Unsupervised Topic Modeling of Reddit Posts for Mental Health Analytics

Aishwarya Sharma, Mansi Singh, Nitin Ramesh, Priya Theagarajan, Rukmini Sruthi Sundaraneedi
aishwarya.sharma@ufl.edu, mansi.singh@ufl.edu, nitinramesh@ufl.edu,
ptheagarajan@ufl.edu, r.sundaraneedi@ufl.edu

Abstract

Topic modeling is a widely used statistical modeling technique for uncovering latent topics in textual data. It is an unsupervised approach capable of identifying clusters and abstract topics in unstructured data. This paper presents a hybrid approach to analyzing mental health-related text data from Reddit using unsupervised topic modeling techniques. We utilized the Reddit Mental Health Dataset with over 100,000 posts and comments to explore the impact of lifestyle factors on mental health outcomes. We implemented three models - Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and a hybrid BERT (Bidirectional Encoder Representations from Transformers) and LSA model on the preprocessed data. We evaluated the model performances using coherence scores such as C_V and U_Mass which are widely used metrics for evaluating the effectiveness and performance of topic modeling techniques. This evaluation indicated the semantic coherence of the discovered topics and how BERT-LSA outperforms the two baseline models. Our research findings also aim to highlight the prevalence of mental health conditions among individuals with certain lifestyle choices and provide insights that could inform better prevention and intervention strategies for mental health.

1 Introduction

Mental health is a critical issue that affects millions of people worldwide. It has significant implications for both individuals and societies, making it a major public health concern. In recent years, social media platforms have provided a valuable source of data for understanding mental health-related language and experiences. With advances in natural language processing (NLP) techniques, researchers can now analyze large volumes of textual data from social media and gain insights into the prevalence, symptoms, and causes of mental health disorders. Our research aims to explore the potential of NLP techniques for analyzing mental health-related language on Reddit, a popular social media platform. Specifically, we will investigate the impact of lifestyle factors on mental health outcomes using unsupervised topic modeling techniques. By examining this relationship, we hope to contribute to the development of more effective mental health prevention and intervention strategies.

Several studies have investigated the potential of NLP techniques for analyzing mental health-related language on social media platforms such as Twitter, Facebook, and Reddit. For instance, (De Choudhury et al., 2016) conducted a study on depression-related language on Twitter and found that users who posted about depression tended to use more negative language and express less social engagement compared to those who did not post about depression. Similarly, (Park et al., 2021) used machine learning techniques to analyze suicide-related language on Reddit and identified certain linguistic features as predictive of suicidal ideation. Research on the impact of lifestyle factors on mental health outcomes has suggested that physical exercise, diet, and sleep can play a significant role. (Huang et al., 2018) found that regular physical exercise was associated with a reduced risk of depression, anxiety, and stress, while (Jacka et al., 2017) found that a healthy diet was associated with a reduced risk of depression. Despite these studies, there is a lack of research that has specifically investigated the relationship between lifestyle factors and mental health outcomes using NLP techniques applied to social media data. This research paper aims to address this gap by exploring the potential of NLP techniques for analyzing mental health-related language on Reddit and examining the association between lifestyle factors and mental health outcomes.

Our research aims to investigate the correlation between mental health concerns and lifestyle

choices using a preprocessed dataset from Reddit. To achieve our goal, we employed various NLP techniques such as Sentence-BERT embeddings, K-means clustering, and LSA dimensionality reduction to identify language patterns that correspond to different mental health concerns and lifestyles. We evaluated different topic modeling methods, including LDA, LSA, and BERT-LSA, using two coherence measures - C_V coherence and U_mass coherence. Our results showed that the BERT-LSA model outperformed the other two models, demonstrating that the most probable words in each topic generated were semantically similar and highly correlated with each other. Furthermore, we conducted a correlation analysis to find a correlation between mental health topics and lifestyle topics obtained by the BERT-LSA model based on the cosine similarity metric (Lahitani et al., 2016). Our findings revealed that several mental health issues had a strong correlation with specific lifestyle choices.

2 Related Work

The field of mental health analytics has seen increasing attention in recent years as mental health disorders continue to be a major global health concern. Natural Language Processing (NLP) techniques have emerged as a promising avenue for understanding and analyzing mental health-related language in social media. In this section, we discuss related work in unsupervised topic modeling of mental health-related text data on Reddit.

Numerous research works have utilized topic modeling methodologies to examine mental health-related textual data gathered from social media platforms. For example, (Pavlova and Berkers, 2020) conducted a study that focused on mapping mental health discourse on Twitter. Mental health discourse in social media contains conversations and idea exchanges about mental health issues, such as awareness, prevention, treatment, and support. The researchers explored the cultural power mechanisms that influence the prevalence of certain mental health topics over others. Their methodology included approaches such as topic modeling, sentiment analysis, and panel data regression analyses. To identify the most prevailing themes within mental health discourse, they utilized the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) method. LDA aims to discover underlying structures in the text by clustering words that co-occur more frequently than would be anticipated by random chance.

Additionally, (Paul and Dredze, 2014) demonstrated the feasibility and effectiveness of using topic modeling for analyzing health-related topics in social media data. The study aims to understand what health topics are commonly discussed on Twitter by analyzing self-reported health statuses across millions of users.A statistical topic model called the Ailment Topic Aspect Model (ATAM) was used to automatically infer health topics in Twitter messages. ATAM discovered clusters of Twitter messages that correlate with seasonal influenza, allergies, exercise, and obesity-related geographic survey data in the United States. Their results demonstrate that it is possible to automatically discover topics that have statistically significant correlations with ground truth data, without requiring historical data or extensive human supervision.

While previous studies have demonstrated the potential of unsupervised topic modeling techniques in mental health analytics, there aren't many studies that relate the impact of lifestyle factors on mental health outcomes and the complex nature of this interaction Our project aims to fill this knowledge gap by using NLP techniques, especially topic modeling, ultimately leading to better prevention and intervention strategies for mental health.

3 Proposed Approach

3.1 Datasets

The dataset selected for this research was required to meet certain criteria. It required unprocessed social media-based textual data which would allow us to extract details on lifestyle choices as well as infer mental health concerns. The "Reddit Mental Health Dataset" (Low et al., 2020) with over 100,000 posts and comments provided the right balance of both as well as the automatic subdivision of data via 18 subreddits. This segregation allowed verifying if the data in the subreddits contained enough mental health and lifestyle data. Thereby selecting the right data for our task of finding some correlation between the two topics. The data pre-processing conducted proved to be an important step in the experiment. Fine-tuning the data was carried out by techniques such as stop word removal, lemmatization, special character cleansing, filtering foreign language content, etc. This allowed us to greatly reduce the noise, improve the accuracy and capture the critical points in the data. This extracted data provided detailed scope for investigating the

prevalence of mental health conditions with certain lifestyle choices.

3.2 Methodology

The Reddit data collected needed to be pre-processed to reduce noise and remove irrelevant information which could affect the accuracy and quality of the model. This also helped make the model more efficient and faster. As part of our data pre-processing, we removed punctuation and converted our data-set to lowercase, after which we tokenized our data and removed stop words from it, after which we lemmatized it to group together inflected forms of a word.

We then implemented basic LSA and LDA (Stevens et al., 2012) models on the data, to observe the list of generated topics and found correlations between them. We thought that a more sophisticated model could be applied by taking advantage of BERT to further improve on our results, so we came up with a hybrid approach discussed in detail in the following sections. Next, we proceeded to compare our model performances using our evaluation metrics which we chose to be C_V (Confirmation Measure of Coherence) and U_Mass (Uniform Mass) coherence scores. Figure 1 gives a broad overview of our applied methodology. C_V coherence score is based on the concept of sliding windows, where it calculates the co-occurrence of word pairs within a fixed-size window in the corpus. It also incorporates a normalization factor to account for differences in word frequencies. Higher C_V coherence scores indicate more semantically coherent topics.U_Mass coherence is based on the idea of document co-occurrence, which measures the frequency of word pairs appearing together in the same document. The U_Mass coherence score is calculated using the logarithm of the conditional probability of two words occurring together, divided by the probability of the two words occurring independently. Higher U_Mass coherence scores also indicate more semantically coherent topics.

The reason we chose these scores to evaluate our models is that evaluating the quality of the discovered topics is essential to assess the performance of the topic modeling algorithms. Since there is no ground truth or labeled data in unsupervised learning, coherence measures like C_V and U_Mass are used to estimate the semantic coherence of the discovered topics. These coherence measures consider the co-occurrence of words within a topic and their distribution across documents in the corpus. Hence, these measures help to evaluate how well the topics are interpretable and meaningful, which is crucial for the usability of the topic modeling results.

3.2.1 Models

The models that we have used are LSA, LDA, and BERT-LSA for topic modeling. We have coupled LSA and LDA with .corr() function from the pandas library to find inter-topic correlation. For the BERT-LSA model, we have applied LSA on BERT to reduce the dimensionality of the sentence embedding and to generate topics. We then used k-means clustering on the generated topics to further cluster them. Our hybrid approach including BERT and LSA has been described in detail in section 3.2.2. The reasons for which we chose LSA and LDA for our use case are as follows

- Unsupervised nature: Since mental health discussions on Reddit can cover a wide range of topics and sentiments, it might be difficult to manually label the data or create a predefined set of categories. LSA and LDA are unsupervised techniques that can automatically identify latent topics in the data without the need for labeled training data.

- Scalability: Both LSA and LDA can handle large volumes of text data, making them suitable for analyzing the massive amounts of content generated on Reddit. Semantic understanding: LSA employs singular value decomposition (SVD) to reduce the dimensionality of the term-document matrix, capturing the semantic relationships between words and documents. LDA, on the other hand, uses a generative probabilistic model to represent the relationships between words, topics, and documents. Both methods can uncover hidden semantic structures in the data, which can be helpful for understanding the context and content of mental health discussions on Reddit.

- Interpretability: The topics generated by LSA and LDA can be interpreted as clusters of words that often appear together in the same context. By examining these topics, researchers and practitioners can gain insights into the most common themes, concerns, and sentiments expressed in mental health-related Reddit posts.

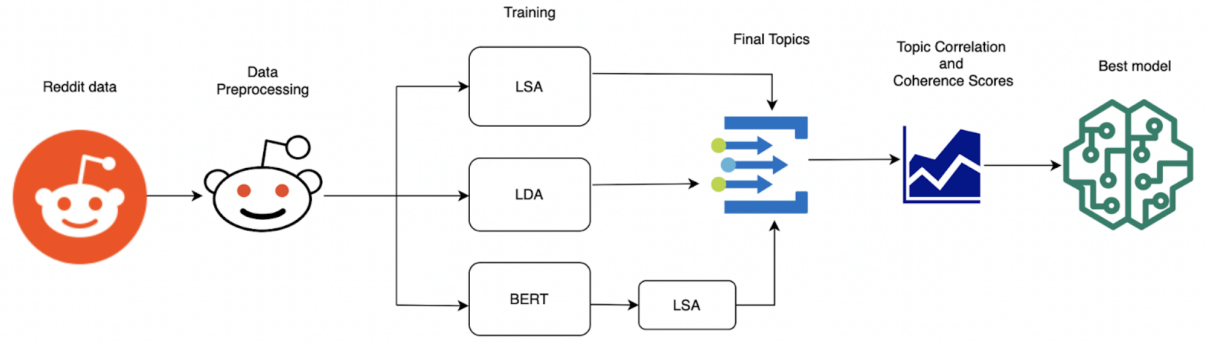

Figure 1: Applied methodology

- Customizability: Both LSA and LDA allow users to adjust the number of topics to be discovered, enabling a fine-grained or coarse-grained analysis of the data depending on the specific needs of the mental health analytics project.

All these models were then trained on subreddits relating to various mental illnesses and lifestyle choices in the attempt to find out which mental health issue is most likely correlated to which lifestyle choice using triangulation by gathering results from these three techniques.

3.2.2 Experiments

To investigate the correlation between mental health concerns and lifestyle choices using Reddit preprocessed dataset we used three approaches. Our goal was to identify the different patterns of language used over social media that correspond to different mental health concerns relating to different lifestyles. To achieve this we used a combination of natural language processing techniques including Sentence-BERT embeddings, K-means clustering, and LSA dimensionality reduction. The main idea behind using these techniques was to capture each post's semantic meanings, which proved useful in the clustering analysis steps. The Sentence-BERT model was used to embed each post into a high-dimensional vector space. This resulted in a vector representation of each post, which was then used as an input for the clustering and dimensionality reduction using K-means and LSA.

For our experiments, we used a pre-trained BERT-based Sentence transformer(Reimers and Gurevych, 2019), specifically, we used the "bert-base-nli-mean-tokens" model, which use siamese and triplet network (Reimers and Gurevych, 2019) structures to derive semantically meaningful sentence embeddings. For further clustering the related topics together we used a widely used unsupervised clustering algorithm, K-means clustering, which partitioned the data into K-clusters based on the similarity of the data points. We experimented with different values of K (i.e. from 2-11) to find the optimal number of clusters. Finally, we used LSA or Latent Semantic Analysis to extract latent or hidden topics from textual data. The technique includes dimensionality reduction using Singular Value Decomposition (SVD). For our study, we used the TruncateSVD from scikit-learn library to reduce the embeddings' dimensionality while preserving the semantic meaning of the textual data. The resulting matrix from TruncateSVD represents the documents in terms of the underlying topics. Finally, the top topics were extracted and the correlation matrix was generated between the LSA components using the cosine similarity metric. Cosine similarity is a measure of similarity between two vectors of an inner product space, it is used to measure the similarity between two topics or a document.

4 Results and Analysis

4.1 Evaluation of Topic Modeling Methods (Mansi)

We evaluated the three topic modeling methods: LDA, LSA, and BERT-LSA to determine the best approach for our task. The effectiveness of the methods was measured using two evaluation measures - C_V coherence and U_mass coherence. The C_V coherence score is used to measure the degree of semantic similarity between the top N words in a topic. The U_Mass coherence score captures the coherence based on the log-likelihood of the corpus, given the topics and their corresponding

words.

Models	C_V score	U_mass score
LDA	0.3364230493	-4.651980784
LSA	0.4438596489	-2.111011644
BERT-LSA	0.5744928124	-1.207974741

Table 1: Evaluation of topic models using C_V score (higher the better) and U_mass score(higher the better)

We implemented the LDA and LSA models using the genism (Řehůřek and Sojka, 2010)library. The models were trained on their default parameters on a sample dataset and calculated the C_V and U_mass coherence scores. The coherence scores were also calculated for our BERT-LSA (0.57) model and it was observed that the C_V coherence score for the model was higher than LSA (0.44) and LDA (0.33) indicating that the most probable words in each topic generated were semantically similar and are highly correlated with each other. In conjunction with C_V coherence score we also used U_mass coherence score for evaluation and BERT-LSA(-1.20) outperformed LDA(-4.65) and LSA(-2.11) giving a higher score than the two baseline models.

4.2 Correlation Analysis

In order to find a correlation between mental health topics and lifestyle topics obtained by the BERT-LSA model we generated a correlation matrix using the cosine similarity metric and generated heatmaps that depict the most and least correlated topics from the dataset. We generated 10 topics from the dataset and manually categorized them into Mental health topics and Lifestyle topics, Table 2 shows the respective keywords associated with each topic.

From our preliminary analysis, interesting facts were revealed about the nature of our dataset. After analyzing the related topics and clusters to identify the common themes we observed that a number of mental health issues had a strong correlation with certain lifestyle choices. For example, in the heatmap shown in Figure 2, it was observed that posts related to depression and negative emotions(Topics 3. 5. 9) were correlated with lifestyle choices(Topics 2, 6, 1) such as substance abuse and social relationships.

In Figure 3 we can identify the correlations between lifestyle choices(Topics 3, 2) such as work-life and sleep schedule, and how it correlates to some mental health conditions(Topics 9, 10) like depression, anxiety, and panic disorders. It was also observed that positive lifestyle choices(Topic 1) mentioning holidays and vacations had low cosine similarity scores with topics related to mental health issues(Topics 5,9,10)

5 Discussion

A comparative study and evaluation of the three models provided some information about their pros and cons. The LDA approach uses a probability distribution over words to generate topics. The LDA implementation was easy to implement and interpret, but it fell short of capturing semantic nuances and latent topics in complex datasets. LSA on the other hand uses a matrix factorization approach which overcomes the limitations of the LDA model. However, it does not work well with sparse texts and is unable to capture nuanced relationships between latent topics. Our evaluation found that BERT-LSA was able to overcome the limitations of the two baseline models. It has been shown to capture latent semantic relationships by generating sentence embeddings and syntactical relationships by the LSA matrix factorization method. BERT-LSA combines the power of BERT, a state-of-the-art language model, and the traditional topic modeling technique LSA.

The results of our unsupervised topic modeling analysis reveal that there are two most prevalent in our dataset: Mental health issues and Lifestyle choices. The mental health issues deal with topics related to depression, anxiety, and eating disorders as well as discussions about seeking help and engaging with therapists and medication. The lifestyle choices categories include topics related to hobbies, exercise, work life, and school life. It was observed that Topics such as Substance abuse and addiction show a high correlation with Topics mentioning Depression, Medication dependency, and seeking help. This could be interpreted as people dealing with negative emotions and depressive thoughts being more likely to engage in substance abuse and unhealthy lifestyle choices like the consumption of alcohol and other kinds of drugs. Similarly, topics mentioning anxiety and panic disorders have a higher correlation with topics having high mentions of family problems, stressful work-life balance, and unstable sleep schedules. Overall, our findings suggest that discussions related to mental health and lifestyle on social media are

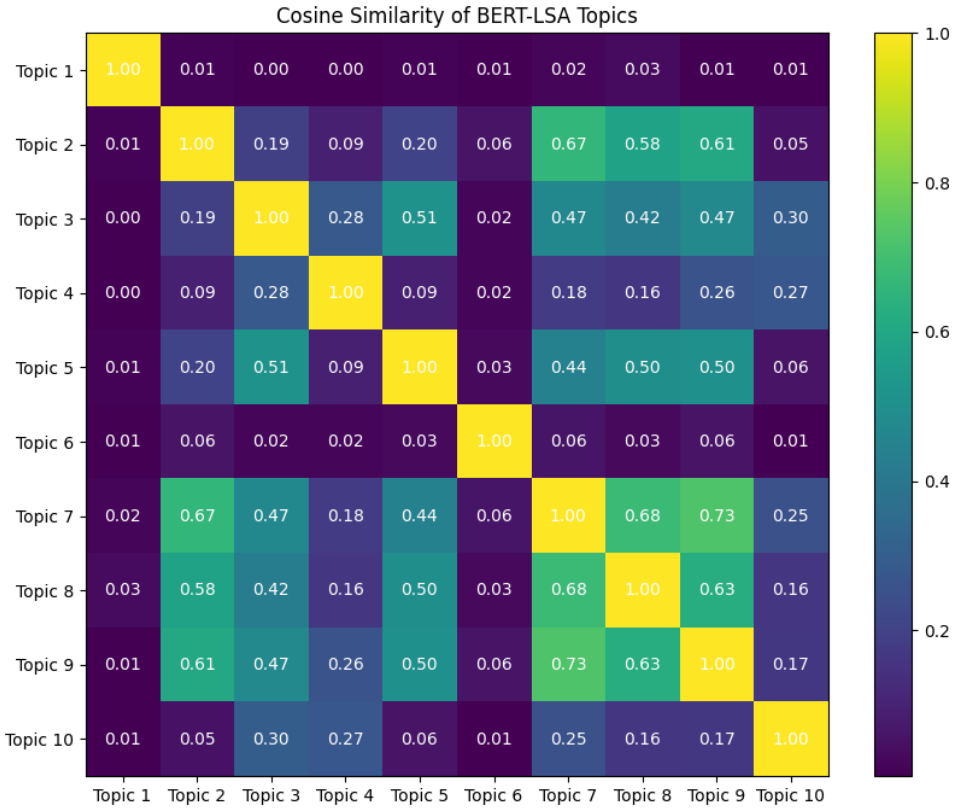

Figure 2: Heatmap showing the correlation between the topics shown in Table 2

Mental Health topics	**Lifestyle topics**
Topic 9: help, doctor, medication, depression, psychiatrist	Topic 2: weed, feel, people, know, crisis, smoking, alcohol, addiction, make
Topic 3: anxiety, medication, take, med, taking, doctor, antidepressant, year, effect, week	Topic 4: pain, cancer, doctor, symptom, blood, health, test, week, hospital, back
Topic 5: anxiety, like, feel, get, heart, feeling, day, attack, time, sweat	Topic 6: car, driving, drive, train, get, accident, noise, bus, road, driver
	Topic 7: like, get, feel, anxiety, time, know, work, job, really, want
	Topic 1: girl, relationship, date, toxic, dating, ex, guy, girlfriend

Table 2: Example of the basic topics generated by the BERT-LSA model from a dataset containing posts and comments from sub-reddits such as /mentalhealthadvice, /fitness, /schoollife

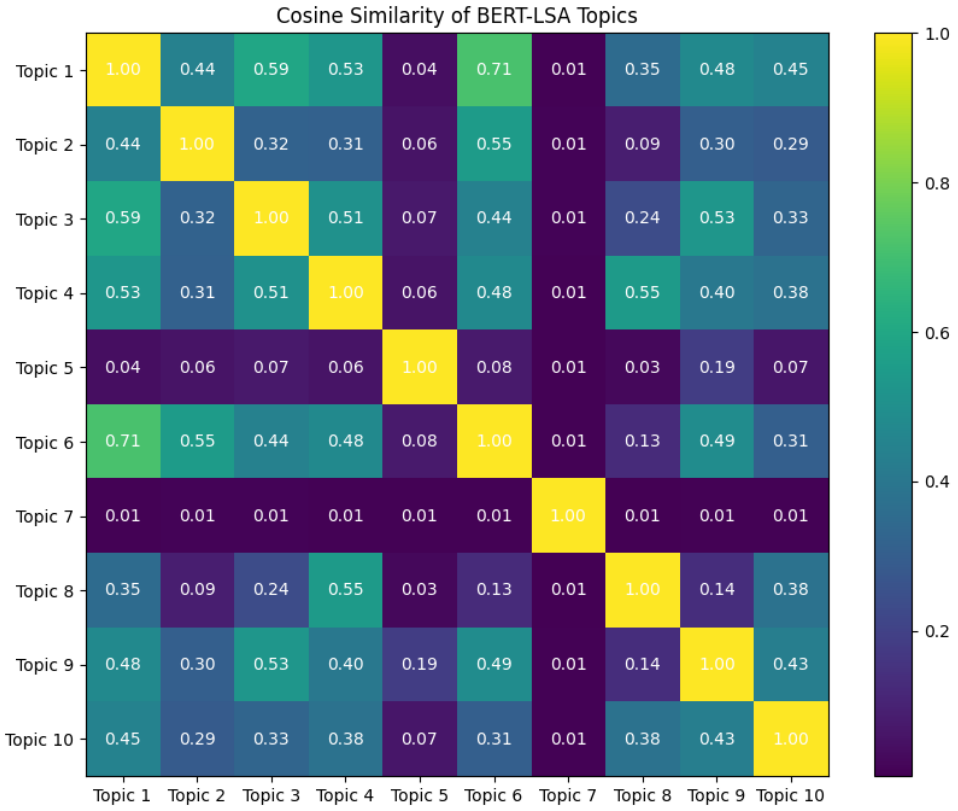

Figure 3: Heatmap showing the correlation between the topics shown in Table 3

Mental Health topics	**Lifestyle topics**
Topic 5: mental, health, people, experience, illness, question	Topic 1: feel, day, sun, vacation, know, time, off, holidays
Topic 9: help, mental, depression, anxiety, therapy, depression, need, get	Topic 2: something, game, sleep, time, headache, night
Topic 10: anxiety, panic, disorder, diagnosed, attack, doctor, medication	Topic 3: work, get, time, want, job, family, home, stress, sleep
	Topic 6: feel, like, people, know, really, even, life, deep, thought
	Topic 8: mom, dad, family, fights, went, started, divorce

Table 3: Example of the basic topics generated by the BERT-LSA model from a dataset containing posts and comments from subreddits such as /mentalhealth, /divorce, /relationships

closely intertwined and have a complex relationship. However, it is important to note further research is needed to fully understand the relationship between mental illnesses and lifestyle.

6 Limitations and Future Work

While our research project identifies the correlation between mental illness and certain lifestyles, it is important to recognize that correlation does not imply causation. The presence of a connection does not necessarily indicate that only lifestyle factors are the cause of mental health issues, as there may be various other factors that could contribute to it. Investigating these causal relationships falls beyond the scope of our project. Since topic modeling is based on the frequency of words within a given document, it may not fully capture the contextual nuances that affect the meaning of the text.

In our research project, we identify the correlation between mental health concerns and lifestyle choices using textual data derived from social media platforms, such as Reddit. Given the global usage of social media platforms, our research can be expanded to include foreign languages, enabling the analysis of geographical location and cultural sentiments' impact on the relationship between mental health concerns and lifestyle choices. Additionally, the effectiveness of mental health and lifestyle interventions can also be assessed by tracking changes in social media conversations over time. To further improve the results, we can incorporate data from additional sources, such as health records and wearable devices. Wearable devices can provide a plethora of data regarding health and lifestyle and allow for more granular correlation.

Our work also holds great potential for extension into various domains beyond mental health. By generalizing the project, valuable insights and understanding can be gained in multiple areas. The hybrid approach has the potential for being generalized for test cases other than mental health. Although, the choice of topic modeling techniques and BERT models would largely depend on the nature of the data and the research question. Topic modeling can be applied to Reddit posts and comments in political subreddits to analyze public opinion, identify trending political issues, and study the evolving political landscape. By studying the discussions in educational subreddits, topic modeling can reveal trends in educational resources, teaching methodologies, and student experiences. This could be valuable for educators and policymakers to make informed decisions.

7 Conclusion

In conclusion, our analysis successfully shows that all three approaches were beneficial in extracting topics corresponding to mental health issues and lifestyle choices. However, the choice of model will depend on the nature of the dataset, in our case BERT-LSA performed significantly better than LDA, and LSA. The hybrid approach was better at capturing the semantic and contextual nuances in the large and complex datasets, which the other two approaches fell short of. The BERT-LSA model was effective in capturing the correlation between the two categories. By applying unsupervised topic modeling to Reddit posts for mental health analytics, researchers, clinicians, and other stakeholders can identify prevalent themes, detect emerging trends, and monitor the effectiveness of interventions or public health campaigns. This information can be invaluable for designing targeted mental health interventions, improving existing resources, and raising public awareness about mental health issues.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Ting Huang, Lu Wang, Xin Yu, Xiaobo Qu, and Jianrong Liu. 2018. Exercise intervention and sleep quality in depressed middle-aged and older adults: A randomized controlled trial. *Behavioral Sleep Medicine*, 16(4):394–405.

Felice N Jacka, Adrienne O'Neil, Rachelle Opie, Catherine Itsiopoulos, Sue Cotton, Mohammedreza Mohebbi, David Castle, Sarah Dash, Cathrine Mihalopoulos, Mary Lou Chatterton, et al. 2017. A randomised controlled trial of dietary improvement for adults with major depression (the 'smiles'trial). *BMC medicine*, 15(1):1–13.

Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online

essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6.

Daniel M. Low, Laurie Rumker, Tanya Talker, John Torous, Guillermo Cecchi, and Satrajit S. Ghosh. 2020. Reddit mental health dataset.

Sohyeon Park, Seoyoung Yoon, Jiyoung Ryu, and Jungyun Kim. 2021. Predicting suicidal ideation on reddit: A machine learning approach. *Journal of medical Internet research*, 23(3):e23061.

Michael J. Paul and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PLoS ONE*, 9(8):e103408.

Alina Pavlova and Pauwke Berkers. 2020. Mental health discourse and social media: Which mechanisms of cultural power drive discourse on twitter. *Social Science Medicine*, 263:113250.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 952–961.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. pages 45–50.