, feel, well, wish, feel...
2	2.0	0.8250	day, tired, hour, sleep, work, cut, night, feeling, appreciate, big	[next, week, week, week]
3	3.0	0.9374	work, job, online, pay, much, learn, part, move, program, question	[sc, private, employer, legally, request, personal, cell_phone, record, personal, cell_phone, wo...
4	4.0	0.8749	study, fail, watch, video, eat, tip, science, area, check, activity	[low, snack, meal, low, snack, meal]
5	5.0	0.1260	cantact, monitoring, donut, fygz, unschoole, igcse, pedagogy, tenure, classcraft, nyc	[shelly, shelly, shelly, shelly, shelly, shelly, shelly, shelly, shelly, shelly, shelly, shelly,...
6	6.0	0.7133	cantact, monitoring, donut, fygz, unschoole, igcse, pedagogy, tenure, classcraft, nyc	[count, twentythree, twentyfour, twentyfive, twentysix, twentyseven, twentynine, thirtyone, thir...
7	7.0	0.9583	year, time, student, life, live, friend, really, end, give, anymore	[feel, alone, one, really, care, else, parent, find, care, find, look, parent, really, care, hel...

Figure 8: Topics in Reddit dataset. Here the Topic Num is the index of a topic. The keywords column has the topic keywords. The topic Perc Contrib explains the total contribution of that particular topic throughout the corpus. The represntative text shows the most relavent post for that topic. We found topics that are related to online education, job, employment and addiction