



**UNIVERSITÀ  
DEGLI STUDI  
DI PADOVA**



**DIPARTIMENTO  
DI INGEGNERIA  
DELL'INFORMAZIONE**

**DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE**

**CORSO DI LAUREA MAGISTRALE IN  
ICT FOR INTERNET AND MULTIMEDIA  
'CYBERSYSTEMS'**

**“LINGUISTIC ANALYSIS OF AGENCY IN ONLINE DISCUSSIONS ABOUT  
POSTPARTUM”**

**Relatore: Prof. Tomaso Erseghe  
University of Padova**

**Laureanda: Selen Arslan  
2004968**

**Correlatore: Prof. Marta Witkowska  
SWPS University of Social Sciences and Humanities**

**ANNO ACCADEMICO 2023 – 2024**

THIS THESIS IS DEDICATED TO MY FAMILY AND TO THE PEOPLE WHO HAVE SUPPORTED  
AND ENCOURAGED ME DURING MY EDUCATION.

# Abstract

The research focuses on understanding the relationship between agency and the emotion score of being positive and negative, as well as the similarity score for the term depression within expressions related to postpartum on social media platforms, such as Reddit and Twitter. The research project is executed through a series of detailed and structured steps, including data scraping, preprocessing, and subsequent analysis with a range of tools and methods. The results of the project are presented through various graphs and accompanied by qualitative explanations. Overall, this research sheds light on the significance of utilizing data analytics in social media networks to determine the association between agentic language and emotion scores, providing valuable insights.

# Contents

ABSTRACT	v
LIST OF FIGURES	viii
LIST OF CODES	xi
LISTING OF ACRONYMS	xii
1 INTRODUCTION	1
2 DATA SCRAPING	3
2.1 Twitter	4
2.1.1 Twitter Developer Platform and Twitter API V2	4
2.1.2 Tweepy	5
2.2 Reddit	6
2.2.1 Pushshift	6
3 DATASET	8
3.1 Twitter	9
3.2 Reddit	13
3.2.1 Pre-Matching	15
3.2.2 Post-Matching	16
4 PREPROCESSING	21
4.1 Superficial Cleaning	22
4.2 Subsentence for Reddit Post and Comment	24
4.3 Deep Cleaning	25
5 TOOLS AND METHODS	31
5.1 LIWC	32
5.2 BERTAgent	32
5.3 Grammatical Calculation	33
5.4 BERTopic	34
5.4.1 Topic Model	35
5.4.2 Reducing Outliers	39
5.4.3 Term Similarity Score	39



5.4.4	Creating New Labels . . . . .	40
5.5	Normalization . . . . .	41
5.5.1	Document Based Normalization . . . . .	41
5.5.2	Topic Based Calculation and Normalization . . . . .	42
5.6	Coloring . . . . .	42
5.7	Visualization . . . . .	43
5.7.1	Document Based Visualization . . . . .	43
5.7.2	Topic Based Normalization . . . . .	47
<b>6</b>	<b>APPLICATION OF THE TOOLS AND METHODS</b>	<b>48</b>
6.1	Twitter . . . . .	49
6.2	Reddit Post . . . . .	63
6.3	Reddit Comment . . . . .	74
6.4	Correlation between Calculated Values . . . . .	85
<b>7</b>	<b>CONCLUSION</b>	<b>88</b>
	<b>REFERENCES</b>	<b>89</b>

# Listing of figures

3.1	Fist Version of Twitter Dataset. . . . .	10
3.2	Distribution of Tweets per Month. . . . .	11
3.3	Tweet Contributions from Individual Users . . . . .	12
3.4	Zoomed-In View of Tweet Contributions from Individual Users. . . . .	12
3.5	Postpartum Subreddits . . . . .	13
3.6	Reddit Post Dataset (post-match) . . . . .	17
3.7	Top Subreddits in Post Dataset . . . . .	17
3.8	Distribution of Posts per Month . . . . .	18
3.9	Contribution of Each Unique User to Post Dataset. . . . .	18
3.10	Reddit Comment Dataset (post-match). . . . .	19
3.11	Distribution of Comments per Month (post-match) . . . . .	20
3.12	Unique Users vs. Collected Comments . . . . .	20
4.1	Structure of Preprocessing Steps . . . . .	22
4.2	Word Count Comparison for Reddit Post, Comment and Twitter. . . . .	24
4.3	Relationship between Preprocessing Steps and Analysis . . . . .	28
4.4	Twitter dataset after Preprocessing. . . . .	29
4.5	Reddit Post Dataset after Preprocessing. . . . .	30
4.6	Reddit Comment Dataset after Preprocessing. . . . .	30
5.1	Document and Topic Based Analyses . . . . .	41
5.2	Color Scale . . . . .	43
6.1	Original Topic Labels vs. Outliers Reduced Topic Labels for Twitter . . . . .	50
6.2	Term Similarity Score for 'Depression' for Twitter . . . . .	51
6.3	Distribution of Term Similarity Score for 'Depression' for Twitter . . . . .	51
6.4	Normalized and Colored of Term Similarity Score for 'Depression' Distribu- tion for Twitter. . . . .	52
6.5	Distribution of Predicted Whole Mean for Twitter. . . . .	53
6.6	Normalized and Colored of Predicted Whole Mean Distribution for Twitter. . . . .	53
6.7	Twitter Document Visualization . . . . .	54
6.8	Twitter Document Visualization for Depression Topics. . . . .	54
6.9	Twitter Document Visualization of Term Similarity Score for 'Depression'. . . . .	55
6.10	Document Visualization of BERTagent for Twitter. . . . .	56
6.11	Label Evolution for Twitter . . . . .	57

6.12	Topic Based Negative Emotions for Twitter . . . . .	58
6.13	Topic Based Composite Emotion Scores for Twitter . . . . .	58
6.14	Topic Based Positive Emotions for Twitter . . . . .	59
6.15	Topic Based BERTAgent for Twitter . . . . .	59
6.16	Topic Based Term Similarity Score for 'Depression' for Twitter . . . . .	60
6.17	Topic Based Verbs for Twitter . . . . .	60
6.18	Original Topic Labels vs. Outliers Reduced Topic Labels for Reddit Post . . .	63
6.19	Term Similarity Score for 'Depression' for Reddit Post . . . . .	64
6.20	Distribution of Term Similarity Score for 'Depression' for Reddit Post . . . .	64
6.21	Normalized and Colored of Term Similarity Score for 'Depression' Distribu- tion for Reddit Post . . . . .	65
6.22	Distribution of Predicted Whole Mean for Reddit Post . . . . .	66
6.23	Normalized and Colored of Predicted Whole Mean Distribution for Reddit Post . . . . .	66
6.24	Reddit Post Document Visualization. . . . .	67
6.25	Reddit Post Document Visualization of Term Similarity Score for 'Depression'. .	67
6.26	Document Visualization of BERTAgent for Reddit Post . . . . .	68
6.27	Label Evolution for Reddit Post . . . . .	69
6.28	Topic Based Negative Emotions for Reddit Post . . . . .	71
6.29	Topic Based Composite Emotion Score for Reddit Post . . . . .	71
6.30	Topic Based Positive Emotion for Reddit Post . . . . .	72
6.31	Topic Based BERTAgent for Reddit Post . . . . .	72
6.32	Topic Based Term Similarity Score for 'Depression' for Reddit Post . . . . .	73
6.33	Topic Based Verbs for Reddit Post . . . . .	73
6.34	Original Topic Labels vs. Outliers Reduced Topic Labels for Reddit Comment. .	74
6.35	Term Similarity Score for 'Depression' for Reddit Comment . . . . .	75
6.36	Distribution of Term Similarity Score for 'Depression' for Reddit Comment . .	75
6.37	Normalized and Colored of Term Similarity Score for 'Depression' Distribu- tion for Reddit Comment. . . . .	76
6.38	Distribution of Predicted Whole Mean for Reddit Comment. . . . .	76
6.39	Normalized and Colored of Predicted Whole Mean Distribution for Reddit Comment. . . . .	77
6.40	Reddit Comment Document Visualization. . . . .	77
6.41	Reddit Comment Document Visualization of Term Similarity Score for 'De- pression'. . . . .	78
6.42	Document Visualization of BERTAgent for Reddit Comment. . . . .	79
6.43	Label Evolution for Reddit Comment. . . . .	80
6.44	Topic Based Negative Emotions for Reddit Comment . . . . .	82
6.45	Topic Based Composite Emotion Score for Reddit Comment . . . . .	82
6.46	Topic Based Positive Emotion for Reddit Comment . . . . .	83

6.47 Topic Based BERTAgent for Reddit Comment . . . . . 83  
6.48 Topic Based Term Similarity Score for 'Depression' for Reddit Comment . . 84  
6.49 Topic Based Verbs for Reddit Comment . . . . . 84  
6.50 Correlation Graphs for Twitter. . . . . 86  
6.51 Correlation Graphs for Reddit Post. . . . . 86  
6.52 Correlation Graphs for Reddit Comment. . . . . 87

# Listings

5.1	Embedding Model . . . . .	35
5.2	Default UMAP Model . . . . .	36
5.3	Default HDBSCAN Model . . . . .	37
5.4	c-TF-IDF Model . . . . .	38
5.5	Default BERTopic Model . . . . .	39
5.6	Extraction of the Location Information from Modified Version of Visualize Documents . . . . .	44
5.7	Modified Version of Visualize Documents Graph . . . . .	44

# Listing of acronyms

<b>API</b> .....	Application Programming Interface
<b>OAuth</b> .....	Open Authorization
<b>HTTP</b> .....	Hypertext Transfer Protocol
<b>ID</b> .....	Identity Document
<b>POS</b> .....	Part of Speech
<b>NLTK</b> .....	Natural Language Toolkit
<b>NLP</b> .....	Natural Language Processing
<b>LICW</b> .....	Linguistic Inquiry and Word Count
<b>TF-IDF</b> .....	Term Frequency-Inverse Document Frequency
<b>UMAP</b> .....	Uniform Manifold Approximation and Projection
<b>HDBSCAN</b> ....	Hierarchical Density-Based Spatial Clustering of Applications with Noise

# 1

## Introduction

Social media refers to “web-based services that allow individuals, communities and organizations to collaborate, connect, interact, and build a community by enabling them to create, co-create, modify, share, and engage with user-generated content that is easily accessible” [1]. Numerous social networks possess a vast amount of content and linkage data, which can be effectively utilized for analysis. The linkage data primarily comprises the network’s graph structure and the intercommunication between various entities, while the content data entails multimedia data such as images and text present within the network. The vastness and complexity of such networks provide remarkable prospects for data analytics pertaining to social networks [2]. In this research, it was focused particularly on data from social media platforms such as Twitter and Reddit, which support textual content and metadata for the user and content.

The focus of this research is understanding the level of agency with the relation of emotion score of being positive and negative and the similarity score for the term depression in the expressions about postpartum using the social media platforms Reddit and Twitter as data sources. The concept of agency is integral to human cognition and behavior as it pertains to goal-directed actions and is significant for both interpersonal and intrapersonal processes, as well as intergroup and intragroup relations. Therefore, the identification and measurement of agentic language within large textual datasets play a crucial role in analyzing human actions, interactions, and the resulting social dynamics [3]. The primary objective of the research was to investigate negative emotions, particularly depression, and to delve deeper into aspect of agen-

tic language. The reason for selecting postpartum as the topic of study was due to the abundance of data available on social media platforms. Furthermore, the topic exhibits a wide array of subtopics, encompassing both positive and negative aspects, such as newborns, families, happiness, depression, and anxiety.

The research was executed through a detailed and comprehensive planning process, encompassing a series of planned and structured steps. The applied steps thought the research begin by scraping data and creating datasets from Reddit and Twitter, and explained the properties of datasets such as distribution of collected data per month, individual user distribution. Then, a thorough discussion was presented on the preprocessing procedures, outlining the interdependence of outcome of the preprocessing steps with the input of the subsequent tools and methods utilized for analysis. A comprehensive elucidation of all the utilized tools and methods was provided, followed by the depiction of the project's results in the form of graphs, accompanied by qualitative explanations.



# 2

## Data Scraping

The growth of social media platforms, such as Twitter and Reddit, has resulted in a massive increase in the volume of user-generated content, including text, images, and videos [4]. This wealth of data presents a valuable opportunity for use in various fields, particularly in the field of language analysis serve a purpose of specific goals, sentiment analysis, competitive intelligence, trend analysis, topic modelling, text classification [5]. To scrape data from these social discussion platforms data scrapping techniques must be employed.

This chapter aimed to explore the processes involved in data scraping from two popular social media platforms, Twitter and Reddit. Understanding the structures of these platforms, including the forms of posts and comments, is essential in effectively accessing and extracting relevant data. This chapter provides an in-depth examination of the techniques and methods used in data scraping from these platforms, with a focus on the use of the 'tweepy' library for Twitter data scraping and the 'praw' library for Reddit data scraping by Python.

## 2.1 TWITTER

Twitter, being a microblogging platform, enables users to publish short messages, referred to as tweets. Microblogging refers to the sharing of brief messages or updates on social media platforms, with the goal of increasing audience engagement [6]. Each tweet can contain a maximum of 280 characters and can be accompanied by multimedia content, such as images and videos [7]. Additionally, tweets can be retweeted, liked, and replied to by other users, creating a chains of interactions [8].

The hashtag '#' mechanism in Twitter serves as an organizational tool for categorizing tweets, thus enhancing their discoverability through searches. Users utilize hashtags to categorize their tweets based on specific topics. This allows others to follow and participate in conversations related to that particular subject [9]. Additionally, hashtags can act as keywords that can amplify the popularity of tweets by making them trend or easily recognizable for users. The structure of tweets, along with the abundance of information generated on Twitter, make it a valuable source for data scraping.

### 2.1.1 TWITTER DEVELOPER PLATFORM AND TWITTER API V2

The Twitter Developer Platform is a comprehensive suite of tools and resources designed to facilitate the integration of applications with Twitter data [10]. The platform provides access to the Twitter API, which enables developers to programmatically retrieve and manipulate data on the platform, including tweets, user information, and media entities.

The Twitter API provides a wide range of functionalities that allow developers to retrieve tweets based on specific parameters, such as keywords, user information, and date ranges. Additionally, the API provides access to information about specific users, including their tweets, followers, and mentions [10].

The Twitter Developer Platform has undergone continuous development and improvement and provides customizable products that can be tailored to meet the specific needs of businesses and significant research initiatives [11]. Twitter's most recent API for Academic Research is Twitter API V2 and enables developers to gather both current and past publicly available data from Twitter, along with enhanced features and capabilities that facilitate the collection of more accurate, comprehensive, and impartial datasets [12].

### 2.1.2 TWEETPY

The data scraping process on Twitter is performed utilizing the Python library, Tweepy. Tweepy is an open source package that offers a comprehensive and easy-to-use interface for accessing the Twitter API. This library provides a range of methods for retrieving tweets, user information, and other data, making the data scraping process efficient and streamlined. Additionally, Tweepy offers the ability to extract tweets based on specific keywords, user information, and time frames, which allows for refined and targeted data collection.

One of the advantages of using Tweepy is that it transparently handles various implementation details related to the Twitter API, such as data encoding and decoding, HTTP requests, results pagination, OAuth authentication, rate limits, and streams. By utilizing Tweepy, developers are able to focus on the functionality they want to build and avoid spending time dealing with low-level details related to accessing the Twitter API. Overall, Tweepy is a valuable tool for data scraping and analysis on Twitter, providing a comprehensive interface for accessing the Twitter API and handling various implementation details efficiently and effectively.

In this project, the Tweepy library is utilized to gather original tweets that are relevant to the topic of "postpartum" or its hashtag representation '#postpartum'. Original tweets refer to tweets that are created and posted by users, as opposed to retweets which are essentially reposts of another user's tweet [13]. The data collection process is done in a systematic manner to ensure that a well-distributed dataset is obtained. In particular, the tweets are restricted to those that were posted in the year 2021 with a goal of collecting 50 original tweets per day. The language of all collected original tweets is English and there are no restrictions on their location, as original tweets are gathered from all around the world. This approach allows the project to focus specifically on the topic of "postpartum" within a defined time period, thus enabling the research to gain insight into the relevant discussions surrounding the topic.

In conclusion, the use of the Tweepy and the collection of original tweets in English from all around the world, with a focus on the year 2021 and a goal of 50 tweets per day, ensure that the dataset is comprehensive and representative for the global discussions about topic of postpartum.

## 2.2 REDDIT

Reddit is a large online discussion board that offers a unique and comprehensive platform for users to share, comment, and interact with each other over a wide range of topics.

The platform is organized into subreddits, which are specialized forums focused on specific topics. Each subreddit is created and moderated by members of the Reddit community, and users can subscribe to them to view content that is relevant to their interests [14]. The central component of the Reddit platform is the post. Posts can be submitted in the form of links or text and are assigned to a subreddit based on the topic of discussion. Posts can receive feedback in the form of comments and upvotes or downvotes from other users, leading to a ranking system that prioritizes the most popular and relevant content [15]. This structure creates a nested hierarchy of comments, allowing for more in-depth discussions and interactions between users.

Reddit's unique structure and large volume of data generated make it a valuable source for data scraping and scientific study. Despite the benefits of the structure and Reddit's millions of subreddits, hundreds of millions of users, and billions of comments the technical barriers to data acquisition still exist, making it time-consuming to collect and analyze data systematically [16].

In conclusion, Reddit is a well-structured platform that offers a wide range of topics and opportunities for discussion, feedback, and interaction between users [17]. The ability to subscribe to subreddits, the customizable user experience, and the ranking system based on upvotes and downvotes make Reddit a valuable resource for scientific study and data collection.

### 2.2.1 PUSHSHIFT

Reddit provides an API in Reddit Developer Platform, known as the Reddit API, for extracting information from the platform. However, the Reddit API has some limitations. It only allows users to retrieve a limited amount of recent comments or submissions from a few different streams for a subreddit, such as hot, new, top, etc. Due to these limitations, researchers may find the Reddit API insufficient for their needs, particularly when it comes to creating large datasets [18].

Pushshift is a platform that has been gathering, storing, and providing access to Reddit data for researchers since 2015. It is primarily used for analyzing and archiving social media information. The Pushshift Reddit dataset includes real-time updates and historical data dating back to Reddit's inception, making it possible for social media researchers to reduce time spent on

data collection and cleaning [19].

The Pushshift API offers a number of advantages over the Reddit API. Pushshift makes it easier for researchers to ingest large amounts of data. Additionally, the Pushshift API provides full-text search functionality against comments and submissions, as well as aggregation endpoints for summary analysis of Reddit activity. These extended capabilities make the Pushshift API a more attractive option for researchers looking to study Reddit data. In conclusion, the Pushshift Reddit dataset and API offer a valuable resource for social media researchers looking to study Reddit data. By leveraging the capabilities of the Pushshift API through the praw library, researchers can collect, store, and analyze large amounts of Reddit data with ease.

In this project, Pushshift API is utilized through the praw library, also known as Python Reddit API Wrapper. The praw library provides a comprehensive interface for accessing Reddit data, allowing for efficient data scraping for specific subreddits, keywords, and time periods. The posts containing the keyword "postpartum" are gathered, followed by the comments that also contain "postpartum". Both categories are gathered for year 2021 and language is selected as English. There is not a restriction for geographic aspect in search algorithm. The posts in Reddit are identified by post IDs and the comments are identified by parent IDs, which indicate the related post. The post IDs and parent IDs from both datasets are matched, allowing each comment to be associated with its parent post in the final dataset.

# 3

## Dataset

In the field of data analysis, the quality and reliability of the data being used can have a profound effect on the accuracy and validity of the results that are obtained from the analysis. This is why it is of utmost importance to ensure that the dataset used is trustworthy and properly prepared.

The process of collecting data involves consideration of the source and the methods used to collect it. In this study, two different data was gathered from both Twitter and Reddit platforms by using the methods explained in previous chapter. The following step in preparing the dataset for the further analysis involves the filtration and selection of the most relevant columns, in order to retain only the information that is essential to the research question being investigated and to eliminate any extraneous or duplicated data. This not only improves the efficiency of the analysis but also reduces the possibility of errors.

Cleaning and preprocessing the dataset is also a crucial step [20] . This entails the removal of any missing values, outliers, or inconsistent data. This is critical in ensuring that the results obtained from the analysis are accurate and free of errors. The significance of this step cannot be overstated and has a significant impact on the validity of the results.

To summarize, the preparation of a high-quality and reliable dataset is an essential aspect of the data analysis process. The methods used for collecting data, filtering and selecting relevant columns and preprocessing the data must be carefully planned and executed in order to achieve accurate results. This chapter will delve deeper into the properties of the scraped data from the Reddit and Twitter.

### 3.1 TWITTER

The Twitter dataset was collected using the Tweepy library in Python, with the request algorithm designed to gather original tweets containing the keyword "postpartum" or its hashtag representation "#postpartum". The collected tweets were restricted to those written in English and came from locations worldwide, in order to achieve a global representation and focus on tweets written in English. To ensure a well-distributed sample, the number of tweets collected per day was limited to 50 for the entire year of 2021, from January 1st to December 31st.

Initially, the project collected a large number of tweets for the year 2021 without a daily limit, however, it was observed that the Tweepy library tended to prioritize collecting tweets from the beginning of the period, leading to a random distribution of tweets from different days. To address this issue, the decision was made to send a request for each day of the year 2021, with a maximum limit of 50 tweets per day. This approach ensured a more uniform distribution of tweets and reduced the risk of biased results.

In the data gathering process, the Twitter API was utilized through the use of the Tweepy library in Python. The library provides the capability to select the type of information to be collected for each tweet, such as user information, media, etc. To obtain a comprehensive understanding of the collected tweets, the following information was selected for inclusion in the data: author ID, tweet ID, text, creation date of the tweet, and language. The selected information author ID, tweet ID, text, creation date, and language were stored in corresponding columns were named 'author\_id,' 'tweet\_id,' 'text,' 'created\_at,' and 'lang' respectively, to clearly indicate the type of information contained in each column.

- **author\_id** : The unique identifier for the user who posted the tweet [21]. This information is crucial in understanding the distribution of the collected tweets, as it allows for analysis of the diversity of the sample. The aim is to have reliable and well-distributed data, with representation from a diverse range of individuals and not just a single user or group of users.
- **tweet\_id** : The unique identifier for the tweet itself [22], which is important for referencing the tweet in the future and tracking its engagement and spread on the platform.
- **text** : The content of the tweet [22], which is the primary and most important information in the project. This column is the focus of the data preprocessing steps and subjected to various steps in order to prepare it for analysis. As a result, this column is critical to the project's objectives and serves as the foundation for the subsequent analysis.

- **created\_at** : The date and time the tweet was posted [22] , which allows for analysis of the time distribution of tweets.
- **lang** : The language of the tweet [22] , which is critical in ensuring that only tweets written in English are included in the study and aligns with the objective of the project to study English tweets worldwide.

The selection of these columns is crucial in achieving the objective of the project to study English tweets worldwide and provides valuable information for analyzing the discourse and sentiments surrounding postpartum experiences.

As a result of the data collection process, a total of 17,664 tweets with the mentioned information were collected.

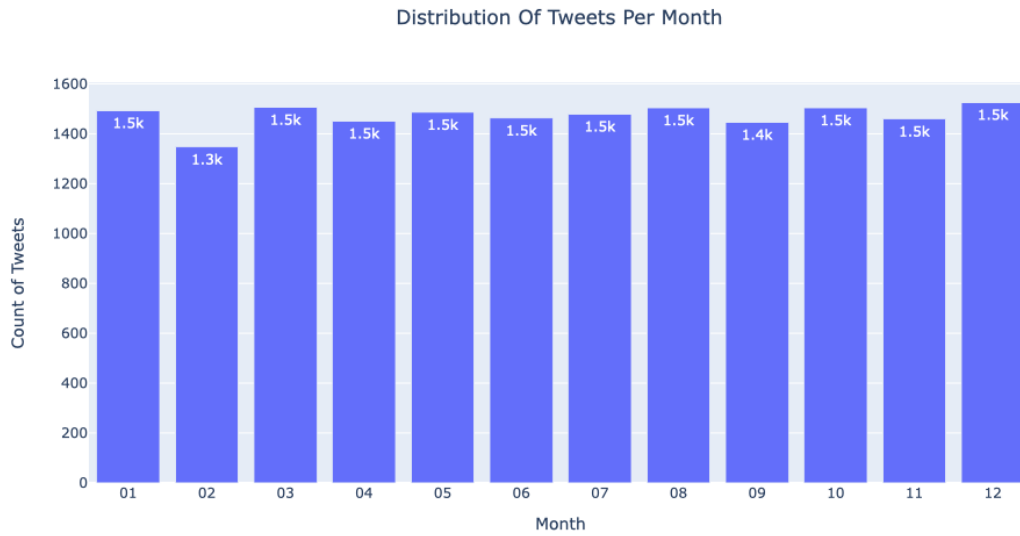
	author_id	tweet_id	text	created_at	lang
0	2861611181	1345157499461246977	Postpartum is not a joke .	2021-01-01 23:58:34+00:00	en
1	178550157	1345156941102919682	@NawfKage postpartum depression? i didnt know ...	2021-01-01 23:56:21+00:00	en
2	1090924279	1345156612185608192	Happy for them. I will say i hope he's better ...	2021-01-01 23:55:02+00:00	en
3	52512334	1345156419528613889	nfs I would hate to be Ari seein this cause ni...	2021-01-01 23:54:16+00:00	en
4	1337201444362063877	1345155775052849153	@Jflexo_ That and postpartum depression 🙄	2021-01-01 23:51:43+00:00	en
...	...	...	...	...	...
17659	935755724	1477022688472305665	@pulte 36wks preggers here and would love to b...	2021-12-31 21:03:45+00:00	en
17660	2603677855	1477021963197591557	@mamasflwrchild I got zero fertility training....	2021-12-31 21:00:52+00:00	en
17661	15255546	1477021747740303365	Curious about your period after giving birth? ...	2021-12-31 21:00:01+00:00	en
17662	3949469128	1477020281919221768	never knew i would be going through postpartum...	2021-12-31 20:54:11+00:00	en
17663	106747111	1477020238713700358	Join us the 2nd and 4th Wednesday of the month...	2021-12-31 20:54:01+00:00	en

17664 rows x 5 columns

Figure 3.1: First Version of Twitter Dataset.

The choice to collect a maximum of 50 tweets per day throughout the year 2021 ensures a well-distributed dataset. As demonstrated in Figure 3.2, the total number of collected tweets per month exhibits a balanced distribution. This can be considered as a reflection of the methodology employed in acquiring the data.





**Figure 3.2:** Distribution of Tweets per Month.

The analysis of the number of users in the dataset was carried out to assess the diversity of the dataset. The data consisted of 17,664 tweets originating from 13,799 unique users, as shown in Figure 3.3. The long-tail distribution of the data demonstrates that a large proportion of the users have only been represented by one tweet in the dataset, while a smaller group of users have contributed a greater number of tweets. Figure 3.4 shows that only 52 users, or 0.37% of the total user base, have contributed more than 10 tweets. The total number of tweets from a single user is relatively small compared to the overall number of tweets in the dataset. This indicates that the contributions are widely distributed across a large number of users, demonstrating good diversity in the dataset.

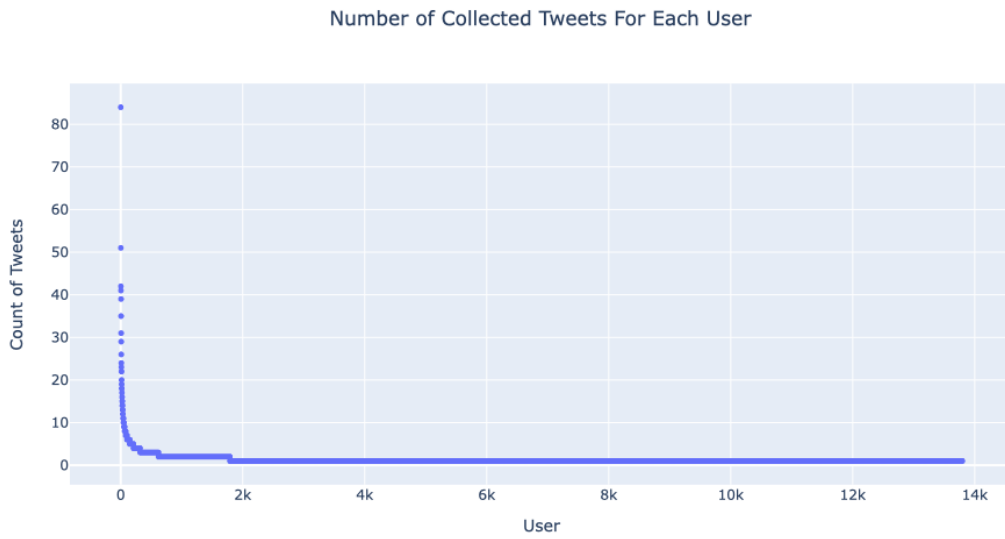


Figure 3.3: Tweet Contributions from Individual Users.

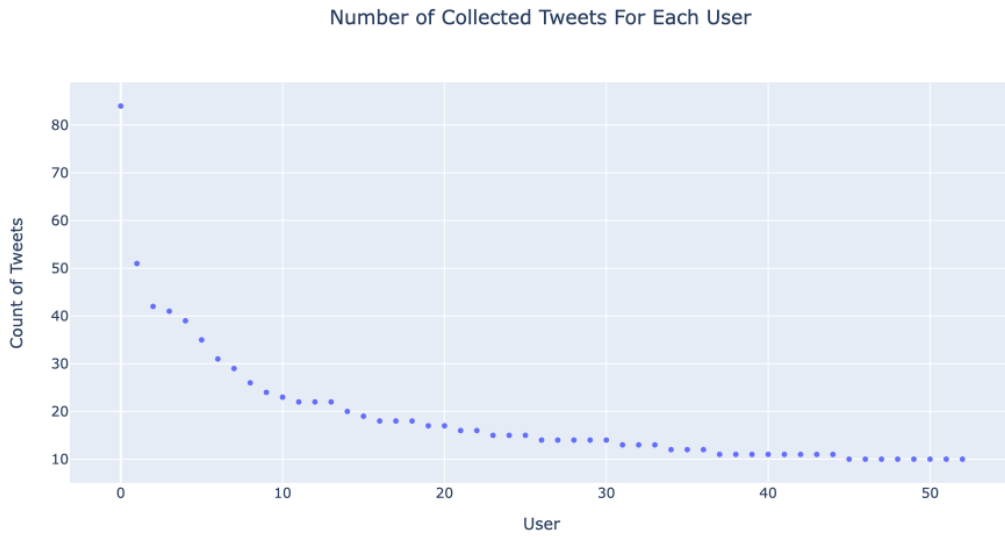
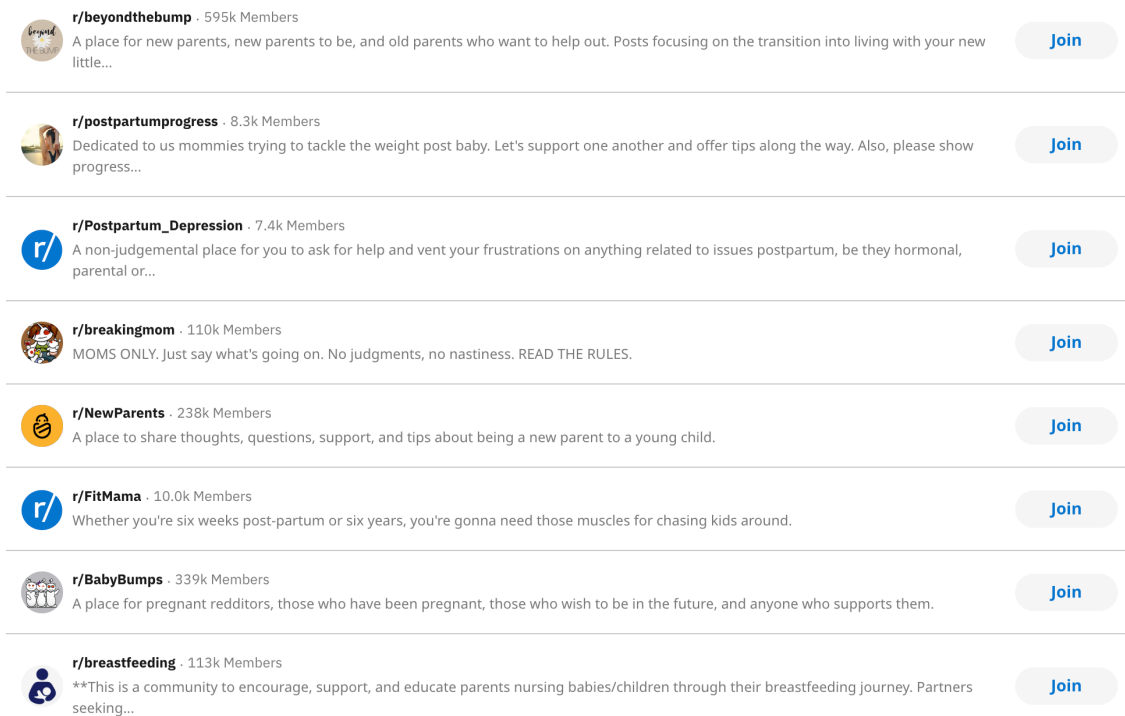


Figure 3.4: Zoomed-In View of Tweet Contributions from Individual Users.

## 3.2 REDDIT

The Pushshift library provides the ability to collect Reddit posts and comments that contain specific keywords or from specific subreddits for a specific period. The data collection process was performed separately for posts and comments from Reddit, in order to ensure a comprehensive analysis of the topic.

There are numerous Reddit communities, also known as subreddits, dedicated to postpartum topics as shown in the Figure 3.5. The term "postpartum" encompasses a range of experiences related to physical changes, emotional well-being, and parenting following childbirth. Given the wide-ranging nature of this topic, it can be difficult to determine an appropriate subreddit for data collection, as many communities on Reddit about postpartum focus on specific subtopics, such as weight loss or depression after childbirth. Furthermore, the high number of members in a subreddit does not necessarily indicate that the discussion is centered on postpartum in a general sense.











	<b>r/beyondthebump</b> · 595k Members A place for new parents, new parents to be, and old parents who want to help out. Posts focusing on the transition into living with your new little...	<a href="#">Join</a>
	<b>r/postpartumprogress</b> · 8.3k Members Dedicated to us mummies trying to tackle the weight post baby. Let's support one another and offer tips along the way. Also, please show progress...	<a href="#">Join</a>
	<b>r/Postpartum_Depression</b> · 7.4k Members A non-judgemental place for you to ask for help and vent your frustrations on anything related to issues postpartum, be they hormonal, parental or...	<a href="#">Join</a>
	<b>r/breakingmom</b> · 110k Members MOMS ONLY. Just say what's going on. No judgments, no nastiness. READ THE RULES.	<a href="#">Join</a>
	<b>r/NewParents</b> · 238k Members A place to share thoughts, questions, support, and tips about being a new parent to a young child.	<a href="#">Join</a>
	<b>r/FitMama</b> · 10.0k Members Whether you're six weeks post-partum or six years, you're gonna need those muscles for chasing kids around.	<a href="#">Join</a>
	<b>r/BabyBumps</b> · 339k Members A place for pregnant redditors, those who have been pregnant, those who wish to be in the future, and anyone who supports them.	<a href="#">Join</a>
	<b>r/breastfeeding</b> · 113k Members **This is a community to encourage, support, and educate parents nursing babies/children through their breastfeeding journey. Partners seeking...	<a href="#">Join</a>

Figure 3.5: Postpartum Subreddits.

In order to address these challenges, it was decided to collect data using the keyword "postpartum" from various locations on Reddit, rather than limiting the data collection to a specific subreddit. This approach provides a greater level of diversity in the data and enhances the reliability of the dataset. In the initial stages of the project, data was collected from two of the most populated subreddits related to postpartum and one subreddit specifically about postpartum depression. However, it was observed that collecting data from these specific subreddits did not result in a well-distributed dataset that captured the various subtopics associated with postpartum. Therefore, it was determined that collecting data using the keyword "postpartum" from various locations on Reddit would provide a more diverse and representative dataset. The objective of this project is to gain a general understanding of the shared ideas about postpartum on Reddit, taking into account the positive and negative aspects of the experiences. The use of the keyword "postpartum" rather than a specific subreddit provides a more varied and trustworthy dataset, and allows for a more representative analysis of the topic.

In this research, the evolution of dataset for Reddit can be broadly categorized into two stages: pre-matching and post-matching. The pre-matching stage involves the collection of large amounts of data, including both posts and comments, from Reddit. This data includes meta information, such as an identifier (ID) for the post in post dataset and a parent identifier (parent ID) for the related parent post of a comment in comment dataset.

The post-matching stage involves the application of a matching algorithm to the collected data. This algorithm matches the post and comment data based on the parent ID and ID, in order to identify and retain only the related information and eliminate any unrelated data. The purpose of this is to analyze the topics of discussion in the posts and the responses in the form of comments to these posts. This approach is motivated by the aim to obtain a clearer understanding of the discourse on the platform and to analyze the interactions between users. By retaining only the related post and comment data, the dataset becomes more focused and relevant to the research question at hand, thereby enhancing the validity and reliability of the results.

### 3.2.1 PRE-MATCHING

#### POST

A total of 16815 posts containing the keyword 'postpartum' that were shared on Reddit between January 1st and December 31st, 2021 were collected. During the quality control process, duplicated posts were identified and removed from the dataset. The duplicated posts were found to be original posts with the same text that were shared by the same users under different subreddits. This phenomenon may have occurred because users wanted to reach a wider audience or receive quicker responses from various communities. However, duplication is not desirable in data analysis, and therefore duplicated data were removed and a randomly selected single post was kept. After removing duplicated posts, the dataset consisted of 11849 unique posts. This dataset served as the base for the mentioned matching algorithm.

#### COMMENT

Total of 70,000 comments containing the keyword 'postpartum' shared in 2021 were collected from the randomly selected subreddits on Reddit platform using the Pushshift library. The large number of collected comments indicates a higher probability of having a greater number of comments associated with the collected posts. The collection of comments in this study was performed randomly based on specific parameters, such as keyword and year, rather than collecting the comments associated with a particular post dataset by adding the post ID to the search algorithm as a parameter. This approach was taken as the latter method would have taken an excessive amount of time, due to the Pushshift library's requirement to search for related comments for each post individually. Given the large size of the post dataset, the decision was made to collect a large number of comments to create a big dataset for comments, which was then utilized for the matching algorithm.

Comment dataset, created with several pieces of information, including the author's username, a unique identifier for the comment (id), the text of the comment, the identifier for the parent object to which the comment is associated (parent\_id), name of the subreddit (subreddit) and creation time of the comment. The parent object can be one of three different types in Reddit, as identified by the "t1", "t2", or "t3" prefix following by the identification number in the parent ID field. These prefixes, as defined by the Reddit API, refer to "Comment" objects, "Account" objects, and "Link" objects, respectively [23]. The first three letters (tx\_) of the parent ID field are removed in further analysis to facilitate the matching process.

### 3.2.2 POST-MATCHING

The process of matching the post id with the parent id of a comment is referred to as the Matching Algorithm in this research. The objective of this algorithm is to link two datasets, namely post and comment, to gather strongly related data. It is important to note that the pre-match versions of the datasets are not secure as the data collected is only a small portion of the vast information source of Reddit, which has a wide range of subtopics. The possibility of the datasets representing different topics, despite having same keyword in their text elements, highlights the need for the Matching Algorithm to create a comprehensive picture of the post-partum topic. Following the implementation of the Matching Algorithm, only the paired data elements are retained. As a result, each post in the dataset is guaranteed to have at least one associated comment.

#### POST

The final dataset, after removing irrelevant posts, consists of 3040 posts shared by 2666 distinct users across 429 subreddits.

The dataset presented in Figure 3.6 pertains to social media activity, containing information about the users and the text they shared on Reddit. It comprises of six columns, which provide specific details about the user and the shared post. The first column, 'author', specifies the name of the user who created the post. The second column, 'id', is a unique identifier assigned to each post, which is essential for tracking its engagement and spread on the platform. The third column, 'body', denotes the shared post. The fourth column, 'subreddit\_id', denotes the id of the subreddit associated with the post while column subreddit denotes the name of the subreddit. Finally, the sixth column, 'created\_at', contains the information of creation time of the post.

	author	id	body	subreddit_id	subreddit	created_at
0	RosyEmmaG	lbms6k	All my homies who have experienced hair change...	t5_2t79l	curlyhair	2021-02-03 13:46:58
1	aspikyplant	lbd2ca	Hi fellow fit pregnant friends! \n\nI'm not re...	t5_372t2	fitpregnancy	2021-02-03 03:15:18
2	estaarr	lbaow7	Hello, I am a FTM and two weeks pp. My body st...	t5_2snqi	NewParents	2021-02-03 01:13:46
3	FrancesRW	lb6z7c	I have bladder prolapse and DR (4-finger width...	t5_2u06v	beyondthebump	2021-02-02 22:29:14
4	maryjaneexperience	lb33vg	Hey parents of actual children and babies! I'm...	t5_2snqi	NewParents	2021-02-02 19:46:30
...	...	...	...	...	...	...
3035	marijuanamama_	pa4wbk	Tips/Recommendations \n\nI'm 35 weeks so almos...	t5_2u06v	beyondthebump	2021-08-23 19:44:33
3036	marcal213	pa4w1x	A little backstory... I've been jumping for qu...	t5_2r4or	Equestrian	2021-08-23 19:44:11
3037	Jennnc213	p9zxxe	I (23f) currently 22 weeks pregnant with my fi...	t5_2xhvq	AmItheAsshole	2021-08-23 15:34:49
3038	lapetiteloup	p9ynyr	Obviously, periods are never really pleasant, ...	t5_2u06v	beyondthebump	2021-08-23 14:19:29
3039	SignificantMeaning24	p9xb9z	Y'all should stop shaming her for saying crate...	t5_4jckkp	ashhventure	2021-08-23 12:49:48

3040 rows × 6 columns

Figure 3.6: Reddit Post Dataset (post-match).

The top 10 subreddits in the dataset, with the highest number of posts, are considered the most crowded and popular subreddits related to postpartum and parenting on Reddit by checking the number of members of the communities in Reddit and represented in Figure 3.7. The size of the subreddits were represented written on the bars. The connection between Reddit post and comment datasets allow for the same inferences to be drawn from the comment data.

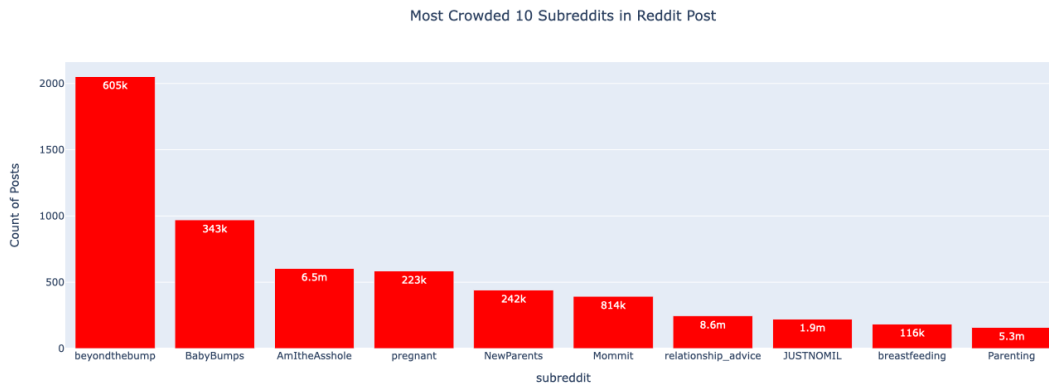


Figure 3.7: Top Subreddits in Post Dataset.

Furthermore, it is crucial to examine the monthly distribution of the collected data for the year 2021. The monthly distribution, as discussed in previous chapters, is an important metric for determining the balance and representation of the data. As shown in Figure 3.8, the distri-

bution of collected posts is well-balanced, providing a diverse representation of users, subreddits, and periods.

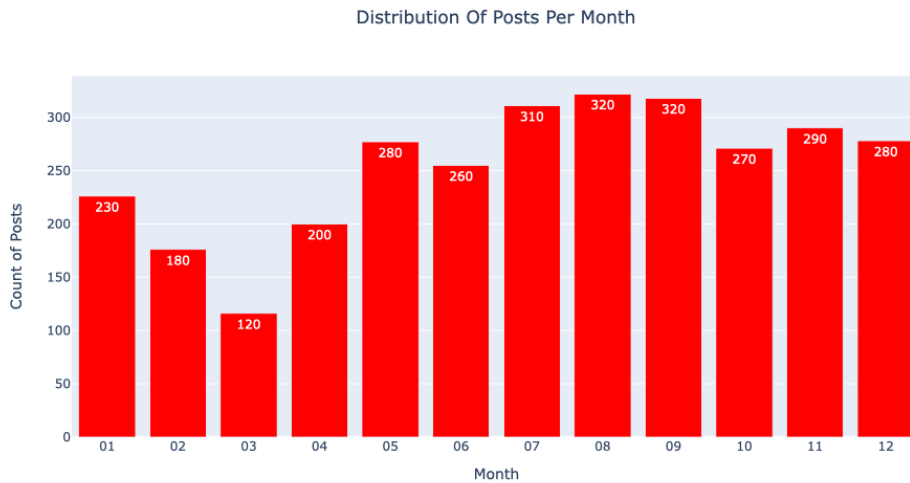
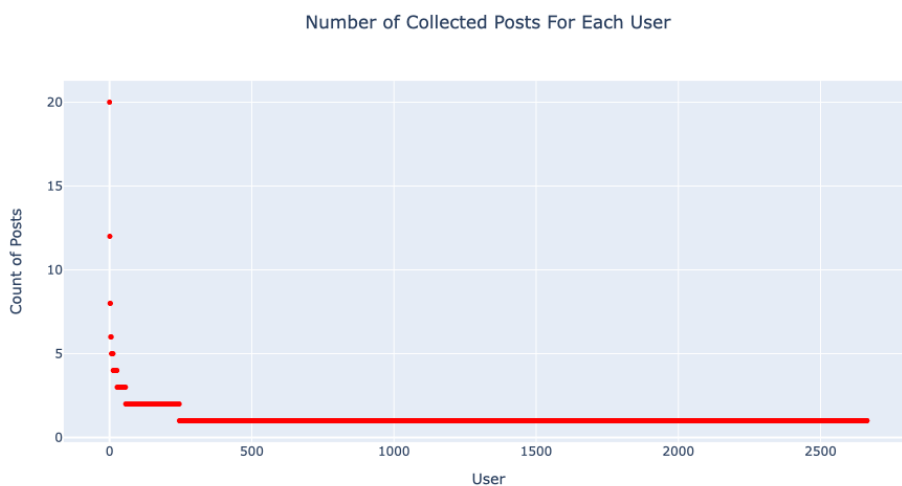


Figure 3.8: Distribution of Posts per Month.

The graph presented in Figure 3.9 illustrates that the contribution of unique users to the dataset follows a long tail distribution, indicating that the majority of users only had one post included in the dataset. Only a small group of users had multiple posts included.





## COMMENT

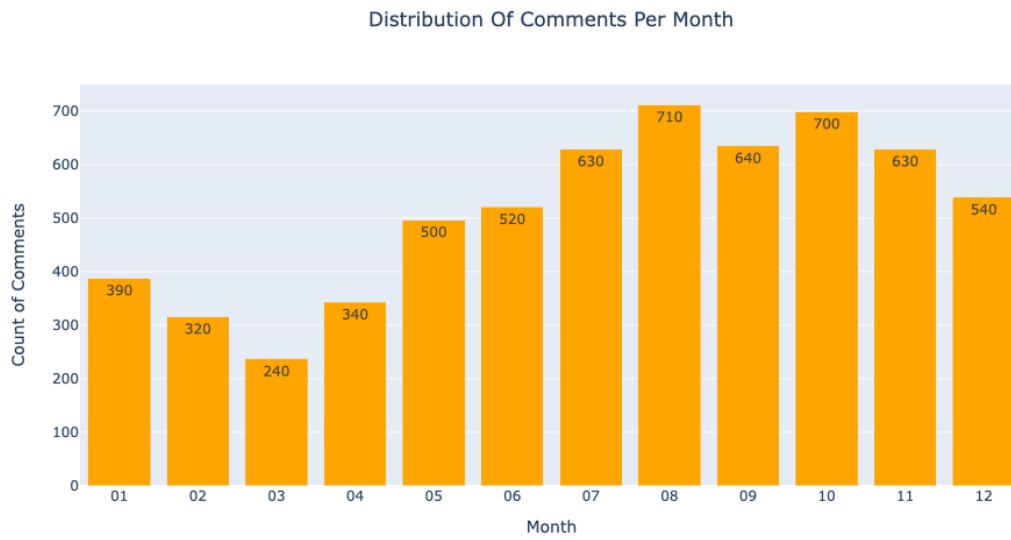
The final comment dataset consists of 6145 comment objects contains the parent post id and presented in Figure 3.10.

	author	id	body	parent_id	subreddit	created_at
0	Serafirelly	hclfqk0	We are currently trying for number two and hav...	pmw6f4	tryingforanother	2021-09-12 20:47:32
1	Crazystafflady	hcpivzx	I have a 15 month age gap between my two.\n\nF...	pmw6f4	tryingforanother	2021-09-13 18:38:04
2	plantkiller2	hclbxq9	We did the natural family planning method unti...	pmhmme	oneanddone	2021-09-12 20:23:13
3	quodestveritas	hcj2zi7	If you've had this BC method before it might n...	pmhmme	oneanddone	2021-09-12 07:29:32
4	No-Consideration-723	hchzwn	Okay I'm in the same boat. About to be 28, had...	pmhmme	oneanddone	2021-09-12 02:00:36
...	...	...	...	...	...	...
6154	SleepingClowns	hmex5v9	I think it's always a struggle to return to se...	r4269g	breakingmom	2021-11-28 17:20:01
6155	D-0ner	hmedlmu	Not likely. If you notice that, tell your TCM ...	r3ur14	ChineseMedicine	2021-11-28 14:40:28
6156	emmahar	hmebwzbz	I had an amazing pregnancy, labour and postpar...	r3rvil	oneanddone	2021-11-28 14:23:33
6157	lovegreenlife	hme7plh	It's the worst. I take natural calm magnesium,...	r3y2st	beyondthebump	2021-11-28 13:40:25
6158	Orunnergirl0	hlypae9	We have a two story home, as well as a basemen...	r1hb2l	pregnant	2021-11-25 00:09:25

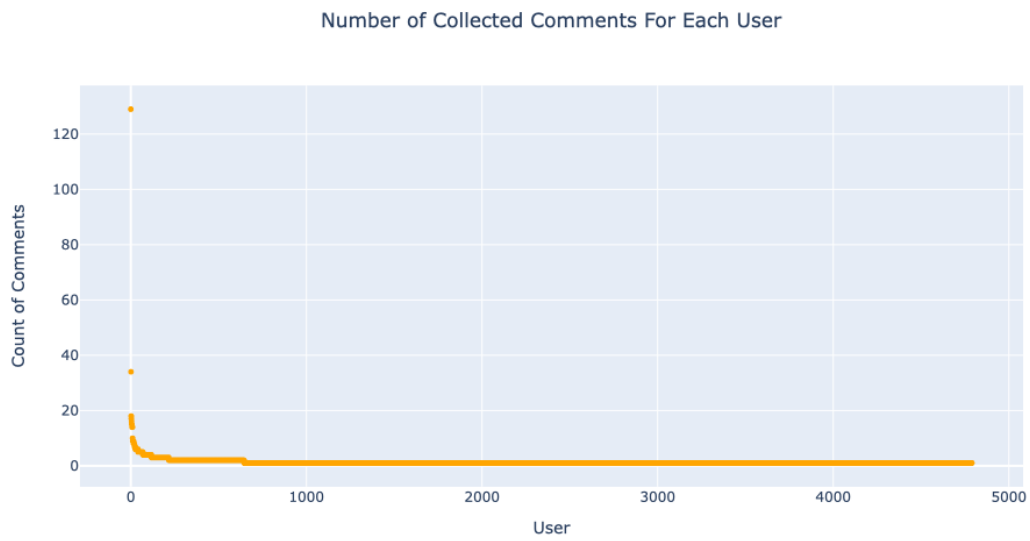
6145 rows x 6 columns

Figure 3.10: Reddit Comment Dataset (post-match).

The comments were created by 4793 unique users in the year 2021. Although a large number of randomly selected comments were collected, the size of the dataset allowed for a sufficient number of matched comments after the reduction process. Similarly to the post dataset, the comment dataset is well-distributed over time and contains a diverse range of users and shown in Figure 3.11 and 3.12.



**Figure 3.11:** Distribution of Comments per Month (post-match).



**Figure 3.12:** Unique Users vs. Collected Comments.

# 4

## Preprocessing

The preprocessing of textual data is a critical step in any analysis that involves working with textual information [24]. It is essential for transforming raw data into a format that can be used as input for subsequent analysis. Preprocessing involves various techniques such as standardization and cleansing, stop word removal, lemmatization [25], which are aimed at removing irrelevant or distracting information, correcting errors, and transforming the data into a suitable format. The objective of preprocessing is to create a high-quality input for the further analysis, which is crucial for obtaining accurate and meaningful results. The quality of the input data directly impacts the effectiveness of the subsequent analysis, highlighting the importance of careful and thorough preprocessing [26].

The preprocessing steps undertaken in this study were divided into two main stages, superficial cleaning and deep cleaning. The reason for dividing the preprocessing into two stages was that the output of each stage was used as input for different algorithms, which required specific types of data inputs. The methods for analyzing the data also differed, and the cleaning steps were selected based on the requirements of the data input for each algorithm.

An extra step which is creating subsentences was applied on Reddit Comment and Reddit Post datasets between the superficial and deep cleaning processes.

The figure 4.1, depicts the preprocessing procedures applied to input datasets, where Reddit Posts are represented by the color red, Reddit Comments by the color orange, and Twitter data by the color blue. The preprocessing entailed superficial cleaning for all datasets, followed by an additional step of subdividing the Reddit Post and Comment datasets into subsentences. The resulting datasets from the Reddit Post and Comment subsentence division, and the superficially cleaned Twitter dataset, were subsequently used as input for deep cleaning.

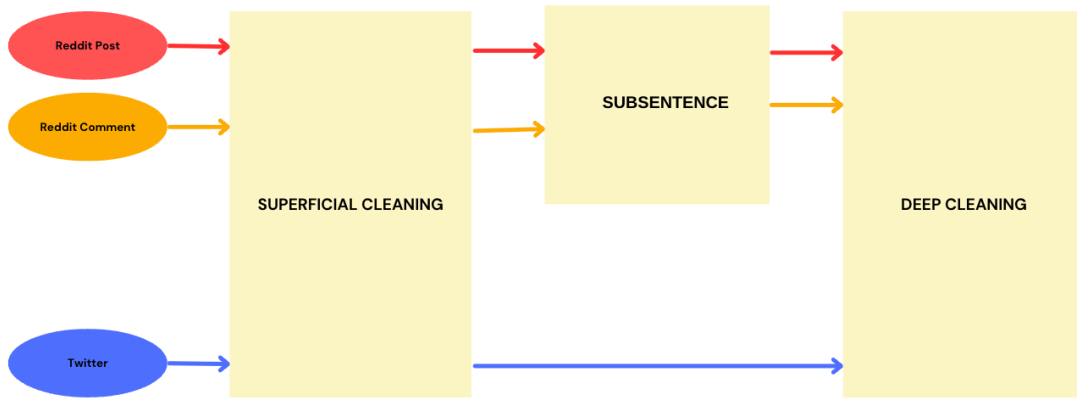


Figure 4.1: Structure of Preprocessing Steps.

## 4.1 SUPERFICIAL CLEANING

The datasets collected from social media platforms such as Twitter and Reddit are produced by actual users, and hence, contain various forms of noise [27]. Unlike scientific paper text data, this data is not clean and consists of different types of noise such as double punctuation, double letters, misspelling, emoticons, special characters, randomly written letters, messages from bots, and moderators. This can create ambiguity for the used algorithms, which may sometimes interpret the non-text data as part of the text data. For example, if a user writes a sentence and includes a website link without any space or punctuation, the algorithm may consider the entire string as text data, including the website link. Similarly, if a user writes a sentence and includes an emoji within a set of parentheses, the algorithm may consider the entire string as text data, including the parentheses and the emoji. It was observed that without cleaning the mentioned noises, the BertAgent and LIWC algorithms which was used for the analysis and will be explained in next section produced different outputs. Additionally, in the

cases explained before the algorithms sometimes eliminated this kind of noise and sometimes consider website links and other non-text values as meaningful data and provide output for them too. Because the collected text data is not written in a standardized way as mentioned before. To use the collected data for further analysis, superficial cleaning steps were applied on the raw texts.

To select the noises to be cleaned in superficial cleaning step, a manual check of the dataset was conducted. The primary rule for selecting the steps was to clean the extra non-text data while not changing the main text. Therefore, the cleaning did not involve lemmatization or stopword removal.

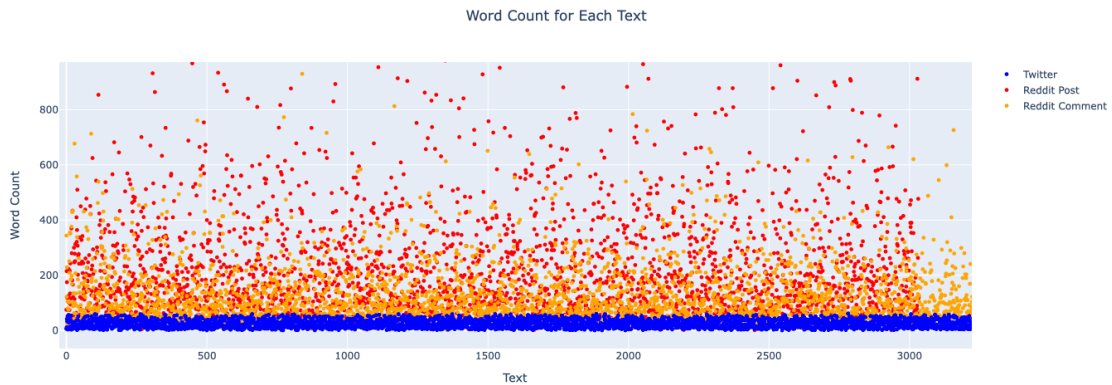
The following steps represent the superficial cleaning process.

1. Fixing the contractions.
2. Removing website links.
3. Removing accented characters.
4. Removing the text inside square brackets which mentions the following link or a warning such as '[read this]'.
5. Removing emoji.
6. Removing moderator messages such as the warnings for the owner of collected post or tweets.
7. Removing hashtags and mentions with punctuation and text such as #keyword or username.
8. Removing double spaces.
9. Removing non-text special words which are based on the text type such as '&#x200B' or '&#x2013'.
10. Removing extra used new lines.
11. Limiting all the repetitions to two characters and removing the extra characters.
12. Removing punctuation except main sentence punctuation,
13. Removing the sentences which represents the rule of the community such as start by '[\*\*(Full Rules)\*\*]'.
14. Removing numbers.

A portion of the collected data was entirely deleted, while in some cases, only parts of the data were removed through the application of the superficial cleaning process. Following the superficial cleaning procedure, the final dataset sizes for the Reddit post, Reddit comment, and Twitter datasets were determined to be 3034, 6016, and 17664, respectively.

## 4.2 SUBSENTENCE FOR REDDIT POST AND COMMENT

The word count of the texts in Reddit comment and post datasets is notably greater than that of the Twitter dataset. This observation is supported by the data presented in Figure 4.2, which illustrates a comparison of the word count for a sample of three datasets. The x axis represents the text and y axis represents the word count of the text. This trend is consistent across all of the datasets examined.



**Figure 4.2:** Word Count Comparison for Reddit Post, Comment and Twitter.

The BERTopic algorithm was used for the topic extraction and will be explained next section, may become less efficient when dealing with longer texts, as these texts may contain various subjects within the same post or comment. To resolve this problem, an approach could involve dividing lengthy documents into smaller units such as sentences or paragraphs [28]. It was determined that each text within the Reddit post and comment datasets would be divided into subsentences using the NLTK sentence tokenizer after undergoing the superficial cleaning process. This approach would allow for more targeted analysis of the text, potentially improving the performance of the BERTopic algorithm. Each text element in the dataset is assigned a main text index, and upon division into subsentences, the order of the subsentences within

the original sentence is stored as the subsentence index for further analysis. It was resulted in final dataset with 44025 elements for Reddit post and 45699 elements for Reddit Comment.

### 4.3 DEEP CLEANING

Deep cleaning is a comprehensive process that involves multiple steps to remove all types of noise from text data and transform it into a cleaned base format. In addition to superficial cleaning, it includes several steps such as lower casing, correcting spellings including removing the punctuation, tokenization, part-of-speech tagging, and lemmatization. The primary objective of this process is to extract the most important information from the data and create a base format that is suitable for input into the BERTopic algorithm. This enables the BERTopic algorithm to effectively identify and group related topics within the data.

Spacy is a popular open-source software library used for natural language processing (NLP) tasks. It offers a range of tools and features, including tokenization, part-of-speech (POS) tagging, stopword removal, and lemmatization. These tools are essential for many NLP tasks, such as text classification, sentiment analysis, and named entity recognition. The decision of using Spacy was based on the literature research and it can be concluded that the quality criterion suggests that Spacy's lemmatization technique, along with its removal of stop words, is more efficient compared to NLTK [29].

In this research, Spacy was utilized for several steps of deep cleaning, including word tokenization, POS tagging, stopword removal, and lemmatization. Word tokenization is the process of splitting a sentence or text into individual words or tokens [30]. POS tagging is Past of Speech tagging and involves assigning a grammatical category to each word in a sentence [30]. It was decided to use the Universal Dependencies part of speech tag set with Spacy. The Universal Dependencies tagset is a standardized set of POS tags that are applicable to many languages. It provides a universal way to label and annotate grammatical categories across different languages, which enables easy interoperability and comparability across different NLP tasks and applications.

List of tags [31]:

- ADJ: Adjective
- ADV: Adverb
- AUX: Auxiliary verb
- CCONJ: Coordinating conjunction
- DET: Determiner
- INTJ: Interjection
- NOUN: Noun
- NUM: Numeral
- PART: Particle
- PRON: Pronoun
- PROPN: Proper noun
- PUNCT: Punctuation
- SCONJ: Subordinating conjunction
- SYM: Symbol
- VERB: Verb
- X: Other

Stopword removal is the process of removing common words such as "the," "a," and "an" that do not carry significant meaning in the text [30]. Lemmatization is the process of reducing words to their base form [30], or lemma, such as converting "running" to "run".

The order in which these tasks are performed can impact the final output of the NLP pipeline. For example, if stop word removal is done before POS tagging, some important words may be mistakenly removed as stop words because their POS tag indicates they are not meaningful. Similarly, performing lemmatization prior to POS tagging implies that the postag information cannot be leveraged to match the tag with the word. Executing lemmatization without utilizing postag information may lead to erroneous outcomes when converting the word to its base form.



The following steps represent the deep cleaning process in this research.

1. Lowercasing,
2. Correcting spellings,
3. Word tokenization,
4. Pos tagging,
5. Stop word removing,
6. Lemmatization,

The aforementioned steps were applied to the subsentence dataset of Reddit comment and post datasets and Twitter dataset.

Figure 4.3 depicts the relationship between preprocessing steps and analysis methods applied to Reddit post, Reddit comment, and Twitter datasets. The schema presents an elaboration of the primary steps outlined in Figure 4.1, including the interrelationships between the output of each step, denoted by colored circles based on the dataset color code. Specifically, the red, orange, and blue circles represent Reddit post, Reddit comment, and Twitter datasets, respectively. The output of the subsentence creation step for Reddit post and comment datasets, as well as the output of the superficial cleaning step for Twitter, served as inputs for the LIWC and BERTAgent algorithms. Additionally, the postagging substep within the deep cleaning process generated inputs for grammatical calculations across all datasets. Finally, the resultant outputs from all cleaning steps served as inputs for the BERTopic analysis method. The mentioned methods will be explained in the next section as mentioned before.

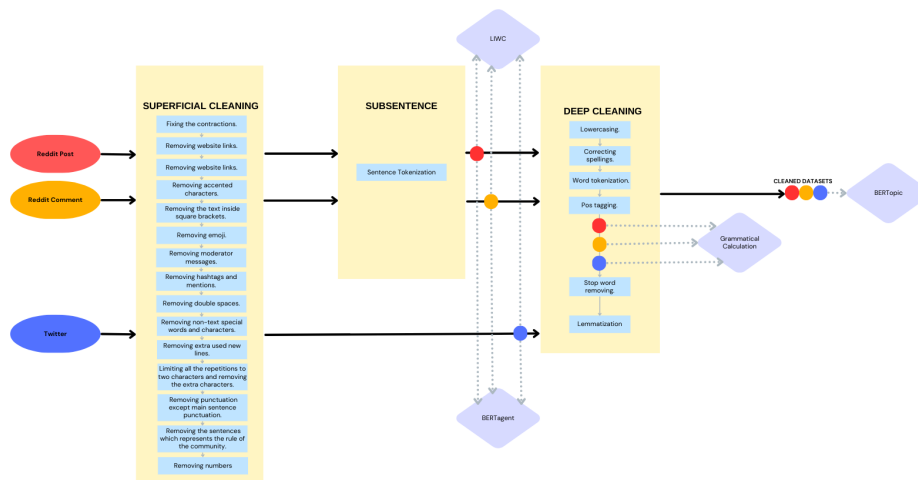


Figure 4.3: Relationship between Preprocessing Steps and Analysis.

Figure 4.4 represents the evolution of the collected texts through the preprocessing steps for Twitter. The represented dataset consists of eight columns, each of which is described as follows:

- **text** : The collected original textual data.
- **after\_superficial\_cleaning** : Superficial cleaning steps applied version of the original data.
- **after\_lowercasing\_spelling** : Lowercased and spelling corrected version of data.
- **token** : Word list created by word tokenization method.
- **token\_pos** : Postag list of created word list.
- **pos\_tag** : List of (word,postag) pairs.
- **stopword\_removed** : Stopword removed version of the data.
- **lemmatized\_text** : Lemmatization applied version of the stopword removed data. The most cleaned version of the data.

	text	after_superficial_cleaning	after_lowercasing_spelling	token	token_pos	pos_tag	stopword_removed	lemmatized_text
0	Postpartum is not a joke .	Postpartum is not a joke .	postpartum is not a joke	['postpartum', 'is', 'not', 'a', 'joke']	['NOUN', 'PART', 'AUX', 'DET', 'NOUN']	[(('postpartum', 'NOUN'), ('is', 'AUX'), ('not', 'PART'))]	postpartum joke	postpartum joke
1	@Nawfkage postpartum depression? i didnt know ...	postpartum depression? i did not know you had ...	postpartum depression i did not know you had ...	['postpartum', 'depression', 'i', 'did', 'not']	['ADJ', 'NOUN', 'PRON', 'AUX', 'PART', 'VERB', '...']	[(('postpartum', 'ADJ'), ('depression', 'NOUN'))]	postpartum depression know baby rude	postpartum depression know baby rude
2	Happy for them. I will say i hope he's better ...	Happy for them. I will say i hope he is better...	happy for them i will say i hope he is better ...	['happy', 'for', 'them', 'i', 'will', 'say', 'i', 'hope', 'he', 'is', 'better']	['ADJ', 'ADP', 'PRON', 'AUX', 'PART', 'VERB', '...']	[(('happy', 'ADJ'), ('for', 'ADP'), ('them', 'PRON'), ('i', 'PRON'), ('will', 'AUX'), ('say', 'VERB'), ('i', 'PRON'), ('hope', 'VERB'), ('he', 'PRON'), ('is', 'AUX'), ('better', 'ADJ'))]	happy hope better pregnancy ari like missed im...	happy hope well pregnancy ari like miss import...
3	nfs I would hate to be Ari seen this cause ni...	nfs I would hate to be Ari seen this because ...	ns i would hate to be ari seen this because ni...	['ns', 'i', 'would', 'hate', 'to', 'be', 'ari']	['NOUN', 'PRON', 'AUX', 'VERB', 'PART', 'AUX', '...']	[(('ns', 'NOUN'), ('i', 'PRON'), ('would', 'AUX'), ('hate', 'VERB'), ('to', 'PART'), ('be', 'AUX'))]	ns hate ari seen nigra kid s baby shower nigra...	hate ari see nigra kid baby shower nigra lea...
4	@Jflexo_ That and postpartum depression 😊	That and postpartum depression	that and postpartum depression	['that', 'and', 'postpartum', 'depression']	['PRON', 'CCONJ', 'ADJ', 'NOUN']	[(('that', 'PRON'), ('and', 'CCONJ'), ('postpartum', 'ADJ'), ('depression', 'NOUN'))]	postpartum depression	postpartum depression
...	...	...	...	...	...	...	...	...
17659	@pulte 36wks preppers here and would love to b...	wks preppers here and would love to buy some p...	was prefers here and would love to buy some po...	['was', 'prefers', 'here', 'and', 'would', 'lo...']	['AUX', 'NOUN', 'ADV', 'CCONJ', 'AUX', 'VERB', '...']	[(('was', 'AUX'), ('prefers', 'NOUN'), ('and', 'CCONJ'), ('would', 'AUX'), ('lo...', 'VERB'))]	prefers love buy postpartum necessities lanalw...	prefer love buy postpartum necessity lanalwell
17660	@mamasflwrchild I got zero fertility training....	I got zero fertility training. I feel rather s...	i got zero fertility training i feel rather su...	['i', 'got', 'zero', 'fertility', 'training', 'i', 'feel', 'rather', 'su...']	['PRON', 'VERB', 'NUM', 'NOUN', 'NOUN', 'PRON', '...']	[(('i', 'PRON'), ('got', 'VERB'), ('zero', 'NUM'), ('fertility', 'NOUN'), ('training', 'NOUN'), ('i', 'PRON'), ('feel', 'VERB'), ('rather', 'ADV'))]	got zero fertility training feel sufficiently ...	get zero fertility training feel sufficiently ...
17661	Curious about your period after giving birth? ...	Curious about your period after giving birth? ...	curious about your period after giving birth ...	['curious', 'about', 'your', 'period', 'after']	['ADJ', 'ADP', 'PRON', 'NOUN', 'SCONJ', 'VERB', '...']	[(('curious', 'ADJ'), ('about', 'ADP'), ('your', 'PRON'), ('period', 'NOUN'), ('after', 'SCONJ'))]	curious period giving birth answer pressing qu...	curious period give birth answer press question
17662	never knew i would be going through postpartum...	never knew i would be going through postpartum...	never knew i would be going through postpartum...	['never', 'knew', 'i', 'would', 'be', 'going']	['ADV', 'VERB', 'PRON', 'AUX', 'AUX', 'VERB', '...']	[(('never', 'ADV'), ('knew', 'VERB'), ('i', 'PRON'), ('would', 'AUX'), ('be', 'AUX'), ('going', 'VERB'))]	knew going postpartum depression shit good fee...	know go postpartum depression shit good feeling
17663	Join us the 2nd and 4th Wednesday of the month...	Join us the nd and th Wednesday of the month a...	join us the nd and th wednesday of the month a...	['join', 'us', 'the', 'nd', 'and', 'th', 'wedn...']	['VERB', 'PRON', 'DET', 'NOUN', 'CCONJ', 'NOUN', '...']	[(('join', 'VERB'), ('us', 'PRON'), ('the', 'DET'), ('nd', 'NOUN'), ('and', 'CCONJ'), ('th', 'NOUN'), ('wedn...', 'NOUN'))]	join nd th wednesday month pm est pm pst prena...	join wednesday month pm est pm pst prenatal ...

Figure 4.4: Twitter dataset after Preprocessing.

Figures 4.5 and 4.6 present the Reddit post and comment datasets, respectively. In addition to the information included in the Twitter dataset, the main text for each post or comment is stored in the "text" column, while the version of the text divided into subsentences is stored in the sub\_sentence column. The order of the main text within the original post or comment is stored in the main\_text\_index column, while the order of the subsentences within the corresponding main text is stored in the sub\_sentence\_column.

	text	after_superficial_cleaning	main_text_index	sub_sentence_index	sub_sentence	after_lower casing_spelling	token	token_pos	pos_tag	stopword_removed	lemmatized_text
0	All my homies who have experienced hair change...	All my homies who have experienced hair change...	0	0	All my homies who have experienced hair change...	all my homies who have experienced hair changes...	['all', 'my', 'homies', 'who', 'have', 'experie...']	['DET', 'PRON', 'NOUN', 'PRON', 'VERB', 'ADJ', ...]	['[all', 'DET'], (my', 'PRON'), (nomes', 'NOUN'), 'NO...']	homes experienced hair changes postpartum shar...	home experience hair change postpartum share e...
1	All my homies who have experienced hair change...	All my homies who have experienced hair change...	0	1	I am getting to the point that I am worried th...	i am getting to the point that i am worried th...	['i', 'am', 'getting', 'to', 'the', 'point', 'worried', 'th...']	['PRON', 'AUX', 'VERB', 'ADP', 'DET', 'NOUN', 'VERB', 'ADJ', ...]	['[i', 'PRON'], (am', 'AUX'), (getting', 'VERB'), 'VE...']	getting point worried limp sad hair hormonal c...	get point worried limp sad hair hormonal change
2	All my homies who have experienced hair change...	All my homies who have experienced hair change...	0	2	My hair is still thick but so soft and very il...	my hair is still thick but so soft and very il...	['my', 'hair', 'is', 'still', 'thick', 'but', 'soft', 'and', 'very', 'il...']	['PRON', 'NOUN', 'AUX', 'ADV', 'ADJ', 'CONJ', ...]	['[my', 'PRON'], (hair', 'NOUN'), (is', 'AUX'), (still', 'ADV'), (thick', 'ADJ'), (but', 'CONJ'), (soft', 'ADJ'), (and', 'CONJ'), (very', 'ADV'), (il...', 'ADJ'), ...]	hair thick soft limp nearly curly handfl year...	hair thick soft limp nearly curly handfl year...
3	All my homies who have experienced hair change...	All my homies who have experienced hair change...	0	3	Ten months postpartum, and I just want to know...	ten months postpartum and i just want to know ...	['ten', 'months', 'postpartum', 'and', 'i', 'just', 'want', 'to', 'know', '...']	['NUM', 'NOUN', 'NOUN', 'CONJ', 'PRON', 'ADV', ...]	['[ten', 'NUM'], (months', 'NOUN'), (postpartum', 'NOUN'), (and', 'CONJ'), (i', 'PRON'), (just', 'ADV'), (want', 'VERB'), (to', 'PART'), (know', 'VERB'), ...]	months postpartum want know	month postpartum want know
4	Hi fellow fit pregnant friends! 'nini'm not re...	Hi fellow fit pregnant friends! i am not reall...	1	0	Hi fellow fit pregnant friends!	hi fellow fit pregnant friends!	['hi', 'fellow', 'fit', 'pregnant', 'friends']	['INTJ', 'ADJ', 'VERB', 'ADJ', 'NOUN']	['[hi', 'INTJ'], (fellow', 'ADJ'), (fit', 'VERB'), (pregnant', 'ADJ'), (friends', 'NOUN')]	hi fellow fit pregnant friends	hi fellow fit pregnant friend
44020	Y'all should stop shaming her for saying cra...	You all should stop shaming her for saying cra...	3032	2	If she left him running a muck in her apartmen...	if she left him running a muck in her apartmen...	['if', 'she', 'left', 'him', 'running', 'a', 'muck', 'in', 'her', 'apartmen...']	['SCONJ', 'PRON', 'VERB', 'PRON', 'VERB', 'DET', ...]	['[if', 'SCONJ'], (she', 'PRON'), (left', 'VERB'), (him', 'PRON'), (running', 'VERB'), (a', 'DET'), (muck', 'NOUN'), (in', 'ADP'), (her', 'PRON'), (apartmen...', 'NOUN'), ...]	left running apartment day shame	leave run apartment day shame
44021	Y'all should stop shaming her for saying cra...	You all should stop shaming her for saying cra...	3032	3	Would you all shame a postpartum mother for be...	would you all shame a postpartum mother for be...	['would', 'you', 'all', 'shame', 'a', 'postpartum', 'mother', 'for', 'be...']	['AUX', 'PRON', 'DET', 'VERB', 'PRON', 'ADJ', ...]	['[would', 'AUX'], (you', 'PRON'), (all', 'PRON'), (shame', 'VERB'), (a', 'DET'), (postpartum', 'NOUN'), (mother', 'NOUN'), (for', 'ADP'), (be...', 'VERB'), ...]	shame postpartum mother depressed knew got pre...	shame postpartum mother depressed know get pre...
44022	Y'all should stop shaming her for saying cra...	You all should stop shaming her for saying cra...	3032	4	No I do not think so!	no i do not think so!	['no', 'i', 'do', 'not', 'think', 'so']	['INTJ', 'PRON', 'AUX', 'PART', 'VERB', 'ADV']	['[no', 'INTJ'], (i', 'PRON'), (do', 'AUX'), (not', 'PART'), (think', 'VERB'), (so', 'ADV')]	think	think
44023	Y'all should stop shaming her for saying cra...	You all should stop shaming her for saying cra...	3032	5	Tajyn knew it would be hard and she is just ex...	talk knew it would be hard and she is just ex...	['talk', 'knew', 'it', 'would', 'be', 'hard', 'and', 'she', 'is', 'just', 'ex...']	['NOUN', 'VERB', 'PRON', 'AUX', 'AUX', 'ADJ', ...]	['[talk', 'NOUN'], (knew', 'VERB'), (it', 'PRON'), (would', 'AUX'), (be', 'AUX'), (hard', 'ADJ'), (and', 'CONJ'), (she', 'PRON'), (is', 'AUX'), (just', 'ADV'), (ex...', 'ADJ'), ...]	talk knew hard expressing hard like	talk know hard express hard like
44024	Y'all should stop shaming her for saying cra...	You all should stop shaming her for saying cra...	3032	6	She is doing a whole lot better than I could w...	she is doing a whole lot better than i could w...	['she', 'is', 'doing', 'a', 'whole', 'lot', 'better', 'than', 'i', 'could', 'w...']	['PRON', 'AUX', 'VERB', 'DET', 'ADJ', 'NOUN', ...]	['[she', 'PRON'], (is', 'AUX'), (doing', 'VERB'), (a', 'DET'), (whole', 'ADJ'), (lot', 'NOUN'), (better', 'ADJ'), (than', 'CONJ'), (i', 'PRON'), (could', 'AUX'), (w...', 'ADJ'), ...]	lot better teaching tricks	lot well teach trick

Figure 4.5: Reddit Post Dataset after Preprocessing.

	text	after_superficial_cleaning	main_text_index	sub_sentence_index	sub_sentence	after_lower casing_spelling	token	token_pos	pos_tag	stopword_removed	lemmatized_text
0	We are currently trying for number two and hav...	We are currently trying for number two and hav...	3033	0	We are currently trying for number two and hav...	we are currently trying for number two and hav...	['we', 'are', 'currently', 'trying', 'for', 'number', 'two', 'and', 'hav...']	['PRON', 'AUX', 'ADV', 'VERB', 'ADP', 'NOUN', ...]	['[we', 'PRON'], (are', 'AUX'), (currently', 'ADV'), (trying', 'VERB'), (for', 'ADP'), (number', 'NOUN'), (two', 'NOUN'), (and', 'CONJ'), (hav...', 'ADJ'), ...]	currently trying number daughter months daught...	currently try number daughter months daughter months ...
1	We are currently trying for number two and hav...	We are currently trying for number two and hav...	3033	1	My pregnancy was a nightmare with having to be...	my pregnancy was a nightmare with having to be...	['my', 'pregnancy', 'was', 'a', 'nightmare', 'with', 'having', 'to', 'be...']	['PRON', 'NOUN', 'DET', 'NOUN', 'SCONJ', ...]	['[my', 'PRON'], (pregnancy', 'NOUN'), (was', 'AUX'), (a', 'DET'), (nightmare', 'NOUN'), (with', 'SCONJ'), (having', 'VERB'), (to', 'PART'), (be...', 'VERB'), ...]	pregnancy nightmare having blood thinkers morn...	pregnancy nightmare have blood thinkers morning...
2	We are currently trying for number two and hav...	We are currently trying for number two and hav...	3033	2	My OB failed to inform me of her induction pro...	my ob failed to inform me of her induction pro...	['my', 'ob', 'failed', 'to', 'inform', 'me', 'of', 'her', 'induction', 'pro...']	['PRON', 'NOUN', 'VERB', 'PART', 'VERB', 'PRON', ...]	['[my', 'PRON'], (ob', 'NOUN'), (failed', 'VERB'), (to', 'PART'), (inform', 'VERB'), (me', 'PRON'), (of', 'ADP'), (her', 'PRON'), (induction', 'NOUN'), (pro...', 'NOUN'), ...]	ob failed inform induction procedures forced L...	ob fail inform induction procedure force induc...
3	We are currently trying for number two and hav...	We are currently trying for number two and hav...	3033	3	My OB was well aware of my mental health issue...	my ob was well aware of my mental health issue...	['my', 'ob', 'was', 'well', 'aware', 'of', 'my', 'mental', 'health', 'issue...']	['PRON', 'AUX', 'ADV', 'ADJ', 'ADP', ...]	['[my', 'PRON'], (ob', 'NOUN'), (was', 'AUX'), (well', 'ADV'), (aware', 'ADJ'), (of', 'ADP'), (my', 'PRON'), (mental', 'NOUN'), (health', 'NOUN'), (issue...', 'NOUN'), ...]	ob aware mental health issue hospitals told d...	ob aware mental health issue hospital tell dra...
4	We are currently trying for number two and hav...	We are currently trying for number two and hav...	3033	4	My Doula was unless and worked with the nurses...	my would was unless and worked with the nurses...	['my', 'would', 'was', 'unless', 'and', 'worked', 'with', 'the', 'nurses...']	['PRON', 'AUX', 'AUX', 'ADJ', 'CONJ', 'VERB', ...]	['[my', 'PRON'], (would', 'AUX'), (was', 'AUX'), (unless', 'ADJ'), (and', 'CONJ'), (worked', 'VERB'), (with', 'ADP'), (the', 'PRON'), (nurses', 'NOUN'), ...]	worked nurses supporting husband forgotten bir...	work nurse support husband forget birth plan L...
45694	We have a two story home, as well as a basemen...	We have a two story home, as well as a basemen...	9048	1	No issues with the stairs, other than having L...	no issues with the stairs other than having to...	['no', 'issues', 'with', 'the', 'stairs', 'other', 'than', 'having', 'to...']	['DET', 'NOUN', 'ADP', 'DET', 'NOUN', 'ADJ', ...]	['[no', 'DET'], (issues', 'NOUN'), (with', 'ADP'), (the', 'DET'), (stairs', 'NOUN'), (other', 'ADJ'), (than', 'CONJ'), (having', 'VERB'), (to', 'PART'), ...]	issues stairs having lug million random toys r...	issue stair have lug million random toy request...
45695	We have a two story home, as well as a basemen...	We have a two story home, as well as a basemen...	9048	2	Oh, and it sucks having to make multiple trips ...	oh and it sucks having to make multiple trips ...	['oh', 'and', 'it', 'sucks', 'having', 'to', 'make', 'multiple', 'trips', '...']	['INTJ', 'CONJ', 'PRON', 'VERB', 'PART', ...]	['[oh', 'INTJ'], (and', 'CONJ'), (it', 'PRON'), (sucks', 'VERB'), (having', 'VERB'), (to', 'PART'), (make', 'VERB'), (multiple', 'ADJ'), (trips', 'NOUN'), ...]	oh sucks having multiple trips bring baby upst...	oh suck have multiple trip bring baby upstairs...
45696	We have a two story home, as well as a basemen...	We have a two story home, as well as a basemen...	9048	3	Otherwise, stairs have not been an issue.	otherwise stairs have not been an issue.	['otherwise', 'stairs', 'have', 'not', 'been', 'an', 'issue', '...']	['ADJ', 'NOUN', 'AUX', 'PART', 'AUX', 'DET', ...]	['[otherwise', 'ADJ'], (stairs', 'NOUN'), (have', 'AUX'), (not', 'PART'), (been', 'AUX'), (an', 'DET'), (issue', 'NOUN'), ...]	stairs issue	stair issue
45697	We have a two story home, as well as a basemen...	We have a two story home, as well as a basemen...	9048	4	I have never given them a second thought, to b...	i have never given them a second thought to be...	['i', 'have', 'never', 'given', 'them', 'a', 'second', 'thought', 'to', 'b...']	['PRON', 'AUX', 'ADV', 'VERB', 'PRON', 'DET', ...]	['[i', 'PRON'], (have', 'AUX'), (never', 'ADV'), (given', 'VERB'), (them', 'PRON'), (a', 'DET'), (second', 'ADJ'), (thought', 'NOUN'), (to', 'PART'), (be...', 'VERB'), ...]	given second thought honest	give second thought honest
45698	We have a two story home, as well as a basemen...	We have a two story home, as well as a basemen...	9048	5	Edit I had two vaginal deliveries and had no p...	edit i had two vaginal deliveries and had no p...	['edit', 'i', 'had', 'two', 'vaginal', 'deliveries', 'and', 'had', 'no', 'p...']	['NOUN', 'PRON', 'VERB', 'NUM', 'ADJ', 'NOUN', ...]	['[edit', 'NOUN'], (i', 'PRON'), (had', 'VERB'), (two', 'NUM'), (vaginal', 'ADJ'), (deliveries', 'NOUN'), (and', 'CONJ'), (had', 'VERB'), (no', 'ADV'), (p...', 'NOUN'), ...]	edit vaginal deliveries pain problems immediat...	edit vaginal delivery pain problem immediately...

Figure 4.6: Reddit Comment Dataset after Preprocessing.

# 5

## Tools and Methods

This section provides in-depth details of the tools and methods employed for the analysis. The algorithms used were selected based on the project's specific requirements to ensure optimal outcomes. The algorithm selection process involved reviewing similar projects and conducting a thorough literature review to identify the most appropriate algorithms for the project.

The collected and cleaned datasets were categorized based on topic labels, and from these datasets, useful information such as emotion score, agency score, and term similarity score were extracted. These extracted data were subsequently utilized in the visualization process, where the documents and topics were colored based on the selected metrics. All the applied methods and tools with the outcomes and related processes will be explained in details.

## 5.1 LIWC

Given the high volume of text based-communication in social media platforms there are distinctive prospects to leverage text for gaining deeper insights into fundamental psychological constructs such as emotions. Recognizing emotions in text is a crucial component in social media analysis, and the Linguistic Inquiry and Word Count (LIWC) is a frequently employed software tool for this purpose [32]. The Linguistic Inquiry and Word Count (LIWC) algorithm is a text analysis tool that quantifies the linguistic and psychological features of written or spoken language. It works by analyzing the occurrence of words and word stems in a given text and classifying them into pre-defined linguistic and psychological categories, such as positive or negative affect, social or cognitive processes, and linguistic dimensions such as pronouns, adjectives, and conjunctions.

LIWC uses a dictionary of words and word stems to categorize text based on their linguistic and psychological properties. The algorithm then calculates the frequency of each category in the text and provides a statistical analysis of the language use, allowing for a deeper understanding of the psychological, cognitive, and social dimensions of the text [33]. LIWC is commonly used in fields such as psychology, sociology, and communication studies, and has been shown to be a reliable and valid tool for analyzing text.

LIWC algorithm was utilized to obtain the negative and positive emotion scores for the Reddit comment, post, and Twitter datasets. These scores were then employed in the topic-based visualization, which will be discussed in the following section. It should be noted that the obtained scores from LIWC algorithm fall within the range of 0 to 100, where higher scores indicate stronger expression of either positive or negative emotions.

The composite emotion score is derived by calculating the difference between the positive and negative emotion scores, serving as a representative measure of both. A low composite emotion score indicates a predominance of negative emotions, while a high composite emotion score signifies a prevalence of positive emotions.

## 5.2 BERTAGENT

BERTAgent is a newly developed tool that uses the BERT-based model to measure the degree of semantic agency in texts. Agency refers to goal-directed actions. It employs advanced machine learning techniques for natural language processing and has been fine-tuned using specific textual data that have been rated by individuals based on the level of agency expressed [3].

The BERTAgent algorithm was utilized to generate a metric for agency which is represented by calculated the outcome of the algorithm called as predicted whole mean score, which was then employed to establish a linkage between the positive and negative scores computed by the LIWC algorithm and extracted composite emotion score. The dataset was represented using the notion that low levels of agency are associated with negative emotions, while high levels of agency are linked to positive emotions. Regarding the composite emotion score, low levels of agency correspond to low composite emotion scores, which in turn indicate high negative emotion scores. Conversely, high levels of agency correspond to high composite emotion scores, which signify elevated positive emotion scores.

### 5.3 GRAMMATICAL CALCULATION

The grammatical calculation refers to the computation of the ratio of non-modal and non-auxiliary verbs in relation to the total number of words in a given text. The purpose of this calculation is to determine the prevalence of main verbs within the text. Modal verbs are a type of auxiliary verbs that express various meanings such as ability, possibility, permission, obligation, and intention. They are used with the main verb to modify its meaning and express the speaker's attitude or stance towards the action being described. The eliminated modal verbs: 'can', 'could', 'may', 'might', 'must', 'shall', 'should', 'will', 'would'.

Auxiliary verbs, on the other hand, are used with the main verb to form various grammatical constructions, such as verb tenses, questions, and negative statements. They do not have inherent meaning on their own but serve to support the main verb. The list of auxiliary verbs: 'am', 'is', 'are', 'was', 'were', 'being', 'been', 'be', 'has', 'have', 'had', 'does', 'do', 'did', 'can', 'could', 'will', 'would', 'shall', 'should', 'may', 'might', 'must'.

The output of the posttagging step yields a set of word-postag pairs that are utilized as input for the grammatical calculation. The methods applied in this calculation include:

- Counting the number of pairs, excluding those that contain the postag "PUNC" (punctuation). This count is assigned as the total number of words.
- Counting the number of pairs that have the postag "AUX" or "VERBS", and the word is not included in the list of modal and auxiliary verbs.
- Dividing the counted verbs by the total number of words calculated.

## 5.4 BERTopic

BERTopic is a natural language processing algorithm that employs the BERT language model for topic modeling tasks. It utilizes a combination of hierarchical clustering and dimensionality reduction techniques to identify topic clusters within a given corpus of text. Basically, BERTopic creates embeddings of documents using pre-trained transformer-based language models. These embeddings are then clustered to group similar documents together. Finally, topic representations are generated based on these clusters [34]. BERTopic is the most pivotal applied analysis method in the project since all the calculated values obtained through the methods and algorithms outlined in this section was utilized to create graphs based on the results of the BERTopic algorithm, which serves as the ultimate outcome of the project.

The topic model consist of several steps. For the three datasets, topic models were created with the same steps but with different parameters. These parameters were chosen according to their suitability for the datasets. The following list consists of the applied topic analysis steps with the elements of the defined topic model and will be explained deeply.

- Creating Topic Model by Defining Following Items:
  - Embedding Model
  - UMAP Model
  - HDBSCAN Model
  - *c*-TF-IDF Model
  - Diversity
  - Number of Topics
  
- Reducing Outliers
- Term Similarity Score
- Creating New Labels

It is important to select of the parameters for topic model elements. As mentioned before, basically all the steps are same but some of elements has the different parameters in it based on the dataset. The ones with the common parameters are number of topics, embedding model, and *c*-TF-IDF Model. UMAP Model, HDBSCAN Model, and diversity has the different input



parameters based on the aspect of the dataset and will be explained in the next subsections. Before mentioning the selected parameters it is important to explain why to use this elements and how the parameters affect the trained model.

### 5.4.1 TOPIC MODEL

#### EMBEDDING MODEL

The first step in building a topic model is to prepare input text data, which is typically referred as documents. In order to do this, the text data must first be transformed into a numerical representation. BERTopic supports several embedding techniques for this purpose such as sentence-transformer, Hugging Face Transformers, Flair, Spacy, Universal Sentence Encoder, Gensim, Scikit-Learn Embeddings, TF-IDF. Among these techniques, sentence-transformer is recommended as it is effective in capturing similarity between documents [35]. BERTopic specifically uses "all-MiniLM-L6-v2" from Hugging Face as its default sentence-transformer model [36], which can map sentences and paragraphs to vectors that are suitable for clustering or semantic search tasks and works well for English documents [37]. The same sentence-transformer model was used for creating embeddings in the Reddit Comment, Reddit Post, and Twitter datasets as shown in the Code Listing 5.1:

Listing 5.1: Embedding Model

```
from sentence_transformers import SentenceTransformer
embedding_model = SentenceTransformer("all-MiniLM-L6-v2")
embeddings = embedding_model.encode(docs=docs) #documents
```

#### UMAP MODEL

Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction algorithm [38] used for visualizing high-dimensional data in a lower-dimensional space. It is a non-linear algorithm that preserves the local and global structure of the data while minimizing the distortion in the low-dimensional representation. UMAP is useful for data exploration, clustering, and classification tasks, and has been successfully applied to a variety of datasets, including text data [39].

In the context of BERTopic applied in this research, UMAP is used to reduce the dimensionality of the embedding vectors called as the reduced embeddings and generated from the text

data, which facilitates clustering and topic modeling. The basic form of an UMAP model can be defined as below:

Listing 5.2: Default UMAP Model

```
from umap import UMAP
umap_model = UMAP( n_neighbors=n_neighbors ,
                   n_components=n_components ,
                   min_dist=min_dist ,
                   metric=metric ). fit ( embeddings )
reduced_embeddings = umap_model.embedding_
```

The different selected parameters results different types of the graphs which may be more local visualization or more global visualization for the same dataset. The outcomes of BERTopic could vary across multiple runs of the same code because of the random and probabilistic aspects of UMAP. However, using custom embeddings allows users to try out BERTopic multiple times with different parameters until the desired topics are obtained [36]. The selection of the parameters was decided after many iterations, taking into account the effects mentioned below for each dataset. Only `n_neighbors` parameter is different for each dataset while the other parameters are same.

**n\_neighbors:** The number of neighbors parameter plays a crucial role in determining how UMAP handles the balance between local and global structures in the data. When `n_neighbors` is set to a low value, UMAP prioritizes local structures and may overlook the bigger picture. Conversely, setting `n_neighbors` to a high value makes UMAP take into account larger neighborhoods for estimating the data's manifold structure, potentially at the cost of losing fine details for the sake of broader information [36]. In the context of the project, using a small value for the `n_neighbors` parameter led to a localized representation of the data. However, this approach was deemed unsuitable for achieving the project's objective of analyzing large-scale Reddit and Twitter datasets.

**n\_components:** The number of components parameter gives an opportunity of selecting the dimensionality of the reduces dimension space to the user [36]. It was decided to set as 2 for all datasets and it is important to mention that BERTopic is used it as 2 in default models.

**min\_dist:** The minimum distance parameter is responsible for regulating the degree of proximity between points in the low dimensional representation generated by UMAP. It dictates the minimum distance allowed between points, directly affecting the level of clustering observed in the embeddings. Lower values of `min_dist` encourage clustering and a finer level of

topological structure, while higher values prevent point packing and prioritize preservation of broader topological structures [36]. This parameter was set to 0.0 for all datasets.

**metric:** This parameter controls how distance is computed [36] and set as 'cosine' for all the datasets.

## HDBSCAN MODEL

HDBSCAN is a density-based clustering algorithm that can effectively identify clusters of high-density data points separated by regions of low density in a high-dimensional space. This makes it a useful technique for analyzing complex and large datasets, where traditional clustering methods may not be suitable. Specifically, HDBSCAN was used as the clustering algorithm for BERTopic, which involves grouping similar documents into topics. In this research, HDBSCAN was combined with UMAP. The integration of these two techniques allowed for the identification of meaningful clusters of data points, based on their similarity, while preserving the data's topological structure. The basic form of an HDBSCAN model can be defined as below:

Listing 5.3: Default HDBSCAN Model

```
from hdbscan import HDBSCAN
hdbscan_model = HDBSCAN(
    min_cluster_size=min_cluster_size ,
    metric=metric ,
    min_samples=min_samples ,
    cluster_selection_method=cluster_selection_method ,
)
```

**min\_cluster\_size:** The minimum cluster size parameter is the key factor in determining the outcome of clustering. Ideally, it should be set to the minimum group size that is considered a cluster. However, its impact may not always be straightforward. The `min_cluster_size` parameter interacts with other clustering parameters and is also influenced by the use of UMAP [40].

**metric:** The metric parameter specifies the distance metric used to calculate the pairwise distances between data points in the clustering process [40]. It was selected as 'euclidean' for all datasets.

**min\_samples:** The minimum samples parameter controls the conservative nature of clustering. A higher value for `min_samples` leads to a more conservative clustering approach, which

results in more data points being identified as noise, and clusters being confined to denser regions. The gradual increase of `min_samples` makes clustering more conservative [40]. In this research, the minimum cluster size for all datasets was set to 1.

**cluster\_selection\_method:** The cluster selection method parameter determines how clusters are selected from the hierarchical clustering tree. For all datasets, it was selected as 'oem' which is the short version of "excess of mass", and this method selects the cluster hierarchy level that maximizes the expected value of the mass of the resulting clusters [40].

#### C-TF-IDF MODEL

BERTopic utilizes an adjusted version of the TF-IDF method, called c-TF-IDF, to accurately represent topics from the bag-of-words matrix. Unlike traditional TF-IDF, c-TF-IDF operates at a cluster or categorical level, taking into account the unique features that differentiate one cluster from another. This approach improves the topic extraction technique by capturing the underlying structure of the data. Although other algorithms can be used, the class-based TF-IDF representation is enabled by default in BERTopic. Nonetheless, the `ctfidf_model` can be explicitly passed to BERTopic, allowing for parameter tuning and customization of the topic extraction process [41]. This model was applied to all datasets in this study and defined as shown below:

Listing 5.4: c-TF-IDF Model

```
from BERTopic.vectorizers import ClassTfidfTransformer
ctfidf_model = ClassTfidfTransformer()
```

#### DIVERSITY AND NUMBER OF TOPICS

Diversity parameter is a value between 0 and 1 with. 0 represents not being diverse and 1 represents being very diverse. Number of topics was used to put a limitation for the number of created topics. It is same for all datasets and selected as 20.

In conclusion, the parameter `n_neighbors` in UMAP model, `min_cluster_size` in HDBSCAN model, diversity defined in topic model creates the differences while clustering and were selected after many iterations for Reddit comment,post and Twitter datasets. Final version of a basic topic model is given below:

Listing 5.5: Default BERTopic Model

```
from BERTopic import BERTopic
topic_model = BERTopic(
    embedding_model=embedding_model ,
    umap_model=umap_model ,
    hdbscan_model=hdbscan_model ,
    ctfidf_model=ctfidf_model ,
    diversity=diversity ,
    nr_topics=nr_topics
)
```

After defining the topic model, topic labels were extracted and used for the further analysis.

#### 5.4.2 REDUCING OUTLIERS

In the context of topic modeling, it is common to encounter documents that do not fit within any of the identified topics, and are therefore labeled as outliers with the value of topic number  $-1$ . However, in certain applications, it may be desirable to minimize the number of outlier documents. To avoid of losing the data in the analysis it was decided to apply the outlier reduction process. BERTopic supports various strategies for reducing the outliers after training a BERTopic model. One approach to reduce the number of outliers is to leverage the embeddings of these documents and compute the cosine similarity between each outlier document's embedding and the embeddings of the identified topics. By doing so, the most appropriate topic embedding for each outlier can be determined, thus reducing the number of documents labeled as outliers. Pre-computed embeddings were used to expedite this process and avoid redundant computation [42].

#### 5.4.3 TERM SIMILARITY SCORE

The similarity score quantifies the degree from 0 to 1 of association between a given topic and the selected term [38]. The term was selected as "depression". This score is calculated for each topic, not individual elements within topics. It should be noted that this process is not sentiment analysis, meaning it does not reflect whether people express positive or negative sentiments about depression. Rather, it simply measures the extent to which the selected term appears in the title. For example, a title with a high similarity score for depression could be one

that conveys positive emotions, even when discussing depression. Alternatively, a title with a high similarity score for depression may still convey negative emotions when discussing depression. This methodology has been employed and analyzed in tandem with LIWC and BERT agent analyses, which will be expounded upon in subsequent sections.

#### 5.4.4 CREATING NEW LABELS

BERTopic allows the developers to create new topic labels. There are several ways to optimizing topic labels. Instead of manually selecting or eliminating labels, the zero-shot classification technique with `bart-large-mnli` can be used for more accurate results by providing a list of candidate labels [43]. Several candidate labels were created for each topic labels. The algorithm assigns new topic labels to related topics based on the match score, which should be higher than selected threshold 0.9. The objective of this process is to create clearer labels by reducing similar words in the same topic label into one representative word, such as using "pregnant" instead of both "pregnancy" and "pregnant". It was decided to not using the term 'postpartum' in the new topic labels. Because all the collected information from Reddit and Twitter is about postpartum.

Following the aforementioned procedures, the gathered data is subjected to document-based and topic-based analysis to facilitate further investigation. The subsequent steps involve normalization of the data and incorporation of relevant color codes to render it suitable for visualization. Figure 5.1 represent the further analyses.

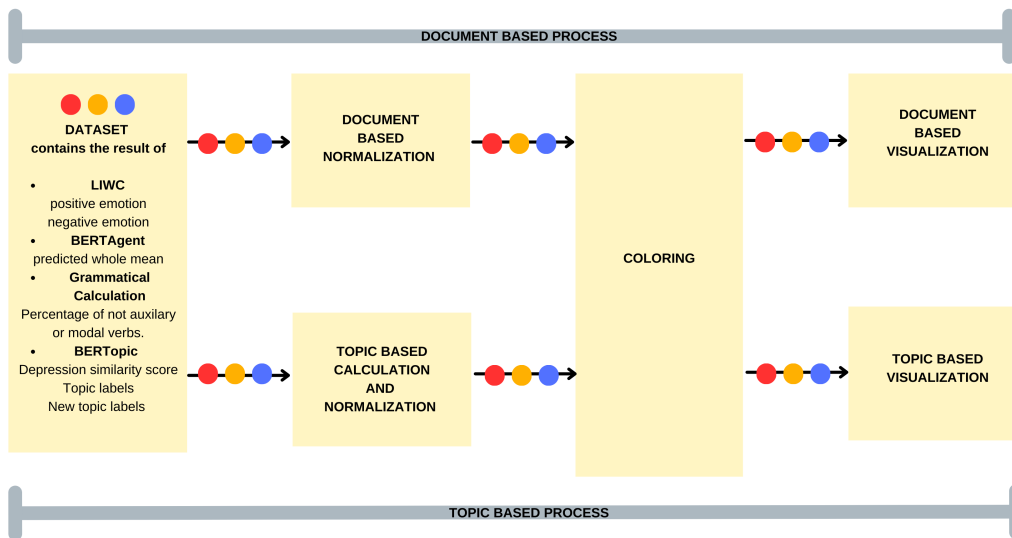


Figure 5.1: Document and Topic Based Analyses.

## 5.5 NORMALIZATION

### 5.5.1 DOCUMENT BASED NORMALIZATION

Document based normalization is a crucial process that involves organizing the predicted whole mean values obtained by applying the BERTAgent algorithm. The predicted whole mean values, which range from -3 to 3, are used to create a document-based graph. To achieve this, the new 20 ranges between -3 to 3 were initially created, and each calculated value was assigned to the appropriate category. However, it was discovered that some categories had no members, while others had few members. To address this issue, the range of predicted mean values was adjusted based on the dense area of the distribution. Values falling outside this range were rolled to the nearest border, and new 20 categories were created based on the adjusted ranges.

In summary, document-based normalization is an essential process that helps to organize predicted whole mean values obtained from the BERTAgent algorithm. The adjustments made to the ranges of predicted whole mean values ensured that all categories had an adequate number of members, thereby enhancing the accuracy of the document-based graph.

### 5.5.2 TOPIC BASED CALCULATION AND NORMALIZATION

The general idea behind the topic based analyses is creating the most clear and explanatory graphs with the calculated results. Some of the calculated values by the applied tools and methods are topic based such as term similarity score for 'depression' and some of them document based such as positive and negative emotions, predicted whole mean value, percentage of not modal and not auxiliary verbs. Different type of normalization methods applied based on the type of the values.

The similarity score for 'depression' is ranges between 0 to 1 and was created based on the topic level. The normalization process was done by applying the following formula:

$$\text{round}(\max(\min((\text{score} - \text{minimum score}) / (\text{maximum score} - \text{minimum score}), 1), 0) * 20)$$

The given formula aims to map an original term similarity score for 'depression' to an integer between 0 and 20, representing the severity of depression. The process first normalizes the original value by scaling it between 0 and 1 based on the minimum and maximum depression values. This normalization step results in a value that represents how much the original value deviates from the minimum and maximum depression values. Then clips the normalized value between 0 and 1, ensuring that the resulting value lies within the specified range. Finally, the clipped value is scaled between 0 and 20 by multiplying it by 20 and rounding it to the nearest integer. The resulting value represents the severity of depression, with higher values indicating greater severity.

The calculated values for document-based features such as positive and negative emotions, predicted mean value, and percentage of non-modal and non-auxiliary verbs were normalized using an average value calculated based on their respective topics. The normalization was achieved by applying the same formula used for term similarity score for 'depression'. The calculated average values were used to set the score, minimum score, and maximum score parameters in the normalization formula. This method allowed for the quantification and comparison of these document-based features across multiple topics, enabling the identification of trends and patterns in the data.

## 5.6 COLORING

The color scheme used in the analysis associated original calculated or normalized values with specific colors based on an ascending sort of the values. The color scale used is shown in Figure 5.2, with yellowish colors representing low values, reddish colors representing middle



values, and purpleish colors representing higher values. It is important to note that the colors were assigned based on the order of the calculated values, rather than their meaning within their respective categories. This method allowed for a visual representation of the values, enabling the identification of patterns and trends in the data.



Figure 5.2: Color Scale.

## 5.7 VISUALIZATION

In this research, a modified version of a graph type provided by BERTopic was used to enable document-based visualization. Bubble charts were employed to present the outcomes in a summary form in topic-based visualization.

### 5.7.1 DOCUMENT BASED VISUALIZATION

Visualizing BERTopic and its derivatives is crucial in comprehending the model's functionality, effectiveness, and areas of applicability. As topic modeling can be subjective, it can be challenging for users to validate their models. Therefore, examining the generated topics and determining their coherence is a vital aspect of addressing this issue [44]. BERTopic supports visualization function which is called as 'visualize documents'. Document-based visualization refers to the process of generating graphs that represent documents with topic information, which are then placed based on their related topic and colored based on the calculated values mentioned earlier. The original version of the visualize documents function is the representation of the documents placed and colored based on their related topics. The function's source code was modified to create a new way of representing the documents. The updated version of the function is now capable of color-coding each document based on specific metrics or topics, while still retaining the original location and topic information provided by the original function.

The visualize document function has been modified and divided into two parts. The first part is responsible for extracting the location information that was calculated by the algorithm and represented in Code Listing 5.6. The second part is responsible for coloring and creating the graph. To use this modified function, the color information for each document must be provided as input, which is represented in Code Listing 5.7.

**Listing 5.6:** Extraction of the Location Information from Modified Version of Visualize Documents

```
import numpy as np
import pandas as pd
import plotly.graph_objects as go
from umap import UMAP
from typing import List
def visualize_documents_coordinates(topic_model,
                                  docs: List[str],
                                  topics: List[int] = None,
                                  embeddings: np.ndarray = None,
                                  reduced_embeddings: np.ndarray = None,
                                  sample: float = None,
                                  hide_annotations: bool = False,
                                  hide_document_hover: bool = False,
                                  custom_labels: bool = False,
                                  width: int = 1200,
                                  height: int = 750):

    topic_per_doc = topic_model.topics_
    # Sample the data
    if sample is None or sample > 1:
        sample = 1
    indices = []
    for topic in set(topic_per_doc):
        s = np.where(np.array(topic_per_doc) == topic)[0]
        size = len(s) if len(s) < 100 else int(len(s) * sample)
        indices.extend(np.random.choice(s, size=size, replace=False))
    indices = np.array(indices)
    df = pd.DataFrame({"topic": np.array(topic_per_doc)[indices]})
    df["doc"] = [docs[index] for index in indices]
    df["topic"] = [topic_per_doc[index] for index in indices]
    df["original_index"]=[index for index in indices]
    # Extract embeddings if not already done
    if sample is None:
        if embeddings is None and reduced_embeddings is None:
            embeddings_to_reduce =
                topic_model._extract_embeddings(
                    df.doc.to_list(), method="document")
        else:
            embeddings_to_reduce = embeddings
    else:
        if embeddings is not None:
            embeddings_to_reduce = embeddings[indices]
        elif embeddings is None and reduced_embeddings is None:
            embeddings_to_reduce =
                topic_model._extract_embeddings(
                    df.doc.to_list(),
                    method="document")
    # Reduce input embeddings
    if reduced_embeddings is None:
        umap_model = UMAP(n_neighbors=10, n_components=2,
                          min_dist=0.0, metric='cosine').fit(embeddings_to_reduce)
        embeddings_2d = umap_model.embedding_
    elif sample is not None and reduced_embeddings is not None:
        embeddings_2d = reduced_embeddings[indices]
    elif sample is None and reduced_embeddings is not None:
        embeddings_2d = reduced_embeddings
    unique_topics = set(topic_per_doc)
    if topics is None:
        topics = unique_topics
    # Combine data and save the original location information
    df["x"] = embeddings_2d[:, 0]
    df["y"] = embeddings_2d[:, 1]
    df_coordinates=df
    return df_coordinates, topic_per_doc
```

### Listing 5.7: Modified Version of Visualize Documents Graph

```

def visualize_documents_graph(topic_model, df_coordinates,
                              df_main: pd.DataFrame = None,
                              color_column_name: str = None,
                              #unique_topics: List[int] = None,
                              docs: List[str] = None,
                              topics: List[int] = None,
                              embeddings: np.ndarray = None,
                              reduced_embeddings: np.ndarray = None,
                              sample: float = None,
                              hide_annotations: bool = False,
                              hide_document_hover: bool = False,
                              custom_labels: bool = False,
                              width: int = 1200,
                              height: int = 750):
    topic_per_doc=topic_model.topics_
    unique_topics = set(topic_per_doc)
    if topics is None:
        topics = unique_topics
    #--- MATCHING ALGORITHM---#
    if df_main is not None: # for modified graph
        df=df_coordinates.copy()
        df_main['original_index']=df_main.index
        df_main['original_index_for_order']=df_main.index
        df_merged=pd.merge(df,df_main,on='original_index')
        df.drop('original_index', inplace=True, axis=1)
        #color_column=df_merged[color_column_name]
    else:
        df=df_coordinates.copy()
        df.drop('original_index', inplace=True, axis=1)
    #####
    # Prepare text and names
    if topic_model.custom_labels_ is not None and custom_labels:
        names = [topic_model.custom_labels_[topic + topic_model._outliers] for topic in unique_topics]
    else:
        names = [f"{topic}_" + "_" .join([word for word, value in topic_model.get_topic(topic)][:3]) for topic in unique_topics]

    ## ORIGINAL GRAPH ##

    # Visualize
    fig = go.Figure()

    # Outliers and non-selected topics
    non_selected_topics = set(unique_topics).difference(topics)
    if len(non_selected_topics) == 0:
        non_selected_topics = [-1]

    selection = df.loc[df.topic.isin(non_selected_topics), :]
    selection["text"] = ""
    selection.loc[len(selection), :] = [None,
                                         None,
                                         selection.x.mean(),
                                         selection.y.mean(),
                                         "Other documents"]

    fig.add_trace(
        go.Scattergl(
            x=selection.x,
            y=selection.y,
            hovertext=selection.doc if not hide_document_hover else None,
            hoverinfo="text",
            mode='markers+text',
            name="other",
            showlegend=False,
            marker=dict(color='#CFD8DC', size=5, opacity=0.5)
        )
    )

```

```

# Selected topics
for name, topic in zip(names, unique_topics):
    if topic in topics and topic != -1:
        selection = df.loc[df.topic == topic, :]
        selection["text"] = ""

    if df_main is not None: # for modified graph

        color=df_merged[color_column_name].loc[df_merged.topic_number == topic]

    if not hide_annotations:
        selection.loc[len(selection), :] = [None, None, selection.x.mean(), selection.y.mean(), name]

    if df_main is not None: # for modified graph
        fig.add_trace(
            go.Scattergl(
                x=selection.x,
                y=selection.y,
                hovertext=selection.doc if not hide_document_hover else None,
                hoverinfo="text",
                text=selection.text,
                mode='markers+text',
                name=name,
                textfont=dict(
                    size=12,
                ),
                #marker=dict(size=5, opacity=0.5)
                marker=dict(color=color, size=5, opacity=0.5)
            )
        )

    else:
        fig.add_trace(
            go.Scattergl(
                x=selection.x,
                y=selection.y,
                hovertext=selection.doc if not hide_document_hover else None,
                hoverinfo="text",
                text=selection.text,
                mode='markers+text',
                name=name,
                textfont=dict(
                    size=12,
                ),
                marker=dict(size=5, opacity=0.5)

                #marker=dict(color=color, size=5, opacity=0.5)
            )
        )

# Add grid in a 'plus' shape
x_range = (df.x.min() - abs((df.x.min()) * .15), df.x.max() + abs((df.x.max()) * .15))
y_range = (df.y.min() - abs((df.y.min()) * .15), df.y.max() + abs((df.y.max()) * .15))
fig.add_shape(type="line",
              xo=sum(x_range) / 2, yo=y_range[0], x1=sum(x_range) / 2, y1=y_range[1],
              line=dict(color="#CFD8DC", width=2))
fig.add_shape(type="line",
              xo=x_range[0], yo=sum(y_range) / 2, x1=x_range[1], y1=sum(y_range) / 2,
              line=dict(color="#9E9E9E", width=2))
fig.add_annotation(x=x_range[0], y=sum(y_range) / 2, text="D1", showarrow=False, yshift=10)
fig.add_annotation(y=y_range[1], x=sum(x_range) / 2, text="D2", showarrow=False, xshift=10)

# Stylize layout
fig.update_layout(
    template="simple_white",
    title={

```

```

        'text': "<b>Documents and Topics",
        'x': 0.5,
        'xanchor': 'center',
        'yanchor': 'top',
        'font': dict(
            size=22,
            color="Black")
    },
    width=width,
    height=height
)

fig.update_xaxes(visible=False)
fig.update_yaxes(visible=False)

return fig

```

### 5.7.2 TOPIC BASED NORMALIZATION

The topic-based representation is a visualization technique that presents the calculated values in the form described by the topic-based normalization process. It utilizes a bubble chart to represent each topic, where the size of each bubble corresponds to the size of the community associated with that topic. The center point of the extracted location information in the document visualization graph was used to position each bubble. To prevent overlap on the graph, some relocation is applied. The calculated values are used to assign a color to each topic. This method enables a comprehensive comparison between the calculated values and provides a summary of the project outcome.

# 6

## Application of the Tools and Methods

This section presents the application of all the tools and methods introduced in the previous section to three different datasets: Twitter, Reddit comments, and Reddit posts. The common parameters for the algorithms were explained in the previous section, and the parameters that affect the grouping of data into topic labels in the topic model were described in detail.

The project outcome was presented through graphs, and the inter-relationships and intra-relationships between the created topics and the graphs created based on different aspects. Qualitative analysis was conducted to further interpret these relationships.

## 6.1 TWITTER

The number of the input documents for the BERTopic model is 17664. After numerous iterations of parameter tuning, it was determined that the BERTopic model yielded the most semantically meaningful clusters for a Twitter corpus consisting of 17,664 input documents when using the following parameter values:

- `n_neighbors` in UMAP: 250,
- `min_cluster_size` in HDBSCAN: 80,
- `diversity` in topic model: 0.5.

All the other parameters were kept as the same as mentioned previous section. After training the topic model with the previously mentioned parameters, all 17664 input documents were assigned to one of the 20 clusters, each labeled with a topic label and a cluster number ranging from 0 to 19 as determined by the algorithm. Any documents that were not assigned to one of these 20 clusters were instead assigned to the outlier cluster, which was labeled with a cluster number of -1 and did not have a corresponding topic label.

The outlier reduction process is implemented using an embedding method to ensure that no documents are left unclustered. Any documents identified as outliers are then assigned to the most appropriate cluster. However, the addition of new documents to a cluster may result in changes to the cluster's size and topic label, as the topic label is a summary representation of the members within that cluster. Thus, the incorporation of new members may cause the cluster to expand and result in a change in its overall topic representation.

The topic model was updated with the newly gathered labels and subsequently used to generate a visualization of the documents. This visualization allows for a more clear understanding of the clustering results by displaying the each document as dot. Figure 6.1 depicts the initial topic labels and the updated topic labels, along with the corresponding number of documents allocated to each topic. The initial classification process had assigned 7520 documents to topic -1, which was an outlier. However, after implementing outlier reduction techniques, these documents were reassigned to the most relevant topics, leading to the elimination of topic -1. Although the updated labels included some synonymous or similar terms, they still shared common representative keywords with the original labels, and there were no significant differences between the two.

topic_number	count_before	count_later	original_topic_labels	ouliers_reduced_topic_labels
-1	7520	0	-1_postpartum_depression_woman_month	This topic does not exist any more.
0	970	1575	0_postpartum_gift_tea_like	0_postpartum_gift_mom_day
1	937	961	1_grow_postpartum_bald_thin	1_hair_loss_postpartum_hairless
2	829	1044	2_feel_postpartum_shit_like	2_feel_postpartum_like_shit
3	797	972	3_depression_postpartum_feel_thing	3_depression_postpartum_feel_like
4	697	902	4_support_postpartum_service_guide	4_support_care_postpartum_help
5	653	745	5_belly_belt_underwear_trainer	5_waist_belt_belly_wear
6	587	879	6_postpartum_week_feel_change	6_body_week_look_postpartum
7	561	1118	7_depression_postpartum_woman_mob	7_depression_postpartum_woman_mother
8	480	904	8_postpartum_pelvic_floor_recovery	8_postpartum_pelvic_recovery_floor
9	471	537	9_month_postpartum_illinois_extension	9_coverage_medical_extend_state
10	388	454	10_lbs_week_postpartum_pregpregnancy	10_weight_lose_lbs_pound
11	359	1549	11_depression_postpartum_symptom_factor	11_depression_postpartum_psychosis_woman
12	338	474	12_health_depression_support_postpartum	12_mental_health_depression_support
13	336	363	13_exercise_fitness_yoga_photo	13_workout_exercise_fitness_gym
14	333	708	14_black_preeclampsia_pregnancy_abortion	14_black_woman_death_pregnancy
15	308	1234	15_postpartum_hospital_delivery_nursing	15_care_postpartum_nurse_baby
16	299	367	16_anxiety_postpartum_depression_feel	16_anxiety_postpartum_fear_feel
17	292	324	17_depression_talk_postpartum_need	17_depression_talk_help_postpartum
18	259	316	18_hormone_emotion_rage_postpartum	18_hormone_cry_rage_emotion
19	250	2238	19_breastfeed_postpartum_start_pregnant	19_postpartum_month_pregnancy_pregnant

Figure 6.1: Original Topic Labels vs. Outliers Reduced Topic Labels for Twitter.

Figures 6.2 and 6.3 display the term similarity score for 'depression's for each topic. The majority of topics in Figure 6.2 have a term similarity score for 'depression' between 0.4 and 0.5. Topics 3, 7, 11, 12, 16, 17 and 18 exhibit the strongest association with the term "depression". It should be noted that most of these topics contain the term "depression" or related words such as "anxiety" and "cry" in their labels.



topic	Count	topic_label	detailed_labels	depression_similarity_score
0	0	1575_0_postpartum_gift_mom_day	0_postpartum_gift_mom_day_baby_like_new_box_go_get	0.4488331994382757
1	1	961_1_hair_loss_postpartum_hairless	1_hair_loss_postpartum_hairless_grow_growth_fall_shed_bald_thin	0.4869668322058158
2	2	1044_2_feel_postpartum_like_shit	2_feel_postpartum_like_shit_go_hard_bad_baby_hate_suck	0.4522059932300161
3	3	972_3_depression_postpartum_feel_like	3_depression_postpartum_feel_like_real_go_know_people_think_talk	0.7667300444500641
4	4	902_4_support_care_postpartum_help	4_support_care_postpartum_help_need_parent_work_new_service_birth	0.44999719328260857
5	5	745_5_waist_belt_belly_wear	5_waist_belt_belly_wear_underwear_clothe_trainer_band_postpartum_fit	0.350878928263511
6	6	879_6_body_week_look_postpartum	6_body_week_look_postpartum_photo_month_like_love_skin_good	0.4536526026029556
7	7	1118_7_depression_postpartum_woman_mother	7_depression_postpartum_woman_mother_help_suffer_download_mental_baby_therapy	0.8666611116790717
8	8	904_8_postpartum_pelvic_recovery_floor	8_postpartum_pelvic_recovery_floor_day_like_go_know_baby_week	0.473029017774046
9	9	537_9_coverage_medical_extend_state	9_coverage_medical_extend_state_expand_health_year_bill_care_maternal	0.4502769026404345
10	10	454_10_weight_lose_lbs_pound	10_weight_lose_lbs_pound_week_gain_postpartum_pregnancy_day_body	0.40663436018496135
11	11	1549_11_depression_postpartum_psychosis_woman	11_depression_postpartum_psychosis_woman_symptom_mental_risk_disorder_experience_baby	0.8718343495881153
12	12	474_12_mental_health_depression_support	12_mental_health_depression_support_maternal_help_study_postpartum_care_dr	0.6880051203897752
13	13	363_13_workout_exercise_fitness_gym	13_workout_exercise_fitness_gym_yoga_run_photo_postpartum_body_week	0.4412865627154632
14	14	708_14_black_woman_death_pregnancy	14_black_woman_death_pregnancy_die_blood_risk_postpartum_maternal_hemorrhage	0.5092043024433028
15	15	1234_15_care_postpartum_nurse_baby	15_care_postpartum_nurse_baby_hospital_birth_delivery_newborn_pregnant_woman	0.46784852205396965
16	16	367_16_anxiety_postpartum_fear_feel	16_anxiety_postpartum_fear_feel_depression_symptom_help_bad_mom_have	0.6995151538797185
17	17	324_17_depression_talk_help_postpartum	17_depression_talk_help_postpartum_need_know_suffer_people_woman_mom	0.7723097838010395
18	18	316_18_hormone_cry_rage_emotion	18_hormone_cry_rage_emotion_postpartum_emotional_reason_bitch_like_hormonal	0.5809625305672812
19	19	2238_19_postpartum_month_pregnancy_pregnant	19_postpartum_month_pregnancy_pregnant_breastfeed_woman_baby_period_week_time	0.4452189902135325

Figure 6.2: Term Similarity Score for 'Depression' for Twitter.

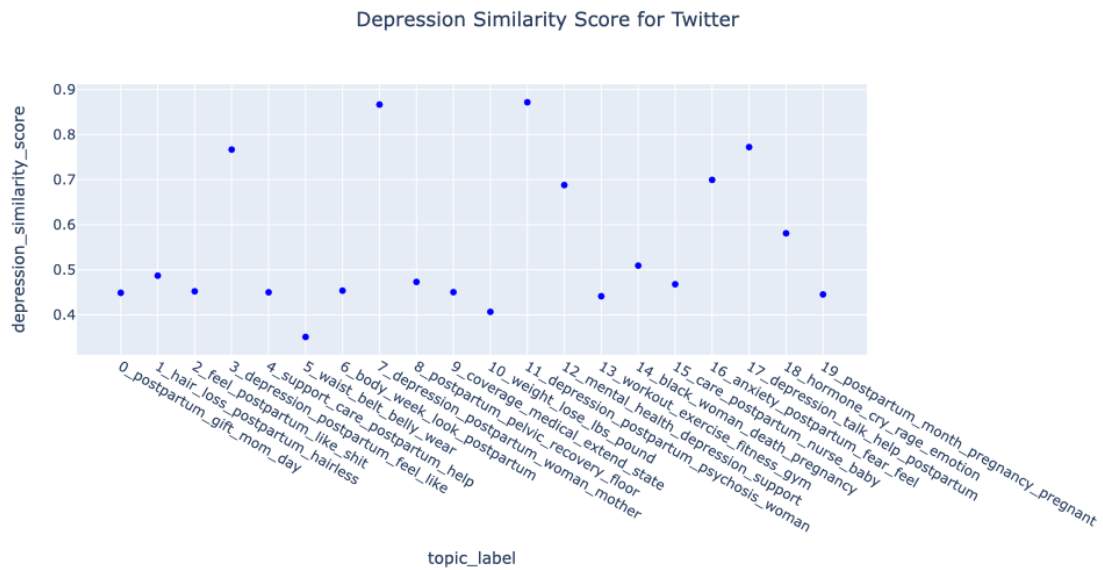


Figure 6.3: Distribution of Term Similarity Score for 'Depression' for Twitter.

The term similarity score for 'depression' was normalized to a range of 0 to 20, as described in the normalization section, and color-coded. Figure 6.4 displays the normalized term similarity score for 'depression' for each topic using the assigned colors, with yellowish hues representing less related topics, orangish and reddish hues representing moderately related topics, and purple and darker hues representing highly related topics. However, topic 5 cannot be observed on the bar graph, despite being assigned the color yellow, due to it has the lowest value observed in Figure 6.3 and its corresponding normalization to a score of 0.

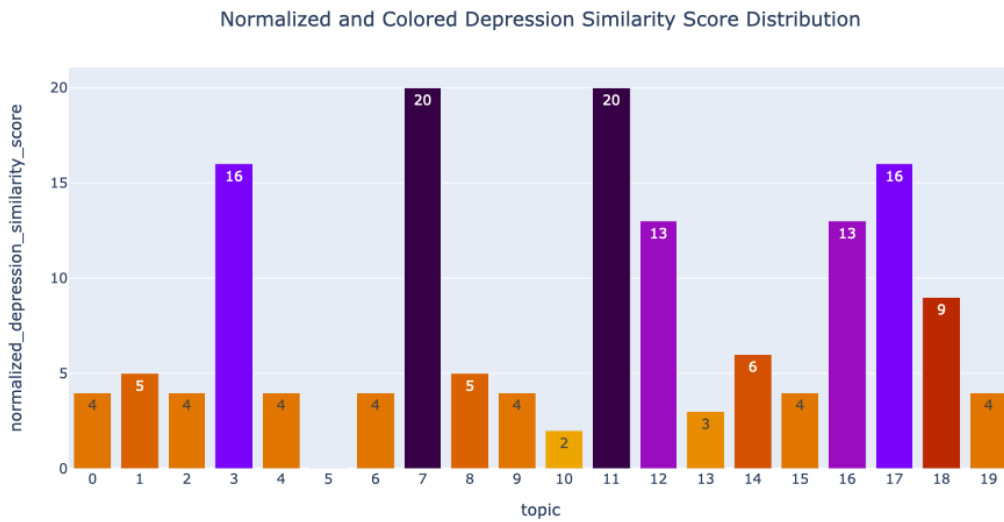


Figure 6.4: Normalized and Colored of Term Similarity Score for 'Depression' Distribution for Twitter.

Figure 6.5 displays the distribution of predicted whole mean value created by the BERTAgent algorithm. The majority of documents yield values between -1.5 and 1.5. Therefore, a normalization range of [-1.5, 1.5] was chosen, and values outside of this range were rolled to the nearest range borders. Figure 6.6 illustrates the distribution of the normalized predicted whole mean values across 20 new ranges created between the selected borders. Each range was assigned a color corresponding to its respective category, which is represented by the color of the bars in the figure.

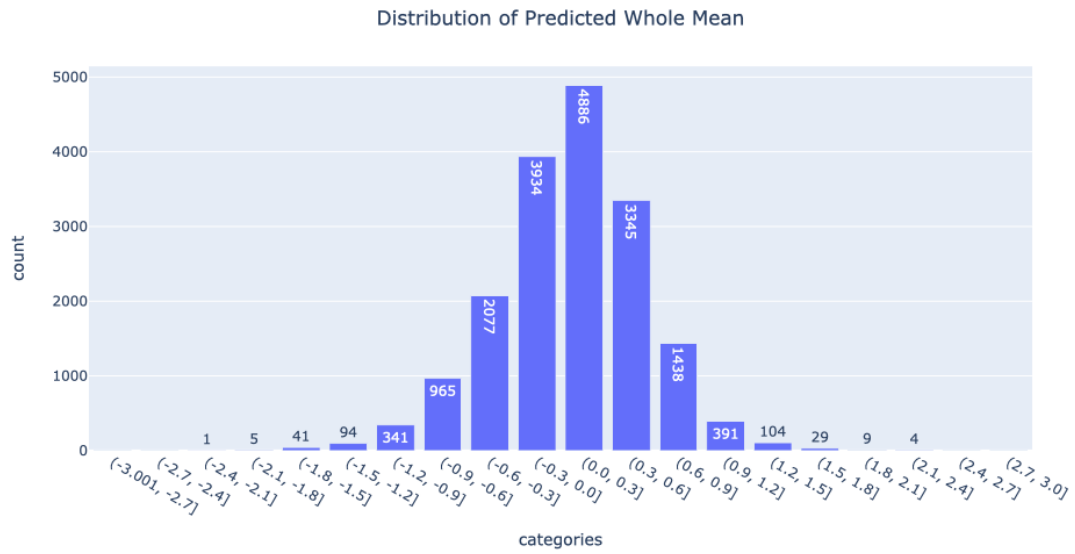


Figure 6.5: Distribution of Predicted Whole Mean for Twitter.

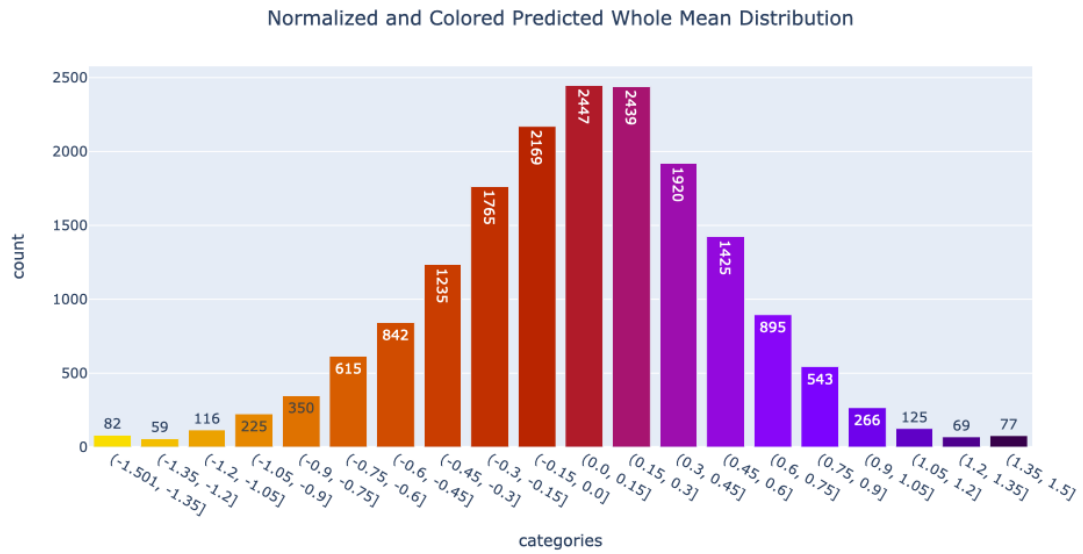


Figure 6.6: Normalized and Colored of Predicted Whole Mean Distribution for Twitter.

### Documents and Topics

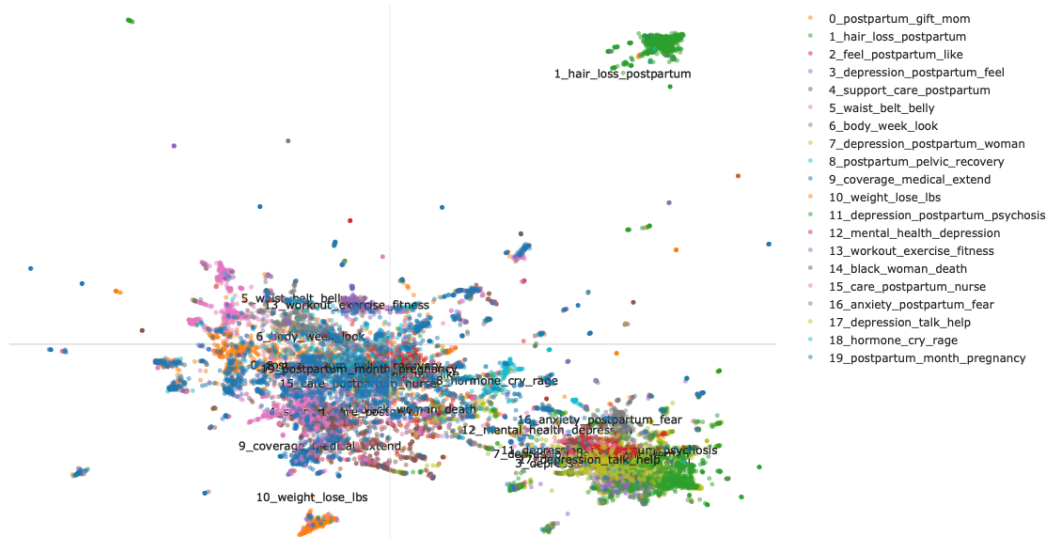


Figure 6.7: Twitter Document Visualization.

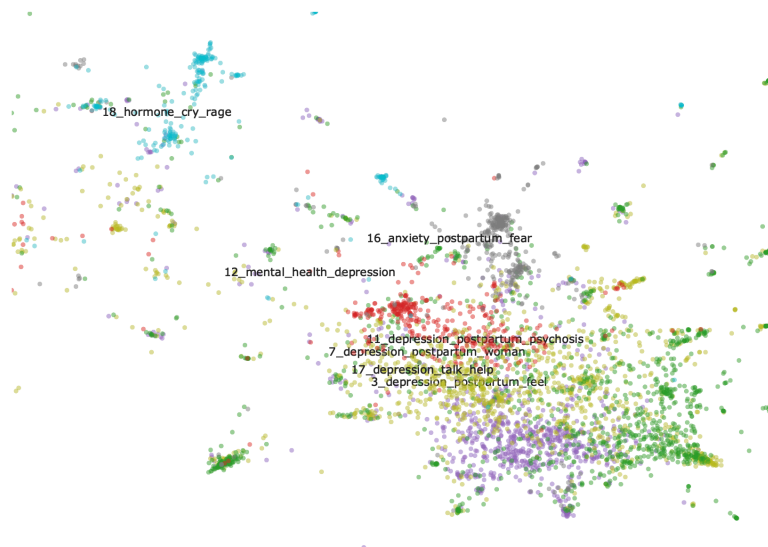
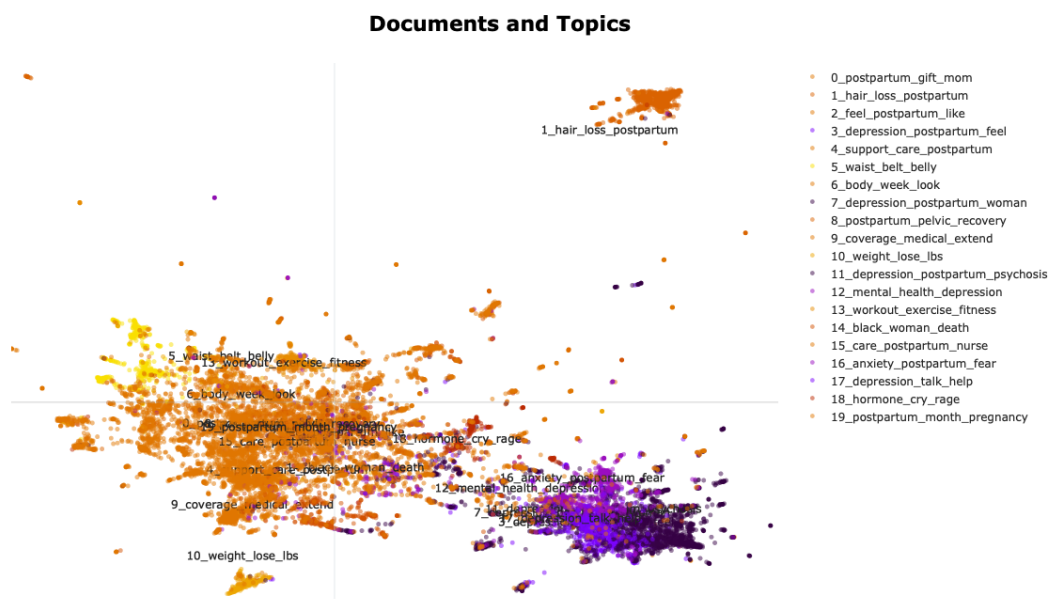


Figure 6.8: Twitter Document Visualization for Depression Topics.

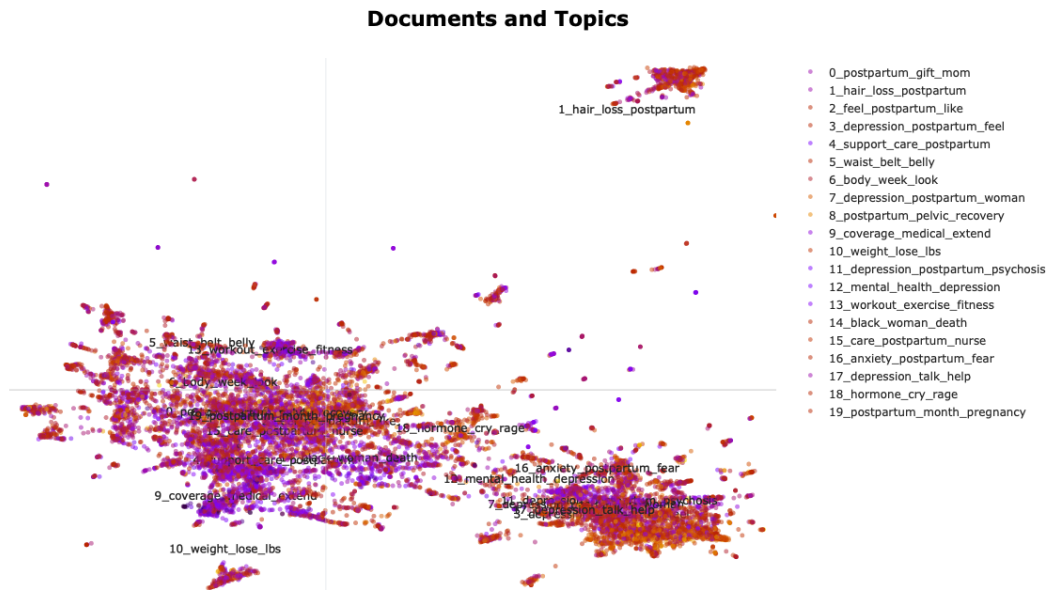
Figure 6.7 represents the document visualization. Each topic has the different colors and the topics which have the high term similarity score for 'depression' were placed on the lower right corner. This can be seen more clear in Figure 6.8 which is the filtered version of the previous graph for topics 3, 7, 11, 12, 16, 17 and 18.

The placement of topic 1 in the upper right corner, which is distant from most of the other topics, raised concerns about its relevance. To evaluate its meaningfulness, it was checked the documents assigned to this topic and found that they exclusively pertained to the hair loss and growth. Based on this analysis, it was decided to retain this topic in the analysis.



**Figure 6.9:** Twitter Document Visualization of Term Similarity Score for 'Depression'.

Figure 6.9 depicts the document visualization for term similarity score for 'depression', created by using normalized term similarity score for 'depression'. As previously discussed, the topics most closely associated with the term "depression" are located in the lower right corner of the graph, which is characterized by majority of purple color. This provides evidence to support the claim that this portion of the graph can be identified as the "depression topics" region in this research. Depression topics region does not represent any sentiment such as negative emotions. It only shows that the topics in that area are about the term 'depression'.



**Figure 6.10:** Document Visualization of BERTAgent for Twitter.

Figure 6.10 illustrates a document-based visualization of the predicted whole mean value obtained using the BERTAgent algorithm. The document clusters were kept in their original coordinates, and each document was color-coded based on its normalized predicted whole mean value can be seen in figure 6.6. The majority of the documents are observed to be colored reddish, which suggests that a significant number of documents have middle-to-high predicted whole mean values. However, it is worth noting that lighter shades can also be observed beneath the reddish colors.

This visualization provides a general overview of the distribution of predicted whole mean values across documents. However, for a more representative analysis of the algorithm outcomes, topic-based analyses were applied.

The new labels used in representing topic-based graphs are a more concise and precise version of the labels employed in earlier analyses. The figure below illustrates the evolution of the topic labels throughout the process, which includes the initial extraction of topic labels, outlier reduction, and creation of new labels.

topic_number	original_topic_labels	ouliers_reduced_topic_labels	new_labels
-1	-1_postpartum_depression_woman_month	This topic does not exist any more.	This topic does not exist any more.
0	0_postpartum_gift_tea_like	0_postpartum_gift_mom_day	gift_box_shop_baby
1	1_grow_postpartum_bald_thin	1_hair_loss_postpartum_hairless	hair loss and growth
2	2_feel_postpartum_shit_like	2_feel_postpartum_like_shit	bad feelings_hard
3	3_depression_postpartum_feel_thing	3_depression_postpartum_feel_like	depression_feel
4	4_support_postpartum_service_guide	4_support_care_postpartum_help	care_birth_support_parenting
5	5_belly_belt_underwear_trainer	5_waist_belt_belly_wear	underwear_belt_belly_fit_trainer
6	6_postpartum_week_feel_change	6_body_week_look_postpartum	body look_photo_good feelings
7	7_depression_postpartum_woman_mob	7_depression_postpartum_woman_mother	depression_maternity_suffer_therapy_help
8	8_postpartum_pelvic_floor_recovery	8_postpartum_pelvic_recovery_floor	pelvic floor_recovery_time
9	9_month_postpartum_illinois_extension	9_coverage_medical_extend_state	coverage_medical_health_bill_care
10	10_lbs_week_postpartum_pregnancy	10_weight_lose_lbs_pound	weight_lose_gain
11	11_depression_postpartum_symptom_factor	11_depression_postpartum_psychosis_woman	depression_psychosis_symptom_experience
12	12_health_depression_support_postpartum	12_mental_health_depression_support	mental health_depression_maternal_support
13	13_exercise_fitness_yoga_photo	13_workout_exercise_fitness_gym	workout_photo
14	14_black_preeclampsia_pregnancy_abortion	14_black_woman_death_pregnancy	hemorrhage_blood_death
15	15_postpartum_hospital_delivery_nursing	15_care_postpartum_nurse_baby	hospital_health care personnel_birth
16	16_anxiety_postpartum_depression_feel	16_anxiety_postpartum_fear_feel	anxiety_fear_symptom_depression_bad feelings
17	17_depression_talk_postpartum_need	17_depression_talk_help_postpartum	depression_talk_sharing experience
18	18_hormone_emotion_rage_postpartum	18_hormone_cry_rage_emotion	hormone_emotion_cry
19	19_breastfeed_postpartum_start_pregnant	19_postpartum_month_pregnancy_pregnant	time_breastfeed

Figure 6.11: Label Evolution for Twitter.

Figures 6.12 to 6.17 depict the results of a topic-based analysis, which was conducted using a normalization process described in the "Topic Based Analyses and Normalization" subsection of the "Tools and Methods" section. Various calculated information, such as positive and negative emotions calculated by LIWC, composite emotion scores, predicted whole mean scores by BERTAgent, term similarity score for 'depression' by BERTopic, and non-modal/non-auxiliary verbs identified through grammatical calculation algorithms, were included in the analysis to create these graphs.

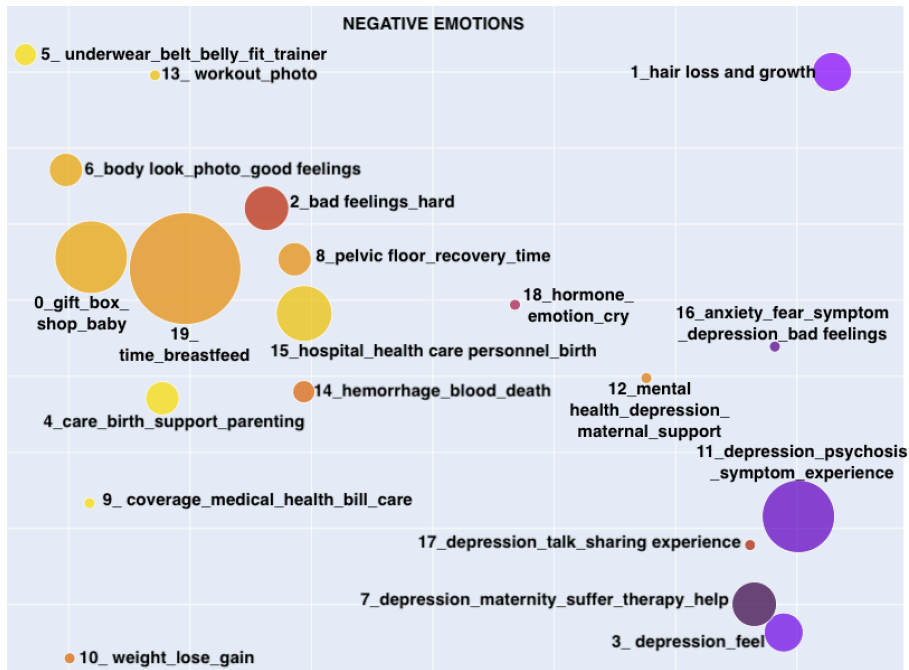


Figure 6.12: Topic Based Negative Emotions for Twitter.

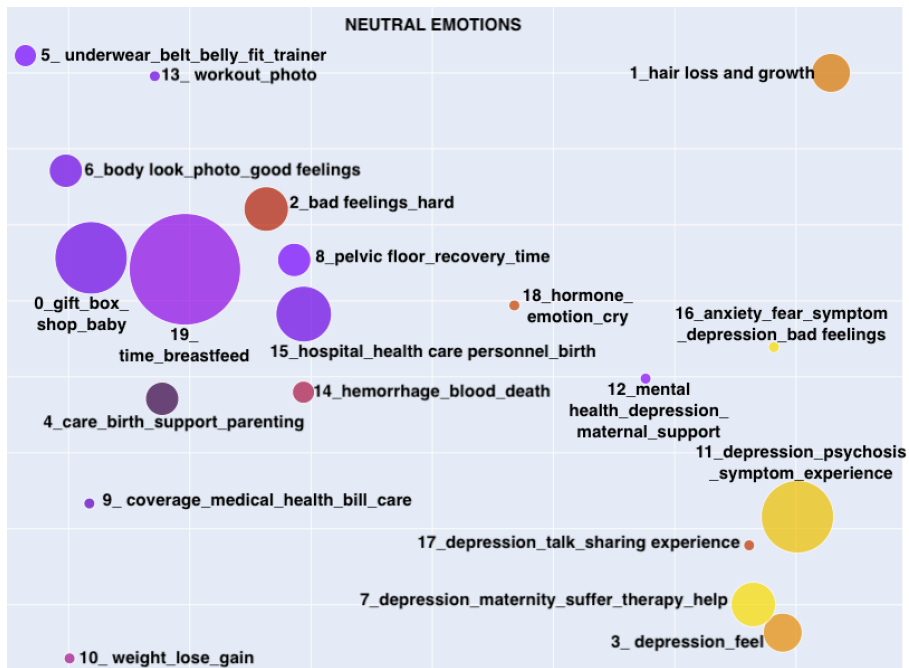


Figure 6.13: Topic Based Composite Emotion Scores for Twitter.



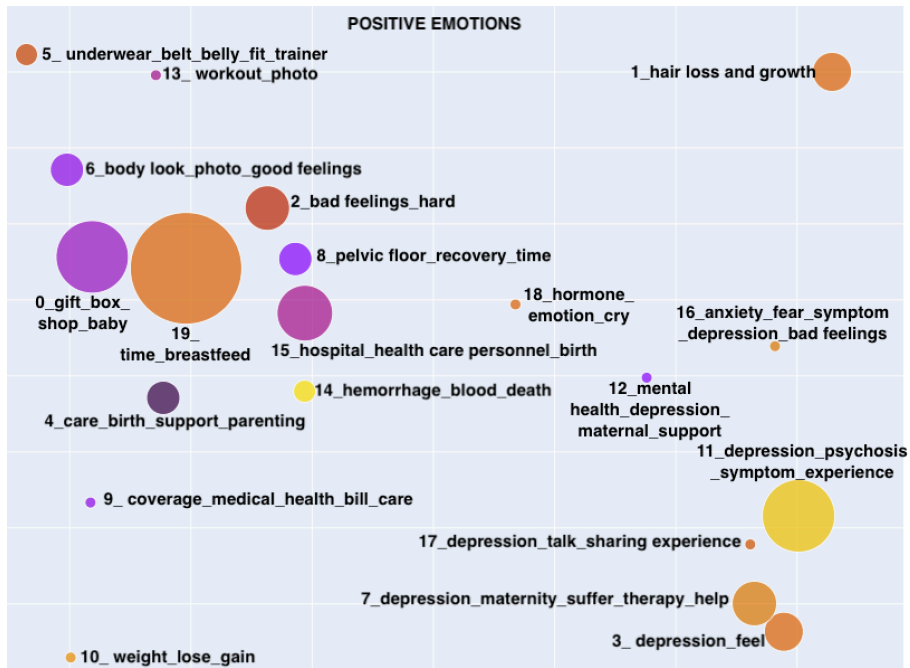


Figure 6.14: Topic Based Positive Emotions for Twitter.

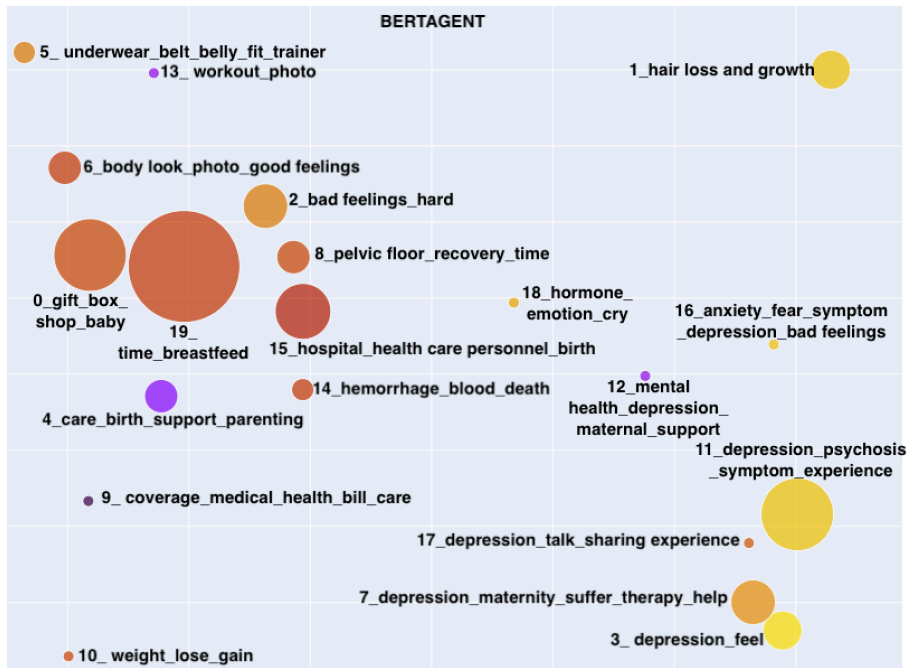


Figure 6.15: Topic Based BERTAgent for Twitter.

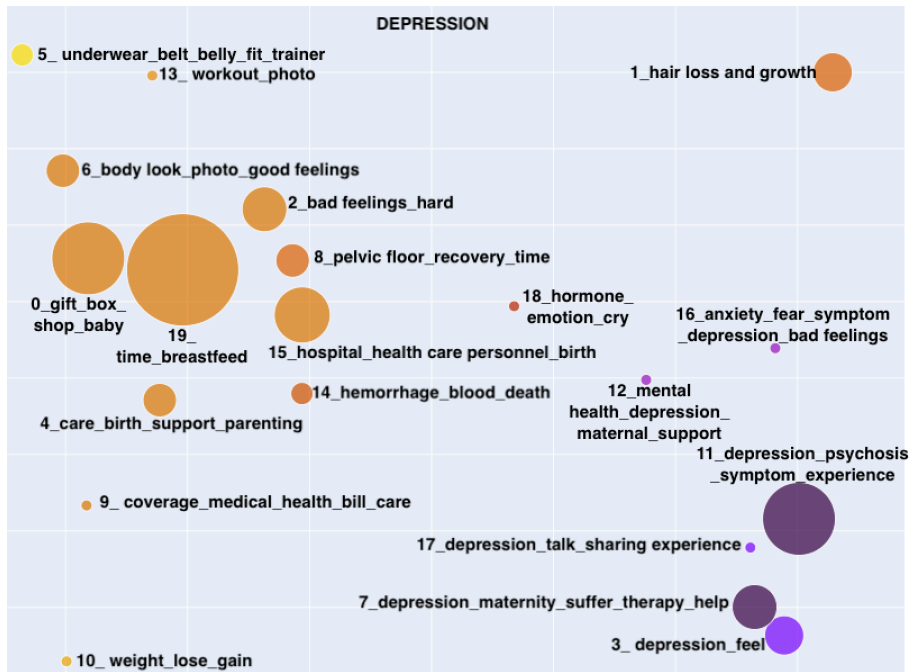


Figure 6.16: Topic Based Term Similarity Score for 'Depression' for Twitter.

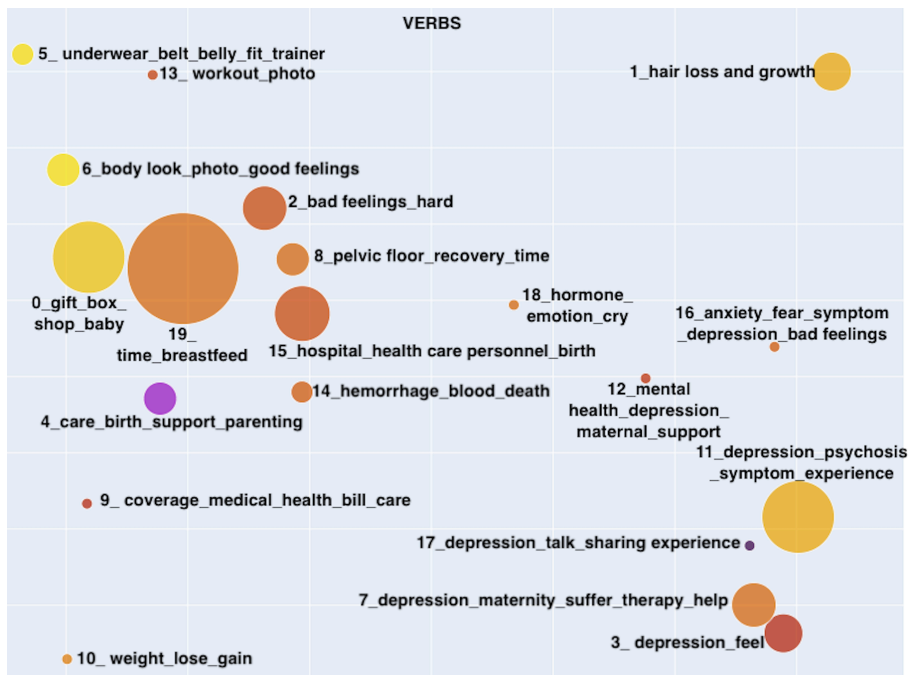


Figure 6.17: Topic Based Verbs for Twitter.

Figure 6.12 uses a color scale that ranges from yellow (low value) to dark purple (high value) to represent negative emotions. The right side of the graph is predominantly colored in dark purple, indicating a high score for negative emotions. Topics 1, 16, 11, 7, and 3 are all colored purple, indicating a high level of negative emotion. Topics 18, 2, and 17 have reddish-dark orange colors, indicating a lower but still relatively high level of negative emotion. The area that covers these topics was referred to as the "negative area" in the research. Only topic 12 has an orange color, which means it has a negative value lower than the middle and is placed close to the negative side of the graph.

In Figure 6.14, which represents positive emotions, the brighter colors are on the right side of the graph, indicating a low value for the mean of being positive in that area and was mentioned as negative area before. Topic 12 has a dark color in that area, indicating a high value of positive emotion. As previously mentioned, topic 12 was orange in the negative emotions graph, indicating a low value of negative emotion while being placed in the negative area. Furthermore, it was colored dark in figure 6.15 for term similarity score for 'depression' and in figure 6.16 for BERTAgent which means high value of relation with the term depression with positive agency. This suggests that topic 12 contains positive expressions about negative subjects in general. It is worth noting that the label for topic 12 contains keywords such as mental health, depression, maternal, and support. Some documents from topic 12:

- 'Did you know that 1 in 7 mothers will experience Post Partum Depression? Mothers and Babies is a 6 week long course designed to help pregnant and new moms navigate the journey of motherhood. Join us for our next session!'
- 'I had a personal trainer I supported back in the day reach out to me because she noticed how my postpartum body affected my mental health. sis said enjoy your birthday then hit me with a she got me'

A comparison between the similarity score for depression graph (Figure 6.15) and negative emotions graph (Figure 6.12) reveals that it cannot be concluded that all topics with a high term similarity score for 'depression' also have a high negative emotion score or vice versa. The lower right side of Figure 6.15 covers topics 3, 7, 11, 12, 16, 17, 18 can be referred to as the 'depression area', as it was previously labeled in the document visualization of this research. The depression area (topics 3, 7, 11, 12, 16, 17, 18) and negative area (topics 1, 3, 7, 11, 16, 17 and 18) have many common topics, but they also exhibit some differences. The examination of topic 12 provides evidence for this claim. Additionally, topic 1, which is labeled with the keywords hair loss and growth, has a high value of negative score while having a middle value

of term similarity score for 'depression'. This suggests that the expressions in that topic are negative but not necessarily related to depression.

Based on the comparison of the BERTAgent results to the deductions presented, it was observed that the right side of Figure 6.15, with the exception of topic 12, exhibited yellowish hues, which can be categorized as the "low agency area" in this study. The low agency area encompassed topics 1, 2, 3, 7, 11, 16, 17, and 18, which coincided with the negative area. Therefore, it can be inferred that the low agency area corresponds to negative emotions, which may or may not relate to depression.

The common topics among areas are 3, 7, 11, 16, 17, 18 and characterized by negative expressions related to depression. For instance, topic 11 is labeled with keywords depression, psychosis, symptom and experience, and some of the documents within this topic are:

- 'My last pregnancy was so agonizing that it caused suicidal prenatal depression and years of postpartum depression and PTSD and chronic pain. And that is what you call convenience. You are an abuser.'
- 'Postpartum depression is seriously the worst'
- 'I am saddened and jealous that my mom is able to tend to everyone else postpartum BUT me. '
- 'The possibility of postpartum depression again makes me sad '

In Figure 6.13, the graph depicting composite emotion scores clearly illustrates two distinct poles. The left side of the graph, characterized by darker colors, denotes a strong presence of positive emotions, while the right side, represented by lighter colors, indicates the presence of negative emotions. These observations were examined through examples and comparisons in the previous analyses.

Figure 6.17 presents a topic-based representation of verbs, and despite being scattered, no discernible pattern could be identified upon comparison with the previous graphs.

## 6.2 REDDIT POST

The BERTopic algorithm was applied to 44025 documents using a set of parameters that were selected after multiple iterations.

- n\_neighbors in UMAP: 200,
- min\_cluster\_size in HDBSCAN: 150,
- diversity in topic model: 0.5.

The objective of the iterations was to obtain the best possible clustering and visualization of the dataset. The resulting topic model contained 21 clusters, out of which 20 clusters were deemed meaningful, and one cluster was designated for the outliers. The outliers were re-assigned to the most related cluster using the reducing outliers algorithm, and as a result, the outlier cluster became an empty cluster. The algorithm then renamed the remaining clusters with the new members, and this can be observed in Figure 6.18.

topic_number	count_before	count_later	original_topic_labels	outliers_reduced_topic_labels
-1	22490	0	-1_feel_time_day_say	This label does not exist anymore.
0	2132	2778	0_bear_newborn_give_feel	0_baby_bear_birth_old
1	1948	2903	1_parent_want_feel_time	1_mom_parent_daughter_child
2	1876	2847	2_pregnancy_week_exercise_yoga	2_pregnancy_pregnant_week_get
3	1581	1988	3_diet_calorie_week_postpartum	3_weight_eat_lose_gain
4	1446	2140	4_pump_formula_feed_latch	4_breastfeed_pump_breast_milk
5	1423	3194	5_postpartum_week_hormone_period	5_postpartum_week_month_baby
6	1367	3427	6_anxiety_postpartum_stress_diagnose	6_depression_anxiety_postpartum_mental
7	1093	2936	7_marriage_say_want_time	7_husband_say_want_time
8	1088	2819	8_feel_upset_look_hate	8_feel_like_feeling_know
9	980	2122	9_pain_section_preeclampsia_postpartum	9_pain_section_hurt_painful
10	959	1818	10_sleep_morning_insomnia_sleeper	10_sleep_night_wake_tired
11	910	2067	11_pad_underwear_dress_fit	11_wear_clothe_pad_underwear
12	792	2956	12_story_experience_expect_tell	12_thank_experience_advice_know
13	740	1606	13_hi_think_post_reason	13_hi_think_hello_post
14	672	2384	14_hospital_midwife_hour_induction	14_hospital_nurse_doctor_baby
15	527	568	15_nan_nederland_kyleena_ammonium	15_nan_calcium_yada_high
16	520	957	16_family_friend_aunt_say	16_family_sister_husband_brother
17	507	1083	17_mil_home_visit_stay	17_mil_house_home_visit
18	489	919	18_love_woman_feel_drive	18_love_sex_relationship_woman
19	485	2513	19_work_go_month_ft	19_week_work_go_day

Figure 6.18: Original Topic Labels vs. Outliers Reduced Topic Labels for Reddit Post.

Figures 6.19 and 6.20 represent the term similarity score for 'depression's for each topic in Reddit Post dataset. Most of the topics has the term similarity score for 'depression' between 0.3 and 0.6 can be seen in figure 6.20. Topic 6 is labeled with the words depression,anxiety,postpartum,mental and has the highest term similarity score for 'depression'.

topic	Count	topic_label	detailed_labels	depression_similarity_score
0	0	2778 0_baby_bear_birth_old	0_baby_bear_birth_old_month_give_newborn_like_want_child	0.4033916688068798
1	1	2903 1_mom_parent_daughter_child	1_mom_parent_daughter_child_kid_mother_want_son_dad_old	0.3683410134666232
2	2	2847 2_pregnancy_pregnant_week_get	2_pregnancy_pregnant_week_get_birth_exercise_baby_month_start_day	0.42313367329384377
3	3	1988 3_weight_eat_lose_gain	3_weight_eat_lose_gain_lbs_pound_food_meal_diet_healthy	0.4004578653168651
4	4	2140 4_breastfeed_pump_breast_milk	4_breastfeed_pump_breast_milk_formula_feed_bottle_supply_exclusively_nipple	0.40625683433605186
5	5	3194 5_postpartum_week_month_baby	5_postpartum_week_month_baby_birth_period_hormone_day_time_like	0.4322945997844319
6	6	3427 6_depression_anxiety_postpartum_mental	6_depression_anxiety_postpartum_mental_health_feel_like_stress_disorder_go	0.8528314463338517
7	7	2936 7_husband_say_want_time	7_husband_say_want_time_talk_wife_tell_go_friend_marriage	0.45978664739060815
8	8	2819 8_feel_like_feeling_know	8_feel_like_feeling_know_look_cry_hate_think_body_want	0.4643969703955391
9	9	2122 9_pain_section_hurt_painful	9_pain_section_hurt_painful_tear_preeclampsia_feel_week_blood_pressure	0.5076790797164811
10	10	1818 10_sleep_night_wake_tired	10_sleep_night_wake_tired_hour_bed_morning_time_day_asleep	0.41633241377975394
11	11	2067 11_wear_clothe_pad_underwear	11_wear_clothe_pad_underwear_dress_stretch_fit_paper_skin_size	0.33954796141708943
12	12	2956 12_thank_experience_advice_know	12_thank_experience_advice_know_story_hope_share_tell_appreciate_expect	0.3438510351182764
13	13	1606 13_hi_think_hello_post	13_hi_think_hello_post_happen_ask_its_hey_question_thought	0.38967770498522625
14	14	2384 14_hospital_nurse_doctor_baby	14_hospital_nurse_doctor_baby_hour_labor_midwife_come_appointment_care	0.4426704845205732
15	15	568 15_nan_calcium_yada_high	15_nan_calcium_yada_high_nanny_mg_ionize_strand_chinese_calculator	0.21623047368865284
16	16	957 16_family_sister_husband_brother	16_family_sister_husband_brother_friend_live_parent_aunt_want_wife	0.4067675995947775
17	17	1083 17_mil_house_home_visit	17_mil_house_home_visit_stay_live_work_room_want_leave	0.384633611593761
18	18	919 18_love_sex_relationship_woman	18_love_sex_relationship_woman_partner_feel_want_drive_find_like	0.48875386771078494
19	19	2513 19_week_work_go_day	19_week_work_go_day_pp_month_year_time_advance_thank	0.36829758446356

Figure 6.19: Term Similarity Score for 'Depression' for Reddit Post.

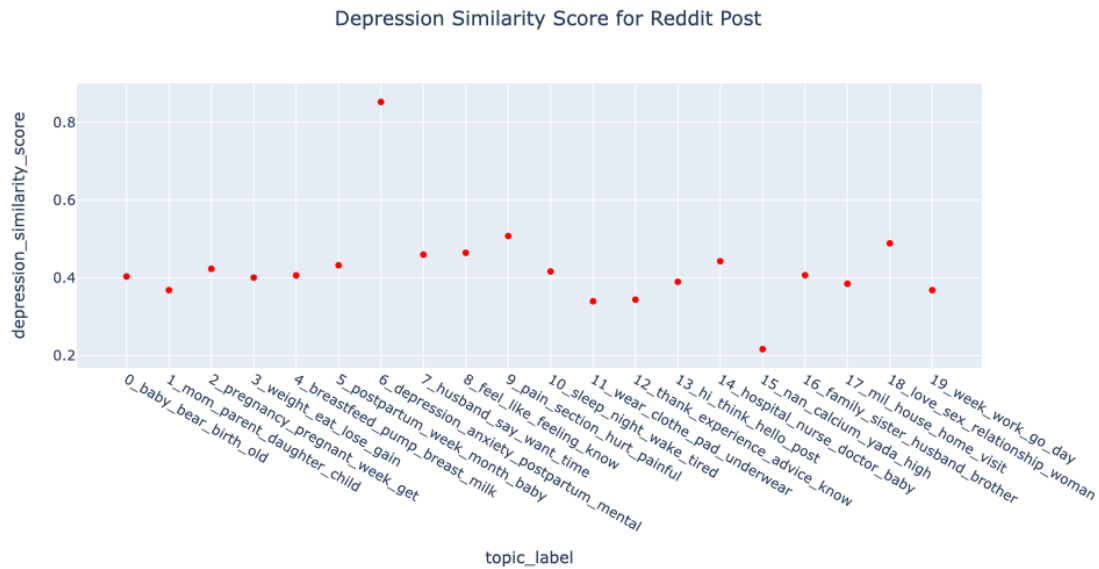
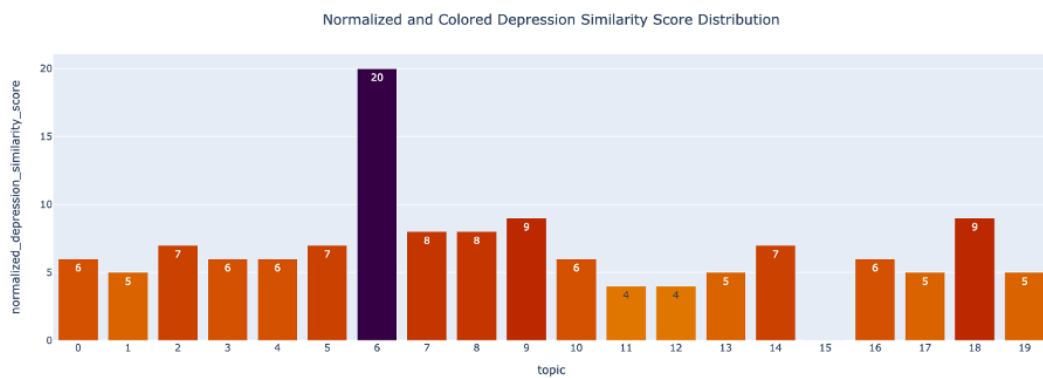


Figure 6.20: Distribution of Term Similarity Score for 'Depression' for Reddit Post.

The term similarity score for 'depression' underwent a normalization process to improve its distribution for visualization purposes. The scores were categorized between 0 and 20, and are presented in Figure 6.21 with corresponding colors. The majority of the scores were within the orange-reddish range, except for one score which was colored dark and corresponds to topic 6. This topic had the highest term similarity score for 'depression'. On the other hand, topic 15 had the lowest value, which was scored as 0 and colored yellow, although it cannot be seen in the bar graph. Overall, the normalization process helped to improve the visual representation of the term similarity score for 'depression'.



**Figure 6.21:** Normalized and Colored of Term Similarity Score for 'Depression' Distribution for Reddit Post.

Figures 6.22 and 6.23 depict the distribution of documents based on the predicted whole mean categories. It was determined that the ranges between -1.5 and 1.5 needed to be cut, and the calculated values had to be redistributed accordingly.

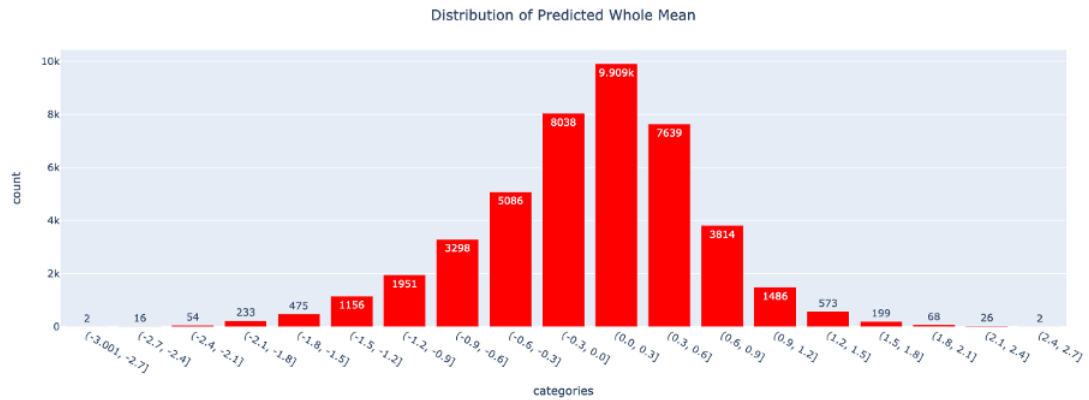


Figure 6.22: Distribution of Predicted Whole Mean for Reddit Post.

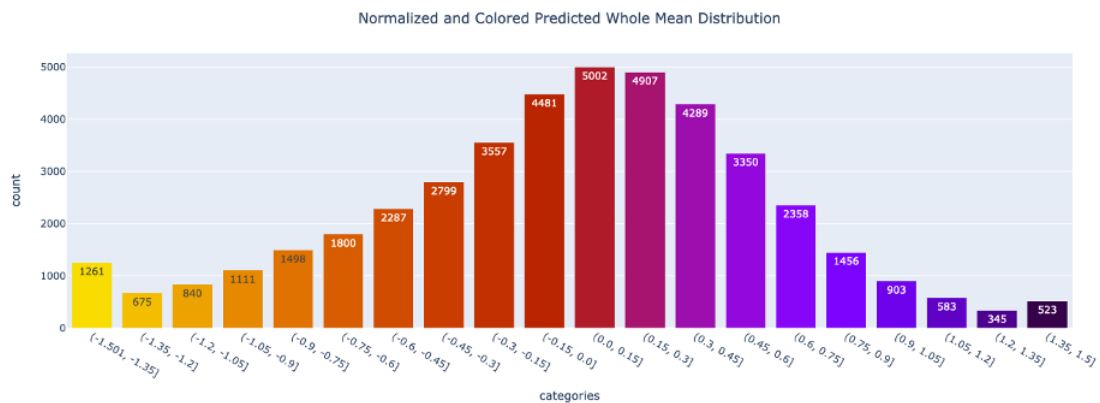


Figure 6.23: Normalized and Colored of Predicted Whole Mean Distribution for Reddit Post.

The document visualization was displayed in Figure 6.24, and it was observed before that only one topic was highly related to the term depression. As a result, there was no depression-specific area in the dataset, and topic 15 was located far away from the other topics. Further investigation revealed that this topic contained NaN values, which were empty values that resulted from the preprocessing steps and some of the non-related data that was collected. Based on this information, it was decided to label this cluster as "NaN" and not consider it for further analysis.

Figure 6.25 is a zoomed-in version of the document visualization for the term similarity score for 'depression'. The results suggest that there is no specific area related to depression in the analyzed Reddit posts.



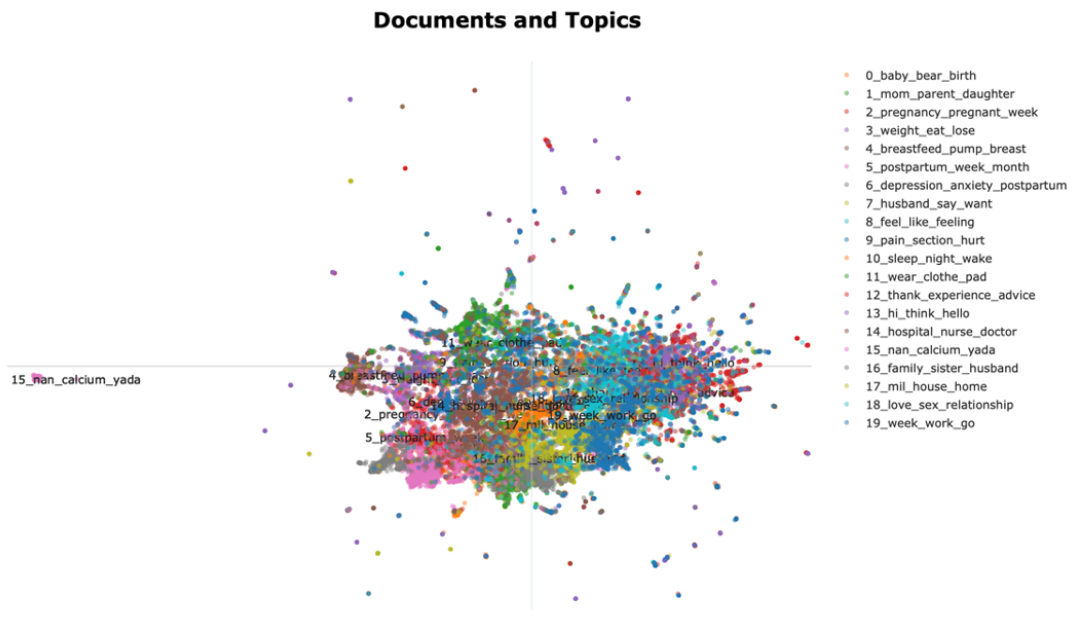


Figure 6.24: Reddit Post Document Visualization.

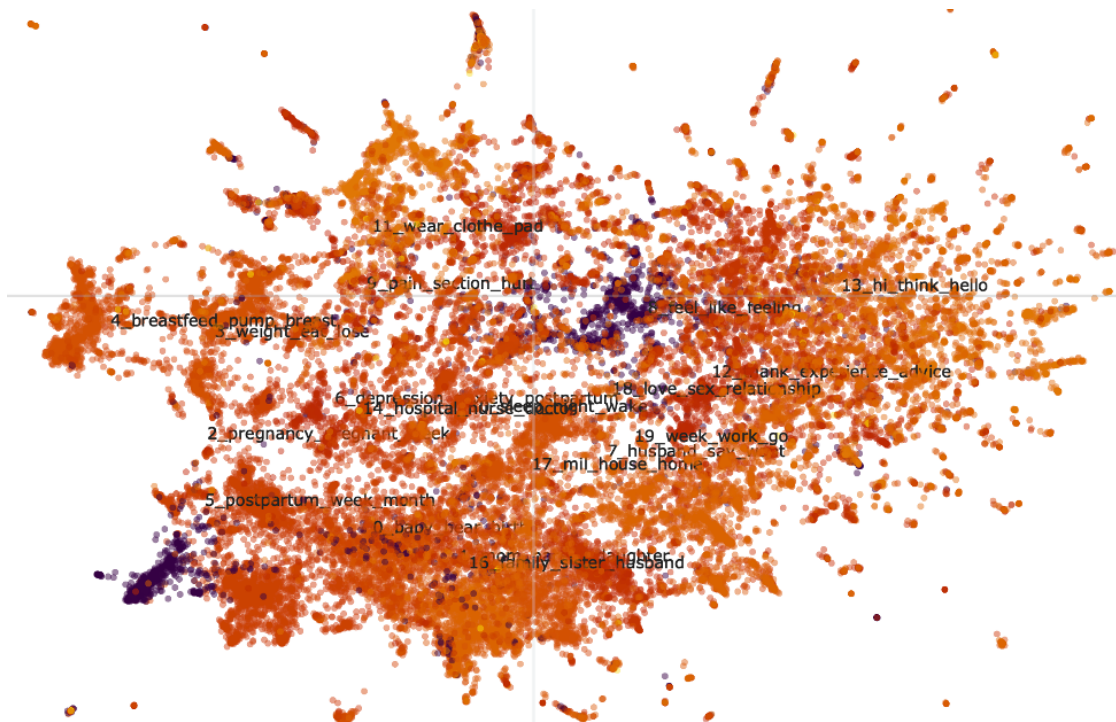


Figure 6.25: Reddit Post Document Visualization of Term Similarity Score for 'Depression'.

Figure 6.26 is a zoomed-in version of the document visualization for the normalized predicted whole mean values. The results indicate that there is no specific area that has a higher or lower agentic value.



Figure 6.26: Document Visualization of BERTAgent for Reddit Post.

The updated labels used in the topic-based graphs were displayed in Figure 6.27, which depicts the evolution of the topic labels. As previously mentioned, topic 15 was labeled as "NaN" due to the presence of NaN values in this topic.

original_topic_labels	outliers_reduced_topic_labels	new_label
-1_feel_time_day_say	This label does not exist anymore.	This label does not exist anymore.
0_bear_newborn_give_feel	0_baby_bear_birth_old	baby_birth_time_newborn
1_parent_want_feel_time	1_mom_parent_daughter_child	relationships_family_members_parent_child
2_pregnancy_week_exercise_yoga	2_pregnancy_pregnant_week_get	pregnancy_birth_time_exercise
3_diet_calorie_week_postpartum	3_weight_eat_lose_gain	weight_loss_diet_eat_body_healthy
4_pump_formula_feed_latch	4_breastfeed_pump_breast_milk	pumping_nipple_breast_feeding
5_postpartum_week_hormone_period	5_postpartum_week_month_baby	baby_birth_time_hormone
6_anxiety_postpartum_stress_diagnose	6_depression_anxiety_postpartum_mental	mental_health_anxiety_depression_bad_feelings
7_marriage_say_want_time	7_husband_say_want_time	marriage_partner_conversation
8_feel_upset_look_hate	8_feel_like_feeling_know	feel_bad_feelings_body
9_pain_section_preeclampsia_postpartum	9_pain_section_hurt_painful	feel_pain_preeclampsia_pressure
10_sleep_morning_insomnia_sleeper	10_sleep_night_wake_tired	sleep_pattern_night_bedtime_time_tired
11_pad_underwear_dress_fit	11_wear_clothe_pad_underwear	clothes_size_pad_fit_underwear_stretch
12_story_experience_expect_tell	12_thank_experience_advice_know	knowledge_talking_thank_advice_experience
13_hi_think_post_reason	13_hi_think_hello_post	greeting_introduction_post_question_asking
14_hospital_midwife_hour_induction	14_hospital_nurse_doctor_baby	hospital_health_care_personnel_appointment_labor
15_nan_nederland_kyleena_ammonium	15_nan_calcium_yada_high	NaN
16_family_friend_aunt_say	16_family_sister_husband_brother	relationships_big_family_family_members
17_mil_home_visit_stay	17_mil_house_home_visit	home_life_mil
18_love_woman_feel_drive	18_love_sex_relationship_woman	relationship_partner_love_sex
19_work_go_month_ft	19_week_work_go_day	time_work_advance_thank

Figure 6.27: Label Evolution for Reddit Post.

The figures from 6.28 to 6.33 depict topic-based graphs for Reddit post. Figure 6.26, specifically, represents the negative emotions graph. The purple color is used to represent topics 6, 9, and 8, which have high scores for negative emotions, while yellowish-orangish colors are used for other topics. Among these purple topics, only topic 6, which is labeled as mental health, anxiety, depression, and bad feelings, is highly related to the term "depression," as shown in Figure 6.32. Therefore, it can be inferred that topic 6 comprises negative expressions about depression. Several examples of topic 6 are provided below.

- 'One week into it I have an intense emotional breakdown and ended up calling triage.'
- 'Is this postpartum depression?'
- 'I have lost focus on literally everything else and it is very alarming.'

Topics 8 and 9 are labeled as "feel, bad feelings, body" and "feel, pain, preeclampsia, pressure," respectively. Although these topics are associated with negative emotions, it cannot be definitively stated that they pertain solely to depression.

Furthermore, the term similarity score for 'depression' graph provides evidence that there is no area in the graph that is strongly associated with the term "depression," as previously mentioned in the document-based graph.

Figure 6.30 displays a color-coded representation of topics based on positive emotion scores calculated using the LIWC algorithm. Topics 12 and 18 are characterized by darker colors, indicating higher positive scores, while other topics are colored in shades of yellow, indicating lower positive scores. The keywords associated with positive topics mostly consist of positive words such as "love," "advice," and "thank." Some examples from these topics were given below:

- 'Does anyone have experiences to share of getting their second dose shortly after delivery?'
- 'Thank you!'
- 'I would appreciate recommendations from people who have experience with who they recommend. '
- 'Luckily I have an extremely supportive partner and he helped me get through those tough times. '
- 'I DO love her, and I work my butt off to provide what she needs and everything I want her to have. '
- 'It is like my body discovered the true meaning of sex and cannot look back. '

Figure 6.29 displays a color-coded representation of composite emotion scores, where brighter colors are used to indicate higher negative scores and darker colors are used for higher positive scores. It is evident from the figure that topics previously identified as negative, namely topics 6, 9, and 8, are depicted with brighter colors, providing evidence that these topics are associated with negative emotions.

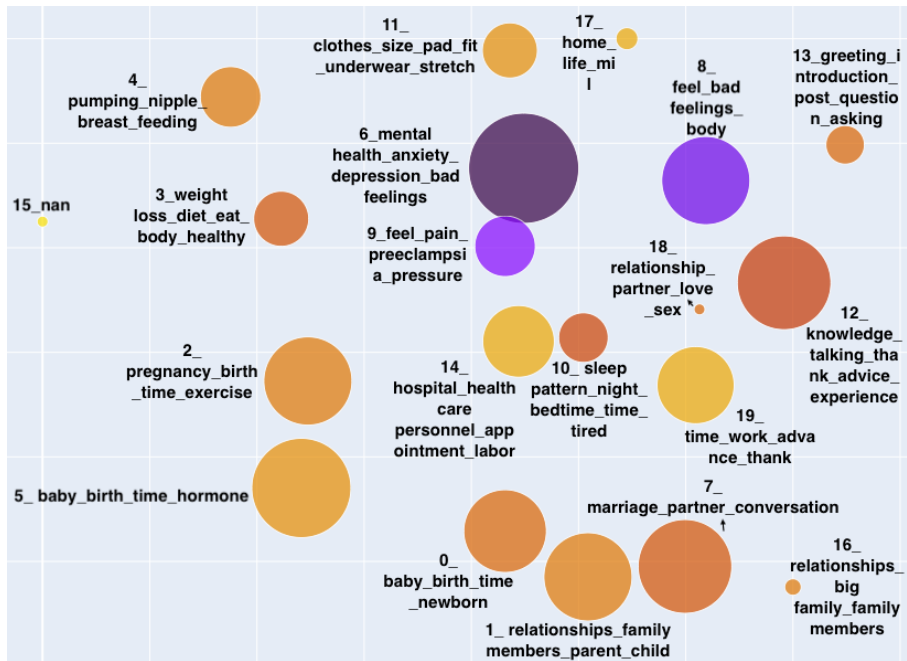


Figure 6.28: Topic Based Negative Emotions for Reddit Post.

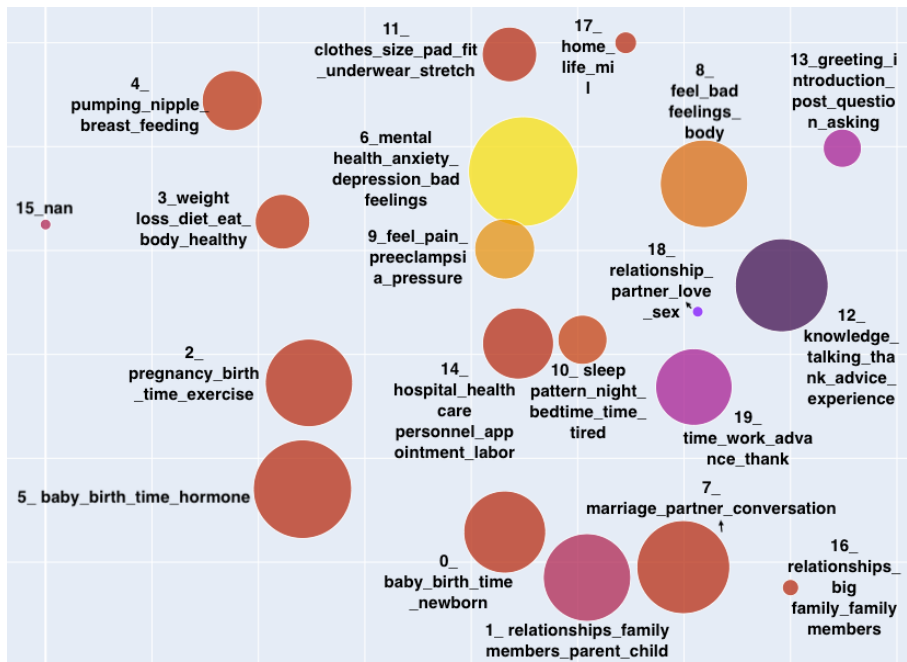


Figure 6.29: Topic Based Composite Emotion Score for Reddit Post.

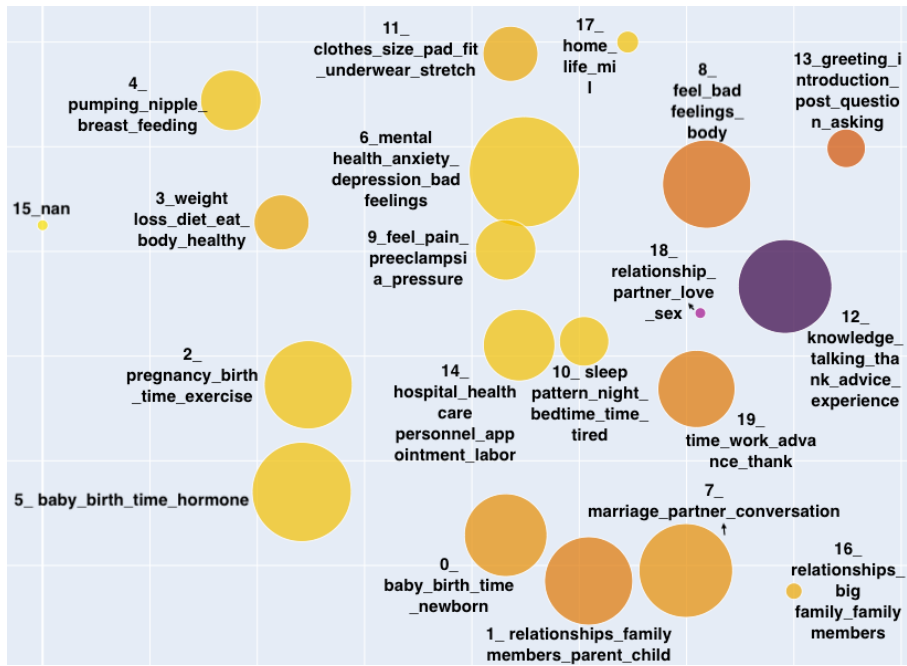


Figure 6.30: Topic Based Positive Emotion for Reddit Post.



Figure 6.31: Topic Based BERTAgent for Reddit Post.

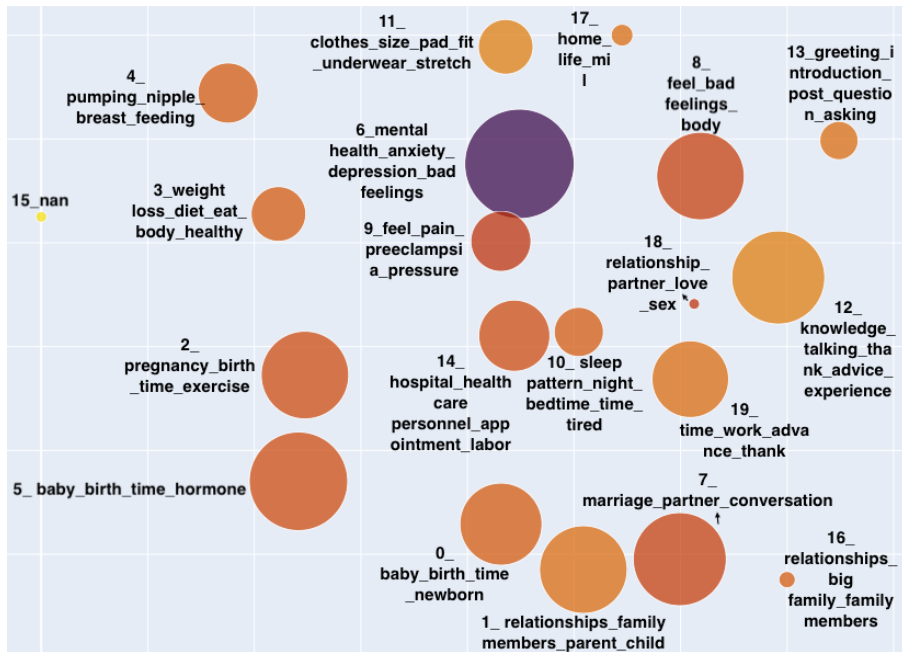


Figure 6.32: Topic Based Term Similarity Score for 'Depression' for Reddit Post.

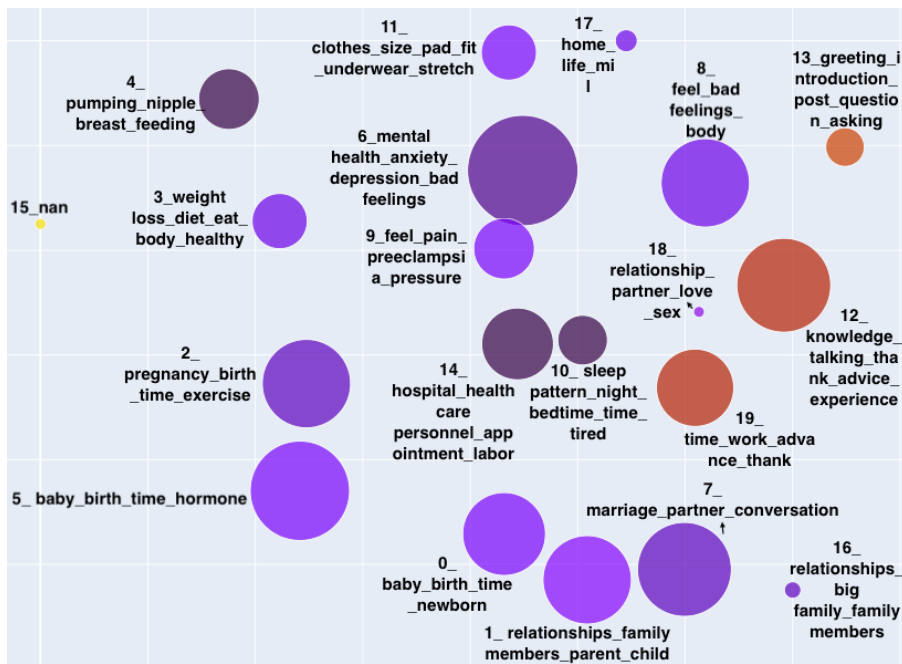


Figure 6.33: Topic Based Verbs for Reddit Post.

### 6.3 REDDIT COMMENT

The input for topic analysis consisted of 45699 subsentences, using the specified parameters:

- n\_neighbors in UMAP: 220,
- min\_cluster\_size in HDBSCAN: 170,
- diversity in topic model: 0.5.

A total of 21 topics were generated through topic analysis, with outliers grouped into a single topic. Following the removal of outliers, a final set of 20 topic labels were assigned and represented in Figure 6.34.

topic_number	count_before	count_later	original_topic_labels	outliers_reduced_topic_labels
-1	21538	0	-1_like_feel_postpartum_child	This label does not exist anymore.
0	4037	7533	0_need_feel_work_try	0_help_need_good_go
1	2484	5953	1_postpartum_midwife_feel_body	1_postpartum_birth_baby_week
2	2047	2617	2_pad_underwear_pack_stretch	2_pad_wear_clothe_underwear
3	1560	2219	3_pain_section_recovery_stitch	3_pain_tear_hurt_section
4	1530	3029	4_anxiety_postpartum_depressed_antidepressant	4_depression_anxiety_postpartum_mental
5	1205	1852	5_pelvic_floor_workout_pt	5_pelvic_floor_exercise_workout
6	996	2074	6_pregnancy_trimester_feel_month	6_pregnancy_pregnant_birth_baby
7	959	1943	7_baby_newborn_month_care	7_baby_kid_newborn_old
8	936	1059	8_lose_pregpregnancy_calorie_postpartum	8_weight_lose_lbs_gain
9	902	1987	9_doctor_ob_appointment_organ	9_doctor_hospital_ob_appointment
10	882	3078	10_feel_like_fault_share	10_feel_like_sorry_feeling
11	880	2403	11_wedding_need_marry_talk	11_husband_wife_family_love
12	873	1526	12_postpartum_ebf_week_go	12_month_pp_nta_period
13	813	1202	13_body_rest_change_exhaust	13_body_break_change_rest
14	801	1718	14_breastfeeding_feed_month_formula	14_breastfeed_breast_milk_feed
15	792	1604	15_deprivation_bassist_nap_room	15_sleep_night_baby_hour
16	770	1468	16_week_postpartum_period_feel	16_week_day_postpartum_year
17	616	1029	17_eat_protein_freezer_chicken	17_eat_food_meal_healthy
18	552	833	18_psychosis_loss_postpartum_month	18_hair_loss_psychosis_postpartum
19	526	572	19_nan_quinn_predominately_nana	19_nan_nanny_quinn_sailor

Figure 6.34: Original Topic Labels vs. Outliers Reduced Topic Labels for Reddit Comment.

Figure 6.35 and 6.36 represents the term similarity score for 'depression' for Reddit Comment dataset. The distribution is similar to the Reddit Post. Most of the topics have the score between 0.3 to 0.6 and only one high term similarity score for 'depression' which is the topic 4 with the score 0.87.



topic	Count	topic_label	detailed_labels	depression_similarity_score
0	0	0_help_need_good_go	0_help_need_good_go_time_think_thing_hope_try_luck	0.36841782034548104
1	1	1_postpartum_birth_baby_week	1_postpartum_birth_baby_week_month_mom_time_like_care_feel	0.4557784236948861
2	2	2_pad_wear_clothe_underwear	2_pad_wear_clothe_underwear_paper_buy_fit_pack_size_stretch	0.3366424373587862
3	3	3_pain_tear_hurt_section	3_pain_tear_hurt_section_heal_recovery_stitch_painful_degree_feel	0.5101056593720892
4	4	4_depression_anxiety_postpartum_mental	4_depression_anxiety_postpartum_mental_health_help_disorder_stress_like_feel	0.8732195870847732
5	5	5_pelvic_floor_exercise_workout	5_pelvic_floor_exercise_workout_pt_walk_core_pregnancy_postpartum_muscle	0.4276832730248496
6	6	6_pregnancy_pregnant_birth_baby	6_pregnancy_pregnant_birth_baby_month_week_get_trimester_period_start	0.4052751321157052
7	7	7_baby_kid_newborn_old	7_baby_kid_newborn_old_child_care_time_month_love_mom	0.3853997622382661
8	8	8_weight_lose_lbs_gain	8_weight_lose_lbs_gain_pound_pregnancy_prepregnancy_month_calorie_loss	0.39293809032379023
9	9	9_doctor_hospital_ob_appointment	9_doctor_hospital_ob_appointment_talk_medication_patient_nurse_need_medical	0.47909977084325917
10	10	10_feel_like_sorry_feeling	10_feel_like_sorry_feeling_know_talk_experience_well_ta_bad	0.4465167271260382
11	11	11_husband_wife_family_love	11_husband_wife_family_love_relationship_need_partner_want_help_talk	0.4450768770384966
12	12	12_month_pp_nta_period	12_month_pp_nta_period_normal_postpartum_ppa_go_time_bleed	0.3907863445375014
13	13	13_body_break_change_rest	13_body_break_change_rest_like_feel_time_exhaust_thing_touch	0.46826547081429354
14	14	14_breastfeed_breast_milk_feed	14_breastfeed_breast_milk_feed_formula_pump_nipple_baby_bottle_supply	0.40602390126511767
15	15	15_sleep_night_baby_hour	15_sleep_night_baby_hour_wake_room_time_bed_deprivation_nap	0.4532406503606169
16	16	16_week_day_postpartum_year	16_week_day_postpartum_year_time_start_get_ago_period_month	0.3310696477874937
17	17	17_eat_food_meal_healthy	17_eat_food_meal_healthy_snack_diet_water_cook_drink_freezer	0.38291289738477957
18	18	18_hair_loss_psychosis_postpartum	18_hair_loss_psychosis_postpartum_fall_lose_grow_month_bald_shed	0.5310027444400557
19	19	19_nan_nanny_quinn_sailor	19_nan_nanny_quinn_sailor_mg_sodium_salt_hypothyroid_number_figure	0.21940525582041764

Figure 6.35: Term Similarity Score for 'Depression' for Reddit Comment.

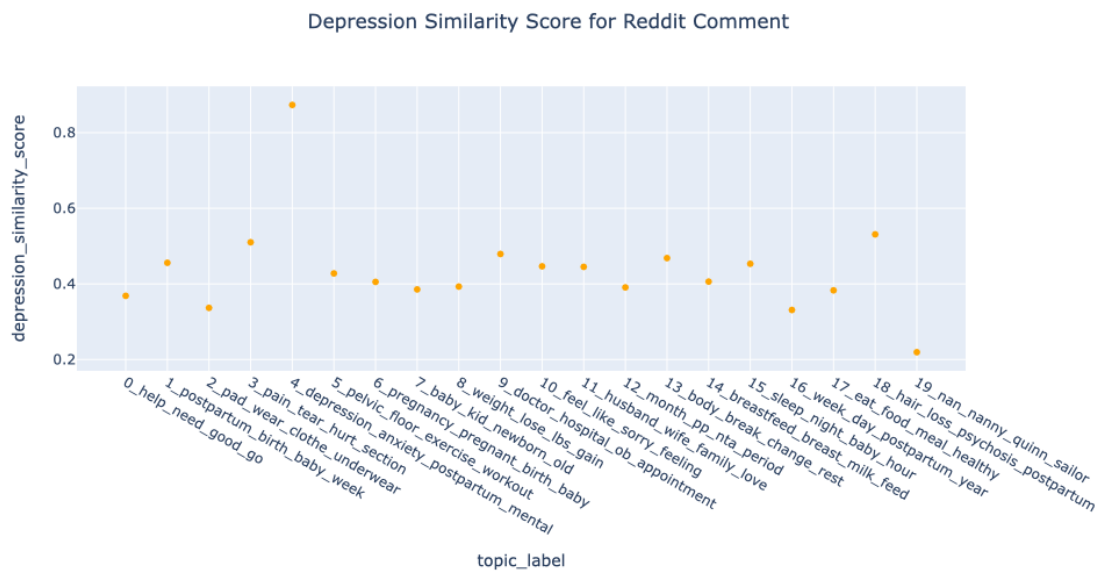


Figure 6.36: Distribution of Term Similarity Score for 'Depression' for Reddit Comment.

Similar to the analysis of Twitter and Reddit posts, the similarity score of depression was normalized and visualized with color in Figure 6.37. The distribution of the predicted mean scores was displayed in Figure 6.38. A threshold of -1.5 and 1.5 was selected to normalize the

data, and the distribution of the resulting normalized predicted mean scores was depicted in Figure 6.39.

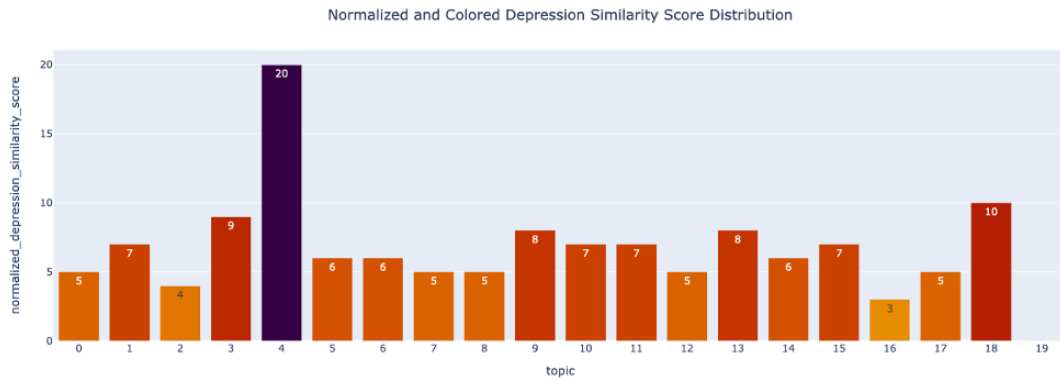


Figure 6.37: Normalized and Colored of Term Similarity Score for 'Depression' Distribution for Reddit Comment.

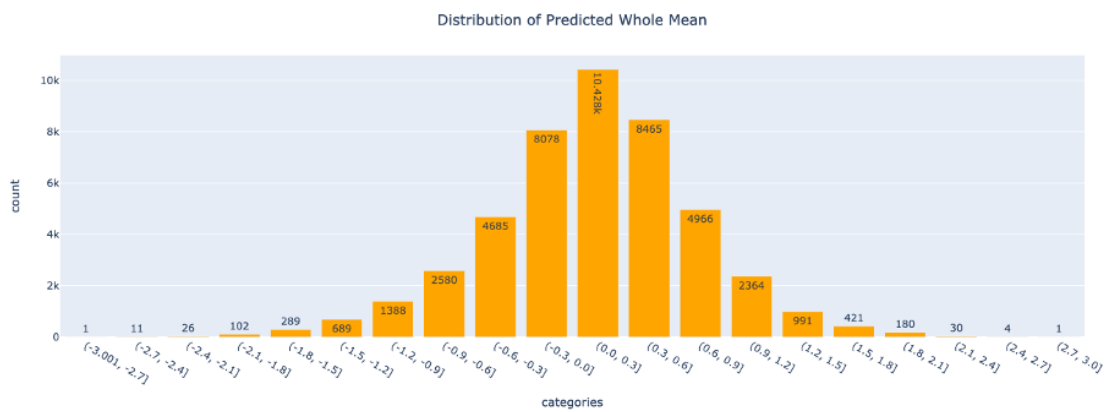


Figure 6.38: Distribution of Predicted Whole Mean for Reddit Comment.

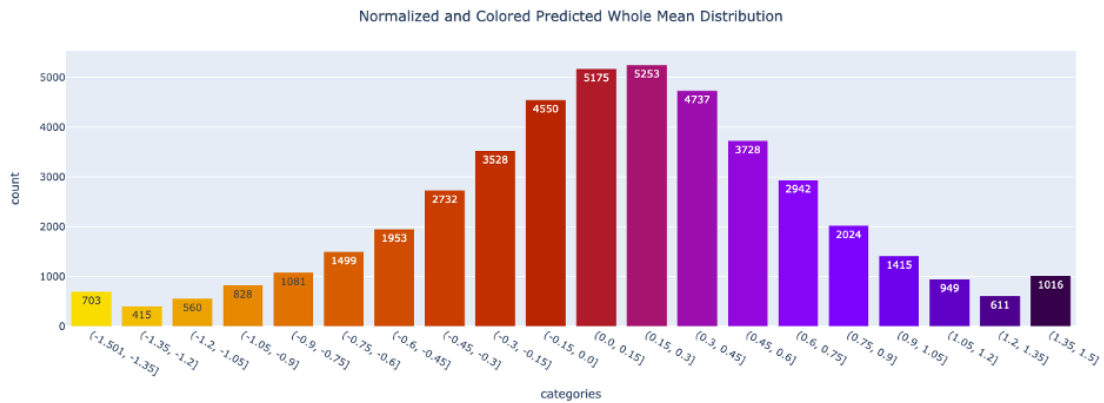


Figure 6.39: Normalized and Colored of Predicted Whole Mean Distribution for Reddit Comment.

Figure 6.40 illustrates the visualization of a Reddit comment dataset in a document clustering context. The clusters generated were examined, and Topic 19 was found to be located far from the other topics. Upon investigation, it was discovered that this topic contained NaN values resulting from preprocessing, as well as several unrelated sentences. As a result, it was decided that Topic 19 would be excluded from further analysis.

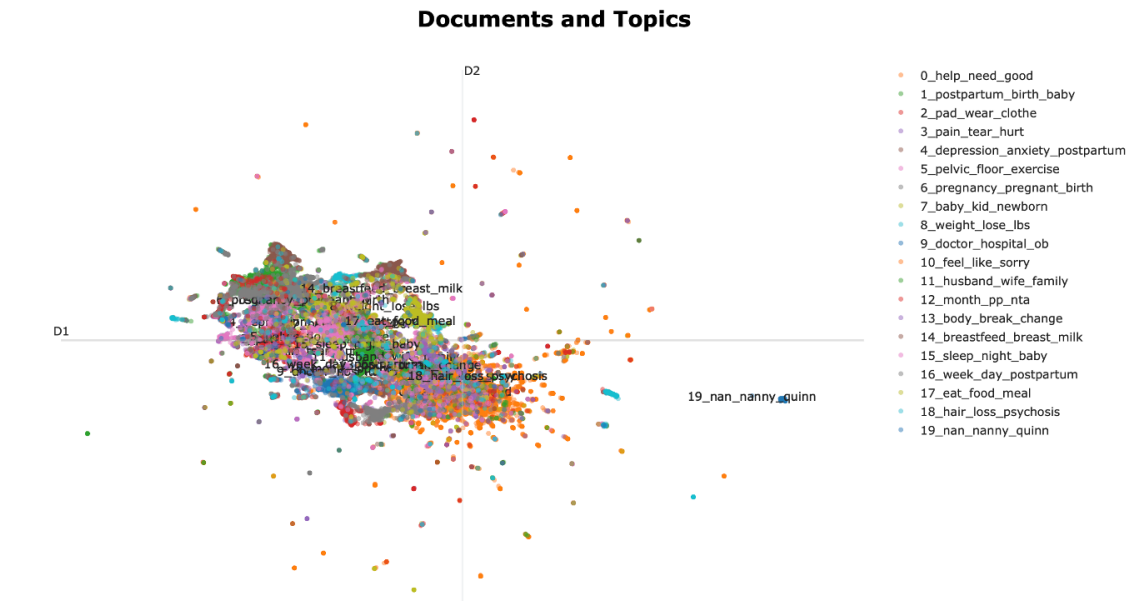


Figure 6.40: Reddit Comment Document Visualization.

The visualization depicted in Figure 6.41 displays a document map that has been color-coded according to the normalized term similarity score for 'depression'. The map does not reveal any

distinct areas that can be classified as depicting depression, as the majority of the colors are in the yellow-orange range. Only a small portion of the map, specifically topic 4, is represented by a darker color.

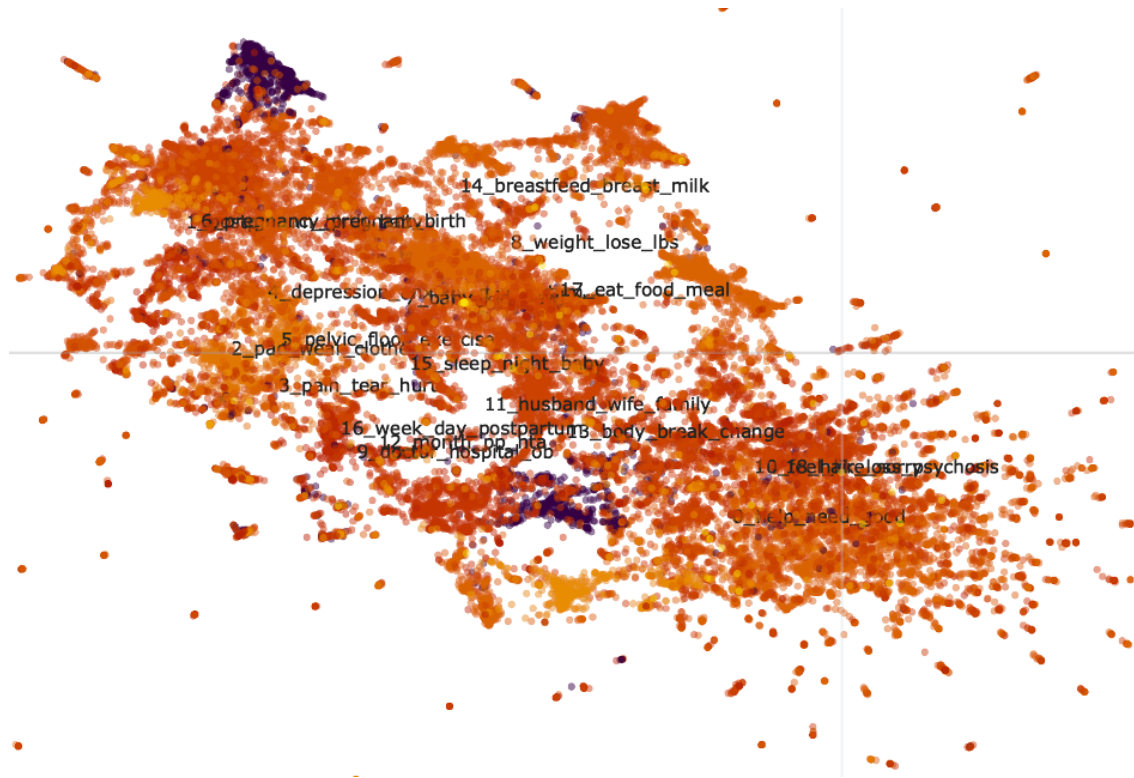


Figure 6.41: Reddit Comment Document Visualization of Term Similarity Score for 'Depression'.

The document visualization shown in Figure 6.42 is color-coded based on the normalized predicted whole mean. Similar to the results obtained from the Reddit and Twitter data, the majority of the colors in the visualization can be discerned, although it is preferable to examine the data using topic-based graphs. The dominant colors in the visualization are dark tones.

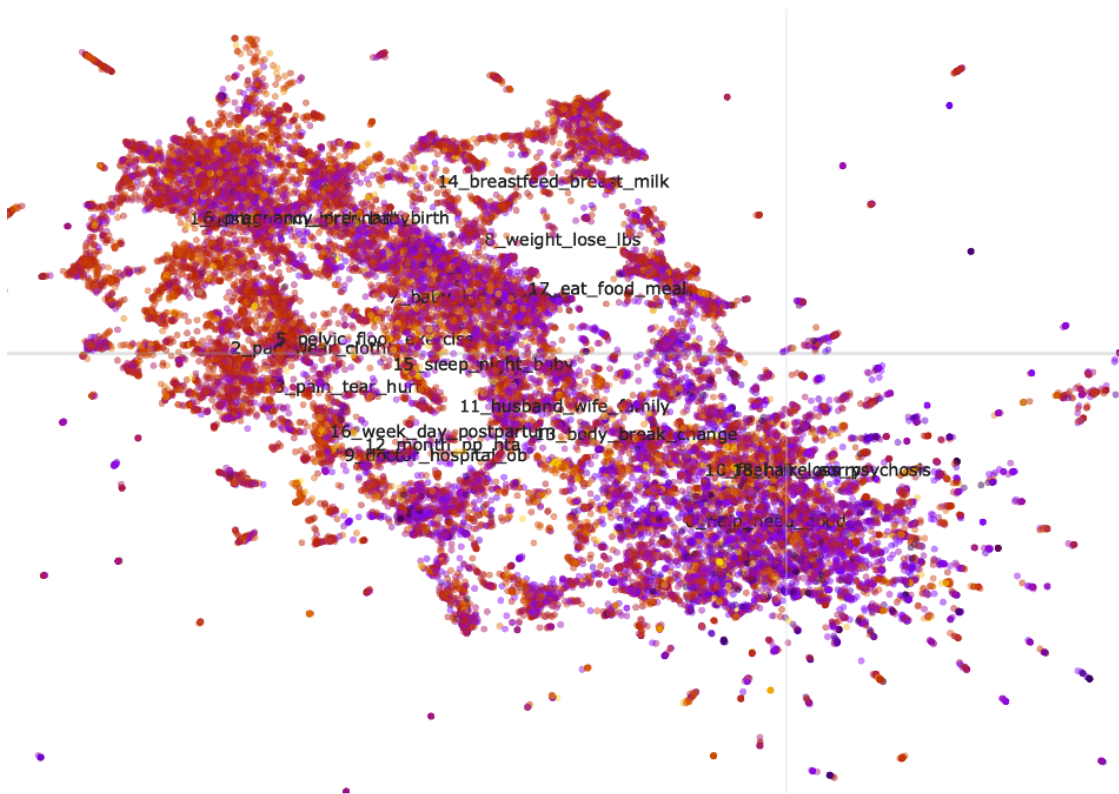


Figure 6.42: Document Visualization of BERTAgent for Reddit Comment.

Figure 6.43 displays the label evolution for a dataset of Reddit comments. The new labels presented in the visualization represent a summarization of the outlier-reduced version of the original labels, and are utilized for the creation of topic-based graphs. It should be noted that in the visualization, topic 19 is labeled as "NaN".

original_topic_labels	outliers_reduced_topic_labels	new_label
-1_like_feel_postpartum_child	This label does not exist anymore.	This label does not exist anymore.
0_need_feel_work_try	0_help_need_good_go	help_need_time_think_luck
1_postpartum_midwife_feel_body	1_postpartum_birth_baby_week	baby_birth_time_maternity_care
2_pad_underwear_pack_stretch	2_pad_wear_clothe_underwear	clothes_size_fit_pad_underwear_stretch
3_pain_section_recovery_stitch	3_pain_tear_hurt_section	pain_recovery_stitch_degree
4_anxiety_postpartum_depressed_antidepressant	4_depression_anxiety_postpartum_mental	mental_health_anxiety_depression_bad_feelings_...
5_pelvic_floor_workout_pt	5_pelvic_floor_exercise_workout	pelvic_floor_exercise_pregnancy_muscle
6_pregnancy_trimester_feel_month	6_pregnancy_pregnant_birth_baby	pregnancy_birth_period_trimester
7_baby_newborn_month_care	7_baby_kid_newborn_old	baby_child_love_maternity_time
8_lose_pregpregnancy_calorie_postpartum	8_weight_lose_lbs_gain	weight_loss_calorie_body_pregpregnancy
9_doctor_ob_appointment_organ	9_doctor_hospital_ob_appointment	hospital_health_care_personnel_appointment_med...
10_feel_like_fault_share	10_feel_like_sorry_feeling	emotions_feelings_mood_experience_talking
11_wedding_need_marry_talk	11_husband_wife_family_love	relationship_partner_help_talk_marriage_love
12_postpartum_ebf_week_go	12_month_pp_nta_period	time_nta_bleeding
13_body_rest_change_exhaust	13_body_break_change_rest	body_rest_time_exhaust_touch
14_breastfeeding_feed_month_formula	14_breastfeed_breast_milk_feed	breastfeeding_nipple_breast_feeding
15_deprivation_bassist_nap_room	15_sleep_night_baby_hour	sleep_pattern_night_time_tired
16_week_postpartum_period_feel	16_week_day_postpartum_year	time_start_get
17_eat_protein_freezer_chicken	17_eat_food_meal_healthy	eat_food_healthy_diet_snack
18_psychosis_loss_postpartum_month	18_hair_loss_psychosis_postpartum	psychosis_hair_loss_and_grow_bald
19_nan_quinn_predominately_nana	19_nan_nanny_quinn_sailor	NaN

Figure 6.43: Label Evolution for Reddit Comment.

Figure 6.44 depicts the negative emotion scores for topics. As illustrated in figure, certain topics were observed to be darker in color, indicating a stronger association with negative sentiment. These topics include topic 4, which contains keywords such as mental health, anxiety, depression, bad feelings, and help; topic 10, which includes keywords such as emotions, feelings, mood, experience, and talking; and topics 2 and 18, which have a pinkish color and are characterized by keywords such as clothes, size, fit, pad, underwear, stretch, and psychosis, hair loss and grow, bald, respectively.

Moreover, as previously mentioned, topic 4 has the highest term similarity score for 'depression', a trend that is also observed in Figure 6.48 where topic 4 is the only dark-colored topic in the depression similarity topic-based graph. Thus, topic 4 is associated with negative expressions about depression. Examples from this topic:

- 'My aunt only had PPD on her first and her fourth child for whatever reason.'
- 'If you cannot find anything there to help you now, at least get on some forums and find

someone to talk to, someone who has been through this depression and came out on the other side.’

- ‘I have a lot of passive anxiety no panic attacks but definitely feeling anxious often without me even realizing it until the past year.’

The analysis of Topic 10 represented with the keywords experience, talking, emotions and feelings shows that there are noteworthy findings regarding the comparison between the negative and positive scores. Both Figure 6.42 and 6.44 show a significant proportion of dark color, indicating high scores in both negative and positive sentiments. Meanwhile, the BERTAgent and composite scores presented in Figure 6.43 and 6.45 exhibit reddish hues, representing moderate levels of agency and neutrality. This suggests that the scores are not entirely leaning towards either positive or negative sentiments. Overall, these results demonstrate the complexity and nuances in the sentiment analysis of Topic 10. Following examples from topic 10 shows that it has positive and negative expressions about emotions and experiences.

- ‘It will get better.’
- ‘Basically know that feeling this way is normal, and treat yourself to whatever may make you feel better!’
- ‘I felt devastated and awful.’
- ‘I was a crying, angry mess who hated everyone and everything around me.’

The topic denoted as topic 0 is the most extensive community and is characterized by the keywords “help,” “need,” “time,” “think,” and “luck.” It is represented by the color purple in both composite and positive emotion graphs, as well as in BertAgent graphs. These findings suggest that this community exhibits predominantly positive expressions.

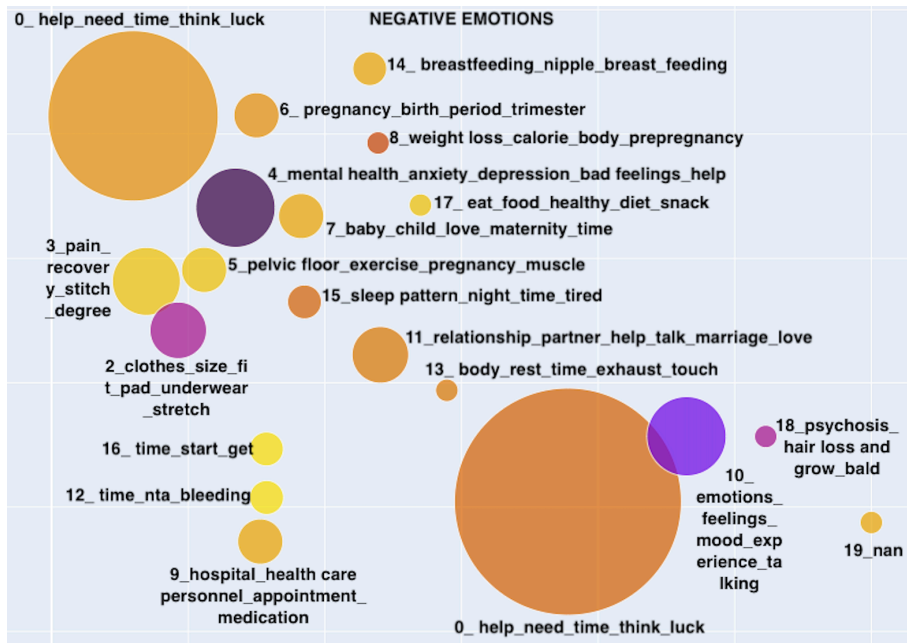


Figure 6.44: Topic Based Negative Emotions for Reddit Comment.

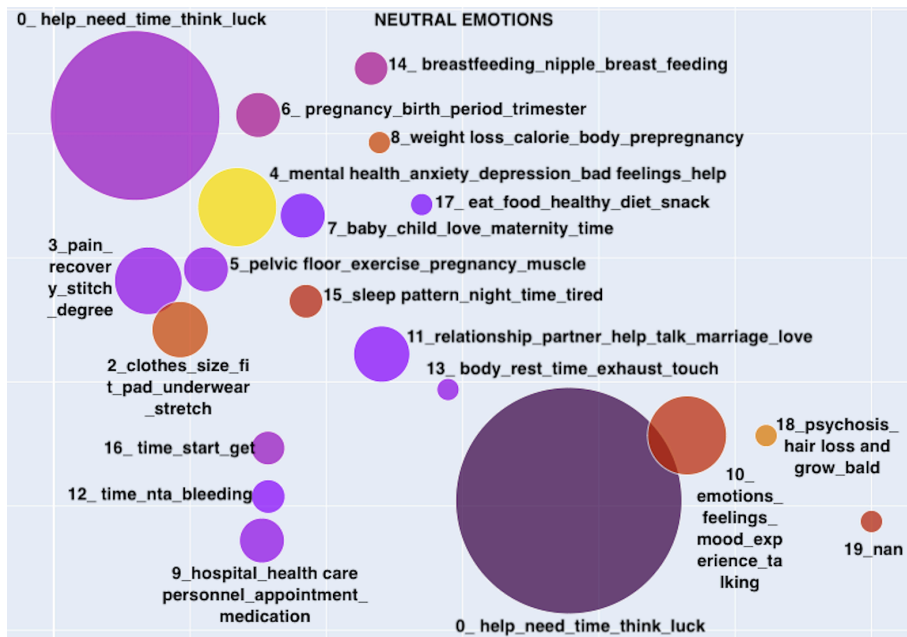


Figure 6.45: Topic Based Composite Emotion Score for Reddit Comment.



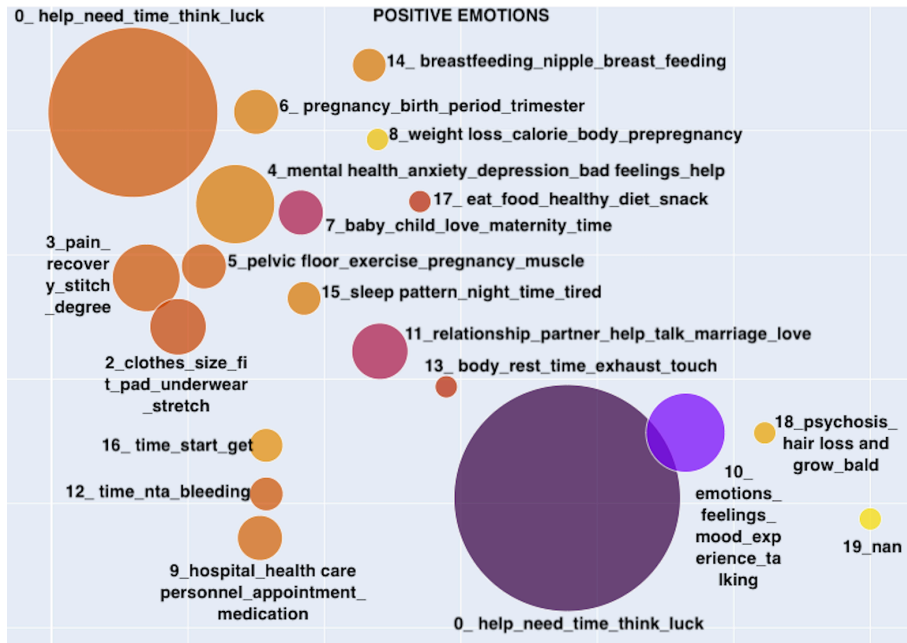


Figure 6.46: Topic Based Positive Emotion for Reddit Comment.

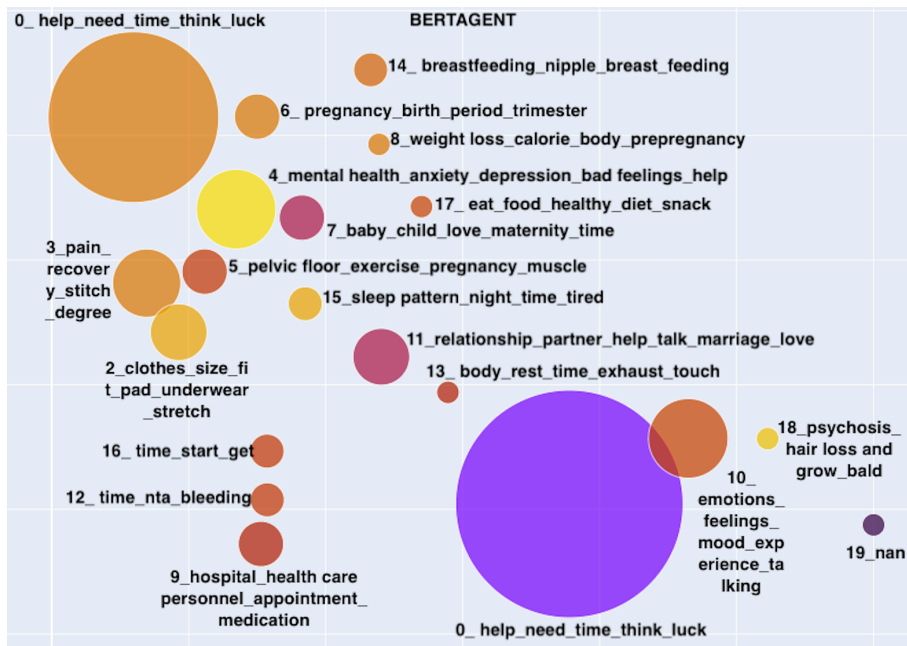


Figure 6.47: Topic Based BERTAgent for Reddit Comment.

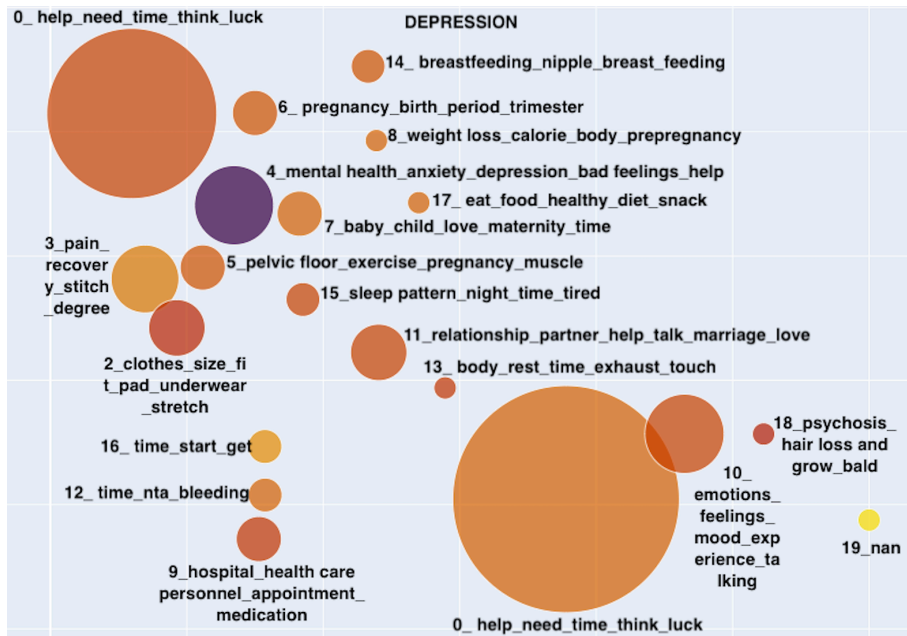


Figure 6.48: Topic Based Term Similarity Score for 'Depression' for Reddit Comment.

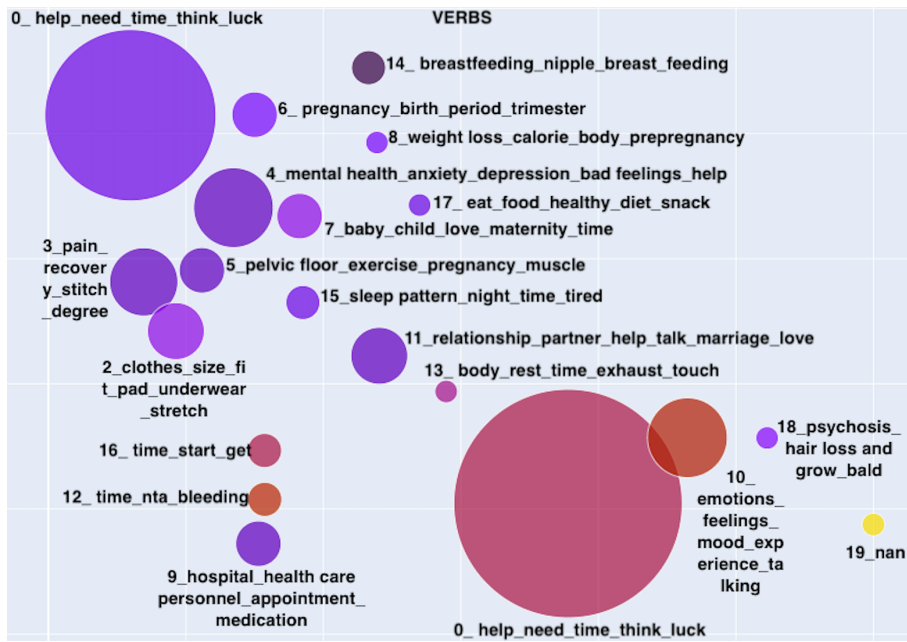


Figure 6.49: Topic Based Verbs for Reddit Comment.

## 6.4 CORRELATION BETWEEN CALCULATED VALUES

Correlation graphs were generated to illustrate the associations between the composite emotion score and predicted whole mean, term similarity score for depression, verbs for Twitter, Reddit posts, and Reddit comments. These graphs provide an overarching depiction of the relationships noted in the bubble graphs. The following figures depict the correlation graphs for the aforementioned variables.

The x-axis represents the composite emotion score, while the y-axis represents the predicted mean of the whole, term similarity score for depression and verbs. The topic label number placed above the dots on graphs. The dot size was determined in proportion to the community size. The variable of size was utilized as the weight parameter when fitting the trendline, and a 95% confidence interval was established.

It was found a positive correlation between the predicted whole mean value and composite emotion score. Specifically, higher values for the predicted whole mean were associated with higher composite emotion scores, reflecting increased agency and positivity. Conversely, it was observed a negative correlation between the term similarity score for depression and the composite emotion score. Specifically, higher term similarity scores for depression were associated with decreased composite emotion scores, indicative of heightened negative emotionality. Finally, It was found no significant correlation between the verbs and the composite emotion score.

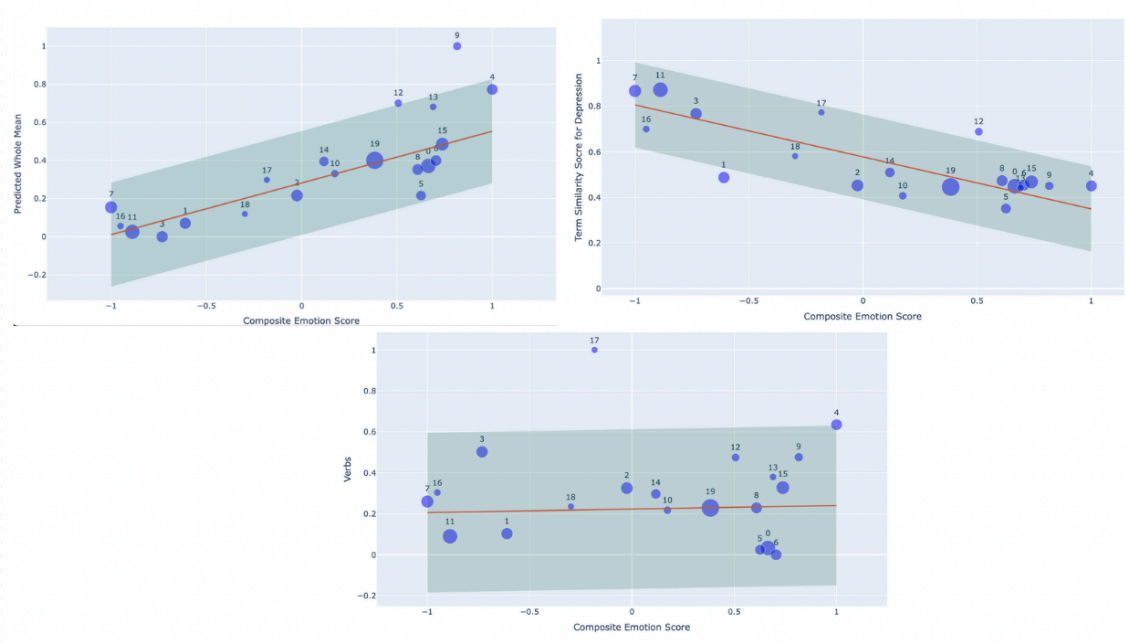


Figure 6.50: Correlation Graphs for Twitter.

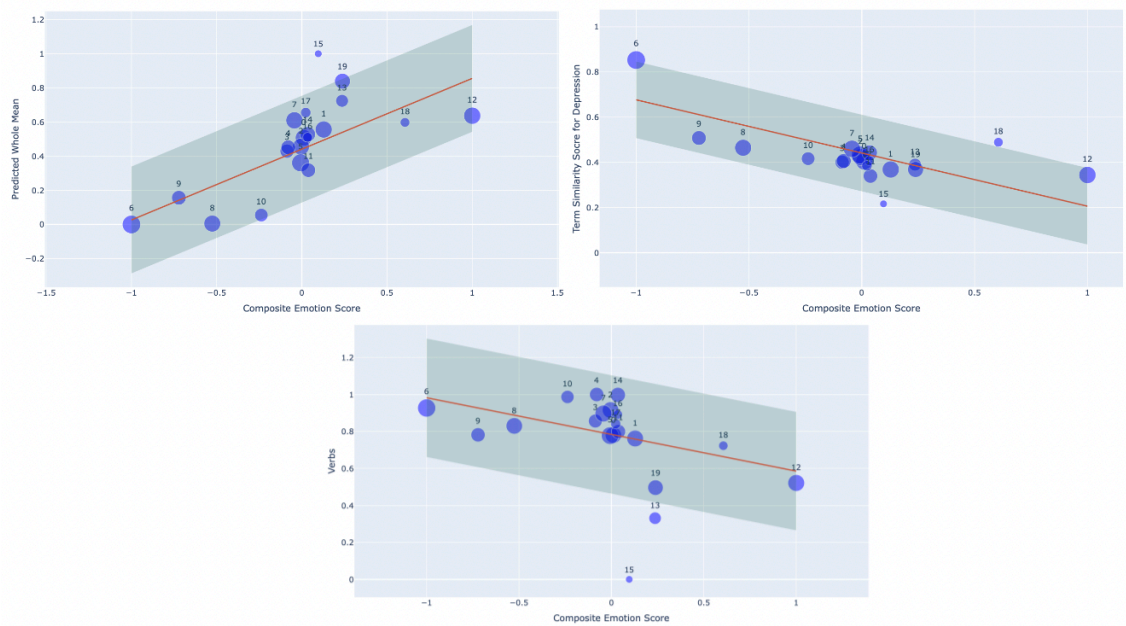


Figure 6.51: Correlation Graphs for Reddit Post.

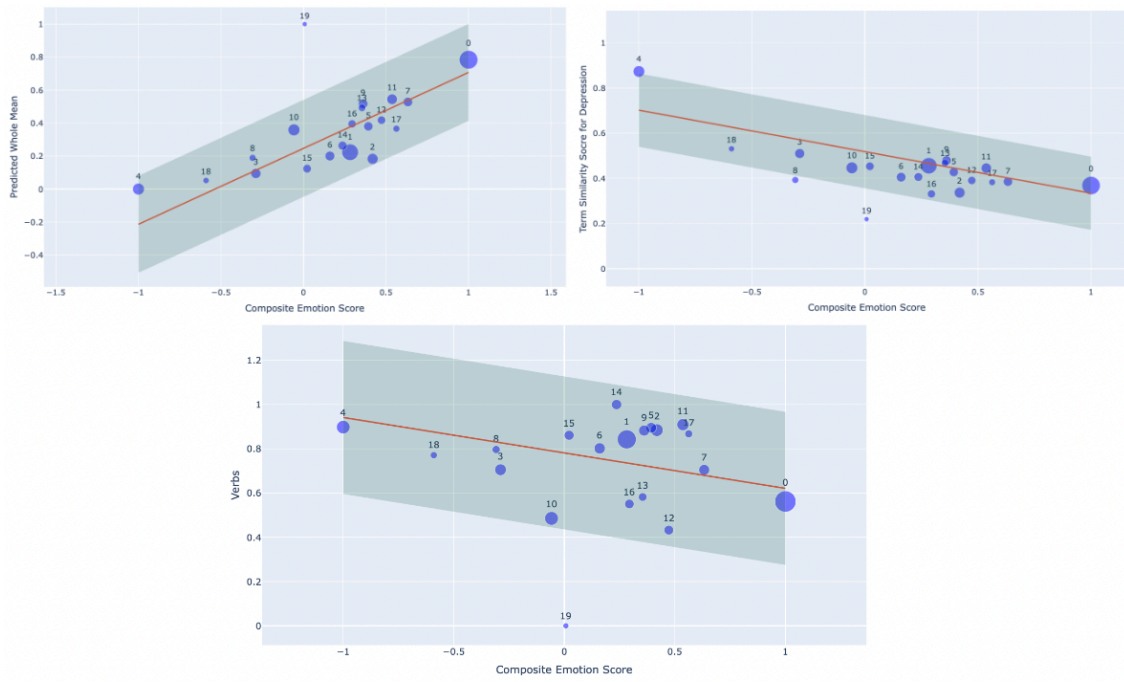


Figure 6.52: Correlation Graphs for Reddit Comment.

# 7

## Conclusion

In conclusion, this research focused on the relationship between agentic language and the emotion score, particularly in relation to the term depression in expressions about postpartum. The concept of agency was crucial in this research, as it pertains to goal-directed actions and is significant for analyzing human actions and interactions. The study demonstrated a detailed and comprehensive planning process, including data scraping, creating datasets, preprocessing procedures, and the application of various tools and methods for analysis such as LIWC, BERTAgent, BERTopic, grammatical calculations, normalization.

The present research's results offer valuable insights into the intricate interplay between negative emotions and the use of agentic language in postpartum expressions gathered from social media platforms such as Twitter and Reddit. The findings' reliability was strengthened through the implementation of normalization procedures, and the results were better visualized through the application of color-coded document and topic-based graphs.

## References

- [1] L. Sloan and A. Quan-Haase, *The SAGE Handbook of Social Media Research Methods*, 2017.
- [2] C. C. Aggarwal, *An Introduction to Social Network Data Analytics*, C. C. Aggarwal, Ed. Boston, MA: Springer US, 2011.
- [3] J. Nikadon, C. Suitner, T. Erseghe, L. Džanko, and M. Formanowicz, “Bertagent: A novel tool to quantify agency in textual data,” *Behavior Research Methods*, 2022.
- [4] M. Injadat, F. Salo, and A. B. Nassif, “Data mining techniques in social media: A survey,” *Neurocomputing*, vol. 214, pp. 654–670, 2016.
- [5] B. Batrinca and P. C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms,” *AI SOCIETY*, vol. 30, pp. 89–116, 2015.
- [6] S. Varsha, S. Vijaya, and P. Apashabi, “Sentiment analysis on twitter data,” *International Journal of Innovative Research in Advanced Engineering*, vol. 2, p. 178, 2015.
- [7] J. Costa, C. Silva, and B. Ribeiro, “Learning in twitter streams with 280 character tweets,” pp. 177–184, 2020.
- [8] N. J. Yuan, Y. Zhong, F. Zhang, X. Xie, C.-Y. Lin, and Y. Rui, “Who will reply to/retweet this tweet? the dynamics of intimacy from online social interactions,” p. 3–12, 2016. [Online]. Available: <https://doi.org/10.1145/2835776.2835800>
- [9] D. Boyd, S. Golder, and G. Lotan, “Tweet, tweet, retweet: Conversational aspects of retweeting on twitter,” pp. 1–10, 2010.
- [10] A. Kumar, D. Chhabra, B. Mendiratta, and A. Sinha, “Potential extensions and updates in social media for twitter developers.” *International Journal of Performability Engineering*, vol. 16, no. 8, 2020.

- [11] I. Dongo, Y. Cadinale, A. Aguilera, F. Martínez, Y. Quintero, and S. Barrios, “Web scraping versus twitter api: A comparison for a credibility analysis,” p. 263–273, 2021. [Online]. Available: <https://doi.org/10.1145/3428757.3429104>
- [12] J. Pfeffer, A. Mooseder, J. Lasser, L. Hammer, O. Stritzel, and D. Garcia, “This sample seems to be good enough! assessing coverage and temporal reliability of twitter’s academic api,” 2022.
- [13] M. Nagarajan, H. Purohit, and A. Sheth, “A qualitative examination of topical tweet and retweet practices,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, pp. 295–298, May 2010.
- [14] K. E. Anderson, “Ask me anything: what is reddit?” *Library Hi Tech News*, vol. 32, no. 5, pp. 8–11, 2015.
- [15] A. Medvedev, R. Lambiotte, and J.-C. Delvenne, “The anatomy of reddit: An overview of academic research,” *Springer Proceedings in Complexity*, pp. 183–204, 05 2019.
- [16] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, “The pushshift reddit dataset,” 01 2020.
- [17] T. Weninger, X. A. Zhu, and J. Han, “An exploration of discussion threads in social news sites: A case study of the reddit community,” p. 579–583, 2013.
- [18] J. Sawicki, M. Ganzha, and M. Paprzycki, “Reddit-tudfe: practical tool to explore reddit usability in data science and knowledge processing,” *CoRR*, vol. abs/2110.02158, 2021.
- [19] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, “The pushshift reddit dataset,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 830–839, May 2020.
- [20] S. Kannan, V. Gurusamy, S. Vijayarani, J. Ilamathi, M. Nithya, S. Kannan, and V. Gurusamy, “Preprocessing techniques for text mining,” *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2014.
- [21] Twitter api v2 data dictionary/ user. [Online]. Available: <https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/user>



- [22] Twitter api v2 data dictionary/ tweet. [Online]. Available: <https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet>
- [23] overview/ type prefixes. [Online]. Available: [https://www.reddit.com/dev/api/overview/type\\_prefixes](https://www.reddit.com/dev/api/overview/type_prefixes)
- [24] R. Sunil, V. Jayan, and V. K. Bhadrán, “Preprocessors in nlp applications: In the context of english to malayalam machine translation,” pp. 221–226, 2012.
- [25] M. Anandarajan, C. Hill, and T. Nolan, “Text preprocessing,” pp. 45–59, 2019.
- [26] S. Gharatkar, A. Ingle, T. Naik, and A. Save, “Review preprocessing using data cleaning and stemming technique,” pp. 1–4, 2017.
- [27] T. Ming, Hsiang, “Research challenges and opportunities in mapping social media and big data,” *Cartography and Geographic Information Science*, vol. 42, no. sup1, pp. 70–74, 2015.
- [28] Document length. [Online]. Available: [https://maartengr.github.io/BERTopic/getting\\_started/tips\\_and\\_tricks/tips\\_and\\_tricks.html](https://maartengr.github.io/BERTopic/getting_started/tips_and_tricks/tips_and_tricks.html)
- [29] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, and A. Chupryna, “Effectiveness of preprocessing algorithms for natural language processing applications,” pp. 187–191, 2020.
- [30] B. Khemani and A. Adgaonkar, “A review on reddit news headlines with nltk tool,” *Proceedings of the International Conference on Innovative Computing Communication (ICICC) 2021*, 2021.
- [31] Universal pos tags. [Online]. Available: <https://universaldependencies.org/u/pos/>
- [32] M. McDonnell, J. E. Owen, and E. O. Bantum, “Identification of emotional expression with cancer survivors: Validation of linguistic inquiry and word count,” *JMIR Form Res*, vol. 4, no. 10, p. e18246, Oct 2020.
- [33] How it works. [Online]. Available: <https://www.liwc.app/help/howitworks>
- [34] M. Grootendorst, “Bertopic: Neural topic modeling with a class-based tf-idf procedure,” 2022.

- [35] Embedding models. [Online]. Available: [https://maartengr.github.io/BERTopic/getting\\_started/embeddings/embeddings.html#sentence-transformers](https://maartengr.github.io/BERTopic/getting_started/embeddings/embeddings.html#sentence-transformers)
- [36] Frequently asked questions. [Online]. Available: <https://maartengr.github.io/BERTopic/faq.html>
- [37] all-minilm-l6-v2. [Online]. Available: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- [38] find\_topics(). [Online]. Available: [https://maartengr.github.io/BERTopic/api/bertopic.html#bertopic.\\_bertopic.BERTopic.approximate\\_distribution](https://maartengr.github.io/BERTopic/api/bertopic.html#bertopic._bertopic.BERTopic.approximate_distribution)
- [39] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2020.
- [40] Parameter selection for hdbscan\*. [Online]. Available: [https://hdbscan.readthedocs.io/en/latest/parameter\\_selection.html](https://hdbscan.readthedocs.io/en/latest/parameter_selection.html)
- [41] c-tf-idf. [Online]. Available: [https://maartengr.github.io/BERTopic/getting\\_started/ctfidf/ctfidf.html](https://maartengr.github.io/BERTopic/getting_started/ctfidf/ctfidf.html)
- [42] Outlier reduction. [Online]. Available: [https://maartengr.github.io/BERTopic/getting\\_started/outlier\\_reduction/outlier\\_reduction.html](https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html)
- [43] Topic representation. [Online]. Available: [https://maartengr.github.io/BERTopic/getting\\_started/topicrepresentation/topicrepresentation.html](https://maartengr.github.io/BERTopic/getting_started/topicrepresentation/topicrepresentation.html)
- [44] Topic visualization. [Online]. Available: [https://maartengr.github.io/BERTopic/getting\\_started/visualization/visualization.html#visualize-topics](https://maartengr.github.io/BERTopic/getting_started/visualization/visualization.html#visualize-topics)