Topic Modeling on Meditation/Mindfulness PubMed Publications

Liang-Yun Cheng, MSE
University of Pennsylvania, Philadelphia, PA

I. Problem Statement/Motivation

Meditation and mindfulness exercises are practices that works on the connection of the mind and body. There are a wide variety of methods with varied level of scientific support on their effectiveness. Generally, meditation is believed to help individuals bring calmness, manage stress from health complications, and enhancing overall wellbeing[1]. Specifically, research has shown mixed results on the benefits of these practices on a variety of health issues, such as stress, anxiety, depression, high blood pressure, pain, sleep quality, substance use disorder, post-traumatic stress disorder, cancer, eating disorder, and attention-deficit hyperactivity disorder[2].

In recent years, the growth in meditation app market suggests an increase in interest in adopting these techniques by the public. The estimated revenue of the mediation app market for 2023 is $4.4 billion, and it is estimated to reach $7.1 billion in 2028, which is growing at 9.8% annually[3]. Similarly, in the research community, there is also an increase in meditation and mindfulness studies. Up until November 2023, there are approximately 2,700 clinical and controlled studies, and 50% of these studies were conducted within the past 6 years. With these macro trends in mind, this study aims to use a natural language processing technique called Latent Dirichlet Allocation (LDA) to conduct topic modeling to understand sub-topic and population that have been investigated in association with meditation and mindfulness practice. Given the potential benefits of mindfulness practice, relative low cost, and mild intrusion on the physical body, it has a potential to be a practice to be supported by public health initiatives.

II. Solution

Latent Dirichlet Allocation (LDA) is a popular natural language processing technique that is used to discover hidden structure, or topics, among a set of documents, or corpus. This is an unsupervised machine learning technique, where no target label is associated with each document. Instead, the algorithm uses probability distributions learned from the data to identify latent groupings among the documents.

There are three main components of an LDA model, which are tokens, documents, and topics. The model makes two major assumptions: 1. documents are made up of topics, and 2. topics are made up of tokens (or words). The model computes the probability and outputs two matrices: probability of topics for each document, and probability of tokens in each topic. These results can then be used to extract the dominant topic for each document, and the top 10 words for each topic, which is then interpreted manually to extract semantic meanings.

The major benefit of LDA is that instead of manually inspecting a large corpus to identify common themes, we can set aside our presumptions and leverage the model to identify interesting hidden structures with a few lines of codes. This is a very useful technique to use to conduct exploratory text mining.

However, there are also major challenges and limitations in this model. First, the number of topics (n) is a hyperparameter. While there are metrics, such as perplexity score and coherence score, that can be used to identify the optimal number of topics, it remains a complex decision for the modeler to make depending on the application. For example, in this current use case, a higher n may not be an issue, since we are generally interested in identifying meaningful sub-groups within the space of meditation and mindfulness research. However, if the application is more targeted, we may want to limit the choice of k to ensure that actionable recommendation can be designed. Also, the coherence of words within each topic produced by the model may also influence a modeler's final choice of n. For example, while n = 10 may yield a better coherence score, while n = 12 may yield more relevant words within each topic despite having a worse score.

Another challenge is that it is expensive to evaluate the true performance of the model. Additional manual annotation is required to examine the model's ability to classify documents into the most appropriate topics. For this study, which is largely exploratory, such validation is not necessary.

III. Method

Dataset:

In this study, three components of research papers on mediation and mindfulness were analyzed: title, abstract, and author names. The data was extracted by the author of this paper using the following steps:

1. Identify relevant articles using PubMed[4] search:

   Search Query: *("meditat*"[Title/Abstract] OR "mindful*"[Title/Abstract]) AND ("clinical trial"[Publication Type] OR "controlled*"[Publication Type] OR "Observational Study"[Publication Type])*

   Additional Filter: Publication Date = 2023/11/30

2. Retrieve publication abstract using metapub:

   Using Python library metapub[5] (version 0.5.5) and PMID obtained in step 1, retrieve article *abstract* from the PubMed database.

Algorithm/Program:

The main Python packages used for this analysis include NLTK for text pre-processing and Gensim for LDA topic modeling.

1. Text Pre-Processing: To reduce noise, the following standard text pre-processing steps were applied to *title* and *abstract* to create the corpus.

   a) Tokenize sentences into individual words
   b) Reduce words into lowercase
   c) Remove stop words (commonly used words, such as 'the' or 'is', but does not add substantial meanings to the sentence)
   d) Concatenate hyphenated words by removing the hyphen (e.g.: self-compassion to selfcompassion)
   e) Remove punctuations and digits
   f) Extract tokens that are noun, verb, or adjective using Part-Of-Speech tags
   g) Lemmatize words into their base forms

   Once *title*, *abstract,* and *authors* fields were cleaned, these fields were concatenated to form the tokens of each document. To further reduce the noise in the corpus, frequent words (words that appear in more than 25% of the documents) and infrequent (words that appear in less than 5 documents) were removed.

2. LDA topic modeling: There are two popular Python libraries that are used to build LDA models, which are scikit-learn's *LatentDirichletAllocation* module and Gensim's *ldamodel.* Gensim's module was chosen to the two advantages:

   a) Hyperparameter: *alpha* and *eta* are two hyper-parameters in the LDA model. *Alpha* captures the a-priori belief of the document-topic distribution. *Eta*, or sometimes refers to as *beta*, captures the a-priori belief of the topic-word distribution. In the Gensim, both parameters could be set to 'auto' and the model would also learn the prior distribution from the corpus.

   b) Coherence model: Gensim provides coherence metric pipeline to better evaluate topic coherence among model output. For this study, UMass coherence score is used, which is calculated as follow:

$$C_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)},$$

   U-Mass coherence score essentially measures the probability of words co-occurring together within the corpus[6].

Workflow Diagram

Figure 2 shows the overview of the process taken to conduct topic modeling analysis. Note that this is an iterative process as shown with the dotted arrows. Interpreting the topics yield insights into potential data pre-processing issues.

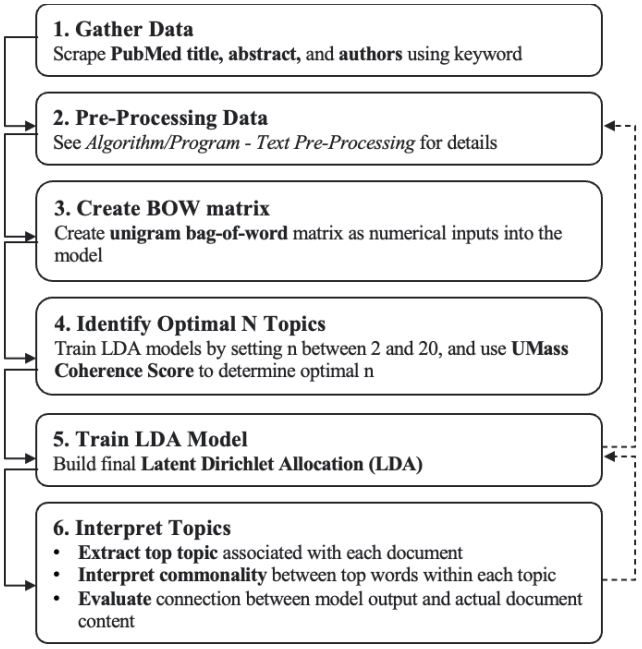

Figure 2. LDA analysis workflow diagram

IV: Results/Discussion:

Sample Size:

The search in PubMed yielded 3,085 articles. After removing articles that did not have abstract and were erroneously retrieved, such as articles with phrases "is mediated by" or "be mindful of", the final dataset contains 2,709 records with *PMID*, *Title*, *Authors*, and *abstract*. The result can be found in appendix: *CHENG_LIANGYUN_Results.xlsx*

Importance of Text Pre-Processing:

While most steps listed in the data pre-processing section were common. Figure 1 (a) and (b) highlights the importance of removing frequent words across documents. In this dataset, these words are often those that are associated with generic research terminology. Interestingly, some trends already emerge from the word cloud post removing frequent words. Some of the topics include pain, women, cancer, and mbsr (Mindfulness Based Stress Reduction (MBSR) therapy.

(a) (b)

Figure 1. Word distribution before (a) and after (b) removing frequent and infrequent words

n-gram Bag-Of-Words

One of the model design choices include whether to include just unigram or n-grams as well. Models were run with unigram and unigram + bigrams. The results for models with bigrams tend to include repetitive tokens. For example, ['cancer', 'breast', 'breast_cancer']. Therefore, the final model was built with unigrams to maximize the variations among the top 10 words within each topic.

Optimal N Topics

LDA models were tested with 2 to 20 topics; each model was initialized with alpha and eta to 'auto'. Each iteration's UMass coherence score was recorded and plotted. The lower the UMass coherence score, the more coherent the words are within each topic. Figure 3 shows the optimal number of topics within the tested range is 20. The processing time for this search is approximately 15 minutes. Please note that there is conflicting information on the internet in terms of whether high or low UMass score indicate better topic coherence. From this study, it seems that the lower the coherence score the better the n. However, more investigation on the metric is warranted for future work.

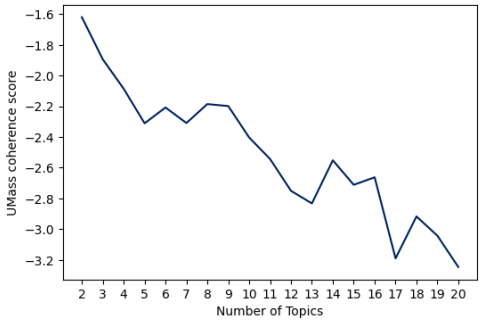

Figure 3. UMass Coherence Score by Number of Topics for LDA model

Final Results

Table 1 summarizes the result of the topics extracted from the corpus using an LDA model with n = 20. Each row is a topic with its top 10 most associated words, and the percentage of documents with that topic as its top topic.

ID	Semantic Topic (by author)	Top 10 Words (by model)	%
6	Breast Cancer	cancer, life, quality, breast, fatigue, survivor, psychological, woman, month, feasibility	11.3%
11	Student stress, Nurse resilience	student, wellbeing, mental, scale, perceive, score, resilience, university, nurse, medical	9.1%
2	* mixed	positive, affect, experience, negative, psychological, state, exercise, mood, condition, wellbeing	7.2%
9	Recurrent Depression	mbct*, depressive, tau, disorder, relapse, scale, score, usual, individual, primary	7.1%
4	Smoking	smoking, mt*, substance, relapse, disorder, alcohol, crave, prevention, cessation, individual	5.4%
7	Yoga, Breathing	yoga, subject, rate, cortisol, breathing, heart, session, decrease, exercise, physiological	5.3%
13	Chronic Pain	pain, chronic, intensity, exercise, opioid, relaxation, low, severity, day, rating	5.2%
19	Women Sexual Health	distress, cbt*, sexual, woman, function, session, emotional, people, feasibility, report	4.9%
3	* mixed	response, emotion, task, condition, emotional, negative, affect, reactivity, brief, mental	4.7%
8	Social Anxiety	mbsr*, psychological, acceptance, disorder, social, act, condition, worry, severity, great	4.7%
15	Weight Control, Diabetes	eat, weight, diabetes, behavior, food, mindful, loss, type, associate, body	4.7%
16	Parent-Child, Veterans	child, parent, adolescent, ptsd*, veteran, tm, disorder, posttraumatic, school, youth	4.4%
18	Caregiver	care, caregiver, randomise, primary, online, registration, cost, session, protocol, mental	4.2%
10	Blood Pressure	blood, pressure, disease, physical, risk, activity, lifestyle, cardiovascular, bp, month	4.1%
14	App, Pregnancy	woman, app, work, perceive, pregnancy, psychological, pregnant, distress, childbirth, employee	4.1%
12	Effect on Brain	brain, functional, mechanism, neural, associate, network, cortex, connectivity, activation, region	3.1%
5	* mixed	skill, emotion, selfcompassion, regulation, compassion, emotional, pilot, anger, session, month	2.8%
0	Elderly, Migraine, ADHD	adult, old, function, attention, performance, memory, headache, executive, task, migraine	2.7%
1	* mixed	ci, mean, score, care, month, v, usual, sd, age, β	2.5%
17	Sleep Disorder	sleep, quality, insomnia, disturbance, index, score, time, scale, psqi, adult	2.5%

Table 1. Topics Identified

Abbreviation:

- mbct: Mindfulness-based cognitive therapy (MBCT)
- mt: Mindfulness training (MT)
- cbt: Cognitive behavioral therapy (CBT)
- mbsr: Mindfulness Based Stress Reduction (MBSR) therapy
- ptsd: Post-traumatic stress disorder (PTSD)
- tm: transcendental meditation (TM)
- bp: blood pressure
- psqi: Pittsburgh Sleep Quality Index (PSQI)

Discussion

Overall, the LDA model is effective in extracting sub-topics that aligned with the application of meditation and mindfulness practice outline by the NIH[7]. Moreover, the model did successfully identify the following new sub-topics:

1) Resilience for healthcare worker and caregiver (topic 11, 18)
2) Stress management for students (topic 11)
3) Recurrent depression (topic 9)
4) Women sexual health issues (topic 19)
5) Social anxiety (topic 8)
6) Patient with diabetes (topic 15)
7) Children with behavioral issues, ADHD, and autism (topic 16)
8) Use of mobile app intervention (topic 14)
9) Women and pregnancy (topic 14)
10) The elderly and their executive functions (topic 0)
11) Migraine (topic 0)

It is worthwhile to notice that some topics contains more than 1 main semantic topics. On the other hand, there were some topics (topic 1, 2, 3, and 5) that had generic words with mixed types of articles. Nevertheless, LDA topic modeling technique successfully discovered applications of mindfulness practice that was not mentioned by NIH. Overall, it is worthwhile to notice that mindfulness practices are associated with helping individuals manage chronic health issues.

Future Work

The results of this study provide a basis in understanding the current research landscape in meditation and mindfulness practice. In the future, it would be interesting to compare topic modeling results produced by BERT models, which does not require extensive text cleaning.

One of the limitations of topic modeling technique is that it does not summarize whether results provide strong evidence in the effectiveness of these practices. With the growing capability of Large Language Model (LLM), it would be helpful to leverage the tool to summarize the impact of meditation and mindfulness practice on the new areas identified above to generate an even more comprehensive understanding of the current evidence for meditation and mindfulness practice.

Reference

[1] Meditation and Mindfulness: What You Need To Know [Internet]. National Center for Complementary and Integrative Health, U.S. Department of Health and Human Services. [cited 2023 Dec 19]. Available from: https://www.nccih.nih.gov/health/meditation-and-mindfulness-what-you-need-to-know

[2] Meditation and Mindfulness: What You Need To Know [Internet]. National Center for Complementary and Integrative Health, U.S. Department of Health and Human Services. [cited 2023 Dec 19]. Available from: https://www.nccih.nih.gov/health/meditation-and-mindfulness-what-you-need-to-know

[3] Meditation Apps – Worldwide [Internet]. statista. [cited 2023 Dec 19]. Available from: https://www.statista.com/outlook/hmo/digital-health/digital-fitness-well-being/health-wellness-coaching/meditation-apps/worldwide#key-players

[4] PubMed [Internet]. National Library of Medicine: National Center for Biotechnology Information. [cited 2023 Dec 19]. Available from: https://pubmed.ncbi.nlm.nih.gov/

[5] Naomi Most. metapub 0.5.5 [Internet]. PyPI: Python Package Index. [cited 2023 Dec 19]. Available from: https://pypi.org/project/metapub/

[6] Enes Zvornicanin. When Coherence Score Is Good or Bad in Topic Modeling? [Internet]. Baeldung. [cited 2023 Dec 19]. Available from: https://www.baeldung.com/cs/topic-modeling-coherence-score

[7] Meditation and Mindfulness: What You Need To Know [Internet]. National Center for Complementary and Integrative Health, U.S. Department of Health and Human Services. [cited 2023 Dec 19]. Available from: https://www.nccih.nih.gov/health/meditation-and-mindfulness-what-you-need-to-know