**BERTopic and Causal Discovery Analysis on Mental Health Dataset: Exploring Factors Affecting Mental Health Issues**

ADVANCED SYSTEM ANALYSIS AND DESIGN

PROJECT REPORT

1114233 AKAR SIMAY

1114614 YOON GO EUN

1. **INTRODUCTION**

Mental health issues such as anxiety and depression are increasingly recognized as significant public health concerns, impacting millions of individuals worldwide. Understanding the factors that contribute to these mental health challenges is crucial for developing effective interventions and support systems. In this study, we aim to explore the potential influences on mental health by analyzing a comprehensive dataset using advanced BERTopic and causal discovery methods.

We utilize the Mental Health Corpus dataset from Kaggle, which comprises 27,977 entries of texts related to individuals experiencing various mental health issues. This dataset provides a rich source of information, capturing a wide range of expressions and sentiments associated with mental health. By leveraging BERTopic, a powerful topic modeling tool, we categorize the texts into distinct topics, each representing a different aspect of mental health discourse. This categorization allows us to identify and quantify the prevalence of different themes within the dataset. Following the topic modeling, we calculate the weights of these topics and integrate this information with the labels indicating whether the comments are considered to be about mental health issues or not. This integrated dataset forms the basis for our causal discovery analysis, where we aim to uncover the relationships and potential causal pathways between the identified topics and mental health outcomes. By employing logistic regression, we further interpret the significance and direction of these relationships, providing insights into how different themes may influence mental health positively or negatively.

The findings of this study have the potential to inform mental health professionals and researchers about the complex interplay of factors affecting mental health. By identifying specific topics that are strongly associated with mental health issues, we can better understand the underlying mechanisms and develop targeted strategies to address these challenges. This research contributes to the broader effort to enhance mental health awareness and improve the well-being of individuals through data-driven insights and evidence-based interventions.

1. **RELATED WORKS**

**2.1. Mental Health**

Text analysis techniques have played a crucial role in mental health research. There have been numerous attempts to assess and predict mental health status using data collected from social media platforms. De Choudhury et al. (2013) conducted a study analyzing Twitter data to detect depression symptoms early, providing clinically significant insights. Additionally, Coppersmith et al. (2014) developed models to automatically detect PTSD, depression, and suicidal tendencies using Twitter data.

Analyzing social media content has become a prominent method for understanding mental health issues. Studies have systematically reviewed the relationship between social media use and mental health, finding that excessive social media use can correlate with increased risks of depression and anxiety, while also noting that these platforms can offer community support for some users (Smith & Haque, 2022). Additionally, research by Buddhitha and Inkpen (2022) investigated the use of AI to identify individuals at risk of mental illness based on their social media posts, utilizing datasets that included self-reported diagnoses and varying levels of suicide risk.

**2.2. BERTopic**

BERTopic is a state-of-the-art technique used to identify topics within text data. This model leverages BERT-based embeddings to enhance topic coherence and enable more accurate topic classification. It is known for its superior interpretability and performance compared to traditional topic modeling techniques like Latent Dirichlet Allocation (LDA). Examples of BERTopic applications include detecting emotions in everyday language or analyzing customer reviews to extract key feedback.

A study applied BERTopic to PubMed articles related to depression, anxiety, and burnout, revealing evolving trends and significant themes within the academic literature on mental health. This approach demonstrates the model’s flexibility and effectiveness in handling complex, multi-faceted datasets (Lezhnina, 2023). Additionally, it has been used to identify interdisciplinary topics in fields like library and information science, helping to uncover how these topics evolve over time (Scientometrics, 2022).

BERTopic has been used extensively to analyze social media content, such as Twitter posts, to detect trending topics and public sentiment. This includes tracking public opinions during events like the COVID-19 pandemic and understanding customer preferences in various industries (SpringerLink, 2023; Frontiers, 2023). The model’s ability to handle the dynamic and noisy nature of social media data makes it particularly valuable for these applications.

**2.3. Causal Discovery**

Causal discovery analysis is used to uncover causal relationships between variables. This technique provides significant insights in mental health research. The work of Spirtes et al. (2000) laid a crucial foundation for constructing causal relationship networks and has been applied to analyze causal relationships across various domains. Recently, there has been an increase in studies combining machine learning with causal discovery to analyze complex causal relationships in mental health data. For instance, Peters et al. (2017) published research using machine learning-based methods to discover causal relationships in psychology datasets.

1. **DATA AND METHODOLOGY**

Figure 1 illustrates the research process. Initially, we perform data processing, followed by the application of BERTopic. After identifying the topics using BERTopic, we calculate their weights. We then integrate this information with the label column in the dataset. Subsequently, we conduct Causal Discovery Analysis o n this integrated dataset. Finally, we interpret our findings on the topics affecting mental health issues using Logistic Regression.

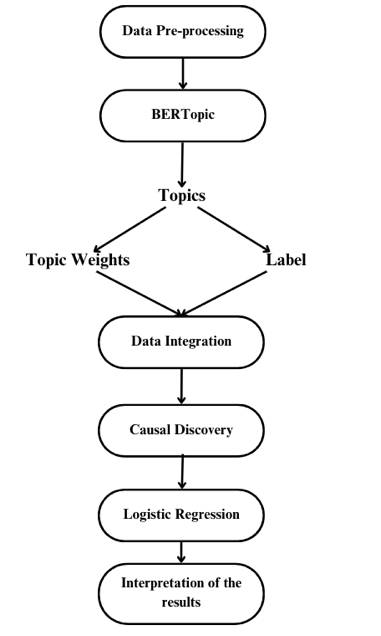
****

Fig.1. Research Process

* 1. **Dataset**

In this study, we utilized the Mental Health Corpus dataset from Kaggle, which has 27,977 entries. This dataset is a collection of texts related to individuals experiencing anxiety, depression, and other mental health issues. It consists of two columns: one with the comments labeled as ['text'] and the other with labels indicating whether the comments are considered poisonous or not, labeled as ['label']. Table 1 displays the distribution of labels in the dataset. As indicated in the table, the data exhibits a balanced distribution among the labels.

Table 1. Dataset Information

|  |  |
| --- | --- |
| **Label** | **Count** |
| 1 | 14139 |
| 0 | 13838 |
| **Total** | 27977 |

* 1. **Data Pre-processing**

First, we converted the entire text to lowercase. During the data cleaning phase, we deleted non-alphabetic characters, single-letter words, numbers, and symbols. We replaced double spaces with single spaces. Afterwards, we applied lemmatization. For the preprocessing part, we finally removed the stopwords using spaCy's stopwords list. Additionally, we added specific words to stopwords list that lacked meaning in the context of the project, were deemed unnecessary, or had their structure altered during lemmatization: ['just', 'wa', 'don', 've', 'ha', 'doe', 'm', 'll', 'didn', 'doesn', 'wan', 'an'].

Then, we checked for any remaining null and duplicate entries in the dataset. There were no null values, but we found and removed 12 duplicate rows. Finally, we used the "langdetect" package to identify and retain only English texts for analysis. There were 2,226 non-English entries, which we removed from the dataset. This streamlined the dataset to 25,739 entries. Figure 2 shows the word cloud of the text tokens.

|  |
| --- |
|  |

Fig.2. Word Cloud of Text Data

Figure 3 shows the 50 most frequent tokens found in the dataset, providing a visual representation of the most common words used in the text.

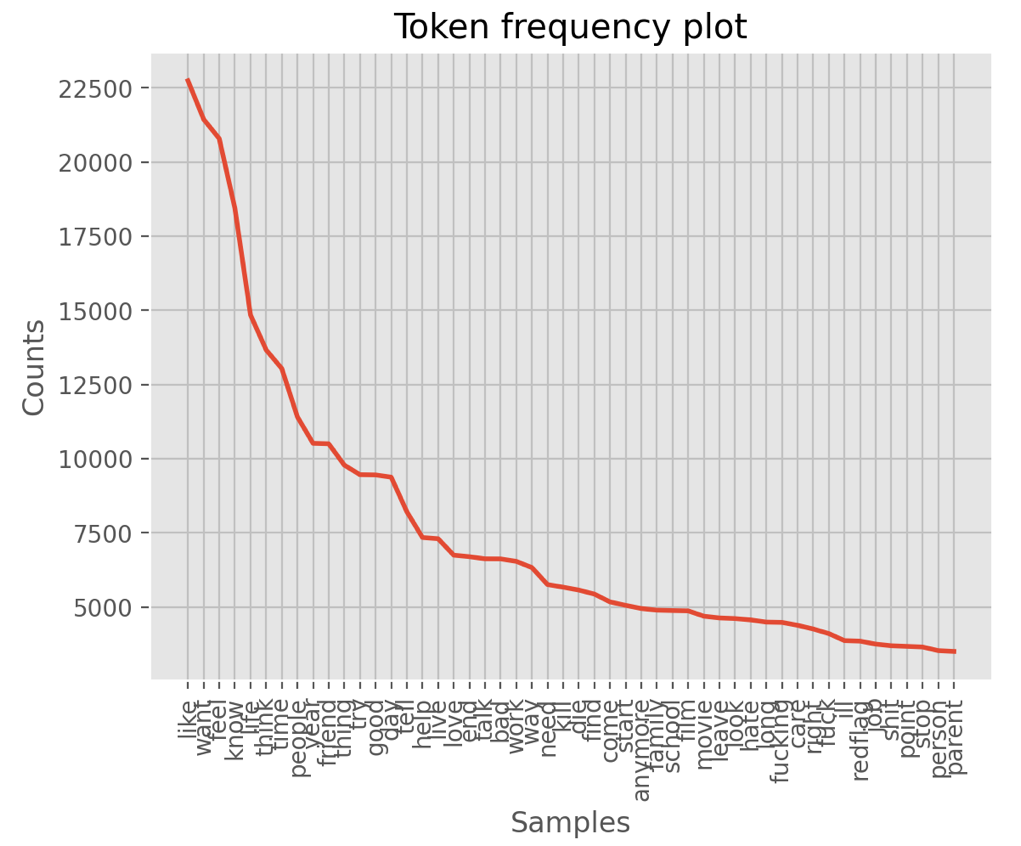


Fig.3. Word Cloud of Text Data

**3.3. BERTopic**

Topic modeling is a method used to uncover the underlying themes of documents by examining the keywords present in extensive collections of literature or textual data (Kim, Lee & Kang 2024). Previous topic models assume that words close to the center of a cluster best represent the cluster and, consequently, a topic. However, in practice, a cluster may not always lie within a sphere around the cluster center, which can render this assumption misleading.

BERTopic, which is one of the topic modeling techniques, utilizes transformers and a class-based variation of TF-IDF to produce coherent topic representations (Grootendorst et al., 2022). Grootendorst et al. (2022) compared BERTopic with LDA, NMF, CTM, and Top2Vec, finding that BERTopic, which incorporates pre-trained transformer-based language models, clustering, and class-based TF-IDF, achieves high topic coherence scores. This high performance is attributed to BERTopic's use of machine learning algorithms such as UMAP and HDBSCAN, in addition to transformers (Atzeni et al., 2022).

We applied BERTopic to analyze hidden subjects and patterns related to mental health data. BERTopic manages its own preprocessing, eliminating the need for extensive data preprocessing when preparing datasets for other models. This is one of the most useful and convenient aspects of BERTopic. Moreover, unlike LDA, BERTopic autonomously determines the number of topics, thus eliminating the need for manual optimization in deciding the number of topics. We specified the UMAP and HDBSCAN models and integrated them into our analysis. Initially, we aimed to identify ten topics. Also, when creating the BERTopic model, we set “calculate\_probabilities” parameter to “True” to calculate topic weights, as we wanted to understand the distribution of topics within our data. This setting allows us to derive probabilities for each topic assignment, giving us insights into how strongly each document relates to each identified topic. Using BERTopic, we processed a total of 25,739 data points.

For each entry, we added the dominant topic number to the data frame. We calculated the weights of each topic for every text and merged this information into the data frame as shown in Figure 4. Topic -1 is specifically reserved for outliers where deriving meaningful interpretations can be challenging, prompting us to exclude texts labeled with Topic -1 from the data frame. In addition, When creating the model initially, we used 10 topics. Based on the results, we applied a threshold and removed topics 6, 7, 8, and 9 along with the documents associated with those topics due to counts less than 100. Afterwards, we normalized the weights of the remaining topics so that their sums equal 1. In the end, the data frame included columns: ‘Document’, ‘Topic’, ‘Topic\_0’, ‘Topic\_1’, ‘Topic\_2’, ‘Topic\_3’, ‘Topic\_4’, ‘Topic\_5’, and ‘label’.

|  |
| --- |
|  |

Fig.4. Data Frame

**3.4. Causal Discovery Analysis**

Causal discovery is a method used to identify cause-and-effect relationships among variables, helping to understand how changes in one variable can influence another. In this project, our goal was to investigate which topic has the most impact on the "label" variable, which represents mental health. Therefore, we conducted causal discovery analysis.

We dropped unnecessary columns, remaining only topic weights and labels in the data frame. A graphical representation based on an adjacency matrix was created using a custom “make\_graph” function. Prior knowledge was then generated using the make\_prior\_knowledge function, where “index = 7” refers to the "label" column, which is the outcome variable. A causal discovery model was set up using LiNGAM. To visualize the causal relationships identified by the model, variable labels were created using the column names from the dataset. The “make\_graph” function represented the graph in the DOT language format, a plain text graph description language used by Graphviz. The graph visualization was then generated using the “make\_dot” function.

Next, we performed logistic regression using the scikit-learn library. Logistic regression is a classification method often used in causal discovery to determine cause-effect relationships between variables. It is particularly useful when the dependent variable is binary, which is why the "label" variable in this project, labeled as 1 or 0, is well-suited for logistic regression. Therefore, logistic regression is an appropriate method when using the "label" variable as the outcome in this project.

1. **RESULT**

**4.1. BERTopic**

The BERTopic results are summarized in Table 2, presenting the seven identified topics along with their representative keywords. Figure 5 illustrates the Word Clouds of the keywords for the topics in the analysis. As previously noted, we excluded four topics (Topic\_6, Topic\_7, Topic\_8, and Topic\_9) due to their counts being less than 100, making them statistically less significant for our analysis. Additionally, Topic\_1, identified as outliers in our analysis, was also excluded from further consideration. After removing the texts associated with Topic -1, and excluding Topic\_6, Topic\_7, Topic\_8, and Topic\_9, the remaining data count in the dataframe was 12,118. Each topic was named based on its representative keywords and relevant documents, as shown in the table.

Table 2. BERTopic Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Topic** | **Count** | **Name** | **Topic Named** | **Representation** |
| -1 | 13512 | -1\_want\_feel\_like\_know | General Feelings | ['want', 'feel', 'like', 'know', 'life', 'think', 'people', 'time', 'friend', 'try'] |
| 0 | 8022 | 0\_like\_feel\_want\_know | Emotion and Thoughts | ['like', 'feel', 'want', 'know', 'time', 'life', 'think', 'friend', 'year', 'people'] |
| 1 | 2928 | 1\_film\_movie\_good\_character | Film and Entertainment | ['film', 'movie', 'good', 'character', 'story', 'great', 'watch', 'scene', 'like', 'play'] |
| 2 | 537 | 2\_wear\_eat\_mask\_like | Diet and Appearance | ['wear', 'eat', 'mask', 'like', 'hair', 'look', 'weight', 'water', 'want', 'milk'] |
| 3 | 397 | 3\_ampxb\_youtube\_draw\_survey | Social Media | ['ampxb', 'youtube', 'draw', 'survey', 'tiktok', 'video', 'link', 'post', 'twitch', 'stream'] |
| 4 | 122 | 4\_racist\_black\_people\_white | Racism and Ethnicity | ['racist', 'black', 'people', 'white', 'trump', 'china', 'racism', 'like', 'spanish', 'race'] |
| 5 | 112 | 5\_math\_test\_drive\_car | Exam Related Stress | ['math', 'test', 'drive', 'car', 'permit', 'driver', 'license', 'physics', 'mailbox', 'crash'] |

|  |  |
| --- | --- |
| Emotion and Thoughts | Film and Entertainment |
|  |  |
| Diet and Appearance | Social Media |
|  |  |
| Racism and Ethinicity | Exam Related Stress |
|  |  |

Fig.5. Word Clouds of Topics

**4.2. Causal Discovery**

Figure 6 illustrates the causal relationships between different topics identified in the mental health dataset and their influence on the label (mental health outcome). The graph was generated using causal discovery analysis, highlighting both the direct and indirect influences of each topic.

The figure reveals several key relationships among the topics and their impact on the mental health label. Topic\_4 (Racism and Ethnicity) and Topic\_5 (Exam Related Stress) both exert influence on other topics. Specifically, Topic\_4 has a direct negative impact on Topic\_3 (Social Media and Surveys), Topic\_2 (Diet and Appearance), and Topic\_1 (Film and Entertainment), while Topic\_5 affects Topic\_2 and Topic\_0 (Emotions and Thoughts) negatively.

Topic\_3 influences Topic\_1 and Topic\_0 negatively and has a smaller effect on Topic\_2. Topic\_2 is influenced by both Topic\_4 and Topic\_5 and affects Topic\_3 and indirectly influences Topic\_0 and Topic\_1.

Topic\_1 is significantly impacted by several other topics and has a direct negative impact on the label. Conversely, Topic\_0, which is influenced by multiple topics, has a strong positive impact on the label.

The label (mental health outcome) is primarily influenced by Topic\_1 and Topic\_0, with Topic\_1 having a moderate negative impact and Topic\_0 having a strong positive impact. The arrows in the figure indicate the direction and strength of these influences, with positive and negative values signifying the nature of the relationship.

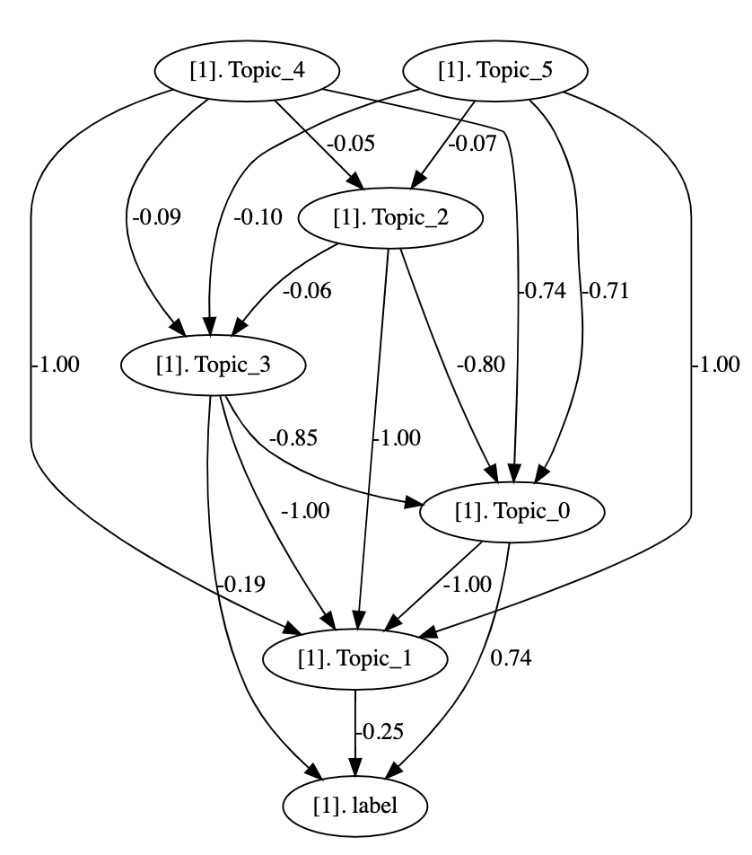


Fig.6. Causal Graph

Table 3 presents the p-values derived from a causal relationship model between independent variables (topic weights) and the dependent variable (label), reflecting their statistical significance and direction of influence. Diagonal cells typically display a value of 0, indicating the lack of direct testing between each topic and itself. Non-diagonal cells signify the statistical significance of relationships: a value of 1 denotes a significant effect of the corresponding topic on the label, while 0 indicates no statistically significant relationship. Interpretations of the cells where significant effects were observed are as follows:

1. **Cell (0, 1):** Indicates that Topic\_0 (Emotions and Thoughts) has a statistically significant impact on Topic\_1 (Film and Entertainment).
2. **Cell (1, 0):** Shows that Topic\_1 (Film and Entertainment) has a statistically significant impact on Topic\_0 (Emotions and Thoughts). This indicates that film and entertainment topics can significantly affect discussions on emotions and thoughts.
3. **Cell (1, 2):** Demonstrates that Topic\_1 (Film and Entertainment) significantly influences Topic\_2 (Diet and Appearance).
4. **Cell (1, 3):** Indicates that Topic\_1 (Film and Entertainment) has a statistically significant impact on Topic\_3 (Social Media).
5. **Cell (1, 4):** Shows that Topic\_1 (Film and Entertainment) significantly influences Topic\_4 (Racism and Ethnicity).
6. **Cell (1, 5):** Indicates that Topic\_1 (Film and Entertainment) has a statistically significant impact on Topic\_5 (Exam Related Stress). This implies that film and entertainment topics can significantly affect discussions about exam-related stress.
7. **Cell (2, 1):** Demonstrates that Topic\_2 (Mask and Appearance) significantly influences Topic\_1 (Film and Entertainment). This suggests that discussions on wearing masks and appearance can significantly impact topics related to film and entertainment.
8. **Cell (3, 1):** Shows that Topic\_3 (Social Media) has a statistically significant impact on Topic\_1 (Film and Entertainment).
9. **Cell (4, 1):** Indicates that Topic\_4 (Racism and Ethnicity) significantly influences Topic\_1 (Film and Entertainment).
10. **Cell (5, 1):** Shows that Topic\_5 (Exam Related Stress) has a statistically significant impact on Topic\_1 (Film and Entertainment).

These findings emphasize the importance of understanding how various topics can influence mental health outcomes, aiding in targeted interventions and content moderation strategies. Therefore, we applied logistic regression to assess both positive and negative effects on mental health issues.

Table 3. Error Independence P-Values

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|  |  | **Topic 0** | **Topic 1** | **Topic 2** | **Topic 3** | **Topic 4** | **Topic 5** | **label** |
| 0 | **Topic 0** | 0. | 1. | 0. | 0. | 0. | 0. | 0. |
| 1 | **Topic 1** | 1. | 0. | 1. | 1. | 1. | 0. | 1. |
| 2 | **Topic 2** | 0. | 1. | 0. | 0. | 0. | 0. | 0. |
| 3 | **Topic 3** | 0. | 1. | 0. | 0. | 0. | 0. | 0. |
| 4 | **Topic 4** | 0. | 1. | 0. | 0. | 0. | 0. | 0. |
| 5 | **Topic 5** | 0. | 0. | 0. | 0. | 0. | 0. | 0. |
| 6 | **label** | 0. | 1. | 0. | 0. | 0. | 0. | 0. |

Using logistic regression in causal discovery is advantageous since it allows for the inference of causal relationships based on the statistical properties of the predictive model. This approach enables us to directly estimate the effects of variables on each other. Table 4 displays the results of the logistic regression model, showing the positive effects (effect\_plus) and negative effects (effect\_minus) of different topics on mental health.

In this analysis, we explored the impact of six topics on mental health issues. Topic\_0, centered on emotions and thoughts, displayed the strongest positive effect (1.245145), indicating a significant association with increased mental health concerns likely stemming from discussions on stressful or anxiety-inducing subjects. Interestingly, Topic\_0 also exhibited a notable negative effect (0.384343), suggesting its potential to alleviate mental health issues at times, showcasing its complex influence. We speculate that Topic\_0's dual effects arise from its diverse content, which includes texts that may exacerbate mental health issues through anxiety-inducing themes, alongside others that offer relaxation and positive emotions.

Conversely, Topic 1 ("Film and Entertainment") demonstrated the highest negative effect (0.623316) among the topics, indicating its significant ability to reduce the likelihood of mental health issues. This effect can be attributed to the positive aspects of escapism and emotional connection that engaging with compelling movie narratives and characters provides. These elements offer temporary relief from real-life stressors, fostering emotional relaxation and social connections through shared interests in films.

However, Topic 1 also showed a moderate positive effect (0.293854) alongside its substantial negative effect. This dual impact underscores that movies can both elevate and increase mental health issues depending on the context. While they offer escapism and entertainment, certain films may trigger intense emotional responses or remind viewers of past traumas, leading to heightened stress and anxiety. Moreover, idealized portrayals in movies can set unrealistic expectations, contributing to feelings of inadequacy or dissatisfaction.

Overall, Topics 2, 3, 4, and 5 had minimal effects on mental health, with both positive and negative impacts being relatively small. As expected, the label variable had no noticeable effect since it is an outcome variable. It's important to understand and balance these impacts to maximize the positive benefits of film engagement and minimize any negative effects on mental well-being.

Table 4. Causal Effects

|  |  |  |  |
| --- | --- | --- | --- |
|  | **feature** | **effect\_plus** | **effect\_minus** |
| **0** | Topic\_0 | **1.245145** | 0.384343 |
| **1** | Topic\_1 | 0.293854 | **0.623316** |
| **2** | Topic\_2 | 0.042458 | 0.045969 |
| **3** | Topic\_3 | 0.231114 | 0.395548 |
| **4** | Topic\_4 | 0.001342 | 0.001339 |
| **5** | Topic\_5 | 0.014876 | 0.014488 |
| **6** | Label (mental\_health\_issue) | 0.000000 | 0.000000 |

**5. CONCLUSION**

In this project, our goal was to identify factors influencing mental health issues and determine their relative significance. We began by applying BERTopic to analyze texts on mental health, extracting and categorizing them into five distinct topics. Subsequently, we calculated the weights of these topics and conducted a causal discovery analysis to explore their effects on mental health.

Using logistic regression, our analysis uncovered significant findings regarding the impact of these topics. Topic\_0, which focuses on emotions and thoughts, exhibited the strongest positive effect, indicating its association with increased mental health concerns, likely stemming from discussions on stressful themes. Interestingly, Topic\_0 also showed a notable negative effect, suggesting its potential to alleviate mental health issues through content that promotes relaxation and positive emotions. Conversely, Topic 1 ("Film and Entertainment") demonstrated the highest negative effect, implying its role in reducing the likelihood of mental health issues through escapism and emotional engagement. However, Topic 1 also displayed a moderate positive effect, underscoring films' dual potential to both alleviate and exacerbate mental health issues depending on their content and context. Topics 2, 3, 4, and 5 had minimal overall effects, highlighting the nuanced impact of different topics on mental well-being.

This project has several limitations. Firstly, determining the number of topics using BERTopic posed challenges. This decision heavily influences the interpretability and coherence of the topics extracted from the mental health-related texts. The subjective nature of topic selection can affect the outcomes and conclusions drawn from the analysis. Secondly, we used an existing dataset from Kaggle. The dataset contained numerous typos and errors that were challenging to detect. Additionally, the dataset lacked timestamp information, which prevented us from analyzing its temporal aspects effectively. These limitations underscore the importance of implementing rigorous data quality control measures in future studies.Thirdly, the dataset was already labeled as about mental health, but the criteria used for labeling were not rigorously defined or verified. This ambiguity raises questions about the accuracy and consistency of the labels applied to the dataset. Exploring how these labels were assigned and considering any discrepancies is essential for ensuring the validity of our analyses and interpretations.

Addressing these limitations is crucial for future research to enhance the robustness and reliability of studies investigating factors influencing mental health issues. By employing alternative topic modeling methods, ensuring rigorous data quality control, and rigorously defining labeling criteria, future studies can build upon our findings. This approach will enable more accurate insights into this critical area of research, thereby contributing to a deeper understanding of the factors impacting mental health.

**REFERENCES**

* Atzeni, D., Bacciu, D., Mazzei, D., & Prencipe, G. (2022). A Systematic Review of Wi-Fi and Machine Learning Integration with Topic Modeling Techniques. Sensors, 22(13), 4925.
* Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. ACL Workshop on Computational Linguistics and Clinical Psychology.
* De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. ICWSM.
* Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure.
* Kim, H., Lee, J., & Kang, S. (2024). Topic Modeling on Patent and Article Big Data Using BERTopic and Analyzing Technological Trends of AI Semiconductor Industry. Journal of Information Technology and Applied Management, 31(1), 139–161.
* Lezhnina, E. (2023). Application of BERTopic to PubMed articles.
* Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press.
* Scientometrics (2022). Identifying interdisciplinary topics and their evolution based on BERTopic.
* Smith, J., & Haque, U. (2022). Social media use and mental health: A global analysis.
* Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, Prediction, and Search. MIT Press.
* SpringerLink (2023). Unsupervised Topic Modeling with BERTopic for Coarse and Fine-Grained News Classification.