Utilizing Machine Learning to Analyze Depressive Language in South Asian College Students

Jaelon Braxton, Jhanvi Kannan, Ali Shah

EPPS 6323.001 Knowledge Mining

Dr. Karl Ho

May 3, 2024

Abstract

This research project aims to develop a machine learning model to analyze depressive language among South Asian college students based on their Reddit posts. Inspired by the need to understand the linguistic patterns and discourse surrounding depression within South Asian college communities, the project initially focused on analyzing keywords extracted from Reddit posts related to depression. However, feedback from our peers and professor led to a refinement of the scope, narrowing the focus to symptoms of Depressive Disorder faced specifically by South Asian college students stemming from 60 Reddit posts. By leveraging R packages such as Tidyverse, tidytext, dplyr, stringr readr, and ggplot2, along with advanced text processing techniques including stemming, stopword removal, and TF-IDF transformation, the project aims to train an unsupervised logistic learning model to evaluate posts and identify depressive language. The research intends to contribute to the development of tools for analyzing mental health on social media, particularly in environments where individuals may express private matters more openly. This paper outlines the methodology, challenges, and expected outcomes of the project, highlighting the significance of understanding and addressing mental health issues among South Asian college students.

Race Related Stress Studies

Each individual possesses a unique method of evolution and self-expression, which is reciprocally met with diverse interpretations. Psychiatric disorders, also termed as mental illnesses, manifest in a multitude of expressions, influenced by individuals' distinct traits, and notably, these disorders may manifest differently among men and women. This inherent variability complicates diagnosis and underscores the formidable challenge of raising awareness about psychiatric disorders. Among these disorders, Depressive Disorder, commonly known as

depression, stands out as a significant focus of our research. Characterized by mood alterations, depression impedes individuals' ability to perform daily tasks, exacerbating their sense of helplessness and fueling a vicious cycle of symptoms(National Institute of Mental Health, 2024). The escalating nature of this disorder underscores its pivotal role in our research endeavor. Depression, often intertwined with stress, is associated with various factors, spanning from situational, health, and environmental contexts. Situations such as the loss of a loved one, academic pressures like finals, interpersonal relationships, financial hardships, and excessive demands can all precipitate stress. Unfortunately, certain environments serve as breeding grounds for stress and emotional despondency, perpetuating a cycle of strain and diminishing resilience.

School serves as a multistructural environment where individuals not only undergo academic tests but also grapple with mental challenges. From a young age, children are thrust into educational systems with the expectation of becoming successful adults equipped with essential knowledge. However, among these learning environments, universities stand out as significant hubs where a considerable number of students experience depressive symptoms. According to the American College Health Association (2020), university students are particularly susceptible to feelings of persistent sadness, loss of interest in activities, disruptions in sleep patterns, and difficulties in concentration. This vulnerability is brought forth by the culmination of academic pressures and newfound independence that characterize the college experience. The transition to college marks a critical period marked by newfound freedom, academic demands, and social adjustments, all of which contribute to heightened stress levels and emotional fragility (Hunt & Eisenberg, 2010). This confluence of factors creates a challenging scenario for students, with tuition costs adding an additional layer of pressure.

Consequently, it is not uncommon for campuses to report high rates of depressive symptoms among their student body.

Moreover, the college experience is characterized by an exploration of identity, which can be both enriching and stressful. Students grapple with defining themselves amidst societal expectations and cultural influences, further creating feelings of stress and uncertainty. Research has indicated that minority communities, including those based on gender, sexual orientation, and ethnicity, are particularly vulnerable to stress-related symptoms (Auerbach et al., 2018). The academic landscape further magnifies these challenges, with diverse communities often facing assumptions and expectations that may not align with their experiences. Thus, the college environment becomes a battleground where different communities navigate issues of identity, representation, and belonging amidst academic pressures and societal norms.

South Asian American college students navigate a distinct interaction with college life, characterized by cultural nuances and clashes with Western customs. The South Asian student community comprises learners from diverse backgrounds spanning India, Pakistan, Nepal, Sri Lanka, Bangladesh, Bhutan, and other countries, resulting in a rich tapestry of cultural diversity. However, this diversity often translates into cultural clashes, particularly regarding concepts of independence and interdependence. While the United States prides itself on individual freedom, Asian cultures tend to prioritize collectivism, with an emphasis on community contributions as seen from tradition stemming from filial piety(Kawamura, K.Y.). This difference in cultural orientation can lead to tensions within South Asian households, where actions perceived as self-centered may be misconstrued as selfishness, especially in contexts where financial decisions are shared among family members. Moreover, the pressure to pursue lucrative careers, such as medicine or law, often stems from familial expectations and societal stereotypes of Asian

Americans as the 'Model Minority'. This expectation of overachievement creates additional stressors for South Asian college students, who may feel compelled to excel academically and participate in extracurricular activities to maintain this perceived standard of success. However, this relentless pursuit of excellence may come at the expense of individual well-being, contributing to symptoms of depressive disorder among South Asian college students.

Introduction

While university life presents numerous challenges for students, campuses often acknowledge the difficulties inherent in this stage of life. Despite the availability of mental health services on college campuses, many students continue to grapple with untreated or unrecognized depressive symptoms, highlighting the pressing need for innovative approaches to mental health assessment and intervention (Mata et al., 2015). Our project seeks to address this need by developing new tools for detecting symptoms of depression within the college community. By harnessing machine learning techniques to analyze text data from Reddit posts, researchers can uncover valuable insights into the prevalence, severity, and nuances of depressive symptoms among college students. Through this research endeavor, we aim to contribute to the growing body of literature on mental health among college-age individuals. By leveraging advanced text processing techniques and unsupervised learning algorithms, we aim to identify patterns and associations within the language used by college students when discussing their mental health experiences. Ultimately, our goal is to shed light on the challenges faced by college students in managing their emotional well-being and to inform the development of targeted interventions and support services tailored to their needs.

Research Question

Is there a significant association between the topics identified through topic modeling of Reddit posts and the prevalence of depressive symptoms among college students in South Asian communities?

Hypotheses

Ho: There is no significant association between the topics identified through topic modeling of Reddit posts and the prevalence of depressive symptoms among college students in South Asian communities.

Ha: There is a significant association between the topics identified through topic modeling of Reddit posts and the prevalence of depressive symptoms among college students in South Asian communities.

Summary of Findings

In the course of our research, we have uncovered several intriguing findings that shed light on the intricate relationship between language, cultural dynamics, and mental health within South Asian college communities. Firstly, our analysis of the top 20 frequently used words revealed a significant emphasis on terms like "college," "parents," and "school," reflecting the central role of educational institutions and familial relationships in discussions about mental health. Notably, the prominence of "mom" as opposed to "father/dad" underscores the maternal influence and caregiving role within South Asian households, highlighting the complexities of familial dynamics in shaping emotional well-being. Moreover, our application of a sophisticated LDA model allowed us to discern two distinct thematic clusters, each offering nuanced insights into the intersectionality of education, emotions, family, and time constraints. The first cluster revolved around themes of education and emotional well-being, while the second cluster centered on familial dynamics and time management. These findings highlight the multifaceted

challenges faced by South Asian college students, emphasizing the importance of understanding cultural nuances in addressing mental health concerns within this demographic. Overall, our research provides a comprehensive understanding of the complex interplay between linguistic patterns and mental health experiences, paving the way for targeted interventions and support services tailored to the needs of South Asian college communities.

Reddit Data Collection and Preprocessing

We chose Reddit as the primary data source for our research project due to its unique characteristics that make it an invaluable platform for studying mental health, particularly within college communities. One key advantage of Reddit is its vast repository of user-generated content, which provides researchers with access to firsthand accounts and personal narratives related to mental health issues. Within subreddits dedicated to topics such as depression, users openly share their experiences, struggles, and coping mechanisms, offering valuable insights into the lived experiences of individuals grappling with mental health challenges, including college students. Moreover, the anonymity and candidness afforded by Reddit encourage users to share their experiences openly, creating a supportive environment where individuals feel comfortable discussing sensitive topics such as depression without fear of judgment or stigma.

Additionally, the candid nature of Reddit discussions allows for a level of authenticity that may be lacking in other social media platforms. Users often share raw and unfiltered accounts of their experiences, providing researchers with a rich and diverse dataset to analyze. This authenticity is particularly important when studying mental health, as it allows researchers to gain a genuine understanding of the challenges and complexities faced by individuals struggling with depression. Overall, Reddit's unique combination of user-generated content, anonymity, and candidness makes it an ideal platform for studying mental health issues,

providing researchers with unparalleled access to the lived experiences and narratives of individuals within college communities.

To gather data for our research project, we utilized an extensive Reddit search strategy, targeting various subreddits related to mental health within South Asian communities. Specifically, we looked into the official subreddits for India, Pakistan, Nepal, Bangladesh, Sri Lanka, and ABC Desis, each known for fostering discussions on diverse topics, including depression. Within these subreddits, we conducted searches using keywords like “college”, “depression”, “burnout”, and “dropout” to identify relevant posts reflecting the experiences and sentiments of individuals grappling with mental health challenges in the context of higher education. Through this systematic approach, we curated a dataset comprising a total of 60 posts, encompassing discussions, personal anecdotes, and reflections pertaining to the intersection of depression and college life within South Asian communities.

Upon identifying the relevant posts, we proceeded to collect the text data by manually copying and pasting the content of each post into an Excel spreadsheet. This process enabled us to compile a comprehensive dataset containing the textual content of the selected Reddit posts. Subsequently, we converted the Excel spreadsheet into a comma-separated values (CSV) file format to facilitate further analysis and processing.

The collected data underwent preliminary preprocessing steps to ensure its suitability for analysis. This included removing any extraneous formatting, such as hyperlinks, formatting tags, or metadata, that may have been present in the original Reddit posts. Filler words were omitted, and contractions were split back up so that they would no longer be shortened for the purpose of text standardization. Additionally, we reviewed the content of each post to ensure consistency and coherence in the text data.

Following data collection and preprocessing, the CSV file containing the Reddit post data was ready for analysis. This dataset served as the foundation for our subsequent text mining and machine learning endeavors aimed at identifying and analyzing depressive symptoms among college students. The availability of this dataset allowed us to leverage advanced text processing techniques and machine learning algorithms to uncover insights into the linguistic patterns and discourse surrounding depression in the context of college life.

Topic Modeling

Drawing inspiration from the methodologies outlined in "Text Mining with R: A Tidy Approach" by Julia Silge and David Robinson, we conducted our analysis by incorporating topic modeling techniques. Topic modeling, a cornerstone of text mining, offers a systematic approach to uncovering latent themes and patterns within textual data. As highlighted in Silge and Robinson's work, topic modeling algorithms, such as Latent Dirichlet Allocation (LDA), hold immense potential for extracting meaningful insights from unstructured text.

In the context of our research project, we apply LDA or similar algorithms to our preprocessed dataset of Reddit posts. Building upon the foundational preprocessing steps described in our methodology, including data collection, text standardization, and stopwords removal, we leverage topic modeling to delve deeper into the nuanced discourse surrounding depressive symptoms among college students.

Silge and Robinson's comprehensive exploration of topic modeling in R underscores the utility of this technique in uncovering hidden structures within textual data. By treating each document as a mixture of topics and each topic as a distribution over words, LDA enables us to identify coherent themes and topics that exist in the Reddit discussions on depression in college settings. Moreover, Silge and Robinson emphasize the interpretability and exploratory nature of

topic modeling, wherein researchers can iteratively refine the model parameters to yield meaningful topics that resonate with the underlying narrative of the data. By adjusting the number of topics, tuning hyperparameters, and assessing model convergence, we strive to uncover the most relevant themes and topics that encapsulate the diverse experiences and perspectives shared by South Asian college students on Reddit.

Furthermore, Silge and Robinson's emphasis on the integration of topic modeling with other text mining techniques, such as sentiment analysis and word embeddings, inspires us to adopt a holistic approach to analyzing depressive language in Reddit posts. By combining topic modeling with sentiment analysis, for instance, we can find not only the prevalent themes but also the emotional notes associated with each topic, providing a richer understanding of the nuanced expressions of depression within the South Asian college community.

Data Preprocessing and Visualization (Code)

Following the data collection process from Reddit, we proceeded to organize the gathered text into a CSV (Comma-Separated Values) file for further analysis. In R, our first step involved importing the CSV file into the program using the read_csv function, as demonstrated below:

```
df <- read_csv("Path_to_CSV")
```

This command facilitated the loading of the CSV file named "data.csv.csv" from the specified directory into the R environment, enabling us to commence our analytical endeavors. Subsequently, our procedural step involved tokenizing the collected data, thereby facilitating the computer's comprehension of the textual information contained within the string variable. Tokenization serves as a pivotal preprocessing technique wherein the raw text is segmented into

individual tokens or units, such as words or phrases, enabling the computer to analyze and interpret the underlying linguistic content effectively. This preparatory step lays the groundwork for subsequent text analysis tasks, empowering the computational framework to extract meaningful insights and patterns from the data.

```
data_tokens <- df %>%
  unnest_tokens(word, Story)
```

Following tokenization, our next step involved leveraging the stopwords function available within the Tidytext package. This function serves the purpose of eliminating non-essential words, commonly referred to as stopwords, from the textual data. Stopwords encompass ubiquitous terms such as "I," "me," "you," and "do," which recurrently appear across documents but contribute minimal substantive information to our analytical objectives. By systematically filtering out these extraneous words, we enhance the relevance and interpretability of the text data, enabling more meaningful analyses and insights to be gleaned from the dataset.

```
# Load stopwords from tidytext
        data('stop_words')
# Removing stopwords
        data_tokens <- data_tokens %>%
                anti_join(stop_words)
```

With the data now meticulously cleaned and preprocessed, we have transitioned to the analytical phase. Our initial step in this endeavor involved identifying the top 20 most frequent keywords within the dataset. By conducting this analysis, we aimed to elucidate the prevailing themes and topics that permeate the corpus, thereby laying the groundwork for more in-depth explorations and insights into the underlying patterns and trends.

```
word_freq <- data_tokens %>%
  count(word, sort = TRUE)
head(word_freq)
```

Following the extraction of the top 20 frequent keywords, we proceeded to visualize this data using the ggplot package. Visualization serves as a powerful tool for synthesizing complex information into digestible insights, enabling stakeholders to gain a comprehensive understanding of the dataset at a glance. Leveraging the capabilities of ggplot, we crafted visually compelling representations that effectively communicate the distribution and significance of these keywords, thereby facilitating informed decision-making and further analysis.

```
ggplot(top_words, aes(x = reorder(word, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(x = "Word", y = "Frequency") +
  coord_flip() +  # Rotate labels for better readability
  theme_minimal()  # Apply a minimal theme
```

In the process of creating our LDA (Latent Dirichlet Allocation) model, we turned to the topicmodels package, leveraging its LDA function to generate clusters from our data. However, prior to modeling, additional data cleaning was imperative. Initially, we transformed our textual data into a format suitable for the model's operation. This involved converting the string data into a document-term matrix, a pivotal step in the LDA modeling pipeline. By structuring the data in this manner, we established a foundational representation that facilitates the identification of latent topics within the corpus, paving the way for subsequent analysis and interpretation.

```
dtm <- tokenized %>%
  count(document_id = row_number(), word) %>%
  cast_dtm(document_id, word, n)
```

After this, we converted our document-term matrix to a matrix that will be used in the LDA model as seen below.

```
# Convert the document-term matrix to a matrix
dtm_matrix <- as.matrix(dtm)
```

After preparing our data into a document-term matrix, we ran our LDA model using the LDA function from the topicmodels package. This step aimed to reveal latent topics within the dataset by estimating topic distributions across documents and word distributions within topics..

```
# Perform LDA
```

```
word_lda <- LDA(dtm_matrix, k = 2, control = list(seed = 1234))
# View the summary of the LDA model
word_lda
word_topics <- tidy(word_lda, matrix='beta')
word_topics
```

And visualized the results.

```
word_top_terms <- word_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

word_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```

Analysis of Findings

In the course of our research, we have uncovered several intriguing findings. Firstly, our analysis revealed the top 20 frequently used words, which are as follows: College, feel, life, time,

parents, school, depressed, people, depression, family, mental, friends, day, health, lot, mom, started, home, classes, job.

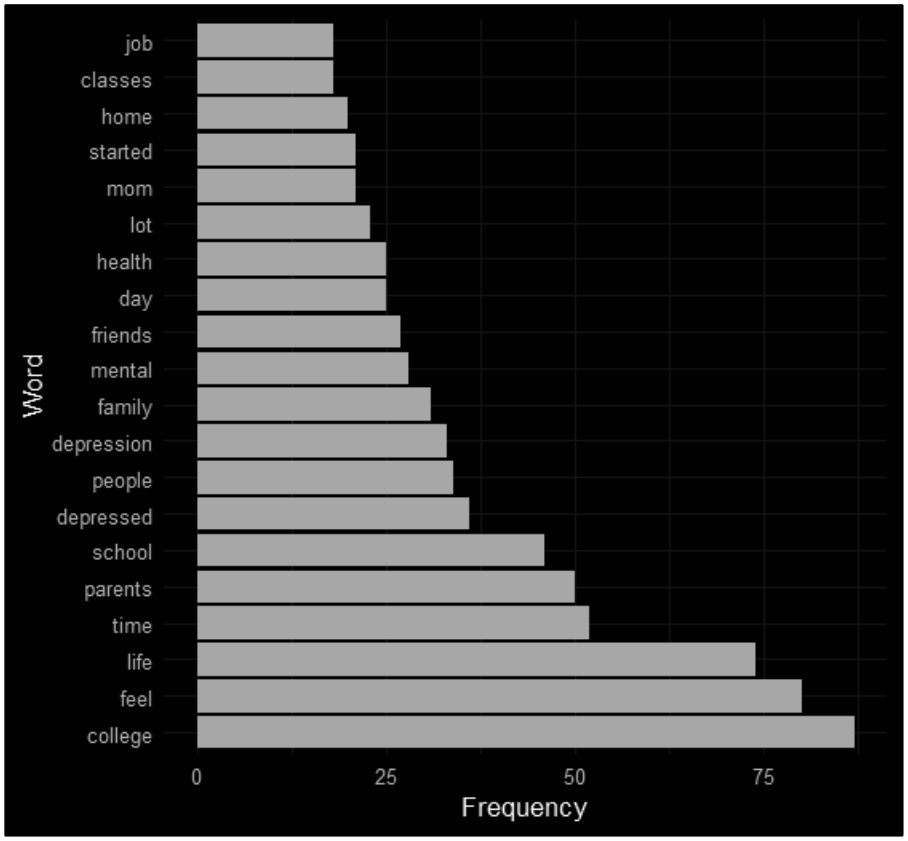

Figure 1

Notably, "college" emerged as the most frequently used keyword. This aligns with the focus of our study, which centers on college-age individuals discussing depression. It's widely acknowledged that the college years can present significant challenges as young adults navigate newfound independence. Another noteworthy keyword is "parents." This finding is particularly relevant, as many South Asians grapple with a cultural disparity in their relationships with their parents. With a significant number of South Asians immigrating to the United States in recent years, the current college generation represents the first fully Americanized cohort. This cultural shift often results in conflicts stemming from differing expectations and values between parents and their children, particularly regarding family dynamics and lifestyle choices.

Similarly, the prevalence of the keyword "school" underscores the ongoing educational context of our study cohort. However, one unexpected observation was the prominence of the keyword "Mom" without a corresponding mention of "father." We postulate that this discrepancy may be attributed to the traditional role of mothers as primary caregivers in many South Asian households. Such roles often lead to heightened tensions and conflicts between mothers and their children, potentially overshadowing paternal influences in discussions of familial dynamics and mental health.

Overall, these findings shed light on the multifaceted challenges faced by college-age individuals, particularly within South Asian communities, and underscore the importance of understanding cultural nuances in addressing mental health concerns within these populations. In addition to our initial findings, we employed a sophisticated LDA (Latent Dirichlet Allocation) model subsequent to data cleansing to glean further insights. This analytical approach enabled us to uncover nuanced patterns within the dataset. After thorough evaluation, we determined that employing a two-cluster solution provided the most meaningful delineation of topics.

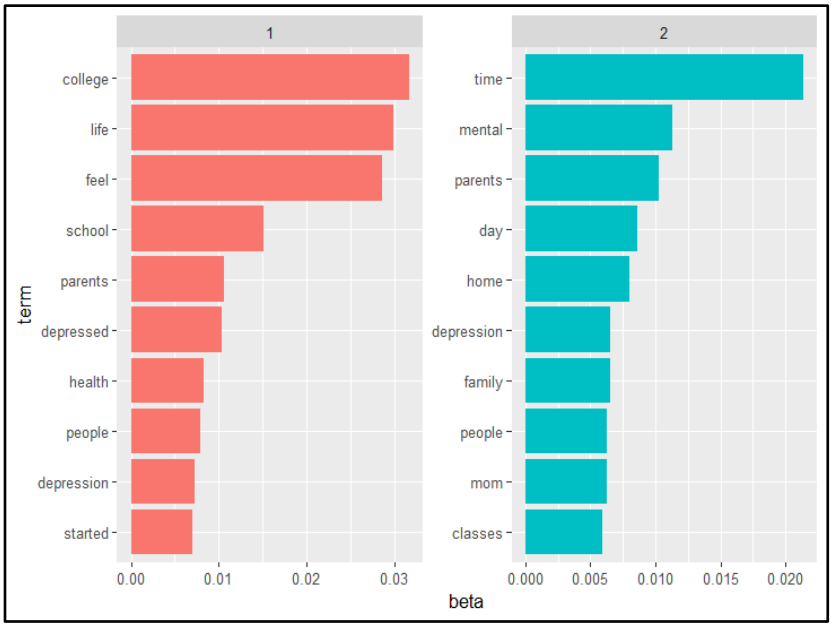

Figure 2

By diving into the intricacies of the data, we observed distinct thematic clusters emerging, each offering valuable insights into the narratives surrounding mental health within the college-aged South Asian demographic. This comprehensive analysis allowed us to discern underlying patterns and thematic connections that might otherwise have remained obscured. Through this iterative process, we identified key themes and topics that resonate deeply within the context of our study. By selecting a two-cluster solution, we aimed to strike a balance between granularity and interpretability, ensuring that the resulting clusters encapsulate the most pertinent aspects of the data while remaining actionable for further investigation.

In essence, the application of the LDA model served as a powerful tool in uncovering latent structures within the dataset, shedding light on the multifaceted nature of the discourse surrounding mental health among South Asian college students. This methodological rigor not only enhances the robustness of our analysis but also lays the groundwork for informed interventions and targeted strategies aimed at addressing the unique challenges faced by this demographic subgroup.

Ultimately, our analysis revealed two distinct clusters, each encapsulating meaningful themes reflective of the complex interplay between mental health and various aspects of the college experience within the South Asian demographic.

The first cluster prominently featured themes centered around "education" and "emotions and health." This thematic convergence was informed by the prevalence of keywords such as "college" and "school," indicative of discussions pertaining to the educational journey of college students. Additionally, words such as "life," "feel," "depressed," and "health" underscored the emotional and psychological aspects of the discourse. This alignment with our data underscores

the prevalence of depression among college-age individuals, manifesting in discussions about academic institutions and personal well-being.

Conversely, the second cluster exhibited themes revolving around "family" and "time," as evidenced by keywords such as "parents," "day," "home," "mom," and "family." This thematic orientation resonates with the intricate familial dynamics prevalent in South Asian communities, wherein familial relationships and obligations often intersect with the time constraints inherent to college life. The relentless pace of college existence, coupled with the pressures of familial expectations, contributes to heightened stress levels among students, precipitating depressive symptoms.

In essence, these clusters offer valuable insights into the multifaceted nature of the college experience for South Asian students, shedding light on the intersectionality of education, emotional well-being, familial dynamics, and time constraints. By discerning these underlying themes, we gain a deeper understanding of the challenges faced by this demographic subgroup, thereby paving the way for targeted interventions and support mechanisms aimed at mitigating the impact of depression and fostering holistic student well-being.

Limitations

Although our model and research methodology was accurate, it was not perfect. Expanding the dataset by increasing the number of posts has the potential to yield more specific and granular clusters through the LDA model. By incorporating a larger volume of text data, the model can capture a wider array of linguistic patterns and thematic variations, thereby enhancing the richness and specificity of the resulting clusters.

However, it's essential to recognize the inherent limitations of the LDA model, particularly in its ability to comprehend the contextual nuances of certain words. While LDA is

proficient at identifying co-occurring terms and uncovering latent topics within a corpus, it may not always grasp the intricate semantic relationships or subtle connotations of individual words. Consequently, the interpretation of LDA results requires careful consideration and may necessitate supplementary qualitative analysis to ensure a comprehensive understanding of the underlying themes.

Moreover, it's imperative to acknowledge the subjective nature of interpreting LDA results. Researcher bias, prior assumptions, and domain knowledge can significantly influence the identification and interpretation of topics extracted from the data. Therefore, robust validation techniques and critical scrutiny are indispensable to mitigate potential biases and ensure the reliability and validity of the findings.

Furthermore, while Reddit serves as a valuable source of data for exploring the experiences and perspectives of South Asian college students, it's essential to recognize its inherent limitations as a sampling frame. Reddit users constitute a specific subset of the broader demographic, characterized by unique demographics, interests, and online behaviors. As such, findings derived from Reddit discussions may not fully encapsulate the diversity and complexity of the entire South Asian college population. Therefore, caution must be exercised when generalizing insights from Reddit to the broader demographic context, and supplementary data sources may be needed to provide a more comprehensive understanding of the target population.

Future Expansion

The rich dataset we've compiled presents numerous opportunities for future exploration and application. One of the most promising avenues lies in utilizing the keywords prevalent among students within the South Asian community. These keywords serve as valuable indicators of underlying thoughts and emotions, particularly those related to depression. By harnessing

these linguistic cues, we can develop sophisticated algorithms and analytical frameworks aimed at identifying early signs of depressive symptoms among college students.

We believe this model has the potential to be scaled for usage in addressing not only this unique community, but also confronting current societal issues and how depressive disorder's online presence is treated. No one should be labeled as a 'Model Minority', and hopefully these word frequencies shed light on the damages such a stereotype can cause. An overall issue within the posts we collected was related to time, and by extension. Pairing this with other words, a story is painted about how the balance between school, parents, and family are major in terms of depressive symptoms. This also could imply that simply our posters were lacking in terms of self care. It demonstrates the issues that result from not taking personal time and furthermore attributing value to personal activities, as a consequence of trying to uphold this unrealistic expectation. From an outside perspective, hopefully studies like this will give more reason for people to step in if they notice individuals who put other needs and expectations ahead of their own. Educators and school administrators can employ similar techniques to monitor student communications and provide targeted support to those in need. Likewise, therapists and mental health professionals can integrate this data-driven approach into their practice, enabling them to identify at-risk individuals more efficiently and tailor interventions accordingly. Furthermore, community leaders and policymakers can leverage these insights to inform the development of proactive mental health initiatives within South Asian communities. By understanding the linguistic patterns associated with depression, they can design culturally sensitive outreach programs and support networks that resonate with the unique needs and experiences of South Asian college students.

On a technology side, with a greater case study, this project could be transformed into being able to detect symptoms of depression in social media users, especially in conjunction with big data. Big data is the enormous amount of info generated by our devices such as smartphones, fitbits, and social media usage that in conjunction can paint a larger picture of human behaviors. The subsequent phase of this project would entail developing a decision tree model capable of leveraging the key words identified to predict the likelihood of an individual experiencing depression based on their statements. This predictive model holds considerable potential for assisting therapists, psychiatrists, and other mental health professionals in the diagnosis and treatment of depression, but at the same time poses a few moral dilemmas.

As AI and machine learning algorithms get more tuned to predict and analyze its users, platforms will be able to make more curated content in terms of individual recommendations, personal content curation, and ads. By integrating information gained from our model, platforms could change how they feed their content, choosing to omit posts that might be triggering and increasing posts that are either gentle in tone, humorous, or entirely unrelated to create more distance from something that may be problematic. Another option could be reminding the user of their current media usage and or promoting posts that deal with mental wellness resources. These options overlook a huge monetary dilemma companies will need to come to terms with in the future deciding how much personal curation will their software be allowed to influence. This could easily become integrated into a tool designed at commercially targeting those demonstrating signs of lesser mentally stability and may result in new methods scam artists might employ to both create and herd more specific audiences. Either futures of social progression and regression are possible, and will ultimately be debated on.

However, by integrating the identified key words into a decision tree framework, our aim for this project would be to create a robust algorithm capable of analyzing textual data and generating predictive insights regarding an individual's mental health status. This tool can serve as a valuable adjunct to traditional diagnostic methods, offering clinicians a data-driven approach to supplement their clinical judgment and enhance the accuracy of their assessments.

Moreover, the decision tree model can facilitate early intervention by identifying individuals at risk of depression, enabling mental health professionals to initiate timely and targeted interventions. By providing actionable insights derived from linguistic cues, this tool empowers clinicians to tailor treatment plans to the unique needs and circumstances of each individual, ultimately improving patient outcomes and enhancing the overall quality of mental health care delivery.

Conclusion

In conclusion, our research has delved into the complex relationship between the topics discussed in Reddit posts and the prevalence of depressive symptoms among South Asian college students. Through our analysis, we uncovered compelling evidence suggesting a significant association between the language used in these online discussions and the manifestation of depressive symptoms. Our examination revealed 'college' as the predominant keyword, highlighting the pivotal role of higher education in shaping mental health dialogues within South Asian communities. Additionally, the salience of terms like 'parents' and 'Mom' underscores the profound influence of familial dynamics and cultural expectations on individuals' emotional well-being. These findings shed light on the multifaceted challenges encountered by college-age individuals, particularly within the context of South Asian cultural norms and societal pressures.

Furthermore, our analysis underscores the importance of understanding cultural nuances and contextual elements in addressing mental health issues among college students. By deciphering the linguistic patterns and discourse surrounding depression, our study enriches the understanding of the hurdles faced by South Asian college students. Moving forward, these insights can inform targeted interventions and support services, fostering a more inclusive and supportive college environment. Our research thus serves as a catalyst for informed initiatives aimed at bolstering mental well-being and resilience within this demographic.

Ultimately, our analysis revealed two distinct clusters, each encapsulating meaningful themes reflective of the complex interplay between mental health and various aspects of the college experience within the South Asian demographic. The prevalence of depression among college-age individuals, manifested in discussions about academic institutions and personal well-being, underscores the need for holistic support mechanisms tailored to address the challenges faced by this demographic subgroup. In essence, these clusters offer valuable insights into the multifaceted nature of the college experience for South Asian students, paving the way for targeted interventions aimed at fostering holistic student well-being. As such, our research contributes to the broader discourse on mental health in higher education, advocating for comprehensive approaches that consider the intersectionality of cultural, social, and individual factors. Through collaborative efforts and evidence-based interventions, we can create a more equitable and supportive environment that empowers South Asian college students to thrive academically and emotionally.

However, it's important to acknowledge the limitations of our study. While our analysis provided valuable insights, it was based on a specific dataset of Reddit posts, which may not fully capture the diversity of experiences within the South Asian college student population.

Additionally, the use of topic modeling techniques, while powerful, relies on assumptions about the underlying structure of the data and may not capture all relevant themes or nuances. Future research could benefit from incorporating a broader range of data sources and methodologies to provide a more comprehensive understanding of mental health among South Asian college students.

Moving forward, it is imperative to recognize the significance of online presence in shaping mental health narratives and well-being. By delving deeper into these digital interactions, we can not only identify potential red flags indicative of mental health concerns but also expedite responses to wellness-based emergencies. Additionally, leveraging online platforms offers a unique vantage point for curating more personalized and supportive feeds through the integration of affective computing technologies. This proactive approach holds promise for fostering increased awareness of mental health issues and facilitating timely interventions tailored to individual needs.

References

American College Health Association. (2020). American College Health Association-National College Health Assessment II: Reference Group Executive Summary Fall 2020. Retrieved from https://www.acha.org/documents/ncha/NCHA-III_Fall_2020_Reference_Group_Executive_Summary.pdf

Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., ... & Kessler, R. C. (2018). WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders. Journal of Abnormal Psychology, 127(7), 623–638. Retrieved from https://pubmed.ncbi.nlm.nih.gov/30211576/

"Depression." National Institute of Mental Health, U.S. Department of Health and Human Services, www.nimh.nih.gov/health/topics/depression#:~:text=Depression%20. Accessed 3 May 2024.

Hunt, J., & Eisenberg, D. (2010). Mental health problems and help-seeking behavior among college students. Journal of Adolescent Health, 46(1), 3–10. Retrieved from https://pubmed.ncbi.nlm.nih.gov/20123251/

Kawamura, K.Y. "Collectivistic Culture." Collectivistic Culture - an Overview | ScienceDirect Topics, 2014, www.sciencedirect.com/topics/psychology/collectivistic-culture.

Mata, D. A., Ramos, M. A., Bansal, N., Khan, R., Guille, C., Di Angelantonio, E., & Sen, S. (2015). Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. JAMA, 314(22), 2373–2383. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4866499/

Lindsey, Billie J., et al. "The Prevalence and Correlates of Depression among College Students." College Student Journal | EBSCOhost, 1 Dec. 2009,

openurl.ebsco.com/EPDB%3Agcd%3A3%3A27953715/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A55492477&crl=c.

Reavley, N. J., McCann, T. V., & Jorm, A. F. (2010). Actions taken to deal with mental health problems in Australian higher education students. Early Intervention in Psychiatry, 4(4), 346–353. Retrieved from https://pubmed.ncbi.nlm.nih.gov/21951839/

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media.