

Categorization of menopause data from social media

Amogh Anil Rao
Trinity College Dublin, Ireland
raoam@tcd.ie

Chirag Saxena
Trinity College Dublin, Ireland
saxenac@tcd.ie

Junwei Yu
Trinity College Dublin, Ireland
yuju@tcd.ie

Vaibhav Srivastava
Trinity College Dublin, Ireland
vsrivast@tcd.ie

1 Abstract

The loss of hormones during menopause is a difficult stage of life that has both physical and psychological effects that might interfere with day-to-day activities. As technology is on the rise, social media platforms and forums can give useful knowledge and social support to people going through menopause, even though medical experts might not always be suited to do so. The study here aims to verify the hypothesis that increased engagement on social media platforms among women undergoing menopause and other section of people have led to enhanced awareness, better access to accurate information, and better support networks which leads to more effective management and coping strategies, both physically and mentally, for women suffering from menopausal symptoms. To explore this hypothesis, the study will analyze discussions around search terms related to menopause support and interaction from Twitter and Reddit, as well as identify language usage variations based on gender, menopausal status, and other clusters using K-means clustering and Latent Dirichlet Allocation (LDA) technique. The study may contribute to a better knowledge of support for people going through menopause by illuminating the terminology used to address menopause in various contexts. **Keywords— Topic Modeling, Clustering, K-Means, Latent Dirichlet Allocation, Twitter, Reddit**

2 Introduction

Menopause is a natural life transition that marks the end of a woman's reproductive years. As mentioned in Rodríguez-Landa et al. (2015), the hormone synthesis stops, which can cause a variety of physical and psychological problems, and it commonly strikes between the ages of 45 and 55. Hot flashes, mood swings, sleep issues, and dry vaginal skin are typical symptoms. Moreover, certain medical disorders including osteoporosis and cardiovascular disease can become more likely throughout menopause.

Although it is a common and natural occurrence, it is still stigmatized and poorly understood in society, as per Short (2017) Menopause remains somewhat taboo, that support is often hard to find for women who need help and there is still a great need to educate both primary and secondary care clinicians in the area of post-reproductive health, the lack of education makes its symptoms more severe on both physical and psychological levels, Han Mi-Jeong (2013) speaks about the signs of depression in middle-aged women going through menopause. In addition, medical personnel does not always have the necessary resources or training to offer menopausal women the help they need, which can make many women feel alone and unsupported. As a result, more women are turning to online forums and services for menopause-related knowledge and support. For menopausal women, the internet has grown to be a key resource for knowledge and assistance. As mentioned in Bresnahan and Murray-Johnson (2002) Internet sources, including websites, blogs, and discussion boards, give access to a wealth of knowledge on menopause, its symptoms, and treatment options. These sites also give women the chance to interact with others who are going through the same symptoms, share their stories, and discuss coping mechanisms. However, the reliability and accuracy of internet resources can vary greatly, and some women might find it difficult to locate the helpful advice and assistance they need.

By examining differences in language use based on factors like the authors' gender, whether the author is experiencing menopause or is interacting with someone who is, and the type of online medium in which the text

appears, this study aims to fill this knowledge gap. The texts were collected as per Asad et al. (2019) from Twitter and Reddit using terms related to menopause support and interaction. By illuminating the terminology used to address menopause in diverse circumstances, this study has the potential to enhance knowledge and support for people going through menopause. It might also help to broader discussions about gender and health by stressing the demand for additional study and assistance for problems with women's health. The results of this study could also be used to guide the creation of therapies and tools that better serve the needs of menopausal women and the people they care about.

Our study aimed to extract texts from Twitter and Reddit using web scraping techniques and identify different categories of the extracted data through text clustering mechanisms as mentioned in the methodology section which is the combination of Ahuja and Dubey (2017) and Suadaa and Purwarianti (2016). The identified categories include individuals supporting menopause, those undergoing menopause, and many more. The paper intends to perform sentiment analysis on the various topics identified by the algorithms by categorizing the data into different groups and identifying the attitudes and opinions of individuals toward menopause. This method provides a comprehensive understanding of the sentiment of people towards menopause and identifies different categories, enabling us to draw insights and recommendations for health practitioners and policymakers.

3 Research Questions

Menopause is a natural phase in a woman's life that is typically seen as taboo and has minimal assistance accessible for people experiencing it causing a lot of discomfort to the person experiencing it. To address this issue and gain a better understanding of menopause from various perspectives, this study poses the following research questions based on our main hypotheses.

1. **Research Question 1:** To gain a better understanding of the societal perspective on menopause and to what extent they can provide assistance to people undergoing menopause.
2. **Research Question 2:** To identify the speaker in literature to obtain a greater comprehension of menopause discussion.

The need for highlighting the above-mentioned is already clearly mentioned in the Introduction section and the mechanism to achieve this can be found below in the Methodology section. We collected data from online platforms, which will be discussed in the methodology section. The authors of the texts on online platforms have diverse backgrounds and the quantities of collected data are adequate for us to investigate these questions and test the hypotheses.

4 Related work

Menopause even though a natural life transition for women is still treated as a taboo and as given in Lee et al. (2015) and Han Mi-Jeong (2013), it is evident that many women face issues in communicating about it and experience lack of understanding. Short (2017) shows the importance of social media platforms to express themselves, which validates our approach for extracting data from social media platforms. These papers helped us confirm the premise of our paper and validate our preliminary approaches.

The paper Wang et al. (2016) speaks about the sentiment analysis mechanism for generalized text extracted from social media and Cho and Kang (2012) presents a statistical text analysis approach for sentiment classification in social media. We use it to identify the sentiment of people towards menopause by analyzing social media data. Asad et al. (2019) gives us the mechanism to extract texts from social media platforms and also some base for sentiment analysis to identify depression amongst people, which is used in our case for menopause. By extracting sentiment and identifying emotions expressed in menopause-related tweets. It acts as a useful tool for sentiment analysis of menopause-related tweets and identifying different categories of people based on their sentiment towards menopause.

Heylighen and Dewaele (2002) explore the idea of contextuality in language and suggest a new empirical measure to evaluate it in their study. The paper's theoretical and methodological focus, though, is more important than the topic of menopause and how it is expressed in language. This helps us identify and categorize the topics from the data found using LDA and K-means on the scrapped data from social media platforms.

Suadaa and Purwarianti (2016) and Ahuja and Dubey (2017) give us deep insights into mechanisms used for

the clustering of texts using LDA, K-means, etc., which have also been used in our case for the clustering of menopause-related texts to identify various categories amongst the extracted texts. To identify different categories of people in our study, we can use text clustering mechanisms mentioned in the methodology section. These mechanisms can group tweets based on common characteristics, such as the gender, age, and geographic location of the users. By analyzing the sentiment of each category separately, we can gain insights into how different groups of people perceive menopause. For example, we may find that women over the age of 50 tend to have a more positive sentiment toward menopause than younger women. By identifying such patterns, we can provide more targeted and personalized support for people going through menopause.

Social media such as Twitter and Reddit, used regularly by more than 1/7th of the world's population as given in Miller (2011) gave social scientists a great opportunity to study human communication. Schwartz et al. (2013) found the linguistic feature is strongly related to personality, gender, and age, which enables us to bifurcate the data. Althoff et al. (2016) found some insights that could help improve counselor training and give rise to real-time quality monitoring and answer suggestion support tools. Golder and Macy (2011) analyzed the data from millions of public Twitter messages, and found that individuals wake up in a good mood that deteriorates as the day progresses. De Choudhury et al. (2013) introduced a social media depression index that may serve to characterize levels of depression in populations with the data gleaned from social media. Menopause is a topic that attracts significant attention online. Platforms such as Twitter and Reddit are excellent forums to gather corpus around menopause, and the users have wide backgrounds such as gender, nationality, and professions. So we can collect data from online forums to investigate our research questions and validate our hypothesis.

Pennebaker et al. (2001) introduced the Linguistic Inquiry and Word Count (LIWC) algorithm, a text analysis tool that calculates the frequency of psychologically meaningful categories in a given corpus. They demonstrated its usefulness in various domains, including social media. Tausczik and Pennebaker (2010) provided a comprehensive overview of LIWC around the applications and limitations. summarized some works related to emotions, personality, and mental health. Golder and Macy (2011) analyzed users' emotions on Twitter during specific periods, revealing that work, sleep, and day length are factors for inference of the emotions. Suadaa and Purwarianti (2016) used LIWC revealed that people's language became the most personal and informal after a breakup. LIWC relies on a predefined dictionary, which may need more manual work and not cover all relevant words and phrases in specific domains and languages. And LIWC is not capable of capturing context-dependent features, while context-dependent features are quite useful for analyzing natural languages. Since LIWC requires a specialized dictionary and we don't have experts to annotate it, we chose not to use LIWC. However, it provided valuable insights for our project.

Blei et al. (2003) introduced the Latent Dirichlet Allocation (LDA) algorithm, a generative probabilistic model that allows for discovering latent topics in large text corpora. With the torrent of data from social media, LDA can find latent topics while not needing a predefined dictionary. Mitchell et al. (2015) created a classifier for classifying users in isolation and found that with the Support Vector Machine classifier, Features extracted from LDA achieve 80.4% accuracy without manual work while LIWC only achieved 68.8%. Hong and Davison (2010) conducted extensive quantitative experiments on standard LDA and demonstrated that topic models learned from aggregated messages by the same user may lead to superior performance in classification problems and topic model features can improve performance in general. LDA offers a solution to LIWC's limitations when investigating menopause, as it doesn't require specific human annotation to identify topics within a corpus. This allows us to explore menopause-related topics.

MacQueen (1967) introduced the K-means algorithm and became a widely used clustering technique in various fields, including text analysis. Steinbach et al. (2000) compared the two main approaches to document clustering, agglomerative hierarchical clustering, and K-means, and found that K-means rely on a more global approach, which effectively amounts to looking at the similarity of points in a cluster with respect to all other points in the cluster and leads to K-means performing better than the hierarchical clustering solution. Suadaa and Purwarianti (2016) provided Indonesian text clustering with labeling by using LDA and Term Frequency-Inverse Cluster Frequency (TFxICF) and found that the LDA algorithm produces documents cluster by topic with cluster quality better than K-Means and Lingo and Word-based LDA generates cluster with better quality than phrase based on LDA. Based on the experiments, they proposed a method that used word-based LDA for clustering and phrase-based TFxICF for labeling. Without prior labels of the corpus, K-Means clustering can enable us to effectively group similar documents together and get some patterns of our data.

5 Methodology

5.1 Data Gathering and Preprocessing

5.1.1 Twitter Data Scrapping

In this study, we aimed to collect tweets related to menopause using the snsrape Python module. We used the TwitterSearchScraper class from snsrape to search for tweets containing the keyword "Menopause" within the specified date range (from 2022-07-01 until 2023-04-12). We set a limit of 50000 tweets to be scraped to minimize data collection time. For each tweet, we extracted the date, ID, content, and username using the corresponding attributes of the tweet object.

After scraping all relevant tweets, we converted the scrapped tweet data to a Pandas DataFrame using the DataFrame() constructor. We specified the columns of the DataFrame to be 'Datetime', 'Tweet Id', 'Text', and 'Username' to match the extracted tweet data. The scraped tweet data was saved as a CSV file. We specified the separator as ',' and encoding as 'utf-8' for compatibility with most text editors and data analysis tools.

Overall, this script demonstrates the use of the snsrape module for Twitter data scraping and Pandas for data manipulation and storage. It can be used as a basis for collecting and analyzing Twitter data on menopause, or other relevant topics.

5.1.2 Reddit Data Scrapping

Social media platforms, such as Reddit, provide a vast source of user-generated data that can be analyzed to gain insights into various topics. In this context, the PRAW (Python Reddit API Wrapper) library is often used to extract data from Reddit. The code presented here demonstrates the use of PRAW to extract data related to the topic of menopause from the Reddit platform. The code uses a Reddit client ID, secret key, and user agent to authenticate the script and access the relevant data. The script first defines the topic to search for and the subreddit to search within. The top 100 posts in the last month are then retrieved for the given topic, and the content of each post is extracted. Finally, the extracted data is stored in pandas data frames for further analysis. The extracted posts and comments are then combined together for further processing.

5.1.3 Data Preprocessing

Preprocessing of text data is a crucial step in the analysis of unstructured textual data, as it assists in removing noise and converting the raw data into a structured format that can be fed into machine learning models. The textual data generally contains noise in different forms such as emoticons, punctuation, text in varying cases, and other irregularities. Therefore, text preprocessing techniques such as tokenization, stop-word removal, stemming, and lemmatization are applied to clean the textual data and make it more suitable for downstream analysis. This process ensures that the text is structured in a way that the machine learning models can understand, enabling accurate and efficient analysis.

Stopwords are frequently occurring words in text that carry little or no meaning in terms of content. Examples of stopwords include "the", "of", "and", "or", "of", and "that". To eliminate these stopwords and other unnecessary elements from text data, the Natural Language Toolkit (NLTK) library is commonly employed. The NLTK library provides approximately 180 stopwords that can be removed from text data to enhance its quality. In our study, we utilized NLTK to remove stopwords, URLs, and punctuation from raw sentences during the text preprocessing stage. This process facilitated the transformation of unstructured text data into a structured and readable format, which is essential for effective sentiment analysis.

Stemming is a process to reduce the word to its root stem for example run, running, runs runned derived from the same word as run. NLTK library is used to stem the words. The stemming technique is not used for production purposes because it is not so efficient technique and most of the time it stems the unwanted words. So we use lemmatization instead. Lemmatization is similar to stemming, used to stem the words into root word but differs in wording. Actually, Lemmatization is a systematic way to reduce the words into their lemma by matching them with a language dictionary.

5.2 Sentimental Analysis

Sentimental analysis is a technique used to examine the emotional tone behind the text. It uses machine learning techniques and can be divided into 3 categories positive, negative, and neutral. Calculating the sentiment score of Twitter and Reddit comments on menopause topics can help the researchers understand the overall attitudes and opinions of people towards menopause and menopausal symptoms. This information can be used to identify the types of support and information that women undergoing menopause are seeking and the kind of language that resonates with them. By analyzing the sentiment of social media comments, the researchers can gain insights into the effectiveness of different coping strategies and support networks and identify areas where more education and support are needed. Overall, sentiment analysis can provide valuable insights into the experiences of women undergoing menopause and help inform the development of more effective support systems.

We used the VADER sentiment analyzer to analyze Twitter tweets and Reddit comments. VADER analyzes the text and gives each word or phrase in the text a polarity score. The score ranges from 0 (neutral) to +1 (extremely positive), with -1 (extremely negative) being the lowest. The words "love" and "feels" would each receive a high positive polarity score from VADER's analysis of this sentence, giving the sentence a score for overall positive sentiment.

5.3 Text Clustering and Topic Modelling

5.3.1 TFIDF

TFIDF stands for Term Frequency Inverse Document Frequency. It is the multiplication of the two terms TF (Term frequency) and IDF (Inverse Term Frequency). TF (Term Frequency) is calculated by the division of the Total number of repetitions of words in sentences and the number of words in sentences. IDF (Inverse Document Frequency) is calculated by the logarithmic division of the number of sentences and the number of words in sentences.

Let's say we have 3 sentences S1: He is the good boy, S2: She is a good girl, S3: Boy and girl are good. Then below TFIDF vector representation.

		Boy -> 2	Girl -> 2	Good -> 3
TF =		S1	S2	S3
	Good	1/2	1/2	1/3
	Boy	1/2	0	1/3
	Girl	0	0	1/3
IDF =	Words	IDF		
	Good	$\log(3/3) = 0$		
	Boy	$\log(3/2)$		
	Girl	$\log(3/2)$		

Figure 1: Example of TFIDF Working

TF-IDF is not used for clustering the text instead it is used differently in the study conducted here. After calculating TF-IDF scores for each term in the corpus, we identified the terms with high TF-IDF scores and filtered out the ones that are common across documents or the ones that do not convey anything meaningful with respect to menopause. Note that TF-IDF was applied after removing the stop words and hence common English words were already removed from the corpus. Some of the words like "car", "weather", "train" etc. which are not related to menopause kept popping up in the topics generated by the trained LDA model. A threshold value of 0.03 was set to filter out the words based on their TF-IDF score and irrelevance to menopause. These words were then removed from the corpus.

5.3.2 N-grams

The n-gram approach is a technique used in text analytics to represent textual data. An n-gram is a contiguous sequence of n items where items either represent a word, character, or any other unit of text. In our study, we employed the n-grams approach to extract features from the corpus. The N-grams approach has proven to be successful as can be seen from the results of the study undertaken by Kosmajac and Kešelj (2020). Data related to menopause cannot be interpreted completely just by extracting single words. We also observed that words like protect characteristic when processed separately as single words do not provide enough context but when considered as a bigram starts to make sense. The same goes for the words brain and fog. The choice of the value of n is an important parameter that can affect the performance of the n-grams approach. For the study, we considered values of n from 2 to 5 and modified the corpus. Then the LDA model was trained for each of these corpora. Distinct and higher coherence scores within a topic occurred for trigrams i.e. for $n = 3$. Therefore, the n-grams method proved effective in identifying patterns and relationships within the text data, and the best value of n was chosen to be 3.

5.3.3 Latent Dirichlet Allocation (LDA)

The Latent Dirichlet Allocation (LDA) is a technique used in statistics for topic modeling. By analyzing the co-occurrence of words within a set of documents, LDA allows for the identification of underlying topics. The LDA model is a probabilistic model that assumes each document is a mixture of a set of topics, and each topic is characterized by a distribution of words.

Mustakim et al. (2021) follows an approach for topic modeling of 1000 medical reports using LDA. Following the same approach in the related field of menopause, we first gathered a corpus of tweets and Reddit comments and preprocessed them as mentioned in the previous section. Applying LDA requires us first to select a set of N topics to extract which was selected based on a grid search of different values of N and finding out the coherence score of each of them. Since we are dealing with Reddit and Twitter data separately, a different number of optimal topics were chosen for both of them based on the coherence scores. The coherence score is an important metric in selecting the optimal number of topics as it is a measure of the degree of semantic similarity between the words in a topic and can be used to verify the quality of the topics. The corpus was also categorized using bigrams and trigrams to capture more important terms which provided more context when words were grouped together in a sequence of two or three. TFIDF was used in a different way to remove commonly occurring words that were not removed using NLTK stopwords. These words occur in a general conversation and did not provide enough context in terms of menopause. Finally, the LDA model was trained on this corpus and according to the optimal number of topics determined by the coherence scores.

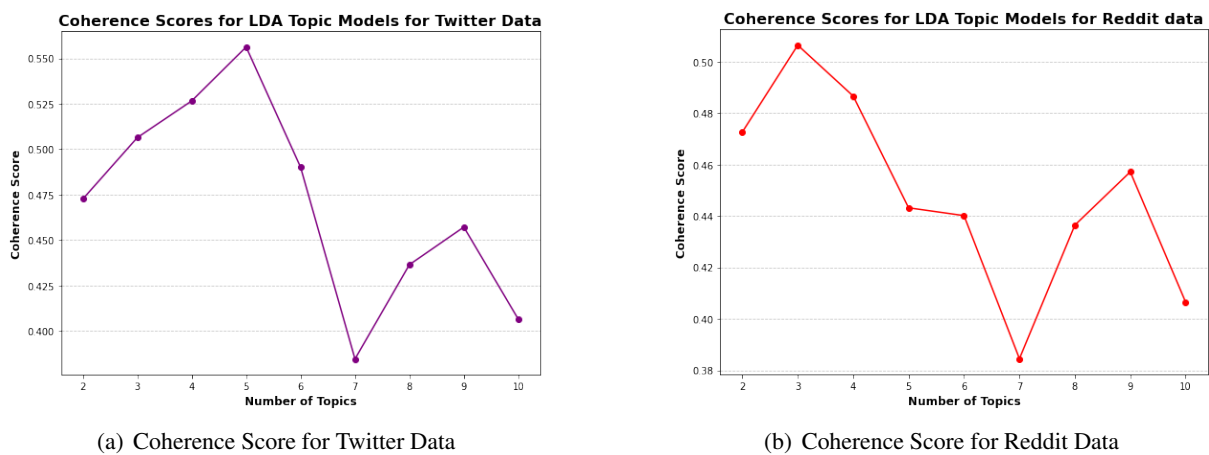


Figure 2: Coherence scores with varying number of topics (2-11)

5.3.4 Word2Vec and KMeans

Word2Vec is a widely used technique in natural language processing that creates vector representations of words known as word embeddings. These embeddings are generated in a high-dimensional space and capture the semantic and syntactic relationships between words. They can be used as features for various NLP tasks such as text

classification, sentiment analysis, and machine translation.

Word2Vec is based on a neural network architecture that utilizes a large text corpus to generate word embeddings. The neural network is trained to predict the context words around a target word. The two primary training algorithms used in Word2Vec are the Continuous Bag of Words (CBOW) and the Skip-Gram models. The CBOW model predicts the target word based on its surrounding context words, while the Skip-Gram model predicts the context words given the target word.

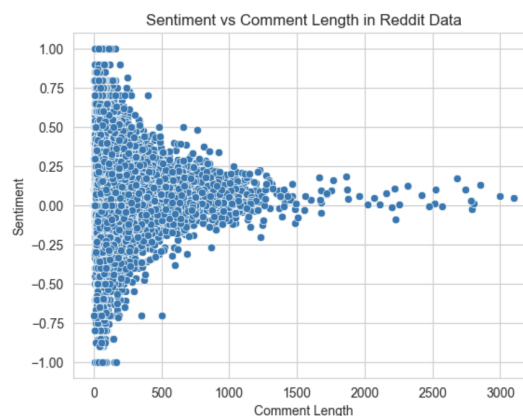
K-means is an unsupervised machine learning technique used to cluster unlabelled data. It is useful when someone wants to divide the data into 'K' unique groups to find underlying patterns. In our case, we need to find a societal perspective on menopause.

Haider et al. (2020) followed an approach to summarize news articles using Kmeans and word embedding techniques in Word2Vec. Following the same approach, Firstly we preprocessed the data. Then we obtained the word embeddings by training the corpus on the neural network using the Word2Vec technique. These are the dense vector representations that capture the semantic information of words. After that, we averaged out all the Word2Vec vectors to give us sentence embeddings. Word2Vec uses a vector representation for each word at a predetermined length; all other elements of the vector are zeros aside from the element that represents the words. Similar-sounding phrases are positioned closer together in space. We used vector size 100 and the Skip Gram technique. Skip Gram generates the embedding by capturing the semantics of surrounding words. It performs better than CBOW (continuous bag of words) which is another technique in Word2Vec. Now we employed the sentence embedding on the Kmeans clustering algorithm. To select the optimal 'K' value, we have used the elbow method. In our research we found out that the best value of 'K' or Topics for Reddit is 4 and for Twitter is also 4.

6 Results and Discussions

6.1 Sentimental Analysis

We have conducted sentiment analysis on comments from both Reddit and Twitter. For Reddit, the average length of comments is 151.07, with a median length of 90 and a range of 3099 characters. The overall sentiment score for Reddit comments is 0.23, which is the mean score across all comments. Additionally, the standard deviation of the sentiment score is 0.5. This means that the sentiment of Reddit comments is quite varied, with some comments expressing very positive emotions and others expressing very negative ones.



(a) Relationship between comment length and sentiment

We found out that there is a relationship between comment length and sentiment of that comment in Reddit data only. Above is the plotted figure depicting the connection between sentiment and comment length. We can clearly see that comments longer than 1,500 characters have a neutral sentiment of around 0. Additionally, we can observe that points around 0 are sparsely distributed, indicating a smaller number of neutral comments in the Reddit data

The same holds true for Twitter data, where the standard deviation of the sentiment score is 0.5 and the overall sentiment mean is 0.13. On Twitter, we found a wider variety of comments, with more positive and neutral

comments than negative ones. Specifically, the breakdown of sentiment percentages on Twitter is 47 percent positive, 31 percent neutral, and 22 percent negative. In contrast, Reddit had 45 percent positive, 14 percent neutral, and 41 percent negative comments.

(b) Most frequent words in Twitter Data

The intertopic distance map was plotted for the 5 topics along with the top 30 words with their frequency of occurrence within a topic. The plot can be seen in Figure 3. Each circle represents a topic, and the arrangement of the circles is based on their similarity or distance from one another. Topics that have a shorter distance between them are more similar to each other while topics that are far apart from each other have a very low similarity between them. Looking at the figure, we can conclude that the 5 topics identified by the LDA model are far apart from each other and none of them intersect, which specifies that these five topics talk about various aspects of menopause.

Table 1: Top 10 Important Words in Each Topic

Topic ID 1	Topic ID 2	Topic ID 3	Topic ID 4	Topic ID 5
woman	hrt	leave	badenoch	support
experience	treatment	work	cause	make
age	affect	health	right_wing	hope
suffer	condition	workplace	minister	problem
need	workplace	mean	political	joy
know	protect_characteristic	fiber	government	insomnia
help	sex	employer	agree	fun
give	someone	struggle	deal	hell
understand	wife	perimenopause	state	shame
man	dr	career	disability	sleep

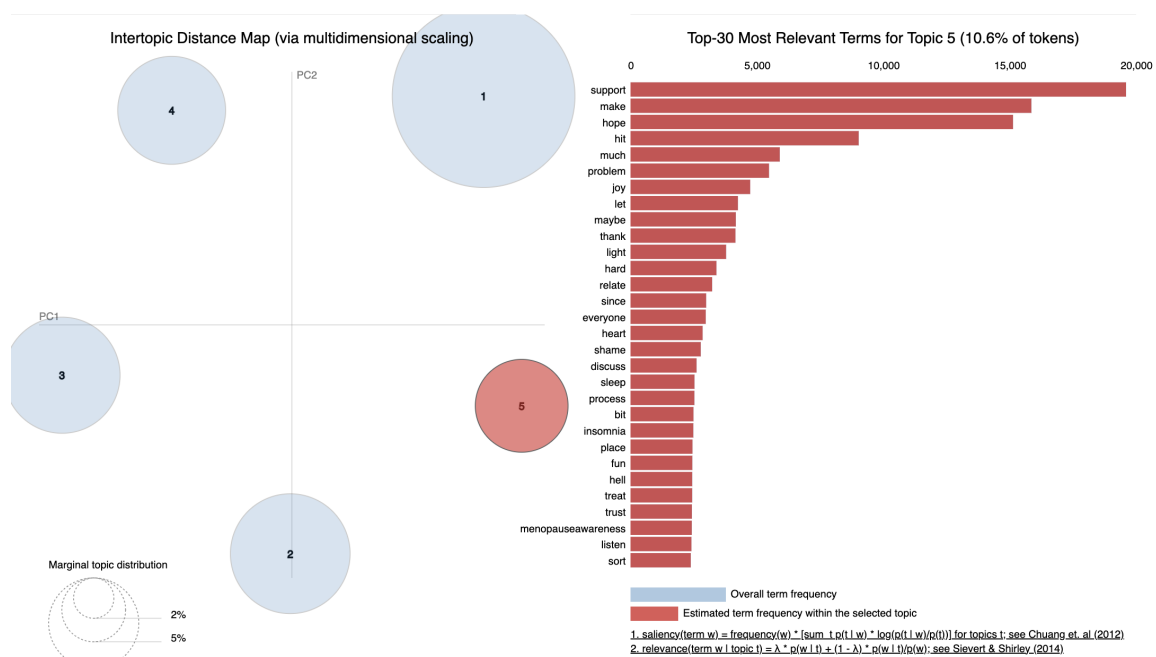


Figure 3: Intertopic distance map and top 30 relevant terms for Twitter Data

To identify what is being talked about in the topic and interpret it, we created a table listing the top 10 important words under each topic. Table 1 shows the top 10 most important words related to menopause under each topic.

Looking at the first topic, we can see the use of words like "woman", "experience", "age", and "suffer" which suggests that the topic is centered around women and their experiences with menopause. The words like "need", "know", "get" etc. show that the topic also involves discussion around seeking information and advice related to women's health. Additionally, words like "help", "give", "understand" etc. suggests that women are seeking help in the form of support and solutions while experiencing difficulties related to menopause. Based on the words in this topic, it is highly probable that the tweets are by women who are or already have experienced menopause.

Based on the words in the second topic, it can be possibly related to hormonal changes, treatment, and the impact of menopause. Usage of words like "hrt", "treatment", "change", "condition" etc. talks about the medical aspects of menopause. The words "affect", "find", "policy", "protectcharacteristic" etc. possibly tells about the discussions on the impact of menopause on women's lives and the need for the right policies to protect them. The topic also includes words like "wife", "sex" etc. which can possibly suggest that menopause may also affect the lives of the partners. Excessive use of medical terms suggests that medical professionals are the ones sharing their experiences online. Additionally, husbands of women undergoing menopause are talking about their experiences with menopause as well.

The words in the third topic mostly relate to words that are used in the workplace. This topic can possibly be labeled as "Workplace and Menopause" since the words in the topic suggest that people are talking about the hardships and challenges faced by women at the workplace undergoing menopause. We can see that the words "leave", "work", "take", "career", "struggle", and "employer" highly point to the possibility of discussions like women may need to take time off the work or struggle with symptoms of menopause while working or menopause serves as a blocker for their career. Based on the words under this topic, it would seem that working women who are undergoing menopause as well as female activists are the ones tweeting about menopause and work life.

The words for topic 4 that are included here as well as more words that come under the topic reveal a very interesting observation. The word "badenoch" refers to the minister Kemi Badenoch who is a Member of Parliament in the UK who recently was involved in a fight for a law related to menopause. The presence of other words like "minister", "rightwing", "political", "government" etc. indicates a focus on political figures and how menopause became an issue recently which was discussed politically. This topic possibly talks about the need for greater attention to menopause-related issues in political and social spheres. Since the topic is mostly focused on political and social issues related to menopause, we can conclude that section of people that are using these words are likely the people who are politically active or interested in social issues related to menopause.

Looking at the words in topic 5, the topic can broadly be classified as "Coping with Menopause Symptoms". Words like "insomnia", "hot flashes" etc. indicate that symptoms related to menopause are being talked about. Usage of words such as "support", "make", "joy" etc. indicate that the tweets are looking for offering support and positivity in dealing with the harsh symptoms of menopause. Words such as "hell", "shame", "hard" etc. suggests how difficult it gets to cope with symptoms of menopause and may affect a person's emotional state. The topic could be representative of women undergoing menopause and their loved ones who are seeking information online to help with menopause symptoms.

6.2.2 LDA for Reddit Data

As in the case of the LDA model for Twitter data, we once again plotted the intertopic distance map along with the top 30 words with their frequency of occurrence within a topic. The plot can be seen in Figure 4. Based on the topic coherence, we selected the optimal number of topics to be 3 for training the LDA model for Reddit data and hence we can see 3 circles that correspond to 3 different topics. The bigger size of the circle in the visualization corresponds to the prevalence of the topic across the corpus. Again, topics that have a shorter distance between them are more similar to each other while topics that are far apart from each other have a very low similarity between them. The three circles are far apart from each other and hence talk about different things regarding menopause.

Similar to what was done for Twitter data, we created a table listing the top 10 important words under each topic to estimate what is being talked about in each topic and possibly interpret it. Table 2 shows the top 10 important words in each of the three topics related to menopause.

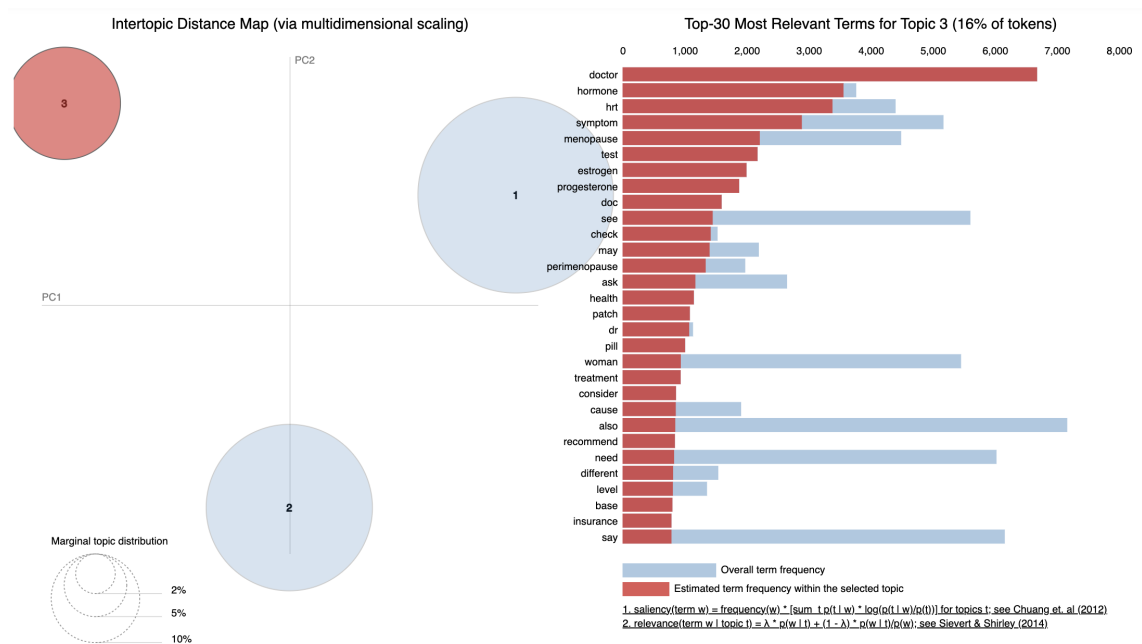


Figure 4: Intertopic Distance map and top 30 relevant terms for Reddit data

The words that are used in topic 1 talk more regarding general discussions related to menopause. Words under this topic include common verbs, adjectives, and nouns used commonly. However, other words like "get", "take", "help" etc. suggest that people are seeking help from the online community for their symptoms. It is difficult to label it accurately because of the usage of common words and the absence of patterns of words. It also becomes challenging to derive which section of people are tweeting here.

Table 2: Top 10 Important Words in Each Topic

Topic ID 1	Topic ID 2	Topic ID 3
think	peri	doctor
know	period	hormone
help	bad	hrt
thank	feel	symptom
love	anxiety	test
female	adhd	progesterone
hard	exercise	estrogen
great	weight	treatment
year	back	prescribe
try	start	cancer

According to the words used in the second topic, the most suitable label for the words that fall under this topic would be "Coping with menopause symptoms and finding relief". The topic was derived based on words like "symptom", "bad", "feel", "anxiety" etc. which suggest the negative symptoms that need to be managed during menopause. There are a certain set of words that also indicate weight gain and physical activity as additional concerns. However, one unexpected word occurs with a high frequency which is "ADHD". It may hint towards women suffering from cognitive changes as well during menopause. These set of tweets are most likely written by women suffering from perimenopause and menopause. It seems like the concerned women are looking for solutions and support.

The set of words in topic 3 refers heavily to medical terms. It talks about medical treatments and tests for menopause symptoms. Words like "doctor", "hormone", "treatment", "HRT", "symptom", "estrogen", "progesterone" etc. refer to the same. It would seem like the section of the population talking here is mostly Doctors or women undergoing menopause and seeking medical attention.

6.2.3 Comparison of conclusions of Twitter and Reddit Data

The data gathered from both platforms contains discussions on menopause-related topics but as we will see the nature and focus of these discussions vary slightly.

Both platforms involve deep discussions related to medical treatments and hormonal changes that women experience while going through menopause. They are also similar when it comes to discussions around coping with menopause symptoms and seeking help from the online community. However, Twitter data has a more diverse range of discussions which is also evident from the higher number of optimal topics. It contains discussions related to the workplace, politics, and social issues as well which are missing from the Reddit data. Additionally, Twitter data involves people other than women undergoing menopause in its data which consists of tweets from medical professionals, husbands of women undergoing menopause, and politically active individuals. On the other hand, Reddit data focuses primarily on women undergoing menopause and looking for help.

All that being said, the goal here is to not compare the data about menopause extracted from Twitter and Reddit. Rather, conclusions made from both these data complement the findings of the research and help us in answering the research questions proposed in the paper. The findings suggest that both Twitter and Reddit provide a platform for women undergoing menopause to discuss their challenges and seek help from the online community. Additionally, Twitter data involves views of people other than women such as their partners, doctors, and politically active people sharing valuable information and providing support. These observations help in answering the first research question which aims to gain a better understanding of the societal perspective on menopause and to what extent they can provide assistance to women undergoing menopause. The second research question aims to identify the speaker in literature to obtain a greater comprehension of the menopause discussion. The findings from both the Twitter and Reddit data reveal that a diverse range of discussions regarding menopause takes place, with Twitter data containing a wider range of topics beyond just menopause symptoms and treatment. We were successfully able to provide a close estimate of the speaker in related topics which can provide a better understanding of menopause discussions and perspectives.

6.3 Topic Modelling using K-Means

As this method is different from LDA, here we performed clustering on two different datasets i.e Reddit comments and Twitter tweets. Our analysis shows that Reddit comments tend to have a more negative sentiment overall, while Twitter has a more positive and neutral view. We analyzed both platforms to get a dynamic overall conclusion. We took a suitable number of topics based on the elbow method, 4 for Reddit and 5 for Twitter. After training each model we will discuss the obtained patterns by identifying the most frequently occurring words within a cluster. We will look at the similarity and dissimilarities of these words and will try to infer the perception of people on menopause on social media.

6.3.1 K-Means for Twitter Data

Again, we will look at the top ten important and frequent words under each topic. Below is table 3 which displays the top ten most important words related to menopause under each topic for Twitter comments.

Table 3: Top 10 Important Words in Each Topic

Topic ID 1	Topic ID 2	Topic ID 3	Topic ID 4	Topic ID 5
get	way	face	woman	menopause
health	speaker	would	go	like
man	within	use	hormone	support
know	shame	hot	life	day
year	interested	long	also	feel
take	positive	bad	change	find
think	politic	laugh	want	skin
make	advocate	estrogen	give	workplace
help	decrease	love	look	free
hrt	inspire	book	heart	mental

The words in topic 1 are related to wellness and health, specifically related to Women's health and hormone replacement therapy (HRT). It suggests that individuals are asking for help related to their health and wellness and may have some previous personal experience with HRT. We also observed the use of personal pronouns like 'I' and 'We' etc which suggests that women that have undergone HRT or going through menopause are talking here.

The words in topic 2 seem to be associated with advocacy and activism, with a focus on positive change and promoting cause. The mention of 'shame' represents that the speaker is trying to decrease shame around the audience on certain topics or issues, and is interested in promoting a positive message. The use of words like interested, advocate, and politic suggests that apart from women talking here, individuals that are interested in politics or socially active are talking about menopause-related issues as well.

The words in topic 3 appear to be related to menopause and the physical changes that women go through during this time. The mention of the word 'estrogen' suggests the women experience hormonal changes and the mention of 'love' and 'laugh' that the speaker may be interested in finding ways to approach menopause with a positive attitude. Here the speaker could be anyone sharing their thoughts on menopause and the physical and emotional changes that women go through during this time.

The words in topic 4 also related to menopause but imply being more focused on the personal experiences of women going through this transition. The mention of 'change' suggests that the speaker is interested in exploring the emotional and psychological changes that women may experience during menopause. In this topic, the speaker is a women audience, interested in understanding the personal experiences of women going through this transition and exploring the emotional and psychological changes that women may face during this time.

The words in topic 5 seem to be focused on support and wellness during menopause. The mention of the word 'mental' suggests that the individuals may be interested in addressing the emotional and psychological challenges that women face and the mention of "workplace" suggests that the speaker may be interested in exploring how menopause can impact women's professional lives. Here the individual could be anyone man or a woman, a healthcare professional, a workplace manager, or a colleague. The main focus of the speaker is to better understand the challenges faced by women during this transition and to promote support and wellness in their personal and professional lives.

After a thorough discussion of the five topics, it can be determined that the heading of topic 1 encompasses "Women's Health and Hormone Replacement Therapy (HRT)." The focus of topic 2 is "Advocacy and Activism," while topic 3 pertains to "Menopause and Physical Changes." Topic 4 centers around "Personal Experiences of Menopause," and finally, topic 5 concerns "Support and Wellness During Menopause."

6.3.2 K-Means for Reddit Data

To identify and interpret what is being discussed in the topic, we created a table listing the top ten important and frequent words under each topic. Table 4 displays the top ten most important words related to menopause under each topic.

Table 4: Top 10 Important Words in Each Topic

Topic ID 1	Topic ID 2	Topic ID 3	Topic ID 4
thing	feel	year	doctor
work	estrogen	time	life
hormone	face	day	month
people	progesterone	menopause	symptom
lot	husband	help	week
something	pain	woman	night
peri	patch	period	body
try	eat	use	age
health	level	thank	hope
blood	cancer	hrt	test

The words in Topic 1 emphasize general health and well-being. The words like 'health', 'blood', and 'try'

may imply that individuals were talking about improving their health. The word 'lot', 'work', and 'people' may imply the discussion of improving their health through changes in their work or social lives. The terms 'hormone' and 'peri' imply that this topic may also be related to menopause and the hormonal changes that occur during this time. The discussion of menopause and hormonal changes might suggest that the conversation could be related to women sharing their experiences or discussing related health issues.

Words from Topic 2 seem to be connected to menopause-related bodily symptoms. Menopause can cause hormonal changes that can cause discomforts like hot flashes and mood swings, and words like 'estrogen', 'progesterone', 'pain', and 'patch' suggest that people may be interested in learning more about these changes. Also words like 'feel', and 'husband' indicates the effect on relationships and social-wellbeing. The presence of words like 'estrogen', 'progesterone', 'feel', and 'husband' could indicate the women sharing their experiences or discussing related health issues. However, this does not guarantee that the speaker is a woman; it could be a man or a person of any gender discussing the topic with a focus on the experiences of others.

The words in Topic 3 seem to be related to 'HRT' (hormone replacement therapy). Words like 'year', 'time', and 'day' suggests that individuals are interested to know when they should start and how long they should use these therapies. The word 'cancer' suggests the effects of using 'HRT' and its associated risks. After analyzing nouns in this topic, it was obvious to say that the speaker is women here concerned about the time and year of menopause occurrence so they can start HRT.

The words in Topic 4 emphasize mostly on menopause. Words like 'doctor', 'life', 'month', and 'symptoms', suggest that individuals may be seeking medical advice and treatment for menopause-related symptoms. The word 'week' may imply that people want to understand the cyclical nature of menopause symptoms. The word 'test' implies that individuals want to know diagnostic tests that can help identify menopause-related health issues. The word 'hope' suggests that individuals might be interested in seeking information and advice on how to control their menopausal symptoms and maintain their standard of living.

After a comprehensive discussion of the four topics, it can be concluded that the heading of Topic 1 pertains to "General Health and Wellbeing." Topic 2 pertains to "Menopause-Related Bodily Symptoms," while Topic 3 covers "Hormone Replacement Therapy (HRT)." Lastly, Topic 4 concerns "Medical Advice and Treatment for Menopause-Related Symptoms."

6.3.3 Comparison of conclusions of Twitter and Reddit Data

Based on the analysis of K-Means clustering on Twitter and Reddit data, the hypothesis that increased engagement on social media platforms among women undergoing menopause and other sections of people has led to enhanced awareness, better access to accurate information, and better support networks which leads to more effective management and coping strategies for women suffering from menopausal symptoms appear to be true.

Both datasets indicated that menopause is an important topic of discussion, with discussions around support and interaction being prevalent. The discussions around menopause on both platforms also highlighted the importance of seeking accurate information and advice, as well as the need for social support and effective coping strategies.

The similarity is that both datasets highlight menopause and women's health as important topics of discussion. Both datasets have topics that are related to hormone replacement therapy (HRT) and menopause-related symptoms. Additionally, both datasets suggest that women may be sharing their experiences or discussing related health issues.

7 Conclusion

This research intends to promote a deeper knowledge of menopause, its influence on society, and the help available to people experiencing it. By examining the language used in conversations about menopause, we can identify the requirements of women going through this stage of life and assist to give the necessary support. We employed two different algorithms i.e. K-means and LDA for identifying different topics within the corpus extracted from Twitter and Reddit. Analyses from both algorithms proved to be consistent and confirmed our hypothesis that menopause discussions are prevalent on social media platforms, with discussions revolving around topics related to medical treatments, coping strategies, and seeking support from the community. We also were able to identify to an extent the speaker in the literature which further confirmed our hypothesis by answering the research question


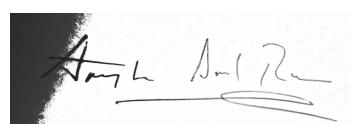

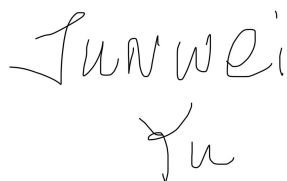
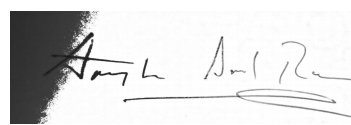


proposed. We observed that apart from women talking about menopause online, loved ones of the women, doctors, and individuals interested in politics also participated in the discussion. The findings suggest that social media platforms provide an avenue for women to discuss their challenges and seek help, which includes access to accurate information and support networks. Through the results of the research, we can conclude that social media platforms play a significant role in providing support and information for women undergoing menopause. Overall, this paper intends to contribute to a better knowledge of menopause and to emphasize the need for increased assistance and services for persons facing this natural life stage.

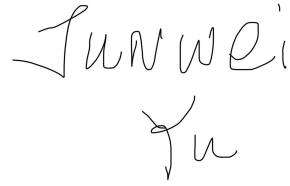
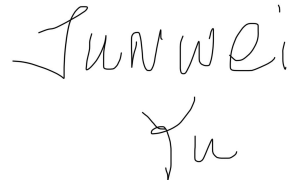


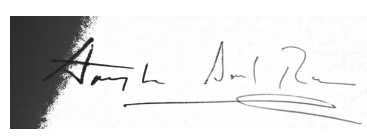
References


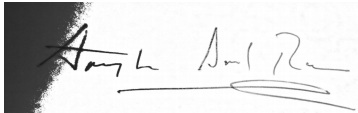

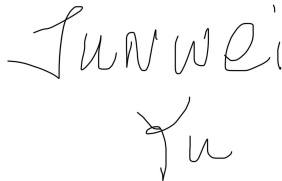
- Ahuja, S. and G. Dubey (2017). Clustering and sentiment analysis on twitter data. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, pp. 1–5.
- Althoff, T., K. Clark, and J. Leskovec (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics* 4, 463–476.
- Asad, N. A., M. A. Mahmud Pranto, S. Afreen, and M. M. Islam (2019). Depression detection by analyzing social media posts of user. In *2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON)*, pp. 13–17.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan), 993–1022.
- Bresnahan, M. J. and L. Murray-Johnson (2002). The healing web. *Health Care for Women International* 23(4), 398–407. PMID: 12148917.
- Cho, S.-H. and H.-B. Kang (2012). Statistical text analysis and sentiment classification in social media. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1112–1117.
- De Choudhury, M., S. Counts, and E. Horvitz (2013). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pp. 47–56.
- Golder, S. A. and M. W. Macy (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051), 1878–1881.
- Haider, M. M., M. A. Hossin, H. R. Mahi, and H. Arif (2020, June). Automatic text summarization using gensim word2vec and k-means clustering algorithm. In *2020 IEEE Region 10 Symposium (TENSYP)*, pp. 283–286.
- Han Mi-Jeong, L. J.-H. (2013). Factors influencing self-identity and menopausal symptoms on level of depression in middle aged women. *kjwhn* 19(4), 275–284.
- Heylighen, F. and J.-M. Dewaele (2002, 09). Variation in the contextuality of language: An empirical measure. Volume 7, pp. 293–340.
- Hong, L. and B. D. Davison (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pp. 80–88.
- Kosmajac, D. and V. Kešelj (2020). Language distance using common n-grams approach. In *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–5.
- Lee, M., B.-c. Koo, H.-s. Jeong, J. Park, J. Cho, and J. Cho (2015). Designing mhealth intervention for women in menopausal period. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pp. 257–260.
- MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297. University of California Los Angeles LA USA.
- Miller, G. (2011). Social scientists wade into the tweet stream.
- Mitchell, M., K. Hollingshead, and G. Coppersmith (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pp. 11–20.

- Mustakim, M., R. Wardoyo, K. Mustofa, G. R. Rahayu, and I. Rosyidah (2021). Latent dirichlet allocation for medical records topic modeling: Systematic literature review. In *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pp. 1–7.
- Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001), 2001.
- Rodríguez-Landa, J. F., A. Puga-Olguín, L. J. Germán-Ponciano, R.-I. García-Ríos, C. Soria-Fregozo, et al. (2015). Anxiety in natural and surgical menopause-physiologic and therapeutic bases. *Durbano F. A Fresh Look at Anxiety Disorders. London: IntechOpen*, 173–196.
- Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9), e73791.
- Short, H. (2017). The role of social media in menopausal healthcare. *Post Reproductive Health* 23(1), 4–5. PMID: 28381100.
- Steinbach, M., G. Karypis, and V. Kumar (2000). A comparison of document clustering techniques.
- Suadana, L. H. and A. Purwarianti (2016). Combination of latent dirichlet allocation (lda) and term frequency-inverse cluster frequency (tfxidf) in indonesian text clustering with labeling. In *2016 4th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–6.
- Tausczik, Y. R. and J. W. Pennebaker (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1), 24–54.
- Wang, Z., C. S. Chong, L. Lan, Y. Yang, S. Beng Ho, and J. C. Tong (2016). Fine-grained sentiment analysis of social media with emotion sensing. In *2016 Future Technologies Conference (FTC)*, pp. 1361–1364.

Statement of Contribution - Group 4

The contributions of each of the group members are mentioned below		
Name	Description	Signature
Vaibhav Srivastava (Chair) -> 22310317	Till Midterm: I have contributed to the preparation of the first draft of the paper by efficiently describing the problem statement, alongwith an overview of methodologies and significance of the study in the abstract section. Majority of the introduction was written by me. The scraping of the tweets was contributed by me and Amogh. The final review of the entire paper was done by me and minor changes were carried out. After midterm: After the midterm I switched to applying LDA to our scraped data and training the LDA model, identifying and labeling topics and interpreting them. I contributed to writing of the paper by writing relevant topics in the methodology section, writing results and discussion sections for LDA and writing the overall conclusion of the paper as well. Final proofreading of paper was also done by me as well. Along with the paper, I also fulfilled my responsibility as the chair and held meeting every week initially to review peer reviews that we had received and switched to frequent meetings near the deadline.	 Vaibhav Srivastava
		 Amogh Anil Rao
		 Chirag Saxena
		 JUNWEI YU
Amogh Anil Rao (Recorder) -> 22306378	Till Mid-term: I have contributed to the preparation of the first draft of the paper by contributing to the parts of the Introduction, Literature Review, and Data Gathering sections of the Methodology. Additionally, I have carefully proofread the document and implemented minor changes where necessary, ensuring that the paper meets the highest professional standards. After Mid-Term: Reading the peer-reviews and	 Amogh Anil Rao
		 Vaibhav Srivastava
		 Chirag Saxena
		Chirag Saxena

	<p>suggesting changes to the paper.</p> <p>Figuring out mechanisms to implement the changes.</p> <p>Finding more research papers to back every possible claim made in the paper.</p> <p>Worked on adding research questions and preliminary conclusions based on the existing results.</p> <p>Proofreading the document and suggested a few corrections.</p> <p>Throughout the course, I have kept track of the moments of all the meetings done by the team.</p>	 <p>Junwei YU</p>
<p>Junwei (Accountant) ->22304206</p>	<p>Midterm: I have contributed to the data collection on reddit with Chirag. And did the text data cleaning, feature extraction and logistic regression part. And analysis and reported the result for the logistic regression.</p> <p>For the final paper: I have contributed to the literature review for the methodology part. I read a paper and summarized it to share at the group meeting. And also in charge of making some changes to the paper to ensure it meets the requirements of the peer review part.</p> <p>As an accountant, I am responsible for recording the time devoted to the project and tracking the progress of each team member.</p>	 <p>Junwei Yu</p>
		 <p>Vaibhav Srivastava</p>
		 <p>Chirag Saxena</p>
		 <p>Amogh Anil Rao</p>

<p>Chirag (Ambassador and Verifier) 22312293</p>	<p>Till midterm:</p> <p>I have contributed to experimentations and explanations in paper involving different methodologies like clustering, TFIDF, Word2Vec. I also reported the results involving the experimentations.</p> <p>After midterm:</p> <p>After the midterm, I continued my research on applying K-means to answer the research question. Additionally, I also added some data preprocessing steps and scraped additional data. I contributed to the writing of the paper by writing the relevant methodology section and writing K-means section in results and discussions.</p>	 Chirag Saxena
		 Amogh Anil Rao
		 Vaibhav Srivastava
		 Junwei Yu