UNDERSTANDING THE IMPACT OF COVID-19 ON ONLINE EATING DISORDER COMMUNITIES ON REDDIT

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Md Al Amin

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

December 2023

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

UNDERSTANDING THE IMPACT OF COVID-19 ON ONLINE EATING DISORDER COMMUNITIES ON REDDIT

By

Md Al Amin

The supervisory committee certifies that this thesis complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Prof. Lu Liu
Chair

Prof. Jen Li

Prof. Pan Lu

Approved:

01/29/2024
Date

Simone Ludwig
Department Chair

ABSTRACT

Like precedent disease outbreaks, COVID-19 has immense implications for health challenges. The social restrictions and disruption in daily activities pose a psychological burden for humans worldwide and may be particularly detrimental for individuals with mental disorders. Psychological stressors stemming from the COVID-19 pandemic and resultant stay-at-home orders may exacerbate Eating Disorder(ED)- related triggers and present a challenging environment for individuals with anorexia nervosa, bulimia nervosa, and binge eating disorder. In this research, we aim to comprehend how COVID-19 has affected individuals with eating disorders through a comparative analysis of data obtained from online communities. We developed a tool for extracting information from well-known social media communities such as Reddit. We collected data spanning two years before and after the declaration of the pandemic from the subreddits r/AnorexiaNervosa, r/BingeEatingDisorder, and r/EatingDisorders. The research presents multi-faceted tasks where we analyze the content of each of the subreddits by applying a strategy that combines topic modeling, social network analysis, and time series modeling for a better understanding of these communities on both content and network levels. Through a comparative analysis, we address the discussion topic changes based on users' content and determine how COVID-19 leads to changes in communication patterns within the communities. Finally, we implement time series models like ARIMA, Prophet, LSTM, and Transformer on daily posts and comments count to forecast users' activities within the subreddit and establish a performance comparison of these time series models. The findings indicate that both the content of users' discussions and the level of communication and online support-seeking related to eating disorders on Reddit underwent significant changes during the pandemic.

ACKNOWLEDGEMENTS

I want to extend my heartfelt gratitude to **Dr. Lu Liu** for placing trust in me, for his unwavering guidance and support, and for providing me with the opportunity to work as a graduate research and teaching assistant in the Department of Computer Science at North Dakota State University.

I am also deeply thankful to the Department of Computer Science at North Dakota State University for affording me the chance to embark on the MS program and for granting me a full tuition waiver.

A special note of appreciation goes out to the **National Science Foundation (NSF)** and **North Dakota EPSCoR** for their invaluable financial support throughout this project.

Finally, I'd like to express my special gratitude to **Dr. Jen Li** and **Dr. Pan Lu** for their roles as supervisory committee members in my oral examination.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Eating disorders (EDs) are serious but treatable mental and physical illnesses that can affect people of all genders, ages, races, religions, ethnicities, sexual orientations, body shapes, and weights. According to the National Eating Disorder Association (NEDA)[1] there are 28.8 million people suffer from clinically significant EDs at some time in their life in the United States only.

The COVID-19 pandemic has even worsened ED communities more than their usual activities. The post-pandemic years fundamentally uprooted and transformed the lives of virtually every individual. No longer could people meet or handshake when that meeting or handshake could be a means of death. Uncertainty and fear surrounding the disease, and lack of consistent and reliable information contribute to rising levels of anxiety and stress among people [38], in the United States alone, 37% of individuals exhibit signs of anxiety and depression [18].

Despite the rising mental impact, the increased risks associated with COVID-19 made traditional support avenues, such as group therapies and individual provider visits, difficult or impossible. This has created a complex challenge for individuals with mental health issues. Individuals fighting EDs are among the most impacted by this, as emerging research is beginning to show.

Given the potential for the COVID-19 outbreak to have devastating consequences on human life, it is critical that we work to understand its negative psychological effects. In this work, we use Reddit, a popular social media platform, to study how COVID-19 has impacted the behavior of people posting in ED forums. We focus on three Reddit sub-forums, referred to as subreddits, which are designed to offer peer support for users who are struggling with specific types of ED. We aim to determine whether there are changes related to the pandemic in online ED forums through a comparative study. To measure this, we build topic models to identify the main topics discussed, social network analysis for users' interaction, and time series models to forecast user's activities over posts and comments during the pandemic.

Our findings include substantially increased rates of posting and commenting in ED subreddits along with strong connectivity between users. A transition of discussion content from usual ED issues to COVID-19-related ED issues. These and other findings provide insights into specific

[1] NEDA

ways in which COVID-19 has not only impacted the behavior of users who discuss ED concerns but also the users who do not have ED problems. To the best of our knowledge, existing research has examined these online support venues and found that during the pandemic years, the specific symptoms of EDs change due to the substantial impacts of the virus on day-to-day life [42]. Some studies only examined the impacts in the first few months of the pandemic which are not sufficient to assess long-term effects. The present study aims to expand the existing analysis by examining data from the first 24 months before and after the declaration of the pandemic, March 2018 - February 2020 and March 2020-February 2022. These date ranges are strategically selected to examine comparative differences between the timelines.

1.1. Research Questions

1.1.1. How Extensively is the Topic of COVID-19 Addressed within the Selected Subreddits?

We anticipate that COVID-19-related subjects will become more prominent in eating disorder (EDs) subreddits as the pandemic unfolds, given its far-reaching effects on individuals' lives [37]. EDs subreddits are particularly oriented toward sharing people's real-life experiences, which reinforces our expectation that discussions related to COVID-19 will be more frequently observed within these subreddits.

1.1.2. Do Conversational Shifts in Eating Disorder Subreddits Differ Amidst COVID-19?

We propose that as a consequence of social distancing and lockdown measures, individuals may have encountered food security challenges. Those who are accustomed to socializing and sharing meals at social gatherings may have experienced heightened feelings of isolation during the pandemic. Factors such as financial instability, the fear of losing loved ones, and abrupt changes in rules and regulations have contributed to increased levels of anxiety and mental stress among the general population. We anticipate that these challenging circumstances will lead to a noticeable rise in discussions related to post-COVID experiences, concerns about food security, seeking support, and family matters. Simultaneously, we expect a decline in discussions related to everyday routines and casual activities like schooling, travel, and sports due to the disruptions caused by the pandemic.

1.1.3. Do Alternations in Social Interaction Patterns in Eating Disorder Subreddits Differ During COVID-19?

We expect that the impact of COVID-19 will be most pronounced among individuals dealing with eating disorders, particularly those who actively participate in eating disorder subreddits more frequently than the general population. Consequently, we expect a substantial surge in posting activity within eating disorder subreddits during the pandemic, as the lockdown measures have prompted people to spend more time online[2]. Additionally, we predict an increase in the level of interaction among users within these eating disorder subreddits, indicating that users will engage with a larger number of fellow users. This expectation is rooted in previous research, which has highlighted seeking social support as a common coping strategy observed during earlier disease outbreaks [9].

1.1.4. Analyze Performance of Time Series Forecasting Models on Post and Comment Data

As mentioned earlier, we anticipate an upsurge in user connectivity through posting and commenting. Consequently, we have employed various time series forecasting models, such as state space models like ARIMA and Prophet, as well as deep learning models like LSTM and Transformer, to predict future activity levels. We have a strong expectation that a self-attention-based Transformer model will outperform other models due to its capability to effectively capture intricate patterns and dynamics from time series data [47].

1.2. Contribution and Outline

Chapter 2 will provide an overview of related works and the theoretical foundations underpinning our research. It will explore the operational principles of the LDA Topic model and various time series models, including ARIMA, Prophet, LSTM, and Transformer.

Chapter 3 will delve into the specifics of our experimental setup and the methodologies employed in our research.

In Chapter 4, we will present and discuss the results obtained from our research.

Chapter 5 will address the limitations of our proposed method and outline potential avenues for further research.

[2]The NY Times

2. RELATED WORK AND THEORETICAL BACKGROUND

Eating disorders have become a prevalent and deeply concerning issue in today's society. These are complex mental health conditions that affect millions of people globally, regardless of age, gender, or cultural background. Understanding the causes, various expressions, and far-reaching consequences of these disorders isn't just an academic pursuit; it's a vital step in developing effective prevention and intervention strategies. In this chapter, we first explore existing research on eating disorders, investigations related to eating disorders on social media, and the impact of COVID-19 on the eating disorders community. The second part of this chapter provides the theoretical framework that informs our research.

2.1. Related Works

2.1.1. Eating Disorders

Eating disorders are severe mental health conditions characterized by unusual eating habits or behaviors related to weight control. Distorted beliefs and attitudes regarding weight, body shape, and eating patterns are central to the development and persistence of these disorders. The nature of these concerns can differ based on gender; in men, for instance, body image issues may revolve around the desire for increased muscularity, while in women, the focus may be more on achieving weight loss [40]. Research has revealed that in 2017, an estimated 16 million individuals globally were affected by anorexia nervosa and bulimia nervosa.

There are six main feeding and eating disorders now recognized in diagnostic systems: anorexia nervosa, bulimia nervosa, binge eating disorder, avoidant-restrictive food intake disorder, pica, and rumination disorder. Thousands of studies have been done on each type of eating disorder with a variety of ages and sex. Jeffrey G *et. al.* [21] address eating disorder among adolescents the Risk for Physical and Mental Disorders During Early Adulthood. Golden *et al* [16] suggested how obesity prevention efforts may lead to the development of an ED in adolescents. Eating disorders and disturbed eating patterns are always associated with mental disorders. Natalie C. *et al* [28] conducted a comprehensive study exploring bidirectional associations between eating disorders and psychiatric disorders.

2.1.2. Studying Eating Disorders via Social Media

Social media platforms are rapidly gaining significance as valuable sources of information pertaining to mental health disorders. Among these disorders, eating disorders represent intricate psychological challenges characterized by unhealthy eating patterns. A growing number of studies have applied computational methods to data collected from social media platforms to characterize behavior associated with mental illnesses and to detect and forecast mental health outcomes (see Chancellor and De Choudhury [8] for a comprehensive review).

Reddit serves as an exceptionally well-suited platform for the examination of eating disorders, primarily due to its semi-anonymous nature, which fosters candid user interactions and diminishes inhibitions related to sharing personal experiences [10]. Furthermore, Reddit features subreddits dedicated to serving as forums for eating disorders discussions, such as r/EatingDisorders, r/AnorexiaNervosa, and r/BingeEatingDisorder. These specialized subreddits enable a more precise analysis of users contending with specific eating disorder conditions.

Several prior studies have concentrated on the characterization of discourse patterns within these Reddit communities. Research endeavors have included the analysis of long-term trends in the usage of topics and word choices [7], the exploration of the relationship between user participation styles and their preferences in topics [13], and the examination of discourse patterns specifically related to self-disclosure, the provision of social support, and anonymous posting [10, 30].

Other studies of Reddit eating disorders communities have aimed to quantify and forecast changes in user behavior. Yousrha *et. al.* [14] analyzed the content and the network of the pro-eating disorders (pro-ED) community and the pro-recovery community on Reddit by applying an approach that combines topic modeling, social network analysis, and sentiment analysis. Ashleigh N *et. al.* [34] presented linguistic measures to find dramatic changes in the lives of those with eating disorders (EDs) during the COVID-19 pandemic.

In addition to analysis focused on the relationship between linguistic measures and EDs, there are a number of studies that have examined social interaction patterns and ED on social media. For example, Yousrha *et. al.* [14] analyzed how different user interaction measures, such as the number of posts and comments authored and received, vary between users who do and do not post on r/Pro-Eating and r/Pro-Recovery. Other existing studies have used similar measures to study the

relationship between user interaction patterns and mental health disorders on Twitter [11,12] and the relationship between user interaction patterns and topic usage on r/depression [13]. Many studies use network analysis methods to measure additional properties of user interactions on social media. Such studies typically create graphs where users are nodes and edges represent interactions between users (e.g., commenting on a Reddit post or re-tweeting a Twitter post). They then characterize social activity by computing metrics such as graph size, density, and clustering coefficient. While some studies analyze per-user interaction patterns by forming a separate graph for each user [11], others use a single social graph to study interactions at the community level [43].

2.1.3. Eating Disorders and COVID-19

Current research on the confluence of EDs and COVID-19 indicates that the impacts are wide-reaching and felt throughout the world, as one might expect [19, 25, 37, 44]. Two main themes have emerged from the widespread impact: First, people are experiencing various barriers (including social, role, and support), and second, there are unexpected benefits for those battling EDs during the pandemic.

The study has indicated that the pandemic has caused wide-ranging social barriers amongst those fighting EDs, including isolation and changes in accountability/responsibility [22], and these social barriers often exacerbate existing negative behaviors. People experiencing stressful life events such as a death in the family or a pandemic—experience a dramatic increase in harmful coping mechanisms and weight/food control behaviors [24]. As a result, people experiencing EDs have reported an increased usage of harmful coping mechanisms.

2.1.4. Social Media and COVID-19

Social media analysis has been used to study the impacts of the COVID-19 pandemic and the spread of the virus. Shen *et al.* [33] used Granger causality tests to show that Weibo posts related to COVID-19 symptoms or a diagnosis could be used to predict case counts up to two weeks ahead of time in China. Ordun *et al.* [29] explored topics and network features in COVID-19 tweets. They studied the propagation of information related to the pandemic and showed a relationship between topics and government press briefings. Gencoglu and Gruber [15] created a causal model involving Twitter activity and sentiment, COVID-19 statistics, country demographic statistics, and government interventions. They found that country Twitter usage, new deaths, new infections, and lockdown announcements all impact COVID-19-related Twitter activity.

2.2. Theoretical Background

To acquire a thorough understanding of the experimental design and configuration underpinning our research project, it is highly beneficial to delve into the foundational technical concepts related to both ***topic modeling*** and ***time series*** analysis. We will commence this exploration by delving into the intricacies of ***Latent Dirichlet Allocation (LDA)*** topic model, followed by a comprehensive examination of the diverse time series models that were thoughtfully incorporated into our research, including the application of ***ARIMA***, ***Prophet***, ***LSTM***, and ***Transformer*** models.

2.2.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA), a technique in the realm of topic modeling, operates as an unsupervised and probabilistic method geared towards the extraction of underlying themes within a given corpus of documents [2]. Within this framework, a **topic** is conceptualized as a probability distribution across a predefined vocabulary. LDA dissects the verbiage present in each document, endeavoring to unveil the joint probability distribution that connects the observable components (i.e., the words within the documents) with the concealed elements (namely, the latent topic structure). Employing a **Bag of Words** approach, LDA abstains from assessing the semantic nuances or contextual meanings of sentences; instead, it focuses its analysis on word frequencies. Consequently, it postulates that the most frequently occurring words within a particular topic essentially encapsulate the essence of that topic. For instance, if one of the topics within a document pertains to **anorexia**, it is reasonable to anticipate that terms like **anorexia**, **bulimia**, and **eating disorder** will manifest higher frequencies compared to non-anorexia-associated documents. The outcome of this process yields a set of distinct topics, with those having close thematic ties grouped together. Subsequently, a probability score is computed for each document each topic, yielding a matrix whose dimensions correspond to the number of topics multiplied by the number of documents.

Text summarization, sentiment analysis, information retrieval, and content recommendation are just a few of the applications that LDA and its derivatives have become increasingly important in recent years. The potential of LDA in interdisciplinary studies has also been investigated. Examples include studying social media data for sociological insights or biological literature for information extraction in healthcare research.

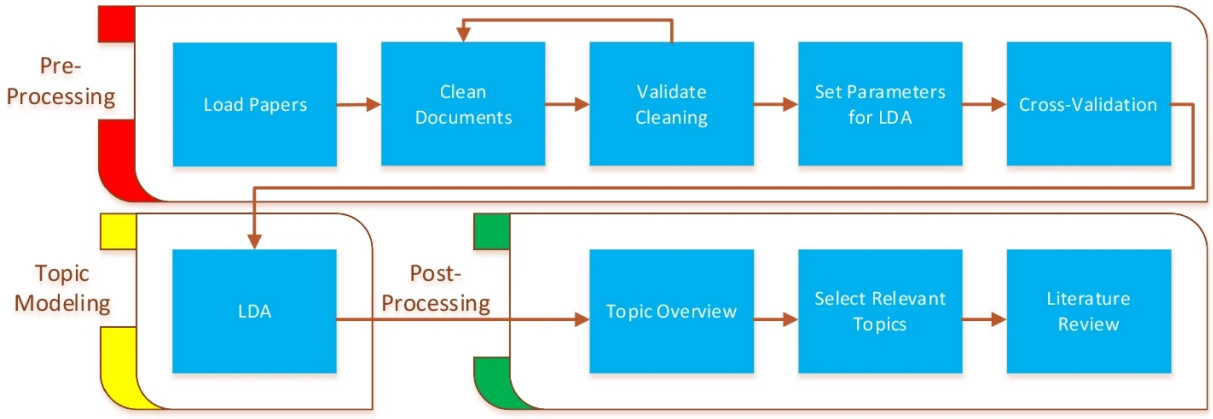

Figure 2.1. The step-by-step LDA topic modeling implementation

The detailed step-by-step implementation of LDA topic modeling that we discuss in the technique chapter is shown in Figure 2.1.

2.2.2. Time Series

Time series analysis is a ***statistical approach*** that focuses on looking at data points that have been collected, recorded, or evaluated throughout a series of time periods. Each individual datum is associated with a unique timestamp or time period inside a time series dataset, and the dataset is systematically arranged in chronological order. Time series data is frequently seen in a variety of industries, including meteorology, finance, economics, and many others, where it is crucial to understand and foresee temporal trends, patterns, and linkages.

In our time series implementation, we have implemented **ARIMA**, **Prophet**, **LSTM**, and **Transformer**.

2.2.3. ARIMA

The Auto-Regressive Integrated Moving Average (ARIMA) technique stands out as a popular approach for simulating dynamic systems. The observed variable $\boldsymbol{x_t}$ is the main focus of the ARIMA framework, and it is expected that $\boldsymbol{x_t}$ may be broken down into three separate components: trend, seasonality, and irregularity. Instead of focusing on each of these elements separately, Box and Jenkins developed the idea of differencing the time series $\boldsymbol{x_t}$, a technique intended to get rid of the trend and seasonal features. Following transformation, this series is handled as stationary time series data and modeled using a combination of its historical values ("**AR**") and the moving average of prior prediction mistakes ("**MA**"). A tuple *(p, d, q)* is commonly used to define an

ARIMA model, where p and q stand for the auto-regressive and moving average component orders, respectively, and d stands for the order of differencing that was used.

$$(1 - \sum_{i=1}^{p} \phi_i L^I)(1 - L)^d = \delta(1 + \sum_{i=1}^{q} \theta_i L^I)\varepsilon_t \tag{2.1}$$

Where L is lag operator, ϕ_i is the moving average part parameter, ε_t is the error term

It's important to remember that ARIMA may be described as a **State Space Model (SSM)** [47], and that standard SSM methods like filtering and smoothing can also be used with ARIMA. It's crucial to recognize that ARIMA adopts a somewhat opaque methodology in that it just considers observed data and avoids analyzing the states of the underlying system. The generalized form of ARIMA model is given in Equation 2.1.

2.2.4. Prophet Model

The Facebook Core Data Science team created **Prophet**, an open-source forecasting tool. It has been specially designed to predict time series data with daily observations exhibiting patterns over a range of temporal dimensions. Prophet has won praise for its user-friendly interface and ability to manage missing data, outliers, and the impact of vacations with ease. Due to this tool's capacity for time series forecasting, research literature has widely explored and used it.

The FB Prophet uses a **Bayesian model** that makes use of curve-fitting methods [20]. It provides parameters that are simple to understand and don't require a lot of time-series data to make reliable forecasts. This method works especially effectively with time series data that clearly show seasonal cycles as important causes. Additionally, it manages planned pauses or holidays in continuous data streams effectively. When it comes to handling missing data, trend deviations, and outlier detection, FB Prophet performs better than other tools. These variances must be addressed in real applications like sales forecasting. It also offers libraries that are simple to use and understand, which improves its usability.

The FB Prophet forecasting is based on an additive regressive model can be formulated as:

$$y(t) = g(t) + h(t) + s(t) + \varepsilon_t \tag{2.2}$$

Where $y(t)$ is the additive regressive model, $g(t)$ is the trend factor, $s(t)$ is the seasonality component, and ε_t is the error term.

The FB Prophet model's implementation is a simple procedure. The first step for analysts is to organize the dataset into a data frame with two distinct columns: *'ds'* for the date stamp in DateTime format and *'y'* for the numerical forecasting measurement. Analysts then build an instance of the ***Prophet()*** class and insert the data frame into it. Analysts can define the preferred forecasting period and continue with the forecasting process after this phase is finished. Multiple columns make up the resulting forecast, with the words *'ds'* and *'yhat'* receiving particular attention. The anticipated values of *'y'* are found in the *'yhat'* column based on the historical record and insights into future trends and seasonality patterns can be gained by graphing the values of *'ds'* and *'yhat'* [32].

2.2.5. LSTM

Long Short-Term Memory, sometimes known as LSTM, is a type of recurrent neural network (RNN). RNNs are a strong subclass of artificial neural networks that have the singular capacity to keep internal memories of input sequences. This makes them particularly well-suited for dealing with sequential data problems, such as time series analysis. Conventional RNNs, on the other hand, frequently struggle with a serious difficulty known as the vanishing gradient problem, which causes sluggish learning or even a total stop in model training [48]. The 1990s saw the introduction of LSTMs, which were specifically created to address this issue. Long-term memory (LSTM) skills allow for successful learning from inputs that are temporally separated from one another, which increases their applicability for modeling long-term data.

An LSTM model is composed of three gates: forget, input, and output. The forget gate decides whether to keep or remove existing information, the input gate determines how much new information will be added to the memory, and the output gate controls whether the existing value in the cell contributes to the output. The architecture of an LSTM model has been shown in Fig 2.2

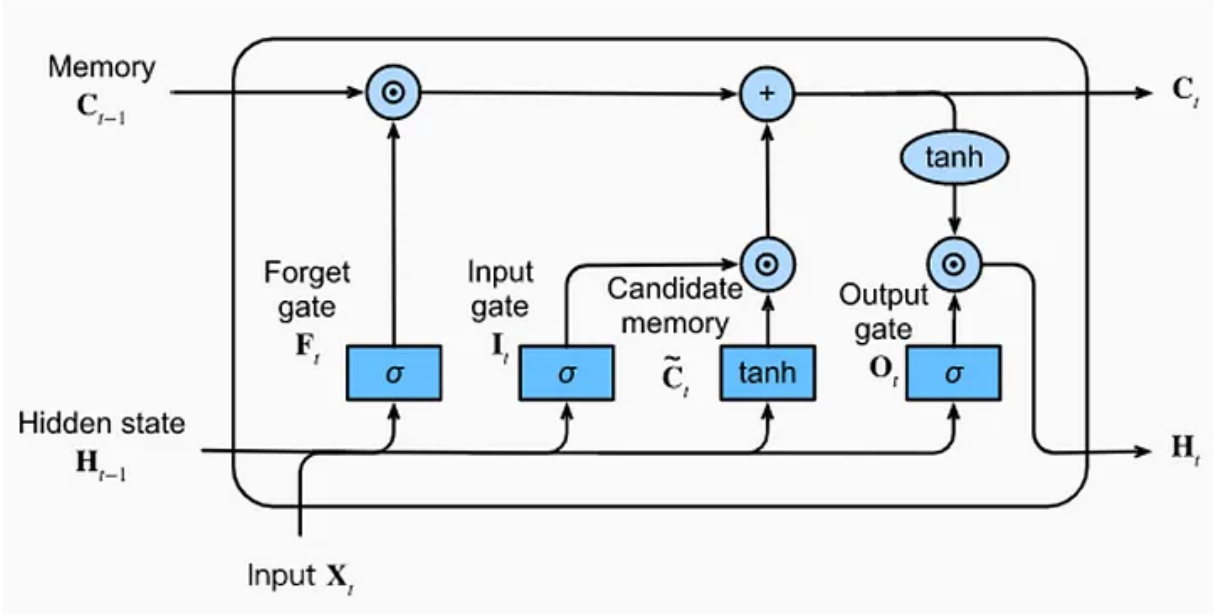

Figure 2.2. Architecture of LSTM unit

- **Forget Gate:** A *sigmoid* function is usually used for this gate to make the decision of what information needs to be removed from the LSTM memory. This decision is essentially made based on the value of H_{t-1} and X_t. The output of this gate is F_t, a value between 0 and 1, where 0 indicates completely getting rid of the learned value, and 1 implies preserving the whole value. This output is computed as:

$$F_t = \sigma(W_{f_h}[H_{t-1}], W_{f_x}[X_t], b_f) \tag{2.3}$$

where b_f is a constant and called the bias value.

- **Input Gate:** This gate makes the decision of whether or not the new information will be added to the LSTM memory. This gate consists of two layers: 1) a *sigmoid* layer, and 2) a "tanh" layer. The *sigmoid* layer decides which values need to be updated, and the tanh layer creates a vector of new candidate values that will be added to the LSTM memory. The outputs of these two layers are computed through:

$$I_t = \sigma(W_{i_h}[H_{t-1}], W_{i_x}[X_t], b_i) \tag{2.4}$$

$$C_t = \tanh(W_{c_h}[H_{t-1}], W_{c_x}[X_t], b_c) \tag{2.5}$$

in which I_t represents whether the value needs to be updated or not, and C_t indicates a vector of new candidate values that will be added to the LSTM memory. The combination of these two layers provides an update for the LSTM memory in which the current value is forgotten using the forget gate layer through manipulation of the old value (e.g. C_{t-1}) followed by adding new candidate value $I_t * C_t$. The following equation represents its mathematical presentation:

$$C_t = F_t * C_{t-1} + I_t * C_t \tag{2.6}$$

where F_t is the result of the forget gate, which is a value between 0 and 1 where 0 indicates completely getting rid of the value; whereas, 1 implies completely preserving the value.

- **Output Gate:** This gate first uses a *sigmoid* layer to make the decision of what part of the LSTM memory contributes to the output [35]. Then, it performs a non-linear *tanh* function to map the values between -1 and 1. Finally, the result is multiplied by the output of a *sigmoid* layer. The following equation represents the formulas to compute the output:

$$O_t = \sigma(W_{o_h}[H_{t-1}], W_{o_x}[X_t], b_o) \tag{2.7}$$

$$H_t = O_t * \tanh(C_t) \tag{2.8}$$

where O_t is the output value, and H_t is its representation as a between -1 and 1.

2.2.6. Transformer

The Transformer design was first presented by Vaswani et al. (2017) [41] in Google's publication, **"Attention Is All You Need,"** in which they proposed a method called **"Self-attention."** The transformer paradigm was specifically suggested for machine translation or sequence-to-sequence translation. However, it has proven to be incredibly effective in other fields including speech recognition, image generation, and time series analysis. Numerous modifications have been suggested over the years to enhance the efficiency of Transformers. However, we concentrated on the fundamental design suggested by Vaswani et al [41]. in our research.

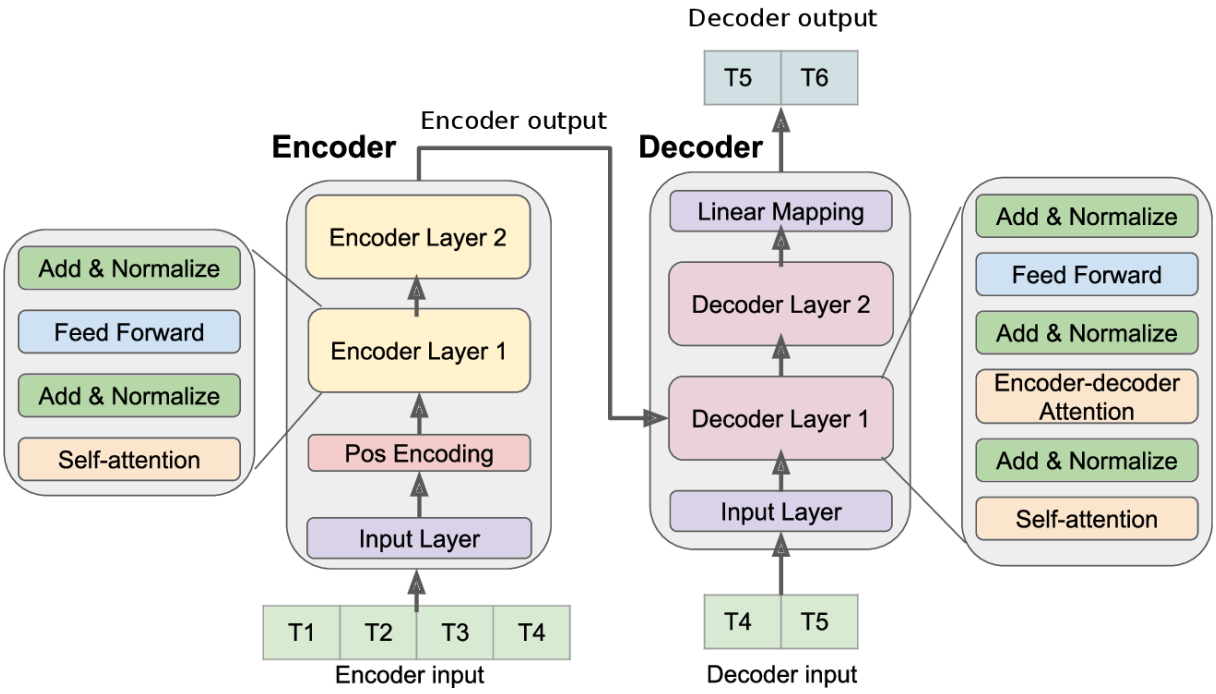

Figure 2.3. Transformer model architecture [47]

Key Components: At its core, the Transformer architecture comprises several key components that enable its effectiveness in capturing dependencies within the sequence of data. Fig 2.3 illustrates the components of typical transformer architecture. These components include:

- **Self-Attention Mechanism:** One of the hallmark features of the Transformer architecture is the self-attention mechanism. This mechanism allows the model to weigh the importance of different positions in the input sequence, facilitating the capture of long-range dependencies. For example, *"The animal didn't cross the street because it was too tired"*. In this sentence, what does *"it"* in the sentence refer to? It refers to *"animal"*. When the model is processing this word, it should be able to associate *"it"* with *"animal"* and NOT *"street"*. The self-attention layer takes care of this and it is very core to why it has been used remarkably in machine translation. Mathematically self-attention can be defined as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.9}$$

Where Q, K, and V are matrices representing queries, keys, and values respectively, and d_k is the dimension of the keys. The softmax operation ensures that the attention score sums to 1, making it a weighted average of the value.

- **Positional Encoding:** Positional encoding describes the order or position of each entity in a sequence so that each position is assigned a unique representation. The transformer model does not contain any recurrence or convolution. In order to make the model able to use the sequence, Vaswani et al [41] injected the relative or absolute position of the tokens in the sequence. For this, the "positional encodings" are added to the input embeddings to process modified input in parallel. The position encoding uses the same dimension as the input embedding d_k. The basic transformer architecture uses sine and cosine functions of different frequencies.

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{\frac{2i}{d_k}}}) \tag{2.10}$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{\frac{2i}{d_k}}}) \tag{2.11}$$

where pos is the position of a entity in the sequence, d_k is the dimension of output embedding space, PE position function mapping a position pos in the input sequence to the positional matrix, and i is the index of output dimension.

- **Encoder:** As shown in Fig 2.3, the encoder block consists of a self-attention layer and a feed-forward layer connected back-to-back with a normalization layer. There are residual connections between each encoder which are commonly used techniques for training deep neural networks and help in training by preserving input representation. The layer normalization operation is also commonly used in neural networks for processing sequential data. It helps with the faster convergence of the model training. The feed-forward layer comprises two linear layers with a ReLU activation function. The output of an encoder block is used as an input to the next encoder block. The input to the first encoder block consists of the sum of word embeddings and positional encoding (PE) vectors.

- **Decoder:** Each decoder block consists of similar layers and operations as the encoder block. The decoder takes as input the encoded representations of the source sequence generated

by the encoder [1]. Inside a decoder, three layers included (1) self-attention layer, (2) feed-forward layer, and (3) encoder-decoder attention layer. There are also residual connections and normalization operations with the intention of encoder. The purpose of having an encoder-decoder attention layer is to allow the decoder to focus on different parts of input tokens while generating the output sequence. This helps the decoder to keep aligned with the source and target sequence.

3. METHODOLOGY

The methodology chapter of this research thesis acts as an essential framework for explaining the methodical technique utilized to look into and address the research issues or hypotheses. In this chapter, we outline the specific methods, techniques, and procedures utilized to collect and analyze data, ensuring the rigor and reliability of our study. This comprehensive study not only provides transparency into our research procedures but also makes it easy for other researchers to evaluate and replicate our study. We begin the chapter by providing an overview research architecture of our study, followed by a description of data collection methods, data analysis techniques, and ethical considerations that guided our investigative journey. We intended to lay a strong foundation for the results and conclusion of this thesis by carefully outlining our methodological decisions, thereby advancing knowledge in the area of study we have selected.

3.1. Research Architecture

In Fig: 3.1, we have shown an overview of the research architecture of our study. We start by extracting data from our target (eating disorders) population. Then we perform data analysis and extract linguistic and network features. With our analyzed data, we implement LDA topic modeling to categorize the best topics discussed in the data, build a graph network, and investigate its properties. We also introduce time series intervention analysis. Finally, we interpret the feature by visualizing the comparison between topics, graph networks, and time series with their pre-COVID and mid-COVID pandemic.

3.2. Data Collection

In the course of this study, we collected data from forums (subreddits) on the website of Reddit.com which had been ranked as the sixth most visited website in the US and ninth worldwide in 2023[1]. We scrapped Reddit posts and comments from March 2018 to February 2022 using the Pushshit API [3]. The World Health Organization as well as the United States declared COVID-19 as a global health emergency on March, 2020[2]. We divided our data into two groups. Two years prior to the declaration of the pandemic (March 2018 to February 2020) as Pre-Pandemic data and two

[1]Reddit.com website ranking
[2]COVID-19 emergency declaration

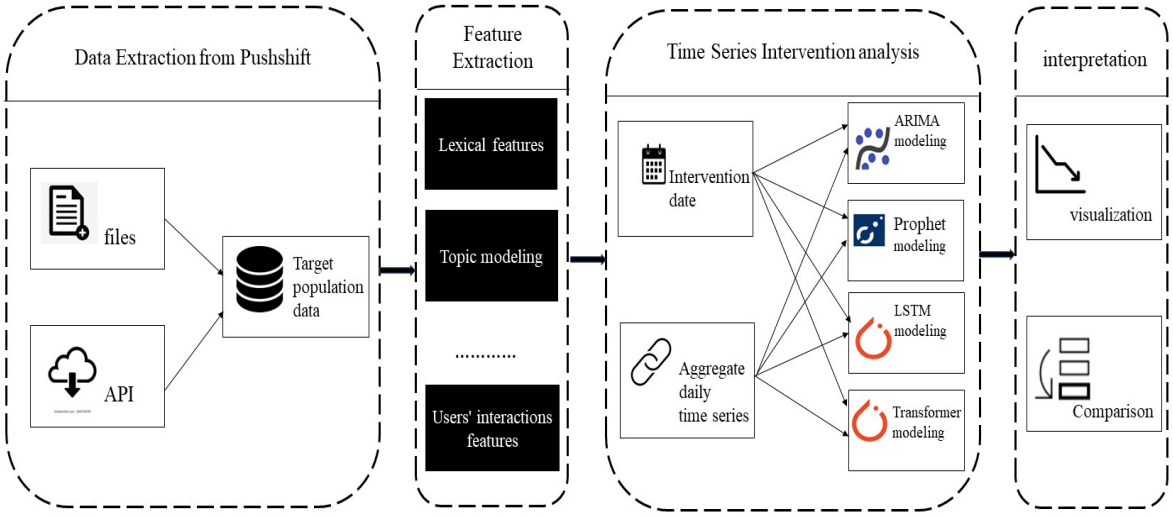

Figure 3.1. A general overview of the research architecture of our system

years after the declaration of the pandemic (March 2020 to February 2022) as Mid-Pandemic data. We collected data from three eating disorders subreddits r/EatingDisorders, r/AnorexiaNervosa, and r/BingeEatingDisorder. There are a few reasons why we have selected these subreddits for our research study. First of all, there was a significant amount of increase number of posts and comments from pre-pandemic to mid-pandemic time. Secondly, overall subreddit group size and active members. Finally, because the subreddits provide support for different eating disorders, their users may have been affected differently by COVID-19.

	r/EatingDisorders	**r/AnorexiaNervosa**	**r/BingeEatingDisorder**	**Total**
Post	2,345	5,125	11,943	19,413
Comment	11,552	19,620	57,864	89,036

Table 3.1. Pre-pandemic subreddit data

Table 3.1 shows statistics on the number of posts and comments made in the three subreddits we chose prior to the epidemic, whereas Table 3.2 shows data for the middle of the outbreak. It is obvious that from the pre-pandemic to the mid-pandemic era, the number of posts and comments on r/EatingDisorders nearly doubled. The number of posts and comments on the subreddits

	r/EatingDisorders	r/AnorexiaNervosa	r/BingeEatingDisorder	Total
Post	3,569	16,283	20,093	39,945
Comment	18,804	73,986	91,374	184,164

Table 3.2. Mid-pandemic subreddit data

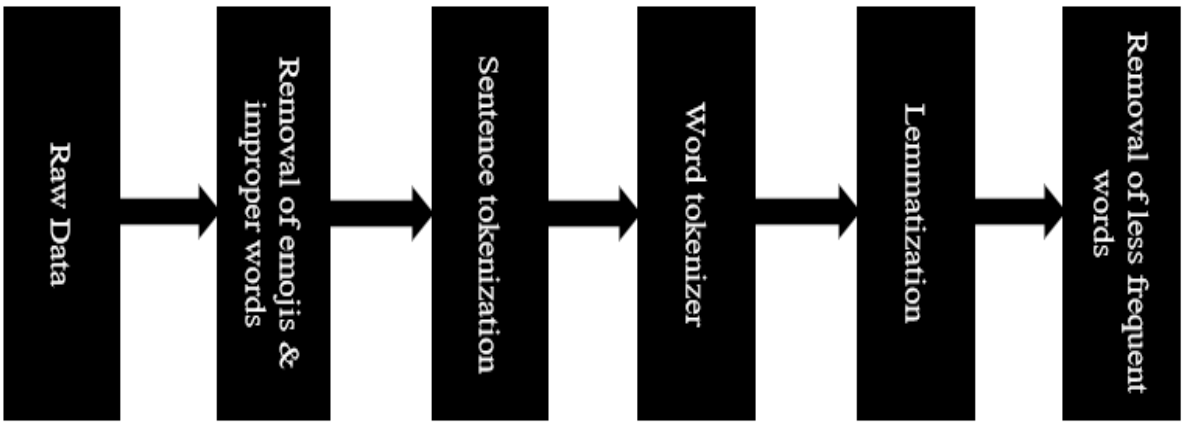

Figure 3.2. Data clearing and preparation steps

r/AnorexiaNervosa and r/BingeEatingDisorder increased by more than a factor of two during the middle of the epidemic.

3.3. Data Cleaning and Preparation

In the primary component of our architecture, as illustrated in Figure 3.1, we utilized the API to collect the data of interest, which was then stored in CSV file format. The subsequent phase of our research entails the essential steps of data cleansing and preprocessing, preparing it for further analysis. For a detailed account of how we managed data marked as "deleted" or "removed" within posts and comments, please consult Appendix 5.2.

Figure 3.2 illustrates the data cleaning and preparation procedures employed in our research project. The sequence of steps we executed encompassed the following:

- **Raw Data:** Pushshift represents a RESTful API [3] that offers comprehensive functionality for Reddit data retrieval, along with the capability to perform robust data aggregations. This API allows us to swiftly access the specific data of interest and uncover intriguing correlations within Reddit content. Our data collection process involves gathering post-related information

through this API, encompassing post title, post body, post ID, user ID, date and time of posting, and the number of comments.

Subsequently, we utilize the official Reddit API, **Python Reddit API Wrapper (PRAW)**, to download all comments associated with a particular post using its unique post ID. This enables us to merge the post and its corresponding comments, creating a comprehensive document for each post, which we commonly refer to as a "document."

- **Removal of Emojis and Improper Words:** As we are aware, Reddit text data is inherently unstructured, containing a variety of elements such as images, videos, emojis, URLs, and inappropriate language. This particular stage assumes a critical role within our data cleaning process, as its primary objective is the elimination of these elements from the dataset. In this phase, we employ a comprehensive set of regular expressions [4] to meticulously detect and subsequently remove irrelevant elements from the textual data.

- **Sentence Tokenization:** Sentence tokenization, a fundamental NLP technique, involves dividing a text or longer piece of content into discrete sentences. This process is commonly employed in natural language processing (NLP) and text analysis to break down text into meaningful units, which are typically sentences. In our data processing, we utilize sentence tokenization to extract significant sentences from our dataset. To accomplish this, we employ the Natural Language Toolkit (NLTK) Python library, which aids in generating these sentences efficiently.

- **Word Tokenizer:** The granularity of a word-level tokenizer is the surface forms of words, i.e., it splits text according to the spaces between words [39]. Word-level tokenization does not require any vocabulary training since one can apply it just by splitting the text with white space characters. In our data processing steps, after sentence tokenization, we apply word-level tokenization and remove the stop words in the English language. To get the significant tokens from our dataset, we use the SpaCy Python library.

- **Lemmatization:** Lemmatization typically involves a systematic approach that relies on vocabulary and morphological analysis of words, primarily aiming to eliminate inflectional

endings while returning the base or dictionary form of a word, referred to as the lemma[3]. This process holds a pivotal role in our data processing pipeline as it standardizes the various forms of a word into a common form. By carefully considering the context, meaning, and part of speech within a sentence, as well as the context of neighboring sentences, lemmatization accurately identifies the lemma of a word. In our study, lemmatization plays a significant role in structuring the dataset by grouping words that share a common root. To achieve this, we employ the SpaCy Python library to generate lemmas from tokens, retaining only those lemmas classified as NOUN, ADJECTIVE, VERB, and ADVERB.

- **Removal of Less Frequent Words:** Our data cleaning pipeline's final step is to get the lemmatized representation of each token located in the related texts. At this point, we build a **term-frequency-inverse-document-frequency (tf-idf)** metric to assess each lemma's significance over the full corpus or dataset in a thorough manner. By emphasizing the inclusion of the most pertinent terms to improve our analytical outcomes, this methodical exclusion of less significant lemmas is a crucial technique targeted at optimizing the performance of our data analysis.

3.4. Data Analysis

The central aim of our research is to investigate the transformation of activity within eating disorder subreddits as it shifted from the pre-pandemic to mid-pandemic periods, employing a comparative study approach. To achieve this objective, we have structured our data analysis into three distinct phases:

(i) Our initial phase involves **topic modeling** applied to both pre-pandemic and mid-pandemic datasets, enabling us to extract valuable insights.

(ii) In the subsequent phase, we perform a comprehensive **social network analysis**, quantifying user activity and interactions within these subreddits.

(iii) Finally, our third phase employs **time series analysis** to discern whether significant changes in activity patterns occurred during the pandemic period.

[3]Lemmatization definition

3.4.1. Topic Modeling

To study the changes in discussion content that occurred during the pandemic, we use Latent Dirichlet Allocation (LDA) topic modeling [5]. The reason behind choosing LDA topic modeling is that it is an unsupervised machine learning method and does not require preliminary classification of the documents. LDA has shown great results in online communities relating to mental disorders [4, 6, 27]. To understand the discussion topics that are common within r/EatingDisorders, r/AnorexiaNervosa, and r/BingeEatingDisorder, we train a single topic model on combined posts and comments from these three subreddits. This provides us with a set of topics, where each topic is defined as a distribution over words. We use this trained model to infer topic distributions for each of the subreddits. This helps us analyze changes in the respective subreddits from pre-pandemic to mid-pandemic times. We ensure that discussions from each of the subreddits are equally represented in our dataset over the timeline. We use the implementation of LDA topic modeling provided in the GENSIM Python Library [31] and train models with $k = 5, 10, ..., 50$ topics. We select a single model to use in our analysis by examining their coherence scores, a measure of the semantic similarity of high-probability words within each topic [26]. As coherence tends to increase with increasing k, we select k as the first local maxima of coherence scores for each of the subreddit data pre-pandemic and mid-pandemic.

In the result chapter, we list the k topics obtained from our topic model along with the highest probability words associated with each topic. We also provide labels that summarize the essence of each topic, which we create by examining their representative words.

3.4.2. Social Network Analysis (SNA)

Social Network Analysis (SNA) is the study of relations between individuals [36]. It is an interdisciplinary descriptive, conceptual, and empirical framework that represents complex systems as networks. The individual who contributes to the group is presented as a node, and the relationship is presented as an edge of the network.

Our research focuses on investigating alterations in social interaction patterns within each subreddit. To accomplish this, we establish a set of user interaction metrics, drawing inspiration from previous research in network analysis as applied to social media platforms. [36, 45, 46].

In order to carry out our Social Network Analysis (SNA), our initial step involves modeling our community as a social interaction graph denoted as $G = (N, E)$, where N represents the individuals who post and comment within the subreddit, and E constitutes the collection of interactions among them. Here, an interaction is defined as a reply or a comment on a comment within the community. In our study, we use an undirected link and remove self-loops in order to maintain the integrity of our metrics calculations. Social network analysis involves a variety of tasks [36]. We then compute twelve metrics commonly used to characterize graphs as our user interaction measures [4], which are described in detail in the Table **??**. We use the NetworkX library [17] to assist with graph creation and metric extraction.

3.4.3. Time Series Analysis

In our research investigation, we apply time series models to predict user activities, specifically focusing on the count of posts and comments. We conduct a performance evaluation of these time series models, drawing inspiration from two distinct prior research approaches. Firstly, in the study by Mrinal Kuram and colleagues [23], they assess the impact of an event on user activity in mental health subreddits. Their method involves using a t-test to compare observations made "before" and "after" a significant event. Secondly, in the work of Neo Wu and the team [47], they employ a Transformer-based time series model and conduct a comparative analysis of its performance in relation to both state space models and deep learning models. We use **ARIMA**, **Prophet**, **LSTM**, and **Transformer** models to forecast user activities through posts and comments count and compare model performance with respect to **ARIMA** model.

Time Series Data: In our analysis, we examine the quantification of daily posts and comments within each subreddit over a specific time frame. This period encompasses two years (731 days) leading up to the COVID-19 pandemic declaration in the United States and two years (730 days) following this declaration. Our approach involves utilizing three distinct levels of data to evaluate user engagement within a given subreddit. The first level is post count, which serves as an indicator of users' tendencies to seek assistance or share information. The second level is comment count, reflecting users' interest in providing support through advice and personal experiences. Lastly, we consider a combined count of both posts and comments, offering a comprehensive overview of the subreddit's overall activity.

Metric Name	Description
Node Count $\lvert N \rvert$	Number of unique users who posted or commented
Edge Count $\lvert E \rvert$	Number of unique users who interacted through a reply to a post or comment
Network Density $\frac{2\lvert E \rvert}{\lvert N \rvert(\lvert N \rvert-1)}$	Number of edge in graph over number of possible edges
Connected Component Count	Number of subgraphs in which all pairs of nodes are connected by an edge
Clustering Coefficient	Measure of the degree to which nodes tend to cluster together
Mean Connected Component Size	Mean number of nodes in a connected component
Mean Shortest Path	Mean distance between each pair of vertices that are connected by a path
Network Diameter	Maximum distance between any pair of nodes within a connected component
Degree Centrality	Importance score based simply on the number of links held by each node
Betweenness	Detecting the amount of influence a node has over the flow of information in a graph
Closeness	Detection of most influential nodes to circulate information efficiently
PageRank	Ranking of the nodes in the graph based on the structure of the incoming links

Table 3.3. Name and description of user interaction measure

Original Search Terms				Additional Terms	
2019-ncov	COVID	mers	SARSCOVID19	corona	rona
2019ncov	COVID-19	sars	wuflu	outbreak	sars-cov-2
coronavirus	COVID19	SARS2	Wuhan	pandemic	virus

Table 3.4. Lexicons of COVID-19 terms

To ensure consistency in our analysis, we apply min-max scaling to all the data, a technique that utilizes the minimum and maximum values within the dataset for normalization. Furthermore, we divide the data into training and testing sets, maintaining a uniform 2:1 ratio for all the models. As a pre-processing step, we employ a seven-day rolling mean to smooth the dataset, effectively filtering out short-term fluctuations and outliers. This enables us to better understand the underlying trends and patterns in the data.

ARIMA Model: The ARIMA model has been chosen for its effectiveness in handling univariate time series data, aligning with the nature of our dataset and research objectives. Following an exploratory data analysis, it has become evident that differencing (integration) is required to achieve stationarity in the data. Consequently, we adopt the ARIMA *(p, d, q)* model, where *'p'* represents the autoregressive order, *'d'* signifies the degree of differencing, and *'q'* indicates the moving average order.

For the implementation of the ARIMA model, we have leveraged the *Statsmodels* library in Python. The model is fitted to the training data, allowing it to learn the necessary parameters for forecasting future values. To optimize the model's performance, an extensive grid search has been conducted, encompassing a wide range of hyperparameters *(p,d,q)*(see in table 3.5 and 3.6).

The aim of this hyperparameter tuning process is to identify the configuration that strikes a balance between the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) scores. This meticulous tuning procedure ensures that the ARIMA model is fine-tuned to the specific characteristics of each dataset, resulting in improved forecasting accuracy.

We evaluate the model's performance using the Root Mean Square Error (RMSE) as the primary metric. The RMSE results are presented in Table 4.10 for the pre-pandemic dataset and in Table 4.11 for the mid-pandemic dataset. These RMSE values are subsequently used for a

comprehensive comparative analysis with other forecasting models, including Prophet, LSTM, and Transformer-based models.

Prophet Model: In our implementation of the Prophet model, we have opted for this specific forecasting tool due to its well-documented effectiveness in handling time series data, especially in situations where distinct seasonal patterns and holidays play a significant role. Prophet, an open-source forecasting tool developed by Facebook, demonstrates particular suitability for datasets characterized by strong seasonal components.

To operationalize the Prophet model, we have leveraged the *prophet* library available in Python. The model's setup configuration is outlined in Equation 2.2. Here, we define the trend using $g(i)$, allowing us to model both linear and non-linear trends in our time series data. We carefully select the appropriate trend component that best aligns with our use case.

The seasonality of our data is captured by $s(t)$, and we approximate it using a Fourier series. Given our practice of smoothing the data on a weekly basis, we opt for yearly seasonality, omitting the optional weekly and daily seasonality components. This approach is well-suited to our dataset's characteristics.

The third term, $h(t)$, pertains to holidays. We have observed that incorporating a list of U.S. holiday dates results in reduced error for the majority of our time series data in the pre-COVID period. This observation is likely attributed to the fact that the Reddit user population is primarily centered in the United States, rendering the inclusion of U.S. holidays a valuable adjustment to our model.

We perform a grid search over the model's hyper-parameters (see in table 3.5 and 3.6) to identify the configuration that yields the best performance. The model is fitted to the training data with optimized hyper-parameters. The fitted model is then used to forecast future value for both pre-pandemic and mid-pandemic periods. The performance of this model has shown in table 4.10 and 4.11.

LSTM Model: The selection of a deep learning-based LSTM model stems from its ability to effectively capture long-term dependencies within sequential data. To prepare the time series data for LSTM modeling, we adopt a supervised machine learning framework. When working with a time series comprising N data points, denoted as $x_{t-N+1}, ..., x_{t-1}, x_t$, and aiming for M step ahead prediction, the transformation entails constructing input (X) and output (Y) sequences. In

Models	Posts	Comments	Combined Posts-Comments
ARIMA	$(p,\ d,\ q) = (6, 2, 1)$	$(p,\ d,\ q) = (0, 1, 0)$	$(p,\ d,\ q) = (0, 1, 0)$
Prophet	changepoint_prior_scale = 0.2 holidays_prior_scale = 0.1 n_changepoints = 50 seasonality_mode = additive	changepoint_prior_scale = 0.5 holidays_prior_scale = 0.5 n_changepoints = 65 seasonality_mode = additive	changepoint_prior_scale = 0.5 holidays_prior_scale = 0.4 n_changepoints = 65 seasonality_mode = additive
LSTM	Neuron = 50 Epochs = 50 Batch Size = 1	Neuron = 50 Epochs = 50 Batch Size = 1	Neuron = 100 Epochs = 50 Batch Size = 1
Transformer	$d_{model} = 512$, Batch Size = 64, Epochs = 200, Attention Type = full, Encoder Number = 4 Decoder Number = 4, Dropout Ratio for each sublayer = 0.2, Learning Ratio = 0.0001		

Table 3.5. Models hyper-parameters summary on pre-pandemic data

Models	Posts	Comments	Combined Posts-Comments
ARIMA	$(p,\ d,\ q) = (0, 0, 0)$	$(p,\ d,\ q) = (6, 2, 0)$	$(p,\ d,\ q) = (0, 1, 1)$
Prophet	changepoint_prior_scale = 0.1 holidays_prior_scale = 0.1 n_changepoints = 30 seasonality_mode = multiplicative	changepoint_prior_scale = 0.1 holidays_prior_scale = 0.1 n_changepoints = 30 seasonality_mode = multiplicative	changepoint_prior_scale = 0.1 holidays_prior_scale = 0.1 n_changepoints = 30 seasonality_mode = multiplicative
LSTM	Neuron = 100 Epochs = 50 Batch Size = 1	Neuron = 100 Epochs = 50 Batch Size = 1	Neuron = 100 Epochs = 50 Batch Size = 1
Transformer	$d_{model} = 512$, Batch Size = 64, Epochs = 200, Attention Type = full, Encoder Number = 4 Decoder Number = 4, Dropout Ratio for each sublayer = 0.2, Learning Ratio = 0.0001		

Table 3.6. Models hyper-parameters summary on mid-pandemic data

this context, the input sequence X encompasses the data points from $x_{t-N+1}, ..., x_{t-M}$, while the output sequence Y comprises the data points from $x_{t-M+1}, x_{t-M+1}..., x_{t-1}, x_t$. It's worth noting that each data point x_t may represent either a single scalar value or a vector containing multiple features. For our specific investigation, we set N to 7 and M to 1 as we prepare the data for input into the LSTM model.

In our quest to optimize the LSTM model's performance, we executed a grid search across the model's hyperparameters, as detailed in tables 3.5 and 3.6, to identify the configurations that yield the most favorable results. After selecting the optimal hyperparameters, we trained the model on the training data. As a regularization technique, we applied a dropout rate of 0.2 to the LSTM layers. The training process employed the Adam optimizer with a learning rate set at 0.02.

Following the model's training, we used it for forecasting future data points, and the computed performance metrics are documented in tables 4.10 and 4.11.

Transformer Model: Our Transformer-based time series forecasting model follows the original Transformer architecture (Vaswani *et. al.* [41]) consisting of encoder and decoder layers.

The encoder comprises an input layer, a positional encoding layer, and a series of identical encoder layers. Initially, the input layer transforms the input time series data into a d_{model}-dimensional vector using a fully-connected network, a crucial step for enabling the model's multihead attention mechanism. To encode sequential information within the time series data, positional encoding employs sine and cosine functions. This is achieved by element-wise addition of the input vector with a positional encoding vector, generating a resultant vector. Subsequently, this vector is passed to the encoder layers. Each encoder layer includes two sub-layers: a self-attention sub-layer and a fully-connected feed-forward sub-layer. After each sub-layer, a normalization layer is applied. Ultimately, the encoder produces a d_{model}-dimensional vector for input to the decoder.

We utilize a decoder design reminiscent of the original Transformer architecture introduced by Vaswani *et. al.* [41]. The decoder consists of the following components: an input layer, four decoder layers that are identical in structure, and an output layer. The decoder's input commences with the final data point from the encoder input. The input layer then transforms this decoder input into a vector with a dimensionality of d_{model}.

In addition to the two sub-layers found within each encoder layer, the decoder incorporates a third sub-layer dedicated to applying self-attention mechanisms over the encoder output. Lastly,

there is an output layer responsible for mapping the output of the last decoder layer to the desired target time sequence.

To enhance the predictive capabilities of our model, we employ look-ahead masking and introduce a one-position offset between the decoder input and the target output within the decoder. This strategic approach ensures that the prediction of a data point in the time series depends solely on the preceding data points.

During the model training process, we follow a similar approach to training LSTM models, treating it as a supervised machine learning task. Specifically, we train the model to predict a single data point in the future based on the information from the previous seven training data points.

The encoder is provided with the input sequence $(x_1, x_2, x_3, x_4, x_5, x_6)$, and the decoder input consists of (x_6, x_7). The primary objective of the decoder is to generate the output x_8. To ensure that the model's attention mechanism doesn't inadvertently focus on data from the future when making predictions, we apply a look-ahead mask. This look-ahead mask restricts the attention mechanism to consider only the data points up to x_7, specifically x_6 and x_7 when predicting x_8.

In other words, when the model is predicting the target data points x_7 and x_8, the look-ahead mask ensures that the attention weights are exclusively placed on x_6 and x_7. This precaution prevents the decoder from gaining information about x_8 from its decoder input, maintaining the integrity of the predictive process. The hyper-parameters while training the model are listed in table 3.5 and 3.6. The performances of the trained model on test data are shown in table 4.10 and 4.11.

4. RESULT AND DISCUSSION

4.1. How Extensively is the Topic of COVID-19 Addressed within the Selected Subreddits?

The primary objective of this study is to gauge the extent of COVID-19-related discussions across these three subreddits. This metric will enable us to assess the potential influence of COVID-19 on our other measurement parameters. Alternatively, this also serves as an indicator of the extent to which users are focused on the pandemic. To approach this research, we draw inspiration from the work conducted by Laura *et al.* [4]. In our analysis, we utilize the COVID-19 terminology provided in Table 3.4 to calculate the proportion of posts and comments per month that reference terms associated with COVID-19.

Result: The findings are presented in Figures 4.1 and 4.2. Figure 4.1 illustrates the percentage of posts related to COVID-19, while Figure 4.2 represents the percentage of comments. It's evident that discussions surrounding COVID-19 are prevalent in all three subreddits. Furthermore, the analysis reveals that these discussions had a gradual onset in January 2020 and significantly increased in March 2020.

Notably, the r/EatingDisorder subreddit exhibits a higher proportion of COVID-19-related discussions in the posts, whereas r/BingeEatingDisorder features a greater prevalence of such discussions in the comments. This indicates that the dialogue concerning COVID-19 occurs across both post and comment sections over the duration of the study period.

Discussion: It's noteworthy that all three subreddits display a gradual rise in COVID-19 discussions starting in January 2020, with peak activity observed in March 2020. Notably, r/EatingDisorder stands out with a higher prevalence of COVID-19-related discussions compared to the other two subreddits. Interestingly, in May and June, there is a further increase in discussions in r/EatingDisorder, which suggests that users in this community may have been particularly affected by eating disorder issues stemming from the impact of COVID-19. Moreover, figure 4.1 illustrates a second pick from January 2021 to May 2021 of the r/EatingDisorder subreddit which strongly supports the second wave impact of COVID-19 on eating disorder people.

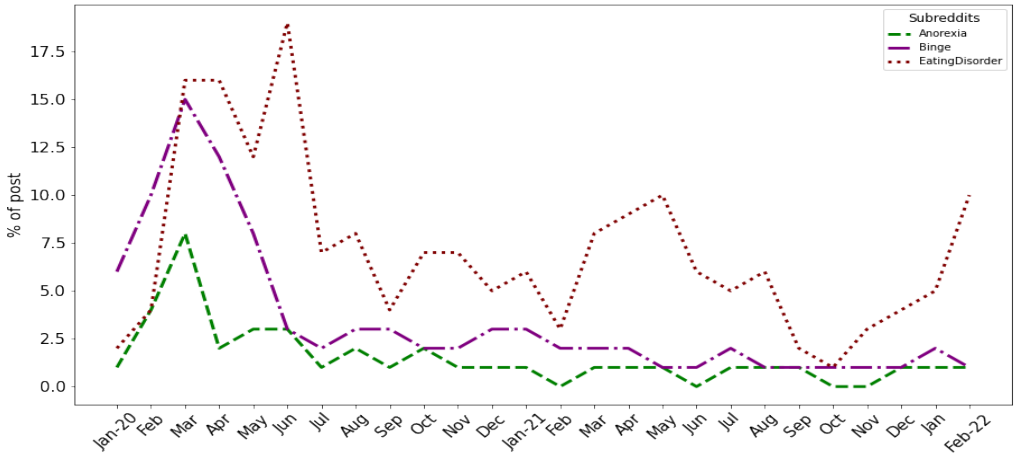

Figure 4.1. Percentage of posts per month mentioned COVID-19 related words

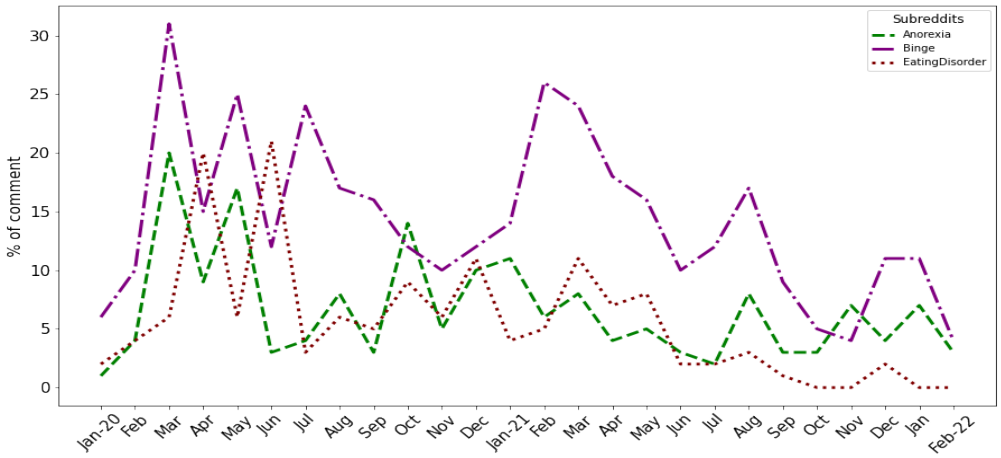

Figure 4.2. Percentage of comments per month mentioned COVID-19 related words

The proportion of posts in r/BingeEatingDisorder and r/AnorexiaNervosa peaks in March and April 2020 before gradually declining over the study period. This trend can be attributed to the initial phases of isolation and lockdown measures, which marked the first experience of such conditions for many individuals. The implementation of social distancing, restrictions on in-person interactions, and the shift to working from home significantly disrupted daily routines and activities. Consequently, these disruptions could have influenced eating patterns, rendering individuals more susceptible to irregular eating behaviors, including overeating.

Figure 4.2 illustrates the proportion of comments related to COVID-19. Notably, the r/BingeEatingDisorder subreddit's comment section exhibits a higher prevalence of discussions concerning COVID-19 compared to the other two subreddits. In contrast, there is a lower percentage of comments mentioning COVID-19 in r/EatingDisorder compared to posts. This observation can be attributed to the fact that when COVID-19 is already mentioned in a post and a conversation is initiated, users in the comment section do not necessarily need to explicitly mention COVID-19, as the context is already established.

When determining the starting point for the COVID-19 period in our research analysis, we selected March 1, 2020, as a meaningful date. On this date, the United States, where a significant portion of Reddit users reside, began to address COVID-19 in a more concerted manner. March 1 closely followed the first reported COVID-19 death in the United States on February 28, and it preceded subsequent developments such as school closures and state lockdowns. Our analysis also reveals that discussions related to COVID-19 on these r/EatingDisorder subreddit reached their peak in March and April, coinciding with the period when the virus started to significantly impact people's lives.

4.2. Do Conversational Shifts in Eating Disorder Subreddits Differ Amidst COVID-19?

Providing a deeper comprehension of the subjects and debates in these communities, we employ the LDA topic model, utilizing both post and comment data. By comparing the results of topic modeling before and during the pandemic, we enhance our understanding of the communities. Additionally, we integrate these findings with various analyses to gain a more comprehensive insight into these online communities.

Result: Tables 4.1 to 4.6 present the outcomes of our LDA topic model. This model was developed using the pre-processed data from both pre-pandemic and mid-pandemic periods for each

subreddit. Detailed information on our approach to constructing this model can be found in Section 3.4.1.

Discussion: Table 4.1 displays the topics that were discussed in the pre-pandemic data of the r/EatingDisorder subreddit. In contrast, Table 4.2 illustrates the topics from the mid-pandemic data. Our coherence analysis revealed the presence of 19 predominant topics in both the pre-pandemic and mid-pandemic datasets of the r/EatingDisorder subreddit. To enhance our understanding of the discussion evolution from the pre-pandemic to mid-pandemic data, we manually assigned topic labels to each of these topics. These labels were based on our observations of the high-probability words provided by the model.

We observe several topics that were discussed both before and during the pandemic period, indicating a degree of continuity in the community's conversations. Topics such as "BODY IMAGE," "RECOVERY NAVIGATION," "SUPPORTIVE RELATIONSHIPS," "EXERCISES," "TREATMENT OPTIONS/RESOURCE," "RESEARCH AND INFORMATION," and "ADOLESCENTS AND ANXIETY/YOUTH MENTAL HEALTH" remained consistent. These topics reflect a common context of discussion, suggesting that the community maintained some shared interests across both time periods.

In contrast, during the mid-pandemic, we noticed a significant increase in discussions related to COVID-19. The "EATING HABIT IN COVID" topic revolved around meal planning, dietary choices, calorie intake, and healthy eating during the pandemic. The "SEEKING SUPPORT COVID" topic centered on seeking help, advice, treatment, support, and recovery strategies from the community during the pandemic. The "POST-COVID EXPERIENCE" topic shared information about individuals' experiences after being infected by COVID-19, including details about taste, weakness, recovery advice, and more.

Furthermore, there were topics like "BALANCED/HEALTHY DIET", "EMOTIONAL STRUGGLES", "SYMPTOMS", and "CALORIE INTAKE". These topics were associated with the aftermath of the COVID-19 pandemic and reflected the community's evolving concerns during that time.

Table 4.3 presents the topics discussed in the pre-pandemic data of the r/AnorexiaNervosa subreddit, while Table 4.4 displays topics from the mid-pandemic data. The coherence scores indicated the presence of 17 dominant topics in both datasets of the r/AnorexiaNervosa subreddit.

To enhance clarity, we manually assigned annotated topic labels to each set of words generated by the model.

It is evident that numerous topics were discussed both before and during the pandemic, indicating a consistent pattern of conversation within the community. Topics such as "POSITIVITY AND HOPEFULNESS/POSITIVITY," "DIAGNOSING ANOREXIA NERVOSA/EXPERIENCING ANOREXIA," "GAME OVER MENTALITY/NEGATIVE PERCEPTION," "PHYSICAL SYMPTOMS/PHYSICAL HEALTH AND SYMPTOMS," "TREATMENT AND SUPPORT," "PERSEVERANCE AND RECOVERY," and "FOOD CRAVINGS/ENJOYING FOOD" remained consistent across both time periods. This suggests that users maintained discussions on these shared interests throughout both time periods.

Nevertheless, there are notable shifts in the topics discussed after the onset of the pandemic. One dominant topic revolves around the rules and regulations of the r/AnorexiaNervosa subreddit, encompassing discussions about subreddit rules, posts, comments, and the actions (such as immediate reporting) taken for violations of community regulations. Another prominent topic centers on "POST-COVID EXPERIENCE," addressing aspects like COVID-19, medications, side effects, pain, and stomach issues. Additionally, there's a focus on "LOCK DOWN AND FAMILY," reflecting discussions related to lockdown measures, family dynamics, and the impact of the pandemic on various aspects of life. These shifts indicate that the community was significantly influenced by the COVID-19 pandemic, with members seeking advice and information about maintaining a "BALANCED DIET" and "EXERCISES" during these challenging times.

Table 4.5 presents the topics that were discussed in the pre-pandemic dataset of r/BingeEatingDisorder, while Table 4.6 displays topics from the mid-pandemic dataset. Coherence scores indicated the presence of 17 dominant topics in the pre-pandemic data and 19 dominant topics in the mid-pandemic data of r/BingeEatingDisorder. Annotated topic labels were assigned to each set of words generated by the model to facilitate understanding and analysis.

It's evident that r/BingeEatingDisorder exhibited similar trends to the other two subreddits. Topics such as "SEEKING HELP/PROFESSIONAL HELP," "COMMUNITY HELP," "MENTAL HEALTH/MENTAL DISORDERS," "POSITIVITY," "INSOMNIA," "EMOTIONS," "TREATMENT," and "ADDICTION/ADDICTIVE BEHAVIORS" displayed consistent discussions across both time periods.

However, Table 4.6 demonstrates that the r/BingeEatingDisorder community was indeed affected by COVID-19 restrictions. The "COVID RESTRICTION" topic refers to lockdown restrictions that individuals experienced during the pandemic. The "DAILY ROUTINE" topic reflects changes in people's daily activities and mobility due to pandemic-related lockdowns. Furthermore, the "FAMILY CONCERNS" topic underscores people's worries about their family members and their eating habits during the pandemic. These findings highlight the pandemic's influence on the community's discussions and concerns.

4.3. Do Alternations in Social Interaction Patterns in Eating Disorder Subreddits Differ During COVID-19?

Our research aims to examine the shifts in social interaction dynamics within three chosen eating disorder subreddits, comparing the pre-pandemic period to the mid-pandemic period. To achieve this, we constructed separate social networks for each subreddit during both time periods. These networks help us measure social interaction, as detailed in table **??**, and we further analyze the alterations in network structure by calculating various centrality metrics.

Result: Tables 4.7, 4.8, and 4.9 present the social interaction metrics for the r/EatingDisorder, r/AnorexiaNervosa, and r/BingeEatingDisorder subreddits, respectively. We observe significant shifts in these metrics across all three subreddits.

There is a noteworthy increase in the number of nodes and edges when transitioning from the pre-pandemic to the mid-pandemic period for all three subreddits. Additionally, there is a substantial rise in the count of connected components for r/AnorexiaNervosa and r/BingeEatingDisorder, while r/EatingDisorder experiences a decrease in this count.

However, we note a decrease in network density, clustering coefficient, and mean shortest path for all three subreddits. Notably, the network diameter remains unchanged for r/EatingDisorder and r/AnorexiaNervosa, except for r/BingeEatingDisorder.

Figure 4.3, 4.4, and 4.5 present the comparison in Betweenness, Closeness, Degree centrality, and Page Rank between pre-pandemic and mid-pandemic social network on each of the subreddits. These figures reveal significant changes in each centrality measure.

Discussion: With the exception of Node Count, all the metrics we measured are related to the connectivity within the social network. The increase in Node Count represents the growth in the number of unique users in the subreddit. Additionally, the rising number of Edge Counts signifies

Annotated Labels	High Probability Words
Recovery and Resilience	recovery recover thought life period relapse healthy thank month gain experience
Body Image and Control Struggles	gain thought fat change back lose control never come keep feeling give
Family and Therapeutic Support	therapy parent therapist thought talk love mother control issue always support
Supportive/Trigger Relationships	friend support trigger love understand relationship care situation partner struggle
Recovery Navigation	healthy easy hope actually advice keep taste hard little struggle sound anxiety
Eating Habits and Medication	meal hungry stomach calorie doctor appetite drink gain water problem advice
Bulimia Recovery	binge-purge stop restrict bulimia cycle control advice gain eating calorie struggle
Vegan Diet and Challenges	vegan diet healthy problem hard recover trigger control health big struggle
Exercises	exercise gym gain lose healthy calorie fat loss underweight goal health hard normal
Meal Planning and Cooking	meal dinner healthy home cook recovery friend snack easy away stop big option
Work Balance	college job struggle life stress healthy habit hard school lose move sure deal hope
Treatment Options	treatment therapist therapy group program inpatient support residential insurance
Mental Health in University Life	university mental health happy recover professional right back worry fight life
Loved One Struggling	love sister guy friend person mom worried problem struggle notice fat skinny thin
Weight Struggling	weigh overweight family life hard hour scale choice problem sorry definitely happen
Depression and Anxiety	depression anxiety develop never hard life stop post serious purge sign disordered
Research and Information	study research question information link seek participate ask experience therapist
Adolescents and Anxiety	child young brother teen kid parent family anxiety struggle advice doctor hospital
Physical and Behavioral Challenges	pain sleep throw binge exercise starve lose calorie issue bulimia smile concerned

Table 4.1. Main result of topic modeling for the pre-pandemic r/EatingDisorder subreddit

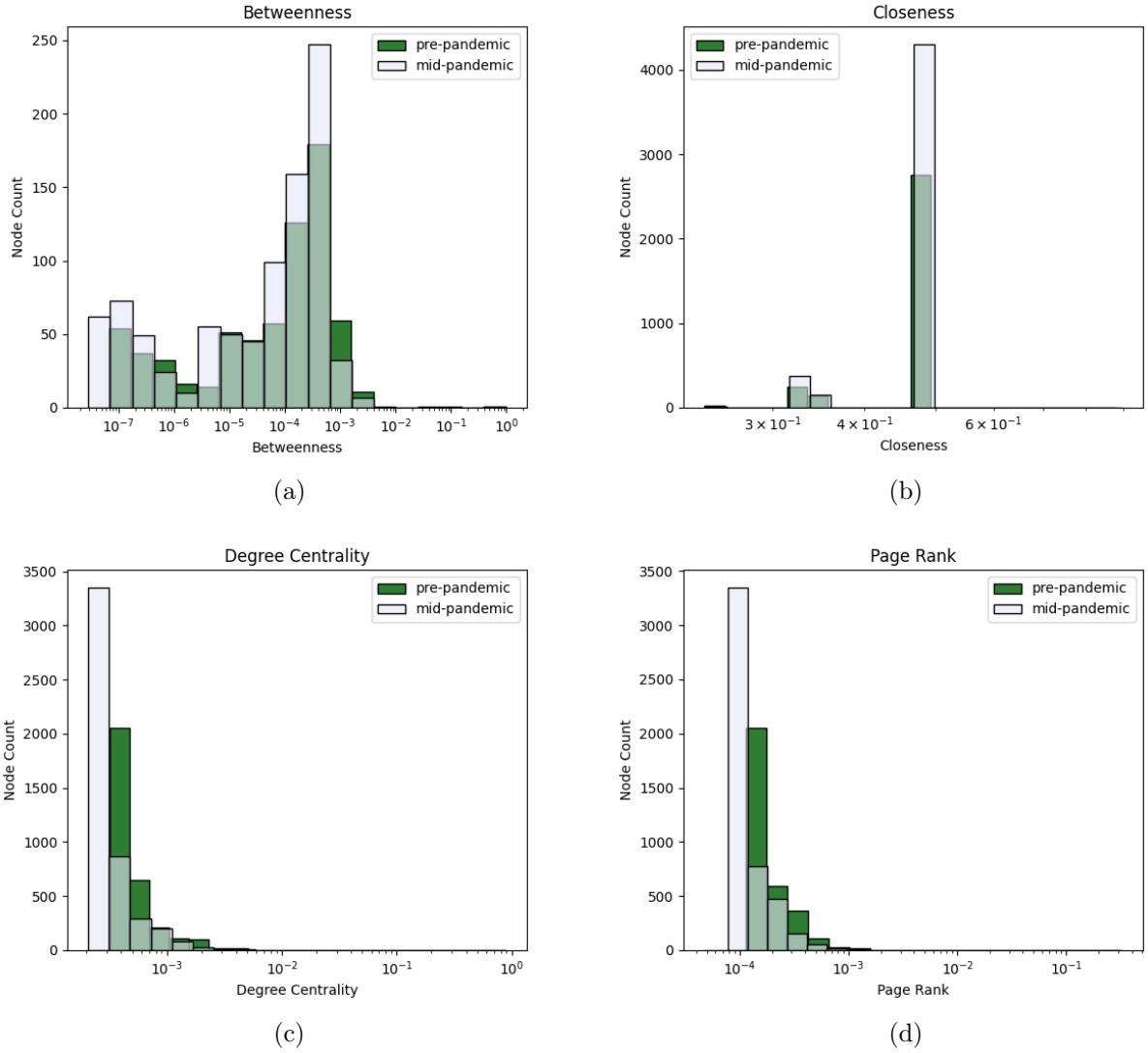

Figure 4.3. Comparison of betweenness, closeness, degree centrality, and page rank between pre-pandemic and mid-pandemic of r/EatingDisorder social network.

increased user engagement within the subreddit, such as seeking help, commenting on others' posts for guidance, sharing experiences, and expressing their feelings. This increase indicates that the COVID-19 pandemic has had a significant impact on a large number of individuals, prompting them to turn to Reddit for discussions related to eating disorders on a larger scale.

Furthermore, the rise in Connected Component Count and Cluster Coefficient, coupled with the decrease in Network Density within r/AnorexiaNervosa and r/BingeEatingDisorder subreddits, indicates that the users in these networks are sparsely connected in comparison to the network size. This suggests that users are primarily engaged in specific discussions and are less likely to post

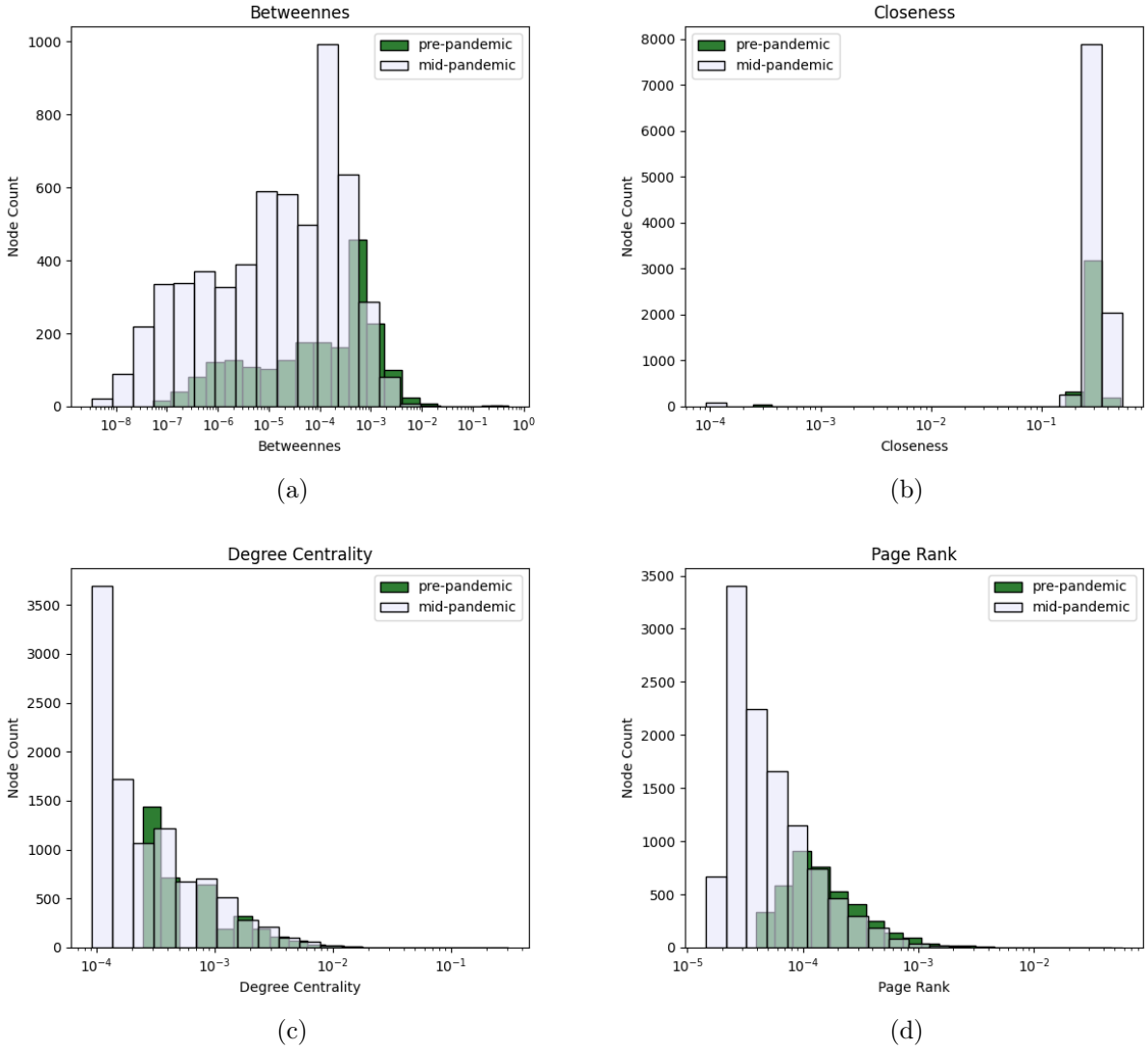

Figure 4.4. Comparison of betweenness, closeness, degree centrality, and page rank between pre-pandemic and mid-pandemic of r/AnorexiaNervosa social network.

unnecessary comments on other users' content. The increase in connected components also suggests a broader range of topics being discussed in these subreddits. This observation is consistent with our findings from topic modeling (as discussed in section 4.2), where we identified a wide array of prominent topics being discussed during both time periods.

The decrease in the Mean Shortest Path observed on each of the subreddits indicates that users are becoming more interconnected and closer to one another within these communities. This finding reflects the increased impact of COVID-19, which has heightened concerns related to eating disorders, compelling individuals to seek assistance and support from online communities.

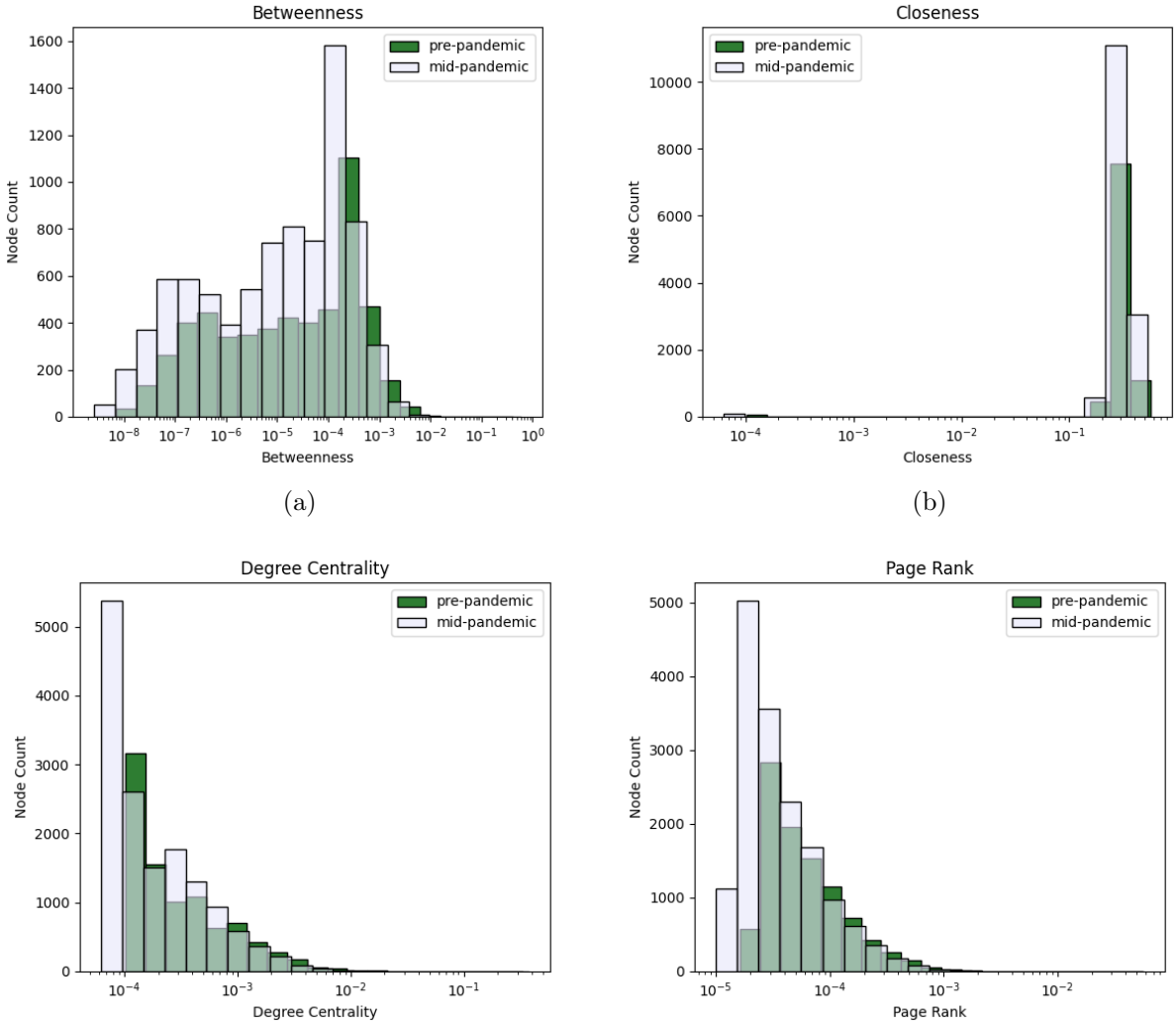

Figure 4.5. Comparison of betweenness, closeness, degree centrality, and page rank between pre-pandemic and mid-pandemic of r/BingeEatingDisorder social network.

In Figures 4.3(a), 4.4(a), and 4.5(a), we observe the betweenness centrality comparisons between the pre-pandemic and mid-pandemic periods for r/EatingDisorders, r/AnorexiaNervosa, and r/BingeEatingDisorder, respectively. The increased Node Count of betweenness values in the mid-pandemic data indicates the growing significance of user connectivity within the social network. Users are now playing a pivotal role in connecting with other users, emphasizing their influence on network interactions.

In Figures 4.3(b), 4.4(b), and 4.5(b), we observe that users have become closer to each other during the COVID-19 pandemic in each subreddit. This is evident from the Node Count

value of closeness centrality, reflecting an increase in user engagement and activeness in seeking social support within online communities during the epidemic period.

Figures 4.3(c), 4.4(c), and 4.5(c) present the degree centrality of r/EatingDisorders, r/AnorexiaNervosa, and r/BingeEatingDisorder. These figures illustrate the increased involvement of new users in each of the subreddits during the COVID-19 pandemic. Notably, both r/AnorexiaNervosa and r/BingeEatingDisorder experienced a higher influx of new users contributing to the community compared to r/EatingDisorders, which corresponds with the Node Count in these subreddits.

PageRank is an algorithm widely used to assess the influence or significance of web pages, notably employed by search engines like Google to rank web pages based on their relevance to search queries. In the realm of social networks, PageRank serves the purpose of identifying influential nodes within the network, essentially nodes that are easily accessible from other nodes. Figures 4.3(d), 4.4(d), and 4.5(d) present the PageRank results for r/EatingDisorders, r/AnorexiaNervosa, and r/BingeEatingDisorder, respectively. The graphs indicate that the number of influential users is relatively low in both the pre-pandemic and mid-pandemic periods, aligning with our earlier observations. It's worth noting that new users have made more substantial contributions to these communities during the epidemic, underscoring the extensive impact of COVID-19 on individuals dealing with eating disorders.

Our findings suggest that users within these communities have heightened their interactions with other users in a more inclusive manner, rather than forming tightly-knit groups that primarily engage with each other.

4.4. Analyze Performance of Time Series Forecasting Models on Post and Comment Data

In this study, our objective is to assess the effectiveness of a time series model based on Transformers using data related to post and comment counts from both the pre-pandemic and mid-pandemic periods. We conduct a comparative analysis of the Transformer-based model with ARIMA, Prophet, and LSTM models. The Root Mean Square Error (RMSE) values and their relative comparisons are presented in Table 4.10 and 4.11 for the pre-pandemic and mid-pandemic datasets, respectively.

We have calculated the relative performance gain concerning the ARIMA model. The comparisons presented in Table 4.10 and 4.11 clearly indicate that deep learning models, on the whole, outperform both the ARIMA and Prophet models. In Table 4.10, the LSTM model exhibits a relative improvement of 39.32% for posts, 41.98% for comments, and 35.04% for the combined count of posts and comments than Prophet model. It's worth noting that we initially expected the Transformer model to deliver superior performance compared to the LSTM model. However, we observe that the Transformer model's performance is slightly lower than LSTM for posts and comments but surpasses it for the combined count of posts and comments.

Table 4.11 provides a similar performance comparison, but this time for the LSTM model on mid-pandemic data. The LSTM model demonstrates a relative improvement of 50.98% for posts, 52.88% for comments, and 50.47% for the combined count of posts and comments when compared to the Prophet model. Notably, the Transformer model's performance remains slightly below that of the LSTM model in this context.

Our results indicate that LSTM demonstrates exceptional performance compared to other models. Deep learning models, such as LSTM, possess the capacity to effectively capture intricate patterns, retain them in memory, and discern dependencies within the data, surpassing the capabilities of linear models like ARIMA and Prophet.

Annotated Labels	High Probability Words
Eating Habits in COVID	meal binge work snack healthy covid restrict diet plan calorie hungry stop keep use full
Supportive Relationship	friend talk sister ask family trigger recovery understand post comment mom deal person
Recovery Navigation	binge stop control back restrict lose healthy end gain life purge month recovery recover cycle
Seeking Support COVID	recovery treatment support recover covid thank struggle hard therapy advice ask understand
Body Image	gain period find experience lose skinny healthy fat hope purge come never hard back less long
Relapse Prevention	relapse recovery trigger friend talk ask partner struggle advice support together care person
Appetite Challenges	hungry stomach hunger appetite meal drink normal doctor recovery struggle put sick
Exercises	exercise work workout gym healthy walk goal hard enjoy focus recovery struggle always use
Balanced/Healthy Diet	healthy health diet vegan scale weigh lose restriction change work behavior stop life loss
Triggers and Support	trigger struggle control friend brother mom home support always talk thank never anxiety
Emotional Struggles	life believe never fault struggle right voice self love blame matter attention thought find
Self-Perception	lose always hair never thin gain skinny hate healthy use sometimes problem grow notice mirror
Post-COVID Experience	covid clothe wear buy like new cook taste thank meat size recovery advice never weak find put
Symptoms	symptom pain stomach issue doctor bulimia problem diagnose sleep experience cause damage
Caloric Intake	calorie count gain healthy fat amount low underweight overweight problem professional exercise
Treatment Resource	therapist treatment resource group program online week call inpatient thank helpful free work
Medication	medication medical nauseous nausea doctor anxiety appetite issue experience month cause
Research and Information	study research question information survey ask individual participate project seek thank
Youth Mental Health	talk seek mental health information parent teacher personal concern available healthy

Table 4.2. Main result of topic modeling for the mid-pandemic r/EatingDisorder subreddit

Annotated Labels	High Probability Words
Body Image & Societal Perceptions	people skinny body think look say thin weight fat always want way hate thing
Support and Communication	help tell say friend talk want ask people make thing need good understand go care
Recovery Challenges	binge time food body need really hard keep stop want try week well scared bad go
Work Life Balance	help work go school know well parent make good think stress thing lot right live try
Diagnosing Anorexia Nervosa	anorexia disorder anorexic weight diagnose doctor people underweight food restrict
Meal Plan/Caloric Intake	food day calorie meal make go dinner good time want snack lunch cook think know
Game Over Mentality	make think know people want fuck lie way see person love fucking time take disorder
Perseverance and Recovery	start body time go recovery long first happen take year slow think recover restrict
Treatment and Support	treatment go help disorder ed therapist need hospital doctor inpatient post place
Positivity and Hopefulness	thank recovery good hope post love well happy relapse much recover proud make
Self-control and Hunger	self control brain hungry hunger starve feeling think thought sleep bed help tip work
Weight Tracking	weight gain lose time weigh go pound healthy scale day start calorie month loss lb
Physical Symptoms	stomach pain hair loss body cold water drink take lot vitamin help sure thing
Negative Perception	hate disgusting look go time fat make know wear clothe fit size see mirror want cry
Food Cravings	food watch make think crave good trigger always miss lot way fast video candy much
Dietary Choices	drink coffee taste make ask water also vegan caffeine sugar ensure good help food
Mental Health Effects	low dream faint anxiety smoke self-harm heart drug doctor anorexia cause problem

Table 4.3. Main result of topic modeling for the pre-pandemic r/AnorexiaNervosa subreddit

Annotated Labels	High Probability Words
Subreddit Rules and Guidelines	subreddit rule harmful post comment op report immediately advice read user moderator
Positivity	recovery recover thank proud deserve well much hope happy love hard body life today
Treatment and Support	doctor help need go therapist know tell disorder talk take support therapy struggle
Perseverance and Recovery	back year start recover go recovery relapse month still ed restrict think long bad period
Balanced Diet/Calorie Intake	food day calorie meal snack hungry much amount lot safe need think full
Relationship and Communication	relationship talk tell find control love think understand need help issue situation sound
Stigma and Community Support	anorexia, disorder ed people person trigger post comment think see sub talk call means
Exercises	exercises work gym weight gain lose body healthy weigh scale much need calorie pound
Experiencing Anorexia	anorexic anorexia disorder underweight overweight control body diagnose healthy valid
Physical Health and Symptoms	symptom doctor drink low body take water cause energy heart work help vitamin blood
Negative Perception	fuck fucking shit people never sometimes second place thing bad cold nice winter world
Post-COVID Experience	covid medication anxiety depression pain bad side effect pain stomach wake night sleep
Lockdown and Family	mom parent dad sister mother mum family lockdown home work school start year food
Enjoying Food	love shower like cookie look warm favorite sweet nice enjoy amazing buy great cake hot
Seeking Advice on Dietary	ask question answer information doctor allow advice vegetarian seem medical subreddit
Digestive and Bloating Issue	stomach bloat constipation tip help try laxative also use water thank suggestion yoga
Anorexia Side Effect	hair loss fall skin dry grow tooth face thin sink take mouth shower look wound

Table 4.4. Main result of topic modeling for the mid-pandemic r/AnorexiaNervosa subreddit

Annotated Labels	High Probability Words
Seeking Help	people say tell talk ask issue problem think understand disorder help see bed thing try problem
Dietary Restriction	meal snack restriction restrict calorie normal hungry hunger body hunger try healthy food much
Mental Health	brain urge thought think mind control give try read thing habit way stop want plan never book
Positivity	good happy well amazing great proud hope thank today start really much today tomorrow week
Self-Improvement	self control improve life think relationship friend live depression year look bed way change issue
Insomnia	night sleep wake bed episode happen urge pm hour morning try work really well think happen
Professional Assistance	therapist therapy find help word disorder recovery really need helpful group try lot thank talk
Calorie Intake	calorie diet weight lose gain calorie pound lb year month week start healthy exercise body
Physical Stress	stomach pain ache bad sick stop think well year purge really much still start happen anymore sorry
Emotion	emotion emotional comfort feeling work find hard stress trigger think watch also eating problem
Negative Perception	fuck fucking shit hate think night much cry night look fat love body man want see thing
Treatment	doctor vyvanse medication psychiatrist adhd prescribe work med start month side effect experience
Expenditure	money dollar buy order spend shop store home place food car work store alone drive take live
Family and Friends	family home house friend week holiday come start leave dinner buy junk candy chocolate cookie
Addiction	addicted addiction addict drug alcoholic drink sugar carb keto diet craving work protein fat
Adolescent Food	young teen child kid baby sister brother mom parent dad mother teacher school vegan donut food
Life Style Changes	quit smoking smoke weed challenge change meme behavior munchie work job travel week month

Table 4.5. Main result of topic modeling for the pre-pandemic r/BingeEatingDisorder subreddit

Annotated Labels	High Probability Words
Struggle and Hope	struggle hard hate bad wish hope want think sorry time try much well stop thing right time
Professional Help	help take doctor therapist therapy professional month tell see need talk try first ask finally work
Community Help	post comment people struggle support talk link share help thank love find need look good helpful
Daily Routine	day work time start back home tomorrow week last know come control next think happen new
Emotions	emotions emotional love sad happy feelings good matter life want thing never look change think
Positivity	good great happy proud amazing well think thank time day today week month keep progress
Diet and Nutrition	diet calorie protein carb high sugar keto carb fat day stomach make amount hungry count think
Insomnia	night sleep asleep think walk try help time thing keep busy really work take tip write urge make
Body and Eating Habits	body physical physically point hunger starve overeat fuel look notice point mental period time
Social Pressure	people friend mom family say tell think see fat make comment understand never always shame
COVID Restriction	covid restrict restriction stop lockdown start book read learn podcast helpful work diet recover
Food Waste	food waste throw away bad guilty always time want stop full literally still normal think control
Weight and Body Image	weigh weight lose gain year month loss pound lb kg time back start healthy scale look fat
Addictive Behavior	addiction addictive emotion feelings thought habits action think brain urge problem learn reason
Eating/Mental Disorders	bulimia anorexia ed purge mental health anxiety disorder issue struggle suffer therapy diagnose
Food and Carvings	buy cookie sugar chocolate sweet pizza bag chip snack cake crave healthy taste candy cheese
Family Concerns	mom dad sister parent person family habit eating watch see unhealthy food perspective grow life
Treatment	vyvanse adhd take medication doctor prescribe med mg work appetite side effect adderall drug
Addiction Recovery	sober away addiction reward money use app uber delivery drive work pay grocery strategy

Table 4.6. Main result of topic modeling for the mid-pandemic r/BingeEatingDisorder subreddit

Metric Name	Pre-Pandemic	Mid-Pandemic
Node Count	3,164	4,839
Edge Count	4,336	6,396
Network Density	0.0009	0.0005
Connected Components Count	2	1
Clustering Coefficient	0.1783	0.1597
Mean Connected Component	1,582	4,839
Mean Shortest Path	2.2499	2.2191
Network Diameter	6	6

Table 4.7. User interaction metrics for r/EatingDisorder social network

Metric Name	Pre-Pandemic	Mid-Pandemic
Node Count	4,079	10,936
Edge Count	9,260	34,266
Network Density	0.0011	0.00057
Connected Components Count	360	710
Clustering Coefficient	0.1365	0.1783
Mean Connected Component	11.3305	15.4028
Mean Shortest Path	3.2629	3.0899
Network Diameter	8	8

Table 4.8. User interaction metrics for r/AnorexiaNervosa social network

Metric Name	Pre-Pandemic	Mid-Pandemic
Node Count	9,733	15,953
Edge Count	28,875	46,050
Network Density	0.00060	0.00036
Connected Components Count	611	1183
Clustering Coefficient	0.1604	0.1619
Mean Connected Component	15.929	13.4852
Mean Shortest Path	3.1525	3.13307
Network Diameter	8	9

Table 4.9. User interaction metrics for r/BingeEatingDisorder social network

Time Series Model	Posts (RMSE)	Comments (RMSE)	Total Posts and Comments (RMSE)
ARIMA	0.709 (0%)	0.567 (0%)	0.585 (0%)
Prophet	0.117 (-83.50%)	0.131 (-76.90%)	0.117 (-80.0%)
LSTM	0.071 (-90.10%)	0.076 (-86.59%)	0.076 (-87.01%)
Transformer	0.09134 (-87.11%)	0.0977 (-82.76%)	0.062 (-89.40%)

Table 4.10. Summary of Time Series Models on pre-pandemic posts and comments count

Time Series Model	Posts (RMSE)	Comments (RMSE)	Total Posts and Comments (RMSE)
ARIMA	0.685 (0%)	0.758 (0%)	0.754 (0%)
Prophet	0.102 (-85.10%)	0.104 (-86.28%)	0.105 (-86.07%)
LSTM	0.050 (-92.70%)	0.049 (-93.54%)	0.052 (-93.10%)
Transformer	0.0715 (-89.56%)	0.0959 (-87.35%)	0.0917 (-87.83%)

Table 4.11. Summary of Time Series Models on mid-pandemic posts and comments count

5. CONCLUSION

In the field of Natural Language Processing (NLP), techniques such as topic modeling, text summarization, etc. are incredibly powerful for understanding large volumes of unstructured text data. Topic modeling, specifically, delves into the underlying structures present within a collection of documents, revealing the hidden themes or subjects that run through the text. By identifying these latent topics, it provides a method for organizing, categorizing, and analyzing extensive datasets in a more comprehensible and manageable manner.

In our research study, we carried out a comparative analysis using Latent Dirichlet Allocation (LDA) topic modeling to assess the alterations in discussion content in response to the COVID-19 pandemic. Additionally, we established a social interaction network to investigate interaction metrics, focusing on how users engage within the subreddits. Although our study is focused on eating disorders, the system we developed is versatile and can be applied across various domains to explore the impact of a stimulus on any community.

Primarily, we observed a significant surge in the number of users across all the subreddits following the declaration of COVID-19 as a pandemic. These subreddits transformed into hubs for sharing information related to the COVID-19 pandemic, new guidelines, and seeking assistance and support. Concerned users increased their participation in these forums by sharing valuable insights and guidance.

We noticed noteworthy changes in the discussion content within the various subreddits. In r/EatingDisorders, there was a predominant focus on topics such as social support, coping with eating habits during the COVID-19 pandemic, post-COVID experiences, and resources like telemedicine during the lockdown. The r/AnorexiaNervosa subreddit saw considerable discussions on changes in subreddit guidelines, seeking support for recovery, post-COVID experiences, lockdown-related issues, food, and family matters. In the r/BingeEatingDisorder subreddit, there was a notable shift in discussions toward disruptions in daily routines, changes in eating habits, issues like insomnia, seeking community assistance and support, and discussions related to mental health disorders. These discussion topics vividly underscore the intensified impact of COVID-19 on individuals dealing with eating disorders.

5.1. Limitations

While this study has made valuable contributions, there are some limitations that should be acknowledged. First, we focused exclusively on Reddit as the social media platform under analysis, which means our findings may not offer a comprehensive representation of all online social media support groups. Additionally, while Latent Dirichlet Allocation (LDA) is adept at identifying topics within a corpus, interpreting and assigning meaningful labels to these topics can be challenging. LDA-generated topics often tend to be abstract and require human judgment for accurate interpretation.

Furthermore, we made the assumption that most Reddit users are based in the USA, as demographic data was unavailable. The impact of COVID-19 on eating disorders can differ significantly by region, so generalizing our findings may not be appropriate without taking regional disparities into account.

5.2. Future Work

Analyzing the influence of COVID-19 on eating disorders opens up numerous possibilities for future research. One avenue could involve delving into social support groups on various social media platforms and exploring posts and comments to gain a deeper understanding of the conversations related to the impact of COVID-19 on eating disorders. An interesting avenue for research would be to compile and analyze control data from COVID-19 discussion subreddits. This could involve studying a cross-section of users, including those who have posted in both eating disorder (EDs) and COVID-19 discussion subreddits, as well as users who were active in EDs subreddits a year ago but are no longer engaged in those communities. By comparing and contrasting these groups, researchers might uncover valuable insights into the changing dynamics of online discussions related to eating disorders during the pandemic, as well as the evolving interactions between these subreddits and the broader COVID-19 discourse. Additionally, it's worth considering the increased mortality rates during the pandemic, which have heightened fear and anxiety levels on a global scale. There's a potential research area that could investigate the connection between the fear of death and the prevalence or exacerbation of eating disorders. This could shed light on how these psychological factors interrelate during challenging times like the COVID-19 pandemic.

REFERENCES

[1] Sabeen Ahmed, Ian E. Nielsen, Aakash Tripathi, Shamoon Siddiqui, Ravi P. Ramachandran, and Ghulam Rasool. Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing*, July 2023. doi:10.1007/s00034-023-02454-8.

[2] Claus Boye Asmussen and Charles Møller. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1), October 2019. doi:10.1186/s40537-019-0255-7.

[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839, May 2020. doi:10.1609/icwsm.v14i1.7347.

[4] Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. Understanding the impact of COVID-19 on online mental health forums. *ACM Transactions on Management Information Systems*, 12(4):1–28, September 2021. doi:10.1145/3458770.

[5] Andrew Y. Blei, David M. Ng and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. URL: https://dl.acm.org/doi/pdf/10.5555/944919.944937, doi:10.5555/944919.944937.

[6] Bradley Carron-Arthur, Julia Reynolds, Kylie Bennett, Anthony Bennett, and Kathleen M. Griffiths. What's all the talk about? topic modelling in a mental health internet support group. *BMC Psychiatry*, 16(1), October 2016. doi:10.1186/s12888-016-1073-5.

[7] Dante Chakravorti, Kathleen Law, Jonathan Gemmell, and Daniela Raicu. Detecting and characterizing trends in online mental health discussions. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, November 2018. doi:10.1109/icdmw.2018.00107.

[8] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1), March 2020. doi:10.1038/s41746-020-0233-7.

[9] Qian Hui Chew, Ker Chiah Wei, Shawn Vasoo, Hong Choon Chua, and Kang Sim. Narrative synthesis of psychological and coping responses towards emerging infectious disease outbreaks in the general population: practical considerations for the COVID-19 pandemic. *Singapore Medical Journal*, 61(7):350–356, July 2020. doi:10.11622/smedj.2020046.

[10] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):71–80, May 2014. doi:10.1609/icwsm.v8i1.14526.

[11] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137, August 2021. doi:10.1609/icwsm.v7i1.14432.

[12] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, 2014. doi:10.3115/v1/w14-3207.

[13] Johannes Feldhege, Markus Moessner, and Stephanie Bauer. Who says what? content and participation characteristics in an online depression community. *Journal of Affective Disorders*, 263:521–527, February 2020. doi:10.1016/j.jad.2019.11.007.

[14] Yousra Fettach and Lamia Benhiba. Pro-eating disorders and pro-recovery communities on reddit. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications &amp Services*. ACM, December 2019. doi:10.1145/3366030.3366058.

[15] Oguzhan Gencoglu and Mathias Gruber. Causal modeling of twitter activity during COVID-19. *Computation*, 8(4):85, September 2020. doi:10.3390/computation8040085.

[16] Neville H. Golden, Marcie Schneider, Christine Wood, Stephen Daniels, Steven Abrams, Mark Corkins, Sarah de Ferranti, Sheela N. Magge, Sarah Schwarzenberg, Paula K. Braverman, William Adelman, Elizabeth M. Alderman, Cora C. Breuner, David A. Levine, Arik V. Marcell, Rebecca O'Brien, Stephen Pont, Christopher Bolling, Stephen Cook, Lenna Liu, Robert

Schwartz, Wendelin Slusser, , and and. Preventing obesity and eating disorders in adolescents. *Pediatrics*, 138(3), September 2016. doi:10.1542/peds.2016-1649.

[17] Aric Hagberg, Pieter J. Swart, and Daniel A. Schult. Exploring network structure, dynamics, and function using networkx. URL: https://www.osti.gov/biblio/960616.

[18] Fengyi Hao, Wanqiu Tan, Li Jiang, Ling Zhang, Xinling Zhao, Yiran Zou, Yirong Hu, Xi Luo, Xiaojiang Jiang, Roger S. McIntyre, Bach Tran, Jiaqian Sun, Zhisong Zhang, Roger Ho, Cyrus Ho, and Wilson Tam. Do psychiatric patients experience more psychiatric symptoms during COVID-19 pandemic and lockdown? a case-control study with service and research implications for immunopsychiatry. *Brain, Behavior, and Immunity*, 87:100–106, July 2020. doi:10.1016/j.bbi.2020.04.069.

[19] Emily A Holmes, Rory C O'Connor, V Hugh Perry, Irene Tracey, Simon Wessely, Louise Arseneault, Clive Ballard, Helen Christensen, Roxane Cohen Silver, Ian Everall, Tamsin Ford, Ann John, Thomas Kabir, Kate King, Ira Madan, Susan Michie, Andrew K Przybylski, Roz Shafran, Angela Sweeney, Carol M Worthman, Lucy Yardley, Katherine Cowan, Claire Cope, Matthew Hotopf, and Ed Bullmore. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *The Lancet Psychiatry*, 7(6):547–560, June 2020. doi:10.1016/s2215-0366(20)30168-1.

[20] Bineet Kumar Jha and Shilpa Pande. Time series forecasting model for supermarket sales using FB-prophet. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, April 2021. doi:10.1109/iccmc51019.2021.9418033.

[21] Jeffrey G. Johnson, Patricia Cohen, Stephanie Kasen, and Judith S. Brook. Eating disorders during adolescence and the risk for physical and mental disorders during early adulthood. *Archives of General Psychiatry*, 59(6):545, June 2002. doi:10.1001/archpsyc.59.6.545.

[22] Mohsen Khosravi. The challenges ahead for patients with feeding and eating disorders during the COVID-19 pandemic. *Journal of Eating Disorders*, 8(1), September 2020. doi:10.1186/s40337-020-00322-3.

[23] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext &amp Social Media - HT '15*. ACM Press, 2015. doi:10.1145/2700171.2791026.

[24] Katie Loth, Patricia van den Berg, Marla E. Eisenberg, and Dianne Neumark-Sztainer. Stressful life events and disordered eating behaviors: Findings from project EAT. *Journal of Adolescent Health*, 43(5):514–516, November 2008. doi:10.1016/j.jadohealth.2008.03.007.

[25] Paulo P. P. Machado, Ana Pinto-Bastos, Rita Ramos, Tânia F. Rodrigues, Elsa Louro, Sónia Gonçalves, Isabel Brandão, and Ana Vaz. Impact of COVID-19 lockdown measures on a cohort of eating disorders patients. *Journal of Eating Disorders*, 8(1), November 2020. doi:10.1186/s40337-020-00340-1.

[26] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL: https://aclanthology.org/D11-1024.

[27] Markus Moessner, Johannes Feldhege, Markus Wolf, and Stephanie Bauer. Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, 51(7):656–667, May 2018. doi:10.1002/eat.22878.

[28] Natalie C. Momen, Oleguer Plana-Ripoll, Zeynep Yilmaz, Laura M. Thornton, John J. McGrath, Cynthia M. Bulik, and Liselotte Vogdrup Petersen. Comorbidity between eating disorders and psychiatric disorders. *International Journal of Eating Disorders*, 55(4):505–517, January 2022. doi:10.1002/eat.23687.

[29] Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. 2020. URL: https://arxiv.org/abs/2005.03082, doi:10.48550/ARXIV.2005.03082.

[30] Umashanthi Pavalanathan and Munmun De Choudhury. Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, May 2015. doi:10.1145/2740908.2743049.

[31] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[32] Christophorus Beneditto Aditya Satrio, William Darmawan, Bellatasya Unrica Nadia, and Novita Hanafiah. Time series analysis and forecasting of coronavirus disease in indonesia using ARIMA model and PROPHET. *Procedia Computer Science*, 179:524–532, 2021. doi:10.1016/j.procs.2021.01.036.

[33] Cuihua Shen, Anfan Chen, Chen Luo, Jingwen Zhang, Bo Feng, and Wang Liao. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland china: Observational infoveillance study. *Journal of Medical Internet Research*, 22(5):e19421, May 2020. doi:10.2196/19421.

[34] Ashleigh N. Shields, Elise Taylor, and Jessica R. Welch. Understanding the conversation around COVID-19 and eating disorders: A thematic analysis of reddit. *Journal of Eating Disorders*, 10(1), January 2022. doi:10.1186/s40337-022-00530-z.

[35] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, December 2019. doi:10.1109/bigdata47090.2019.9005997.

[36] Lei Tang and Huan Liu. Graph mining applications to social network analysis. In *Managing and Mining Graph Data*, pages 487–513. Springer US, 2010. doi:10.1007/978-1-4419-6045-0_16.

[37] Jet D. Termorshuizen, Hunna J. Watson, Laura M. Thornton, Stina Borg, Rachael E. Flatt, Casey M. MacDermod, Lauren E. Harper, Eric F. van Furth, Christine M. Peat, and Cynthia M. Bulik. Early impact of scpCOVID/scp-19 on individuals with scpself-reported/scp eating disorders: A survey of ˜1, 000 individuals in the united states and the netherlands. *International Journal of Eating Disorders*, 53(11):1780–1790, July 2020. doi:10.1002/eat.23353.

[38] Julio Torales, Marcelo O'Higgins, João Mauricio Castaldelli-Maia, and Antonio Ventriglio. The outbreak of COVID-19 coronavirus and its impact on global mental health. *International Journal of Social Psychiatry*, 66(4):317–320, March 2020. doi:10.1177/0020764020915212.

[39] Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahi

nuç, and Oguzhan Ozcelik. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21, March 2023. doi:10.1145/3578707.

[40] Janet Treasure, Tiago Antunes Duarte, and Ulrike Schmidt. Eating disorders. *The Lancet*, 395(10227):899–911, March 2020. doi:10.1016/s0140-6736(20)30059-3.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL: https://arxiv.org/abs/1706.03762, doi:10.48550/ARXIV.1706.03762.

[42] L. Vuillier, L. May, M. Greville-Harris, R. Surman, and R. L. Moseley. The impact of the COVID-19 pandemic on individuals with eating disorders: the role of emotion regulation and exploration of online treatment experiences. *Journal of Eating Disorders*, 9(1), January 2021. doi:10.1186/s40337-020-00362-9.

[43] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, February 2017. doi:10.1145/3018661.3018706.

[44] Yilin Wang, Peijing Wu, Xiaoqian Liu, Sijia Li, Tingshao Zhu, and Nan Zhao. Subjective well-being of chinese sina weibo users in residential lockdown during the COVID-19 pandemic: Machine learning analysis. *Journal of Medical Internet Research*, 22(12):e24775, December 2020. doi:10.2196/24775.

[45] Hywel T.P. Williams, James R. McMurray, Tim Kurz, and F. Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, May 2015. doi:10.1016/j.gloenvcha.2015.03.006.

[46] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*. ACM, April 2009. doi:10.1145/1519065.1519089.

[47] Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion. Deep transformer models for time series forecasting: The influenza prevalence case, 2020. URL: https://arxiv.org/abs/2001.08317, doi:10.48550/ARXIV.2001.08317.

[48] Anita Yadav, C K Jha, and Aditi Sharan. Optimizing LSTM for time series prediction in indian stock market. *Procedia Computer Science*, 167:2091–2100, 2020. doi:10.1016/j.procs.2020.03.257.

APPENDIX. A NOTE ABOUT REMOVED OR DELETED POSTS AND COMMENTS

In our text-based analysis, we have chosen to exclude posts and comments where the author or text is labeled as "[removed]" or "[deleted]." However, in our post count and user interaction analysis, we make an effort to include posts and comments even if the text has been deleted, as long as the data is accessible. This decision is informed by our observation that data from Pushshift, both the files and API, is not consistently scraped from Reddit at the same time it was initially posted. This discrepancy in timing could potentially lead to fluctuations in the metrics that don't accurately reflect user activity.

Nonetheless, there are certain metrics for which it's not feasible to consider deleted content, such as adding users with no username to the user graph or counting words in deleted comments. In these cases, we exclude the deleted data from our analysis.