Suicide Ideation Detection from Social Media Using Language Models: Data Augmentation and Interpretability

Hamideh Ghanadian

Thesis submitted to the University of Ottawa in partial
fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science and Artificial Intelligence

Department of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Supervisors:
Professor Hussein Al Osman
Dr. Isar Nejadgholi

Examining Committee Members

The following served on the Examining Committee for this thesis proposal. The decision of the Examining Committee is by majority vote.

External Examiner: Amine Trabelsi
Assistant Professor,
School of Computer Science,
University of Sherbrook

Internal Member: Diana Inkpen
Professor,
School of Electrical Engineering and Computer Science
University of Ottawa

Shervin Shirmohammadi
Professor,
School of Electrical Engineering and Computer Science
University of Ottawa

Olga Baysal
Associate Professor,
School of Computer Science,
University of Carleton

Supervisor(s): Hussein Al Osman
Associate Professor,
School of Electrical Engineering and Computer Science

Isar Nejadgholi
Adjunct Professor,
School of Electrical Engineering and Computer Science

Abstract

Early detection of suicide is a vital research area that holds great potential for facilitating early prevention and interventions by mental health professionals. With accurate and reliable detection of suicide ideation, targeted interventions can be developed to reduce suicide rates and provide better support for at-risk individuals. While traditional methods of identifying individuals at risk of suicide have primarily relied on clinical assessments and crisis hotlines, the ubiquity of social media platforms has opened new avenues for early detection and intervention, as many individuals at risk of suicide might express suicidal ideation in their social media interactions. However, developing models for suicide detection on social media is a challenging area of research, primarily due to ethical and practical issues in data collection and annotation.

In this work, we investigate the potential and limitations of Large Language Models in addressing data quality and accessibility issues in suicide detection on social media. First, we explore the capabilities of the state-of-the-art generative LLMs as Zero-shot or Few-shot alternatives to classifiers trained with annotated datasets. Our evaluations of the ChatGPT system underscore the limitations of this model in detecting suicide notes and highlight the necessity of high-quality training datasets for fine-tuning specialized classifiers for this task.

Then, we turn to assess the quality of existing datasets collected from social media. With this assessment, we seek to uncover the extent to which social media datasets mirror or diverge from conventional psychological understandings about suicide-related topics. We ground our evaluation of the datasets in established psychological literature by identifying risk factors linked to suicide, such as mental health challenges, relationship conflicts, and financial distress. Employing a guided topic modelling technique, we identify the distribution of mentions of risk factors in existing datasets. Our results demonstrate that while surface-

level risk factors such as depression and anxiety dominate the topics of these datasets, more stigmatized topics such as racism, immigration challenges or sexual orientation prejudices are completely absent in these datasets. These results highlight the necessity of creating more diverse datasets that cover the risk factors related to under-represented social groups.

Next, we focus on addressing the topic coverage issues in training datasets. Acknowledging that the sensitivity surrounding suicide-related data poses challenges in accessing diverse real-world examples, we introduce an innovative strategy that leverages the capabilities of generative AI models, such as GPT models, Flan-T5, and LLama2, to create synthetic data for suicidal ideation detection. Our data generation approach is grounded in social factors extracted from psychology literature and aims to ensure coverage of essential information related to suicidal ideation. Our comparison of synthetic and real data shows that synthetic data is more balanced in terms of risk factor coverage, is not significantly different from real data in terms of complexity and readability, and is significantly less diverse in terms of the vocabulary used.

We then study the impact of psychology-grounded synthetic data on both the performance and the internal representations of suicide-detection models. We first leverage the generated synthetic data as standalone training data and as an augmentation source for fine-tuning BERT-family models for suicidal ideation detection. Our results show that synthetic datasets generated across multiple large language models enable strong generalization to real-world data, achieving an F-score of 82% when evaluated on held-out real samples. Moreover, augmenting this synthetic data with only 30% of the real dataset yields models that outperform those trained exclusively on the full real dataset, demonstrating a cost-effective strategy for improving performance while mitigating topic imbalance.

Finally, we examine how topic-aware data augmentation influences the internal representations learned by these models. Using sparse autoencoders and geometric analyses, including

UMAP projections and cosine-distance measurements, we analyze whether psychologically meaningful risk factors are encoded as more distinct and separable directions in the models' latent spaces. Our findings indicate that augmentation not only improves predictive performance but also leads to more structured and interpretable internal representations, with several previously under-represented risk factors becoming more clearly encoded. Together, these results highlight the value of combining synthetic data generation with representation-level analysis to develop more reliable and transparent models for suicidal ideation detection.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, *Dr. Hussein Al Osman*, for his continuous support, patience, and guidance throughout my doctoral journey. His calm presence, thoughtful advice, and genuine care created an environment in which I could grow both academically and personally. He was always willing to listen, to encourage, and to guide me with clarity and generosity. I am especially grateful for his trust in me and for the steady confidence he showed in my abilities, which gave me strength during challenging moments and helped me move forward with purpose and resilience.

I am profoundly thankful to my co-supervisor, *Dr. Isar Nejadgholi*, whose support I felt from the very beginning of my journey to its completion. She played a crucial role at the start of my education in University of Ottawa by introducing me to Dr. Al Osman and helping shape the path that followed. Throughout these years, her kindness, professionalism, and strong scientific background have been a constant source of reassurance and inspiration. Her encouragement made a lasting difference, and I am deeply grateful for her belief in me and her continuous support from start to finish.

I would like to express my heartfelt gratitude to my parents for their unconditional love, belief, and encouragement throughout my life. Their support has been a constant source of strength, even in moments when I doubted myself. I also thank my brothers, for being part of my foundation, for all of the support they've provided me over the last several years, and for all of the incredible strength they've forced me to see in myself.

My deepest appreciation goes to my son, *Mohammadhossein*. His presence transformed my life in the most profound way, bringing meaning, motivation, and joy to every step of this journey. This work, and the journey behind it, is inseparable from the love and purpose he brought into my life.

I would also like to thank my friends and colleagues for their kindness, encouragement, and support throughout this journey. Their presence and understanding made this path lighter and more meaningful.

And finally, "All the praises are to Allah, the Lord of all the worlds."(Quran 1:2)

Dedication

To my parents,

and to my son,

Mohammadhossein—

my brightest light.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACE Automatic Concept-based Explanations 11, 89

AUG Augmented Dataset 99, 102, 104, 149

CNN Convolutional Neural Network 17

GLLM Generative Large Language Model 8, 10, 16, 30, 63, 64, 74, 85, 149

GLUE General Language Understanding Evaluation 20, 39

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise 53, 54

LLM Large Language Model 11, 20, 67, 71, 110, 111, 115

NLP Natural Language Processing 1, 2, 4, 10, 11, 13, 14, 16, 17, 19, 21, 31, 61–64, 80, 82, 85, 110, 111

PR Precision Recall Curve 34

ReLU Rectified Linear Unit 100

RLHF Reinforcement Learning from Human Feedback 16

RNN Recurrent Neural Network 29

SAE Sparse Autoencoder 88, 92, 98, 99, 106

SHAP Shapley Additive Explanations 11, 89

TCAV Testing with Concept Activation Vectors 11, 21, 88–90, 99

TF-IDF Term Frequency-Inverse Document Frequency 52, 54, 56, 57

TTR Type-Token Ratio 76

UMAP Uniform Manifold Approximation and Projection 8, 53, 54, 56, 91, 97, 102, 108, 113

UMD University of Maryland Reddit Suicidality Dataset 18, 25, 32, 38–40, 57, 62, 63, 66, 71, 72, 74, 75, 78, 82, 97, 99–101, 104–106, 149

WHO World Health Organization 1, 2

Chapter 1

Introduction

1.1 Motivations

Suicide is a global public health concern with profound implications for individuals, families, and communities. The World Health Organization (WHO)[1] has declared that the alarming rise in suicide rates in recent years underscores the urgent need for effective suicide prevention strategies. Understanding the underlying social phenomena that lead to suicide is a critical step in developing effective prevention strategies. This knowledge informs the development of targeted interventions and policies aimed at reducing suicide rates and providing better support for at-risk individuals. While traditional methods of identifying individuals at risk of suicide have primarily relied on clinical assessments and crisis hotlines, the ubiquity of social media platforms has opened new avenues for early detection and intervention[1].

In recent years, there has been a growing interest in using Natural Language Processing (NLP) techniques for suicide prevention [2, 3]. Vioulès et al. [4] used text data extracted

[1] The World Health Organization (WHO)

from Twitter to detect linguistic markers of distress and other risk factors, these systems can help identify individuals with a risk of suicidality and provide early interventions to prevent such incidents. NLP techniques, therefore, offer a promising avenue for suicide prevention efforts, enabling more proactive and effective interventions to support those in need. Researchers have developed suicide detection systems to analyze and interpret social media data, including text data. Social media platforms are becoming a common way for people to express their feelings, suffering, and suicidal tendencies. One of the most effective methods recommended by the WHO for preventing suicide is to obtain information from social media and report suicidal ideation to healthcare providers to enable early identification, assessment, and follow-up with affected individuals[5]. Hence, the datasets collected from social media offer a window into individuals' digital lives, providing rich data for research and mental health support. Researchers and data scientists leverage these datasets to analyze text, images, and interactions, seeking patterns and linguistic cues associated with suicidal ideation. By identifying users who may be at risk, the models trained with these datasets enable early interventions and offer valuable insights for public health professionals. Understanding the themes that resonate with those who may be in distress allows for more empathetic and tailored responses, ultimately contributing to the development of effective suicide prevention strategies. However, certain themes are discussed less on social media due to societal stigma and the specific demographic of social media users. Therefore, datasets used for training suicidal ideation detection models do not include all the relevant topics and themes associated with suicide. The lack of important topics and risk factors in social media datasets used for suicidal ideation detection poses a significant challenge in the development of effective mental health AI models. These datasets often fall short of representing the full spectrum of issues and underlying factors related to suicidal ideation. As a result, several potential issues may arise:

- **Limited Generalization:** The model trained on such a dataset may not generalize well to real-world scenarios. It might be biased or overly focused on the specific topics in the training data and perform poorly on cases that involve different themes or factors related to suicide.

- **Under-representation:** The model may under-perform on the test data of certain subgroups or demographics not adequately represented in the training dataset. Certain topics that may not be present in the training data are typically associated with specific subgroups. Hence, the absence of such topics affects the model's performance on the data of these subgroups. For example, victims of racism might be less openly expressing the cause of their suicidal thoughts. This can lead to false negatives, where the algorithms do not identify racialized individuals who need help due to the under-presentation of this topic in training datasets.

- **Biased Outputs:** If the dataset contains biases or skewed thematic information, the model might inadvertently perpetuate these biases. For instance, if the dataset predominantly reflects negative stereotypes or stigmatization of certain topics related to suicide, the model may reinforce these biases in its predictions. Introducing posts from diverse topics makes the dataset more representative of a broader range of perspectives and sentiments. This diversity can counteract the impact of any existing biases in the original dataset, providing a more balanced and comprehensive view. For example, if in a dataset, LGBTQ+ community issues predominantly appear in high-risk suicidal examples, the models might misidentify members of this community as individuals at risk, ignoring the context of their discussions.

Generative language models open new venues to address the data accessibility and quality in this field. First, these models can be used as Zero-shot and Few-shot classifiers

without the need for extensive training data. Second, they can be used to generate synthetic training data. Synthetic data generation offers a viable solution to mitigate the data availability limitation by creating artificially generated data that closely resembles real-world data. Synthetic data generation can be instrumental in machine learning applications as it addresses many challenges of real data collection and annotation. Below, we review a list of common challenges in data collection that can be managed through synthetic data generation.

- **Data Scarcity:** In many NLP tasks, such as mental health-related applications, relevant data may be limited due to privacy concerns or the complexity and cost associated with manual annotation. Synthetic data generation allows researchers and practitioners to overcome data scarcity issues and augment the limited amount of publicly available data [6].

- **Data Diversity:** NLP models trained on limited data may suffer from poor generalization and performance when exposed to diverse and previously unseen examples. Moreover, certain topics can be undermined or overlooked in real data due to being less discussed. This can happen for several reasons. For example, certain topics may be stigmatized and considered too sensitive or taboo, making people hesitant to discuss them openly. This could include subjects related to mental health, addiction, discrimination, or social issues that carry societal stigmas. Additionally, topics relevant to marginalized or minority communities may receive less discussion due to systemic biases, unequal representation, or limited platforms for their voices to be heard. Also, some topics may be highly specialized or complex, requiring specific expertise or background knowledge to engage in meaningful discussions. Encouraging diverse perspectives and actively seeking out less-discussed topics can contribute to a more comprehensive and nuanced understanding of real-world

issues. Synthetic data generation can help enrich the training data by introducing a wider variety of linguistic patterns, sentence structures, vocabulary, and topics. This, in turn, improves the model's ability to handle variations in natural language and increases its robustness [7].

- **Privacy Preservation:** Suicide detection tasks often involve sensitive information. Generating synthetic data allows researchers to create representative samples that preserve the privacy of individuals while maintaining the statistical properties and distribution of the original data. [8]
- **Annotation Cost:** Suicide detection is a complex task, and high-quality annotation can only be performed by experts and trained annotators, which can be costly [9]. Synthetic data generation addresses the data annotation issue by using targeted data generation so that each generated example is pre-labeled with a specific category.

In this research, we first evaluate the performance of general-purpose generative language models in this task. Given that these models are exposed to a wide variety of topics and datasets, we explore their potential and limitations in classifying social media posts reflecting suicidal ideation.

Then, we integrate insights from established psychological literature to pinpoint key risk factors associated with suicide. Using these insights, we conduct a thorough analysis of suicide-related datasets gathered from social media and utilize topic modelling techniques to identify their prevalent risk factors and topics. The objective is to evaluate the alignment between social media discussions and traditional psychological literature. In the field of psychology, practitioners and researchers employ a combination of quantitative and qualitative research methods to understand the factors contributing to suicide [10, 11]. By understanding these factors, they identify statistical trends and appreciate the

individual narratives and contextual factors that contribute to suicide risk. Psychologists, in collaboration with other professionals, continue to explore these various research methods to shed light on the complex interplay of biological, psychological, and social factors contributing to suicidal ideation and behaviors.

Additionally, the research explores the coverage of topics in synthetic datasets as a potential means of data augmentation. Augmentation of real-world data with meticulously generated and labeled synthetic data is one of the techniques used to address the issues mentioned above. This solution has captured more attention in recent months with the ubiquitous emergence of generative models, which makes high-quality data generation more feasible than ever. This work aims to assess and compare the diversity of real and synthetic datasets. We ground our assessments in a comprehensive set of social factors associated with suicide in the psychology literature.

Despite recent advances in large-scale language models and their promising performance in suicide ideation detection, most existing approaches treat these models as black boxes, focusing primarily on predictive accuracy. In high-stakes domains such as mental health, however, performance alone is insufficient. Models that influence decision-making must provide insight into why a prediction is made, which risk factors are being activated, and whether these factors align with established knowledge. Without interpretability, it is difficult to assess whether a model is relying on meaningful psychosocial signals or on spurious correlations present in biased or incomplete datasets.

Therefore, interpretability plays a complementary and integrative role in this thesis. Beyond improving detection performance through topic-aware data augmentation, this work seeks to understand how suicide-related risk factors are internally represented within neural language models. By analyzing model representations, we can evaluate whether augmentation strategies genuinely increase coverage of psychologically grounded risk factors,

or merely improve surface-level performance metrics. In this sense, interpretability bridges the gap between data-centric interventions such as synthetic augmentation and theory-driven suicide research, ensuring that improvements in model performance correspond to meaningful and responsible representations of suicide risk. This perspective positions interpretability not as a standalone objective, but as a necessary component for validating, contextualizing, and governing the use of generative language models in suicide ideation detection.

1.2 Contributions

The main contributions of this study are as follows:

- We investigate the performance of a widely used general-purpose generative language model, GPT3.5, in assessing the degree of suicidality in Reddit posts. We use this model in the Zero-shot and Few-shot Learning modes and compare that with two transformer-based models, ALBERT and DistilBERT, fine-tuned with task-specific labeled datasets. Our results indicate that while GPT3.5 shows promise in suicide risk evaluation via Zero-shot learning, fine-tuned ALBERT demonstrates superior performance. This comparison suggests that while general-purpose language models might be a viable option in detecting suicidal ideation when no training data is available, the development of specialized datasets and fine-tuned models is critical for creating reliable models.

- We conducted a literature-informed analysis of social and psychological factors associated with suicidal ideation and used these factors as an analytic framework to evaluate existing social media datasets. Specifically, we identified a set of nineteen commonly reported risk-related themes from prior psychology and clinical research and assessed their coverage

in three real-world social media datasets using guided topic modeling. This analysis revealed substantial gaps in topic coverage, with several clinically relevant risk factors being underrepresented or absent in the real-world datasets.

- We created synthetic and augmented datasets by incorporating our extracted suicidal risk factors into synthetic data generation using three Generative Large Language Models (GLLMs), namely, GPT-3.5 Turbo, Flan-T5, and Llama2. The objective of dataset generation was to enhance the topic diversity of the datasets to encompass a larger number of known risk factors associated with suicidal ideation. Moreover, the synthetic dataset was annotated by two experts to provide a gold-standard synthetic dataset. We fine-tuned a state-of-the-art transformer-based large language model, ALBERT, to compare the performance of these datasets in training a model. Our results show that synthetic data is a promising approach to addressing the data issues in this field.

- We investigated the internal reasoning processes of suicide detection models through a mechanistic interpretability framework. Building on the linear representation hypothesis and the theory of superposition, we applied dictionary learning using sparse autoencoders to uncover the extent to which psychological risk factors emerge as coherent, monosemantic directions in the model's latent space. Using UMAP visualizations and cosine-distance measurements, we quantified the separability of concepts such as anxiety, family issues, racism, and financial stress from non-risk text. Our findings demonstrate that topic-aware synthetic augmentation not only improves predictive performance but also leads to clearer, more structured internal representations, revealing how models encode and distinguish psychosocial concepts.

1.3 Research Publications

- Paper 1:(Conference) Ghanadian, H.; Nejadgholi, I.; and Al Osman, H.."ChatGPT for Suicide Risk Assessment on Social Media: Quantitative Evaluation of Model Performance, Potentials and Limitations". Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2023). (The contents of this paper will appear in Chapter 4)

- Paper 2: (Journal) Ghanadian, H.; Nejadgholi, I.; and Al Osman, H.. "Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models." IEEE Access (2024). (The contents of this paper will appear in Chapter 6)

- Paper 3:(Journal) Ghanadian, H.; Nejadgholi, I.; and Al Osman, H.. "Improving Suicidal Ideation Detection in Social Media Posts: Topic Modeling and Synthetic Data Augmentation Approach". JMIR Formative Research (2025) (The contents of this paper will appear in Chapter 5)

- Paper 4:(Conference) Ghanadian, H.; Nejadgholi, I.; and Al Osman, H.. "Beyond Accuracy: Interpreting Topic Coverage in Suicide Ideation Detection Models". (To be submitted).

1.4 Organization of the Thesis

The road map for the rest of this thesis is outlined below:

Chapter 2: The Background and Related Works section provides a comprehensive review of methods for detecting suicidal ideation, encompassing both traditional approaches

and those based on Natural Language Processing (NLP). It delves into a detailed examination of various techniques, including topic modeling, synthetic data generation, and the explainability methods of machine learning models.

Chapter 3: Datasets utilized in this study are comprehensively introduced and elaborated upon to provide a thorough understanding of their characteristics, composition, and relevance to the research objectives.

Chapter 4: An Evaluation of GPT-3.5 Turbo, a widely recognized GLLM, is presented with the objective of investigating its limitations and potential in detecting suicidal ideation using a real-world dataset.

Chapter 5: Assessment of Topics in Social Media Datasets extracts the most frequently discussed subjects based on the risk factors from the literature. Moreover, we draw upon established psychological literature to identify core topics linked to suicide, such as mental health challenges, relationship conflicts, and financial distress.

Chapter 6: A Synthetic Data Generation Approach is presented that leverages the capabilities of generative AI models, such as GPT-3.5 Turbo, Flan-T5, and Llama, to create synthetic data for suicidal ideation detection. Our data generation approach is grounded in social factors extracted from psychology literature and aims to ensure coverage of essential information related to suicidal ideation.

Chapter 7: Interpretability of LLMs builds the theoretical and methodological foundation necessary to analyze how suicide-related concepts are internally represented within transformer models. This chapter introduces the linear representation hypothesis and the phenomenon of superposition, which explains why models often entangle multiple psychological concepts within shared latent directions. It then motivates the use of

dictionary learning—specifically sparse autoencoders as an appropriate tool for uncovering monosemantic, interpretable features from the residual stream. In addition, this chapter reviews concept-based explanation methods such as TCAV, ACE, and ConseptSHAP, clarifying that these methods quantify *sensitivity*, whereas our work focuses on *separability*, and studies whether risk-factor concepts emerge as distinct geometric directions.

Chapter 8: The Conclusion synthesizes the thesis evaluation of LLMs, topic analysis, synthetic data generation, and mechanistic interpretability into a unified perspective on how psychological insight, data augmentation, and model transparency can jointly improve suicidal ideation detection. This chapter highlights the broader implications of the work and articulates how combining domain knowledge with interpretability-driven modeling leads to safer and more reliable mental-health NLP systems.

Chapter 2

Background and Related Work

2.1 Suicidal Ideation and Related Social Factors

Suicidal ideation refers to the presence of thoughts, considerations, or plans related to self-harm or suicide, and its accurate identification is critical for early intervention and the prevention of self-harm. Within psychology, extensive research has been conducted to understand the underlying psychological, social, and environmental factors contributing to suicidal thoughts and behaviors.

A key focus of research pertains to the exploration of risk elements linked to suicidal ideation. Several investigations have explored how psychological factors, including depression, anxiety, a sense of hopelessness, and feelings of low self-worth, influence the emergence of suicidal thoughts [12, 13]. These studies investigate the strong association between suicidal thoughts and conditions like depression [14, 15], bipolar disorder [16], borderline personality disorder [17], and substance abuse [18]. Through a comprehensive exploration of the interactions between these conditions, researchers strive to create tailored interventions

that can effectively address the distinct challenges experienced by individuals challenged with suicidal ideation [19, 20]. Furthermore, environmental elements, such as a history of trauma, social isolation, and the availability of lethal methods, have been recognized as potential risk factors [21, 22, 23].

The psychology literature offers valuable insights into the diverse processes and factors that contribute to suicide risk. Psychological theories and frameworks such as the interpersonal theory of suicide [24], the cognitive model of suicidal behavior[25], and the social-ecological model[26] provide a theoretical foundation for understanding the complex interplay between individual vulnerabilities and environmental factors.

Extensive research within the field of psychology, focusing on topics related to suicidal ideation, has played a pivotal role in deepening our comprehension of the factors at play. This research aims to untangle the origins, risk factors, and social factors intertwined with suicidal thoughts, with the ultimate goal of formulating impactful prevention measures and mitigating the worldwide toll of suicide. In Section 5, we outline the social factors that have been explored in the literature as pertinent aspects of suicidal ideation.

2.2 Suicidal Ideation Detection Using NLP

In recent years, NLP research has increasingly recognized the significance of addressing the task of suicide detection within text data [2, 3]. Given the proliferation of online communication and social media platforms, analyzing text data for signs of suicidal ideation has become a vital avenue for early intervention. This interdisciplinary intersection between psychology and NLP provides a unique opportunity to harness the power of language processing to support mental health and well-being.

Numerous scientific investigations have demonstrated the significant influence of reciprocal connectivity within social networks on an individual's propensity for suicidal ideation. These studies indicate that the quality and extent of one's connections in online social spaces can serve as important predictive factors in understanding and assessing the risk of suicidal thoughts and behaviors. Hsiung et al. [27] analyzed the changes in user behavior following a suicide case that occurred within a social media group. Jashinsky et al. [28] highlighted the geographic correlation between suicide mortality rates and the occurrence of risk factors in tweets. Their results indicate a strong correlation between state Twitter-derived data and actual state age-adjusted suicide data. Moreover, they concluded that X platform (Twitter) may be a viable tool for real-time monitoring of suicide risk factors on a large scale. Colombo et al. [29] conducted a study focused on the examination of tweets containing expressions of suicidal ideation, with a specific focus on the behavioral patterns exhibited by users in their interactions within social networks.

There are a vast number of research techniques that investigate the risk factors of suicide ideation. For instance, clinical methods examine the resting state of heart rate [30] and event-related initiators such as depression [31] as suicidal indicators. Traditional methods use questionnaires, electronic health records, and face-to-face interviews to assess the potential risk of suicide [32, 33, 34].

In recent years, NLP researchers have started to analyze users' posts on social media websites to gain insight into language usage and linguistic clues of suicidal ideation [35, 36]. Using NLP techniques, suicide-related keyword dictionaries and lexicons are manually built to enable keyword filtering [37].

Several studies have explored mental illness detection from social media text [38], motivated by the growing role of online platforms in reflecting users' personal experiences.They highlight the challenges of identifying mental health conditions from noisy and unstructured

social media data, particularly under limited labeled data settings. They evaluate several deep neural network architectures for detecting depression from Twitter posts and show that model performance is strongly constrained by data scarcity [39]. Moreover, Skaik et al. investigate the detection of depressive signals from Twitter by leveraging personal narratives from users with self-reported diagnoses. They demonstrate that machine learning and deep learning models trained on such data can achieve strong predictive performance and produce prevalence estimates consistent with national statistics[40].

The related analysis contains lexicon-based filtering [41], topic modeling within suicide-related posts [42], transformer-based models, and unsupervised learning [43]. In line with this field of research, we examine the use of the GPT3.5 model, as a widely used general-purpose generative model, for this task, where no labelled data (Zero-shot setting) or a small labelled dataset (Few-shot setting) is available.

Topic modeling has proven to be a powerful tool for analyzing vast amounts of text data from various sources, including social media, forums, healthcare records, and other online platforms, where individuals may express their thoughts, emotions, and experiences related to suicide. Kumar et al. employed topic modeling to investigate the Werther effect on social media [44]. This effect describes the increased rate of completed or attempted suicides following the depiction of an individual's suicide in the media, typically a celebrity. They demonstrated that discussions about self-deprecating behavior, suicidal tendencies, and disengagement from social interactions are more commonly observed in the aftermath of celebrity suicides.

Moreover, Chishima et al. [45] studied the categorization of participants' narratives of daily experiences during the COVID-19 pandemic into different thematic domains using topic modeling. Then, they explored the associations between these thematic categories and the mental health status of the participants. The findings of this study unveiled

that individuals in Japan who articulated concerns regarding economic impacts, physical symptoms, and disinfection-related items during the pandemic exhibited diminished levels of life satisfaction, elevated depressive symptoms, and experienced more negative effects.

In this work, we use guided topic modeling, where a set of predefined topics are explored in a dataset. With guided topic modelling, we assess the extent to which the suicide-related themes and topics are covered in existing datasets.

2.3 Generative Large Language Models

Generative Large Language Models(GLLM) have proven valuable in a range of NLP tasks, including assessing individuals' mental states through their online interactions, particularly on social media platforms[46].

ChatGPT is a state-of-the-art artificial intelligence (AI) Chatbot developed by OpenAI [47] that has gained widespread attention for its ability to generate human-like text. The original GPT model was trained on a massive corpus of text data, including books, articles, and web pages, using an unsupervised learning approach. The model's performance on a range of language tasks has since been surpassed by newer models, including GPT-2 [48] and GPT-3 [49], which have larger training datasets and more sophisticated architectures. However, the GPT3.5 model has been fine-tuned on large datasets of conversation data, including social media posts, customer support interactions, and chatbot logs [50]. GPT3.5 differs from prior models as it employs Reinforcement Learning from Human Feedback (RLHF). Unlike supervised learning methods that depend on pre-existing training data, in RLHF, the response generated by a model to a given input is evaluated by a human reviewer. The feedback obtained from the evaluator is used to train the model using reinforcement learning, with the objective of maximizing the reward received [51].

Several recent studies have explored the effectiveness of GPT3.5 in a variety of settings, including chatbots and virtual assistants. One study created a corpus named Human ChatGPT Comparison Corpus by collecting a set of question-and-answer datasets covering various domains such as finance, medicine, and psychology [52]. They conducted a comparative analysis between the responses generated by ChatGPTand those provided by humans to investigate the distinguishing features of ChatGPT's responses. In [53] ChatGPT was employed to produce a simplified version of a radiology report, which was then evaluated for quality by radiologists. Another study investigated the proficiency of ChatGPT in answering questions related to the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams [54]. They found that ChatGPT performed similarly to a third-year medical student.

Bang et al. [55] proposed a framework for evaluating interactive GPT3.5 language learning models using publicly available datasets. They evaluated GPT3.5 using 23 datasets covering 8 different NLP tasks, such as summarization, machine translation, sentiment analysis, question answering, etc. They reported that GPT3.5 outperforms large language models with Zero-shot Learning such as InstructGPT, NLLB-200 and XLM-R LARGE on most tasks and even outperforms fine-tuned models on some tasks.

Yang et al. [56] conducted extensive research to assess GPT3.5's capabilities in mental health analysis and emotional reasoning, encompassing a battery of five distinct tasks. They also delved into exploring diverse emotion-based prompting strategies alongside investigating the application of generative models for crafting elucidations regarding GPT3.5 's decision-making processes, with the overarching objective of achieving interpretable mental health analysis. The findings from their experiments revealed that GPT3.5 exhibited superior performance in mental health analysis compared to conventional neural network-based methods such as Convolutional Neural Network (CNN) and Gated Recurrent Unit (CNN).

In this work, we evaluate GPT3.5's capabilities in detecting suicide notes and generate synthetic training data for this task.

2.4 Suicidal Ideation Detection Datasets

Social Media Datasets:

Several datasets have been collected from social media platforms to serve as a resource for creating suicidal ideation detection systems. These datasets encompass a wide range of information collected from various social media sources, including *Twitter*, *Reddit* and other user-generated content. Sinha et al.[57] created a manually annotated dataset from *Twitter* using a lexicon of suicidal phrases and a lexicon along with the social engagement data associated with real-time and historical tweets. The resulting dataset consists of 34,306 tweets with two labels, *Suicidal* and *Non-Suicidal*. Gaur et al.[58] collected and annotated a 5-label Suicide Risk Severity Assessment dataset from Reddit, which includes *Suicidal Ideation (ID)*, *Suicidal Behavior (BR)*, *Actual Attempt (AT)*, *Suicide Indicator (IN)* and *Supportive (SU)* categories. This dataset is extracted from *SucideWatch*[1] subreddit and has been annotated by four practicing clinical psychiatrists, ensuring the accuracy and reliability of the annotations. The dataset comprises 500 posts, carefully selected to represent a diverse range of content related to suicidal ideation.

Another widely referenced dataset in the field of suicidal ideation detection is the University of Maryland Reddit Suicidality Dataset(UMD) [59],[60]. The UMD dataset is a collection of Reddit posts and comments created by individuals who expressed suicidal thoughts or behaviors. The dataset contains over 100,000 posts and comments collected

[1] SuicideWatch subreddit

from various subreddits, including those related to mental health and suicide prevention, such as *Depression*[2] and *SuicideWatch* subreddits. The data was collected over several years and includes the content of the posts and comments, as well as the location and timing of the posts. Researchers have extensively used the UMD dataset to develop and test NLP algorithms and machine learning models to identify and analyze patterns in online communication related to suicide risk [61]. Ji et l.[62] proposed a method for improving text representation through the incorporation of sentiment scores based on lexicon analysis and latent topics. Additionally, they introduce the use of relation networks for the detection of suicidal ideation and mental disorders, leveraging relevant risk indicators. Ji et al. [63] utilized two pre-trained masked language models, MentalBERT and MentalRoBERTa, specifically designed to support machine learning in the mental healthcare research field. The authors assess these domain-specific models along with various pre-trained language models on multiple mental disorder detection benchmarks. The results show that utilizing language representations pre-trained in the mental health domain enhances the performance of mental health detection tasks, highlighting the potential benefits of these models for the mental healthcare research community.

Synthetic Dataset Generation:

To overcome the limitations of real-world data availability, NLP researchers have explored the use of synthetic datasets for several applications. For example, He et al.[64] utilized language models to generate synthetic unlabeled text. They introduced the Generate, Annotate, and Learn framework that leverages synthetic text for knowledge distillation, self-training, and Few-shot learning purposes. To generate the data, they fine-tune pre-trained language models on relevant datasets with limited examples. The synthetic text

[2] Depression subreddit

is then annotated with soft pseudo labels using the best available classifier for knowledge distillation and self-training. This paper achieves state-of-the-art results for knowledge distillation with 6-layer transformers on the GLUE leaderboard[65].

Bonifacio et al.[66] presented an effective approach to leveraging LLM in retrieval tasks, resulting in significant improvements across various datasets. Instead of directly utilizing LLMs during retrieval, they harness the LLMs' capabilities to generate labeled data using a Few-shot learning approach. Subsequently, they fine-tune smaller retrieval models on this synthetic dataset and employ them to re-rank the search results obtained from a primary retrieval system. They provide a novel method to adapt LLMs for Information Retrieval tasks previously deemed infeasible due to their demanding computational requirements. By shifting the computational burden from the retrieval stage to the generation of synthetic data for training, they make it feasible to exploit the power of LLMs without compromising efficiency. In an unsupervised setting, their approach significantly outperforms recently proposed methods, highlighting its retrieval performance and scalability superiority.

In this work, we assess the topic coverage in existing suicide detection datasets and generate an augmentation set for this task using generative models.

2.5 Interpretability of Models

Explainability methods in machine learning can generally be categorized into local or global explanations. Local explanations aim to explain why a model makes a specific decision, whereas global explanations are meant to explain a model's overall behavior across a broader range of data [67].

Local explanations are threefold. Feature-based explanations attribute model decisions

to important input features. They measure the importance of an input feature for the prediction at the local level [68, 69, 70]. Sample-based explanations attribute model decisions to previously observed samples. These methods analyze how similar instances in the dataset have been classified by the model to explain the decision-making process[71, 72, 73, 74]. Another method is counterfactual-based explanations, which involve generating counterfactual examples, hypothetical instances that are similar to the input data but with slight changes that would result in a different prediction or decision from the model. Counterfactual-based explanations can help humans understand why a model made a particular prediction or decision, identify biases or weaknesses in the model, and provide insights into how to improve model performance [75, 76, 77].

Global explanations can be created by aggregating local explanations or by measuring the sensitivity of models to concepts and features at a global level. Concept-based explanations are a global approach to explainability introduced in computer vision, to explain image classification models [78, 79]. Kim et al. [80] introduced Concept Activation Vectors, which provide an interpretation of a neural net's internal state in terms of human-friendly concepts. They presented TCAV technique, which uses directional derivatives to quantify the degree to which a user-defined concept is important to classification results.

In NLP, concept-based explanations were used to measure the sensitivity of an abusive language classifier to the emerging concept of COVID-related anti-Asian hate speech [81]. They demonstrated that while general classifiers for abusive language can effectively identify emerging explicit abusive expressions, they struggle to detect new forms of subtler, implicit abuse. Moreover, they modified the TCAV approach to quantify the level of explicitness for an unlabeled instance, serving as a measure to inform data augmentation during the enhancement of a general abusive language classifier to include a new kind of abuse. They reported that their method achieved greater accuracy with fewer training instances compared

to the commonly used confidence-based augmentation methods. Nejadgholi et al. [82] utilize a concept-based explanation framework to compute the model's sensitivity to sentiment, a feature previously employed as a salient feature for detecting toxic language. Their findings show that the sentiment information is outweighed by the influence of identity terms when used as input features in the classification task.

More recently, a growing body of work has shifted from post-hoc explainability toward mechanistic interpretability, which aims to understand neural networks by directly analyzing the internal representations and computations that give rise to model behavior. Rather than explaining predictions in terms of input features or example-level similarities, mechanistic interpretability treats the model as a system whose internal states can be studied to reveal how information is represented, combined, and transformed across layers [83, 84, 85]. This perspective seeks explanations grounded in the model's internal geometry and activations, offering a more faithful account of how decisions are formed than external attribution or sensitivity-based methods.

A central hypothesis motivating this line of work is that neural representations are approximately linear in sufficiently expressive feature spaces, but that multiple concepts may be superimposed within shared activation dimensions [86]. Under this view, individual neurons or hidden dimensions are not expected to correspond cleanly to human-interpretable concepts. Instead, meaningful features may be encoded as directions in activation space that can be recovered through appropriate transformations. Recent work has shown that sparse dictionary learning applied to residual-stream activations can uncover such feature directions, yielding representations that are both more interpretable and more aligned with abstract semantic concepts [87, 88, 89]. These approaches provide a principled framework for analyzing how high-level concepts are encoded internally, moving beyond sensitivity-based explanations toward a structural understanding of learned representations.

Chapter 3

Datasets

In this section, we review the specifications of the existing datasets collected from social media that were assessed in this work.

3.1 The University of Maryland Reddit Suicidality Dataset(UMD)

The University of Maryland Reddit Suicidality Dataset(UMD) [59, 60] is collected from the Reddit platform. Reddit is an online website and forum for anonymous discussion on a wide variety of topics. It is made up of millions of collective forums or groups called subreddits, including the *Depression*[1] and *SuicideWatch*[2] subreddits. This dataset is a collection of Reddit posts and comments created by individuals who expressed suicidal thoughts or behaviors. The dataset contains over 100,000 posts and comments collected

[1] Depression subreddit
[2] SuicideWatch subreddit

from various subreddits, including those related to mental health and suicide prevention, such as "*r/SuicideWatch*". The data was collected over a period of several years and includes the content of the post and comments as well as the location and timing of the posts.

This dataset contains annotations at the user level, utilizing a four-point scale to indicate the severity of the suicide risk: (a) *No risk*, (b) *Low risk*, (c) *Moderate risk*, and (d) *High risk*. According to Zirikly et al. [59], the dataset is divided into three subsets, each containing annotations for a distinct task.

Task A focuses on risk assessment and involves simulating a scenario in which an individual is suspected to require assistance based on online activity, such as posting to a relevant forum or discussion (e.g., r/SuicideWatch). The objective of the task is to evaluate the individual's risk level based on their online activity. This task requires minimal data, with each user typically having posted no more than a few times on SuicideWatch.

Task B involves the same risk assessment problem as Task A, but with the added utilization of user posts from sources other than SuicideWatch.

Task C pertains to screening and is designed to simulate a situation where an individual has agreed to social media monitoring, such as a new mother at risk of postpartum depression, a veteran returning from deployment, or a patient recommended by their therapist.

In this study, we utilized the subset designated for Task A. This task necessitates minimal data, with users generally contributing only a limited number of posts on SuicideWatch. Among the 993 labeled users, 496 have made at least one post on the SuicideWatch subreddit. The remaining 497 users serve as control subjects. Since the provided labels are user-level labels, we aggregated each user's posts into a single data point through the concatenation of all the posts made by a particular user. The dataset is divided into 80% training and

20% testing subsets. The GPT-3.5 model evaluation, presented in Section 4, was conducted solely on the testing subset, comprising 172 instances with proportional representation for each label. Table 3.1 presents the class sizes of the UMD dataset in multi-class setting.

Table 3.1: The description of the multi-class UMD Dataset

	No Risk	Low Risk	Moderate Risk	High Risk
UMD Dataset	26.73 %	15.27 %	30.69 %	27.28 %
# of Users	196	112	225	200
Training subset	27.45 %	16.39 %	31.90 %	24.24 %
# of Users	154	92	179	136
Testing subset	24.41 %	11.62 %	26.74 %	37.20 %
# of Users	42	20	46	64

Furthermore, to employ binary classification, we binarize the UMD Dataset. Based on the definition of each class, "*No Risk*" and "*Low Risk*" classes are considered as Non-Suicidal and "*Moderate Risk*" and"*High Risk*" as Suicidal. Table 3.2 presents the description of binarized UMD dataset.

Table 3.2: The description of the training and testing subset of binarized UMD Dataset

Binary Dataset	Non Suicidal	Suicidal
Training Subset	43.84%	56.14%
Number of Users	246	315
Testing Subset	36.3	63.94
Number of Users	62	110

3.2 Knowledge-Aware Assessment Dataset

Gaur et al. [58] developed an annotated gold standard dataset of 493 Reddit users, out of 2181 potentially suicidal users, using their content from mental health-related subreddits

Table 3.3: Detailed Description of the Knowledge Aware Suicidality Assessment Dataset

	Suicidal Ideation (ID)	Suicidal Behavior (BR)	Actual Attempt (AT)	Suicide Indicator (IN)	Supportive (SU)
Multi Class	34 %	15 %	9 %	20%	22 %
# of posts	170	76	45	98	104
Binarized	291			202	
# of Posts	59%			41%	

within the time frame of 2005 to 2016. The Dataset consists of 5 different categories of suicidality, including Suicidal Ideation (ID), Suicidal Behavior (BR), Actual Attempt (AT), Suicide Indicator (IN) and Supportive (SU). Suicidal Ideation (ID) refers to thoughts of suicide, which may involve concerns related to suicide risk factors such as job loss or the end of a significant relationship. Suicidal Behavior (BR) is defined as actions that carry a higher risk, such as self-harm (either current or historical), active planning to commit suicide, or a history of institutionalization for mental health reasons. An Actual Attempt (AT) encompasses any deliberate action that could potentially lead to intentional death. This includes but is not limited to instances where an individual sought help, reconsidered their decision, or publicly expressed thoughts of suicide. The Suicide Indicator (IN) category serves as a classification method to distinguish individuals who use at-risk language from those who are actively experiencing general or acute symptoms. Often, users converse in supportive conversations and share their personal histories while using language from the clinical lexicon. The Supportive (SU) category pertains to individuals engaging in discussions without expressing any history of being at-risk, either in the past or at present.

To binarize this dataset based on the definition of each class, "*Suicide Indicator*" and "*Supportive*" classes are considered as Non-Suicidal and "*Suicidal Ideation*","*Suicidal Behavior*" and "*Actual Attempt*" as Suicidal. The detailed description of the dataset is presented in the Table 3.3.

3.3 2021 Reddit Dataset

In addition to the existing datasets, we collected a new dataset of suicidal social media posts from the Reddit platform using the Reddit API[3]. Specifically, we focused on the "SuicideWatch" subreddit to gather posts related to suicide to analyze the posts and topics discussed within this subreddit.

The initial data collection was conducted on September 11, 2021, with the goal of gathering 2,500 posts published between May 1, 2021, and September 1, 2021. Subsequently, we conducted extensive text preprocessing, which included removing links, eliminating duplicates, handling special characters, filtering out irrelevant and non-informative posts, performing lemmatization, and conducting spellchecking.

After these preprocessing steps, our dataset consisted of 2,052 unlabelled posts from the 'SuicideWatch' subreddit, which we utilized for the purpose of topic modeling in this study.

[3] Reddit API Documentation

Chapter 4

An Evaluation of A GPT Model as A Classifier

In this chapter, we investigate the strengths and limitations of GPT-3.5 model, an advanced language model created by OpenAI [47], as a tool for suicidal ideation assessment from social media posts. The GPT-3.5 model API provides access to a powerful natural language processing tool that can generate human-like text, answer questions, and perform a variety of other language-related tasks. With GPT-3.5 model, developers can build conversational interfaces, Chatbots, and virtual assistants to interact with users and provide informative responses. However, some studies have highlighted the potential risks and ethical concerns associated with the use of GPT-3.5 model and other language models in sensitive domains, such as mental health and suicide prevention [90]. Therefore, it is crucial to carefully evaluate the use of GPT-3.5 model in such settings to better understand its potential and limitations.

To assess GPT-3.5 model's reliability in the suicide prevention task, we will take

two steps: first, we will evaluate GPT-3.5 model's ability to identify the severity level of suicidality, and second, we will compare GPT-3.5 model's performance to fine-tuned transformer-based models that have been trained on human-annotated datasets.

4.1 Suicidality Assessment Using GPT-3.5 model

GPT-3.5 is a state-of-the-art artificial model that is based on the transformer architecture, which has been shown to outperform traditional Recurrent Neural Network (RNN) models in various language tasks, including machine translation, text classification, and dialogue generation.

The original GPT model was trained on a massive corpus of text data, including books, articles, and web pages, using an unsupervised learning approach. The model's performance on a range of language tasks has since been surpassed by newer models, including GPT-2 [48] and GPT-3 [49], which have larger training datasets and more sophisticated architectures. The language model we utilized by ChatGPT is *gpt-3.5-turbo*[1]. Chat models accept a sequence of messages as an input and produce a message generated by the model as an output. Although the chat format is primarily intended for conversations spanning multiple turns, it is equally useful for single-turn tasks that do not involve any conversations. We used the *OpenAI Python library*[2] to access the *ChatCompletion* functionality of the *gpt-3.5-turbo* model through its API. Understanding the various components of the ChatGPT API is essential for maximizing its utility and effectiveness in diverse applications. The key aspects of the ChatGPT API, including the input message, temperature parameter, and inclusiveness rate play crucial roles in shaping the behavior and output of the model.

[1] https://platform.openai.com/docs/models/gpt-3-5
[2] https://github.com/openai/openai-python

Each component offers unique functionalities and capabilities that contribute to the overall performance and user experience of the ChatGPT API.

Input Message: The primary input for the system is the "message" parameter, which must be an array consisting of message objects. This object includes a "role" (either "system", "user", or "assistant") and a "content" (the message content). A conversation can consist of a single message or can extend over multiple pages. We provide a single message to the system which describes the definitions of suicide severity assessment.

In the input message, we use prompt engineering. We initiated the prompt construction process with a simple initial prompt and iteratively refined it through multiple rounds of trial and error. This iterative approach allowed us to gradually evolve the prompt, making necessary adjustments based on the observed outcomes and performance of the model. We drew inspiration from a short course on ChatGPT Prompt Engineering[3] offered by *DeepLearning.AI*. The complete implementation including Zero-shot Learning, Few-shot Learning and the fine-tuned classifiers is available on GitHub[4].

Temperature Parameter: The Temperature value in GLLMs s is a parameter that controls the randomness and creativity of the model's responses. To produce a response to a given input message, the model generates a probability distribution over all possible next words or tokens in the response. The temperature parameter affects the probability distribution over the possible tokens at each step of the generation process.

A high temperature value (close to 1) will result in more diverse and unpredictable responses, as the model samples from less likely tokens in the distribution. This can result in more creative and surprising responses but may also increase the likelihood of generating

[3] ChatGPT Prompt Engineering for Developers
[4] GitHub

nonsensical or irrelevant text. On the other hand, a low temperature value (e.g. 0.1) will result in more conservative and predictable responses, as the model chooses the most likely tokens in the distribution. This can result in more coherent and on-topic responses but may be more repetitive or less attractive. The temperature parameter in GPT-3.5 model allows users to control the balance between creativity and coherence in the model's responses based on their specific needs and preferences.

Inconclusiveness Rate: We define an additional metric, the *Inconclusiveness rate* for further evaluation of GPT-3.5 model in this task. This parameter refers to the proportion of test cases that do not yield a definitive or conclusive result. In other words, it is the rate at which the evidence or information is inconclusive to support a clear decision. To calculate the inconclusiveness rate, we count the instances where GPT-3.5 model reports inconclusive suicidality risk assessments. Then, we divide this number by the total instances in the test dataset and express the result as a percentage. An example of an inconclusive response generated by GPT-3.5 model is, *"As an AI, I cannot provide an assessment of the suicidal risk level for this instance"*.

In this project, we evaluate GPT-3.5 model's ability for Zero-shot and Few-shot Learning to predict the severity of a text's suicidality. Zero-shot Learning and Few-shot Learning are two techniques in NLP that are used to develop models that can perform tasks without explicit training data [91].

4.1.1 Zero-shot

Zero-shot learning refers to the model's ability to generate coherent and contextually relevant responses for tasks it has not been explicitly trained on. Unlike traditional machine learning approaches where models are trained on specific datasets for predefined tasks,

Zero-shot learning allows GPT-3.5 model to generalize its knowledge and apply it to new and unseen tasks. Zero-shot learning enhances the versatility and adaptability of GPT-3.5 model, allowing it to handle a wide range of tasks and domains without the need for extensive task-specific training.

In Zero-shot learning, the model is provided with a prompt or description of the task it needs to perform, along with relevant context or constraints. Even though the model has not been trained on examples of this specific task, it leverages its pre-existing knowledge and understanding of language to generate responses that align with the given task description. The following example shows the input message for Zero-shot Learning. For this project, we utilized the UMD dataset to assess the suicidality of its content. In order to be consistent with the instructions given to human annotators of the UMD dataset, we obtained the definitions of each risk category from Shing et al.[60]. These definitions have been adapted from Corbitt et al. [92] and specify a four-way categorization scheme of risk.

```
res=openai.ChatCompletion.create(model="gpt-3.5-turbo",
messages=[{"role": "system", "content":
"""specify the level of the suicidality of the text based on the
    following criteria:

0= I do not see evidence that this person is at risk for suicide.
1= There may be some factors here that could suggest risk, but I do
    not really think this person is at much of a risk of suicide.
2= I see indications that there could be a genuine risk of this person
     making a suicide attempt.
3= I believe this person is at high risk of attempting suicide in the
    near future. Only say number"""},
```

{"role": "user","content": """*DESIRED INPUT TEXT*"""}],
temperature=0.1)

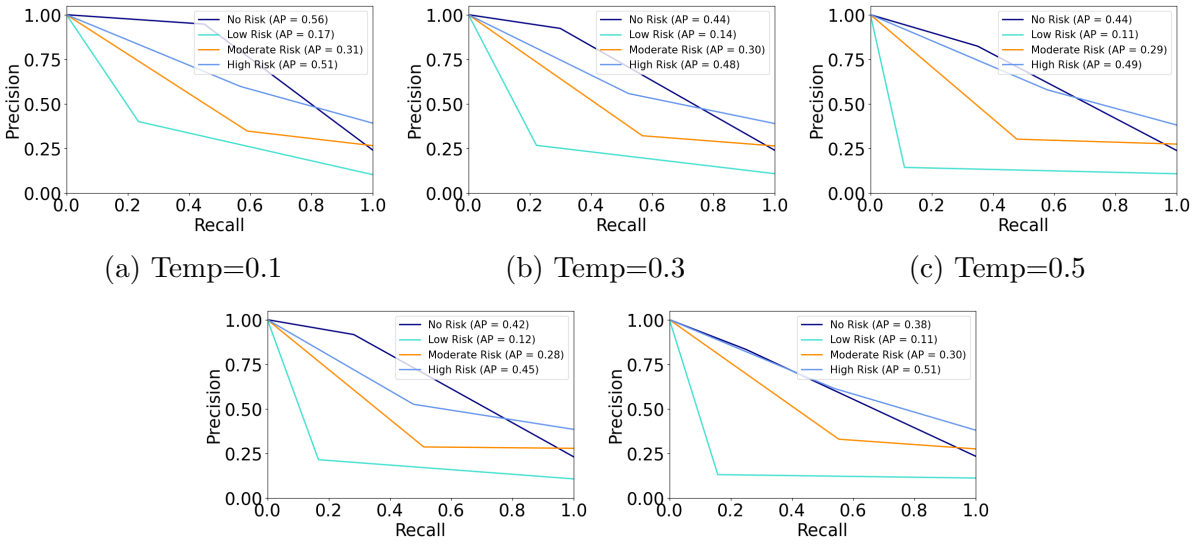
For evaluation, we report four widely-used metrics in this task, accuracy, precision, recall, and F1-score to provide a comprehensive evaluation of the performance of the classification models [93]. For GPT-3.5 model, we also report the *Inconclusiveness rate*. We use five different temperature values to evaluate the impact of temperature on the generated response and report the inconclusiveness rate of GPT-3.5 model at each temperature. The rest of the metrics are used to evaluate the performance of GPT-3.5 model for the instances in which GPT-3.5 model was able to generate a conclusive answer. Table 4.1 presents the performance of the GPT-3.5 model for five different temperature values. To report the results, we rely mostly on the F1-score metric, which provides a balanced measure of a model's precision and recall. Moreover, utilizing the F1-score metric ensures a comprehensive evaluation of the model's performance, considering both its ability to correctly identify positive instances and its ability to avoid false positives in an imbalanced dataset.

Table 4.1: Performance and inconclusiveness rate of GPT-3.5 model for Zero-shot Learning in five different temperature values. The row with the highest F1-score is highlighted.

Temperature	Accuracy	Precision	Recall	F1-Score	Inconclusiveness Rate
0.1	**0.88**	**0.57**	**1**	**0.73**	**2.91 %**
0.3	**0.67**	**0.33**	**1**	**0.50**	**2.32 %**
0.5	**0.67**	**0.22**	**0.67**	**0.33**	**1.71 %**
0.7	**0.64**	**0.27**	**1**	**0.43**	**1.16 %**
1	**0.54**	**0.21**	**1**	**0.35**	**0 %**

As presented in Table 4.1, a higher temperature will result in a more creative and random output but with a greater risk of generating nonsensical or irrelevant text. Conversely, a lower temperature will result in more indecisiveness, but with a lower risk of errors, i.e., the highest F1-score is achieved with a temperature of 0.1. We observed that GPT-3.5 model's

Figure 4.1: Precision-Recall graph of the GPT-3.5 model at different temperature values in Zero-shot setting



(a) Temp=0.1 (b) Temp=0.3 (c) Temp=0.5

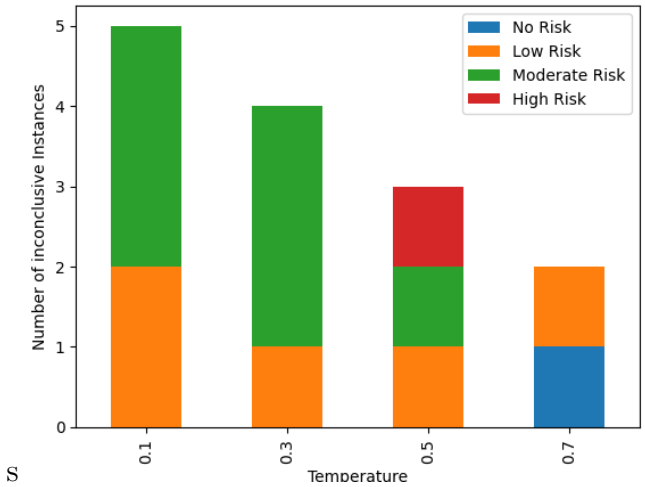
(d) Temp=0.7 (e) Temp=1

inconclusiveness rate (inability to assess the level of suicidality of instances) is 2.91% for a temperature of 0.1, which is the highest rate for all temperature values. As the temperature value increases, the inconclusiveness rate and F1-score decrease.

For further evaluation, we present the Precision-Recall (PR) curve in Figure 4.1 at each temperature. The PR graph displays the trade-off between precision and recall for different thresholds used to classify instances. Moreover, Figure 4.1 shows the impact of increasing temperature values on predicting the suicidality of the text in each class. As the temperature increases, the area under the PR graph declines. In other words, the graph shows lower values for both precision and recall measures. Furthermore, this figure reports that the average Precision of GPT-3.5 model in predicting the *No Risk* and *High Risk* classes is higher, compared to the two middle classes, *Low Risk* and *Moderate Risk*.

The bar chart depicted in Figure 4.2 illustrates which classes are more challenging for

Figure 4.2: Number of instances for which GPT-3.5 model refrains from making a decision, at different temperature values and for classes *No Risk*, *Low Risk*, *Moderate Risk*, and *High Risk*



S

GPT-3.5 model for suicidality assessment. Figure 4.2 shows that at the temperature of 0.1, 3 out of 5 inconclusive instances belong to *Moderate risk* and 2 out of 5 instances belong to *Low Risk* categories.

4.1.2 Few-shot

To implement Few-shot Learning with GPT-3.5 model, we use prompt engineering. The prompt consists of a few examples from the training dataset, and the model is trained to assess the suicidality level of the text based on the given criteria. Similar to Zero-shot Learning, we begin by providing the definitions of each risk category in the prompt, followed by eight training examples and their corresponding labels, with each example and label being separated by a comma and placed in individual paragraphs. The prompt concludes with a request for GPT-3.5 model to provide an assessment based on the given criteria. In this project, the prompt consists of two examples for each category (eight in total)

drawn from the training dataset for the input message and followed by the same assessment question as Zero-shot. The following example shows the input prompt for Few-shot Learning using GPT 3.5-turbo.

```
Criteria="""
    0= I do not see evidence that this person is at risk for suicide.
    1= There may be some factors here that could suggest risk, but I
       do not really think this person is at much of a risk of suicide
       .
    2= I see indications that there could be a genuine risk of this
       person making a suicide attempt.
    3= I believe this person is at high risk of attempting suicide in
       the near future. Only say number.
    """

Examples="""
    Risk=0,Example1:
    Risk=0,Example2:
    Risk=1,Example1:
    Risk=1,Example2:
    Risk=2,Example1:
    Risk=2,Example2:
    Risk=3,Example1:
    Risk=3,Example2:
    """
```

```
prompt = f"""specify the level of the suicidality of the text based on
    the Risk level criteria: ```{Criteria}``` and training examples
    :```{Example}```"""
```

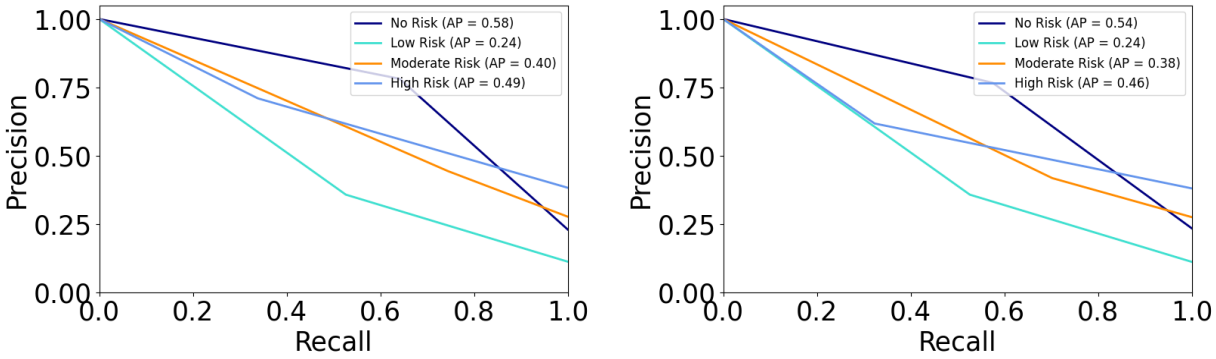
Table 4.2 presents the results of GPT-3.5 model in Few-shot settings at different temperature values. GPT-3.5 model achieves the highest F1-score at the temperature of 0.1. Furthermore, we observed that the inconclusiveness rate of GPT-3.5 model in Few-shot Learning was significantly lower compared to Zero-shot Learning. Additionally, the inconclusiveness rate remained almost constant at different temperature values, indicating that GPT-3.5 model is more confident in generating responses when it is provided with a few examples.

Table 4.2: Performance of GPT-3.5 model for Few-shot Learning in five different temperature values. The row with the highest F1-score is highlighted.

Temperature	Accuracy	Precision	Recall	F1-Score	Inconclusiveness Rate
0.1	0.81	0.67	0.77	0.71	0.58 %
0.3	0.81	0.67	0.77	0.71	0.58 %
0.5	0.76	0.57	0.67	0.65	0.58 %
0.7	0.75	0.56	0.77	0.62	0 %
1	0.75	0.56	0.77	0.62	0 %

Figure 4.3 presents the PR graph of GPT-3.5 model for two extreme temperature values. As presented in Table 4.2, the precision and recall values for temperature values 0.1 and 1 are not significantly different as it is reflected in the PR curve as well. However, the PR curve of two classes, *Moderate Risk* and *High Risk*, slightly improves by decreasing the temperature.

Figure 4.3: Precision-Recall graph of the GPT-3.5 model at two extreme temperature values (0.1 and 1) in a Few-shot settings, for classes *0=No Risk*, *1=Low Risk*, *2=Moderate Risk*, *3= High Risk*.



(a) Temperature=0.1

(b) Temperature=1

4.2 Fine-tuned Transformer based Models

In this study, two transformer-based classifiers were built to determine the level of suicidality in the text. These classifiers were fine-tuned on the UMD dataset to identify language patterns and features that indicate suicidal ideation. We utilize ALBERT [5] and DistilBERT[6] language models to build the classifiers. For implementation, we employed the Huggingface library [94], an open-source library and data science platform that provides tools to build, train, and deploy machine learning models.

ALBERT and DistilBERT are two pre-trained language models from the BERT family of LMs. The BERT model was initially proposed by Delvin et al.[95] as a bidirectional language model pre-trained on a large corpus comprising the Toronto Book Corpus and Wikipedia. The model is named bidirectional because it can simultaneously gather the context of a word from either direction. Unlike the generative models such as GPT-3.5

[5] AlBERT
[6] DistilBERT

Turbo, FlanT5 or Llama, which include a decoder structure, the BERT family of language models are encoder models and can be fine-tuned for specific tasks such as classification tasks.

The ALBERT model was proposed by Lan et al. [96] to reduce memory consumption and increase the training speed compared to BERT. In other words, ALBERT is a more lightweight version of BERT that maintains its high level of accuracy, making it a powerful tool for various NLP applications. The DistilBERT model was proposed by Sanh et al. [97]. The authors reported that it has 40% fewer parameters than BERT and runs 60% faster while preserving over 95% of BERT's performance as measured on the GLUE language understanding benchmark. Both models are designed as lightweight alternatives to BERT, with ALBERT emphasizing parameter efficiency and DistilBERT focusing on knowledge transfer through distillation. Overall, ALBERT, with a smaller number of parameters, shows more efficient performance compared to DistilBERT.

We used the Trainer[7] class from Huggingface transformers[8] for feature-complete training in PyTorch. The hyperparameters were selected based on the default values commonly used in similar studies. The final hyperparameters used in our experiments were Learning Rate= $2e^{-5}$, Batch Size = 4, Dropout Rate = 0.1, and Maximum Sequence Length = 512. The performances of these models on the UMD dataset are presented in Table 4.3 and are compared with the results obtained by the GPT3.5-turbo model.

As presented in Table 4.3, while GPT-3.5 model's performance is comparable to a fine-tuned DistillBERT, it falls considerably short (by 13% for F1-score) compared to a fine-tuned ALBERT model. The results of this experiment, suggest that the ALBERT model reaches promising results with an F1-score of 0.869, outperforming both the DistilBERT

[7] Trainer
[8] Huggingface Transformers

Table 4.3: Comparison of the two transformer-based models with GPT-3.5 Turbo. Fine-tuned ALBERT is highlighted for achieving the highest F-score.

Model	Accuracy	Precision	Recall	F1-Score
AlBERT	0.865	0.861	0.865	0.869
DistilBERT	0.77	0.804	0.771	0.745
Zero-shot GPT-3.5 (temp=0.1)	0.88	0.57	1	0.73
Few-shot GPT-3.5 (temp=0.1)	0.81	0.67	0.77	0.71

and GPT-3.5 models, with F1-scores of 0.745 and 0.73, respectively. While the ALBERT model achieved the highest score among the three models, it should be noted that it is trained on the UMD dataset for the suicidal assessment task specifically. On the other hand, GPT-3.5 model is trained on a large corpus of text data using a self-supervised learning approach for multiple tasks.

4.3 Discussion

This study focuses on the evaluation of the accuracy and quality of responses generated by GPT-3.5 model for the assessment of suicidal ideation levels. The performance of GPT-3.5 model was assessed in Zero-shot and Few-shot Learning scenarios. One of the key advantages of Zero-shot Learning is its ability to generalize to new classes, which is important in scenarios where the number of possible classes is large and difficult to define in advance. GPT-3.5 model achieved an F1-score of 0.73 in Zero-shot Learning, and an F1-score of 0.71 in Few-shot Learning on our test set (temperature=0.1). In Zero-shot Learning, the model is able to leverage its existing knowledge to make predictions for new tasks. This approach can be particularly effective when the model needs to generalize to

a wide range of possible new tasks. On the other hand, Few-shot Learning requires the model to learn from a limited amount of training data for each new task. This approach can be more challenging, as the model has to generalize from a small set of examples and may struggle to identify patterns or relationships that are important for the new task.

In this study, we carried out an experiment to examine the impact of the temperature hyperparameter on the performance of GPT-3.5 model. In Zero-shot Learning, our findings indicate that there is a negative correlation between the F1-Score and the temperature hyperparameter. In other words, as the temperature increases, the model's performance tends to decrease. These results suggest that careful optimization of hyperparameters, such as temperature, is crucial for achieving optimal performance of GPT-3.5 model. In Few-shot Learning, there is still a negative correlation between the F1-Score and the temperature hyperparameter. However, the change in the F1-Score value is subtle, indicating that the impact of temperature tuning on model performance may not be significant.

Another assessment in this study involves examining how the temperature parameter affects GPT-3.5 model's inconclusiveness rate. In sensitive tasks such as suicide risk assessment, it is crucial to have a highly accurate model that can provide reliable predictions. In some cases, it may be preferable for the model to provide an *"I do not know"* response rather than providing unreliable predictions about suicidality. Careless responses from a suicidal assessment model can have serious consequences, including false positives or false negatives, which can harm individuals at risk. Table 4.1 shows that the inconclusiveness rate in temperature 0.1 of Zero-shot Learning is 2.91%, and the F1-Score is 0.73. By increasing the temperature, we have fewer inconclusive instances and yet a lower F1-score over the rest of the responses. Table 4.2 for Few-shot Learning shows that the *inconclusiveness rate* becomes almost constant and smaller among different temperature values because the model has learned to generalize based on the limited number of examples provided during

training and the model is not able to generate as much variation in response because it may overfit to the training examples. As a result, the model may be less prone to generating random or unexpected responses.

Figure 4.2 indicates that the inconclusive instances mostly belong to two middle classes *Low Risk* and *Moderate Risk*. These two classes are subjective due to the vague boundaries of definitions. For example, the Zero-shot Learning model was not able to provide an assessment for the following instance: "*I have ups and downs, I've had them for a long time and I don't know why, since December I've been going to therapy, I've been getting meds too and at first they helped suppress the storm of thoughts that won't let me sleep eat and think, I keep finding myself trying to sleep*". On the other hand, the Few-shot Learning predicts *High Risk* suicidality level for this instance, and the human experts annotated this instance as *Low Risk*. This example clarifies that generating an "I do not know" answer here can be preferable to a wrong assessment.

4.4 Conclusion

Data collection and annotation are essential but expensive processes in supervised machine learning. Obtaining high-quality labels can be costly and time-consuming in sensitive tasks such as suicide detection. Based on our results, one possible approach to reduce the cost and increase the quality of data annotation is to use GPT-3.5 model in an expert-in-the-loop setting. For instance, when no labelled training data is available, GPT-3.5 model can be used to flag samples of suicide notes from large sets of collected data. Human annotators can then review the model's output and provide corrections or feedback to refine the output. However, our results indicate that with a small training set, fine-tuned models can be trained to outperform GPT-3.5 model. Therefore, although ChatPT is a helpful tool in the

early stages of data annotation, it is not reliable enough to replace fine-tuned specialized models.

Chapter 5

Assessment of Topics in Social Media Datasets

In an era dominated by social media conversations, it is pivotal to comprehend how suicide is discussed online. In this research, we draw upon established psychological literature to identify core topics linked to suicide, such as mental health challenges, relationship conflicts, and financial distress. Then, we undertake a comprehensive analysis of suicide-related data sourced from social media, employing a guided topic modelling technique to extract the most frequently discussed subjects based on the risk factors from the literature. Our study seeks to uncover the extent to which social media mirrors or diverges from conventional psychological understandings.

5.1 Social and Psychological Knowledge Extraction

The psychology literature offers valuable insights into the diverse processes and factors that contribute to suicide risk. Psychological theories and frameworks such as the interpersonal theory of suicide [24], the cognitive model of suicidal behavior[25], and the social-ecological model[26] provide a theoretical foundation for understanding the complex interplay between individual vulnerabilities and environmental factors.

This research aims to untangle the origins, risk factors, and social factors intertwined with suicidal thoughts, with the ultimate goal of assessing the coverage of these topics in existing datasets. In this section, we outline the social factors that have been explored in the literature as pertinent aspects of suicidal ideation.

We conducted a scoping review to offer a current and thorough synthesis of psychology studies, aiming to identify topics, themes, and risk factors related to the sensitive and complex issue of suicidal ideation. Scoping reviews serve as a method to assess the scope of literature on a particular subject, providing insight into the available research and offering a broad overview of its focus [98]. Additionally, they highlight the types of evidence that guide practice in the field and examine how the research has been conducted [99]. Our protocol was developed using the scoping review methodological framework proposed by Arksey and O'Malley [100] and further refined by Peters et al. [101].

Research Question: The central research question guiding this review is: *"What are the most frequently reported risk factors associated with suicidal ideation in the psychology and mental health literature?"* By answering this question, the review seeks to provide a clearer understanding of the key factors that contribute to the emergence of suicidal thoughts, ultimately informing future research and interventions aimed at preventing suicide and supporting individuals at risk.

Search Strategy: We scoured prominent academic databases such as *PubMed*, *PsycArticles*, *ScienceDirect*, and *Google Scholar*, employing a systematic approach. Our search strategy involved using identical keywords across all databases: Suicidal Ideation" and Suicide Risk Factors."

Inclusion Criteria: To be eligible for inclusion, studies were required to be peer-reviewed review or systematic review publications that investigated risk factors associated with suicidal ideation. Eligible studies had to be published between January 2014 and August 2024, address all forms of suicidal behavior including ideation and attempts, and focus on the general population.

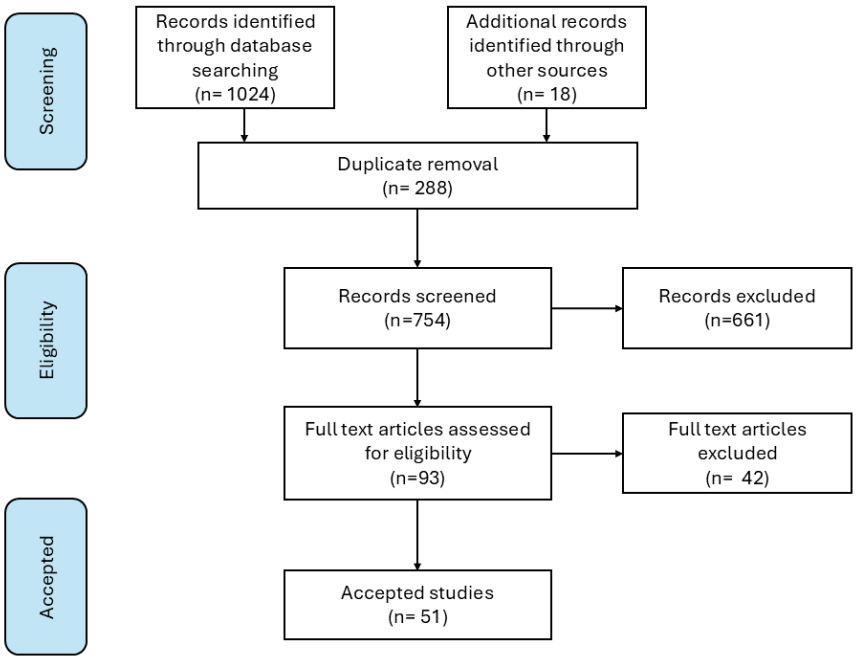
Exclusion Criteria: During the screening phase, studies were excluded based on title and abstract if they were not directly related to suicidal ideation or did not focus on identifying risk factors. Duplicate records were also removed at this stage.

During the full-text assessment phase, studies were excluded if they did not meet the inclusion criteria upon closer inspection. Common reasons for exclusion included: focus on highly specific subpopulations , lack of explicit discussion of suicidal ideation risk factors, non-review articles despite initial classification, and insufficient methodological detail.

Study Selection: All identified publications were initially screened for relevance based on abstract and title. Subsequently, the full text of selected publications was assessed for eligibility. Furthermore, the reference lists of eligible papers were used to identify additional studies. Figure 5.1 presents the details of our scoping review process.

This exploration allowed us to identify a wide array of relevant review papers, forming the foundation of our research. Subsequently, we reported the most common topics among all the selected research articles. Based on our analysis of the literature, the following social and psychological factors were consistently reported in relation to suicidal ideation in

Figure 5.1: Scoping Review Workflow for Extracting Suicide Risk Factors from Psychology Literature



psychology. These topics are not listed in a specific order of importance but represent the consistently reported themes in the literature reviewed:

Mental health disorders and Personality traits: "Depression" is a well-documented and significant risk factor for suicide. The persistent feelings of sadness and emotional pain that characterize depression can lead individuals to contemplate or attempt suicide as a means of escape from their suffering. It's vital to recognize the signs of depression, offer support, and connect individuals to mental health professionals and resources for effective treatment and intervention [102, 103, 104].

"Anxiety" disorders were also commonly associated with suicidal ideation. The chronic emotional distress and physical symptoms associated with severe anxiety can contribute to

the development of suicidal thoughts [105].

"Post-Traumatic Stress Disorder (PTSD)" has been strongly associated with an increased risk of suicidal ideation. Individuals who have experienced traumatic events may struggle with the emotional aftermath, including intrusive memories, hyper-arousal, and avoidance of reminders, which can lead to constant emotional distress and the sense of being overwhelmed. These constant feelings of traumatic memories can contribute to thoughts of suicide [106],[107].

"Bipolar disorder", formerly called manic depression, is a mental health condition that causes extreme mood swings. These include emotional highs, also known as mania, and lows, also known as depression [108]. Suicide attempts and completed suicide are significantly more common in patients with bipolar disorder when compared with the general population [108],[109].

"Schizophrenia" is a severe mental disorder that disrupts both cognitive and social functioning, often resulting in the onset of additional health conditions [110]. Schizophrenia is strongly linked to an increased risk of suicidal ideation. The distress caused by its symptoms, feelings of isolation, and perceived loss of control over one's mind can contribute to a deep sense of hopelessness. Additionally, the stigma and social withdrawal often associated with schizophrenia may exacerbate these feelings, further increasing the risk of suicide [111],[112].

"Borderline Personality Disorder" is a prevalent mental health condition linked to elevated suicide rates, significant functional impairment, frequent co-occurrence with other mental disorders and extensive treatment needs. Long-term outcome studies of patients with borderline personality disorder have documented a high rate of suicide completion [113][114].

The expression and experience of "Anger" have been reported as influential factors in suicidal ideation. Unresolved anger and intense emotional turmoil can drive individuals towards suicidal thoughts and actions. Anger, when left un-managed, can escalate distress and lead to impulsive decisions with dire consequences. Recognizing and managing anger is a crucial facet of suicide detection [102],[115].

"Perfectionism", marked by excessively high standards and self-criticism, has been identified as a psychological factor related to suicidal ideation. The relentless pursuit of unattainable standards can lead to feelings of inadequacy and despair, increasing the risk of suicidal ideation. Recognizing the need for balance and self-compassion is pivotal in addressing this risk factor [102].

Feelings of "Hopelessness", characterized by a pervasive sense of despair and an inability to envision a better future, can be a potent predictor of suicidal behavior. Those burdened by overwhelming hopelessness may see suicide as the only means of escape from their emotional suffering [102],[103],[116].

Substance abuse: Alcohol and drug misuse is a significant risk factor for suicidal ideation and behavior [117]. The connection between substance abuse and suicide is complex, as substances like drugs and alcohol can impair judgment and exacerbate underlying emotional distress. Individuals who struggle with addiction may turn to substances as a means of coping with psychological pain, and when combined with impaired decision-making, this can increase the likelihood of suicidal thoughts and actions [118], [119], [120].

Sociodemographic Status: Prolonged "Unemployment" can erode self-esteem, create financial difficulties, and contribute to feelings of hopelessness, increasing the risk of suicide. Government support and job assistance programs can help mitigate this risk [116].

"Financial hardships" can trigger intense stress, making individuals vulnerable to suicide.

Economic support and resources are instrumental in addressing this risk factor [105], [121].

"Education pressure" including exams and expectations, can lead to emotional turmoil and increased risk of suicidal ideation, particularly among students. Educational institutions must provide resources for coping with academic stress [122], [123].

"Sexual minority" individuals, such as those who identify as LGBTQ+, often face unique stressors related to their sexual orientation or gender identity [124]. Discrimination, prejudice, and stigma can lead to feelings of isolation, rejection, and psychological distress. Research consistently shows that sexual minority individuals are at a higher risk for suicidal ideation and attempts compared to their heterosexual counterparts [125], [126],[127], [128].

Abuse: "Bullying", including physical, verbal, or cyberbullying, has consistently emerged as a significant topic related to suicidal ideation. The experience of bullying can lead to social isolation, low self-esteem, and emotional distress, contributing to the development of suicidal thoughts [105], [129], [130], [131]

"Sexual abuse" is a recognized risk factor for suicide, which includes any sexual activity that occurs without consent, also referred to as sexual assault or sexual violence [132]. Dissociation is a common response to sexual abuse, and higher levels of dissociation have been associated with self-harm, suicidal thoughts, and suicide attempts [133], [134]

Family related issues: "Family-related stressors", such as conflict, dysfunctional dynamics, and poor communication, can significantly impact an individual's emotional well-being and contribute to suicidal thoughts. Strengthening family relationships and providing support to those affected is essential in mitigating this risk factor [116],[104].

Difficulties in "relationships", including conflicts, breakups, and marital dissatisfaction, have been reported as significant topics in relation to suicidal ideation. Relationship

problems can contribute to emotional distress and feelings of hopelessness, leading to thoughts of suicide [116].

"The death of a family members" or friends through death has been reported as a topic associated with suicidal ideation. Grief, feelings of loneliness, and a sense of being unable to cope with the loss can increase the risk of suicidal thoughts [135].

Racism: Studies have consistently highlighted the significant impact of racial discrimination on suicidal ideation. Experiencing racism and racial prejudice can increase the risk of suicidal thoughts [136],[137], [138].

Immigration: The process of immigration, with its cultural adjustments, isolation, and uncertainty, can intensify stress and emotional distress, increasing the risk of suicide among immigrants. Providing support and resources tailored to the immigrant experience is essential [139],[140].

Dementia: Dementia, particularly in its advanced stages, can lead to significant cognitive and emotional challenges. Individuals with dementia may experience confusion, memory loss, and personality changes, which can be distressing for both them and their caregivers. The experience of losing one's cognitive abilities and identity can contribute to feelings of hopelessness and despair, leading to thoughts of suicide [141],[142], [143].

Chronic Physical Problems: Living with chronic physical health conditions can be emotionally taxing, and individuals facing such challenges may be more susceptible to suicidal ideation. Chronic pain, disability, and limitations in physical functioning can erode one's quality of life and lead to a sense of hopelessness [144]. Coping with the constant demands of managing a chronic condition can also contribute to emotional distress [145], [146].

5.2 Topic Modeling

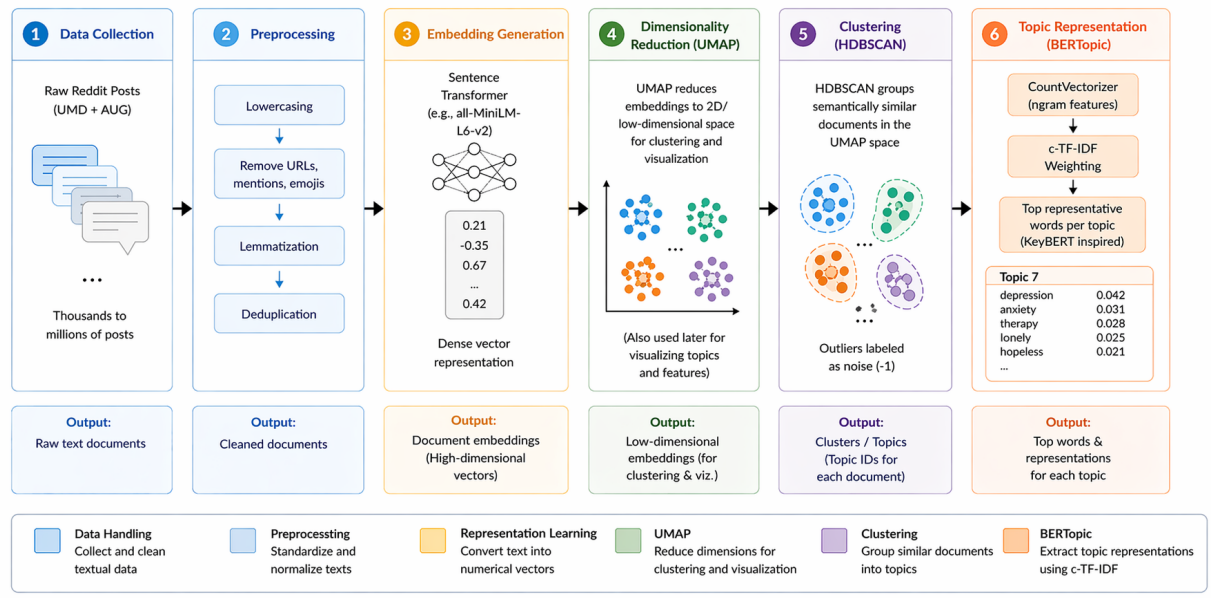
In this section, we present the methodology for conducting topic modeling using BERTopic [1], a state-of-the-art deep learning approach, which we will use to uncover the topic distribution in suicide detection datasets. BERTopic leverages transformer-based language models as embedding models, combined with clustering methods for topic extraction [147]. BERTopic integrates transformer-based techniques with Term Frequency-Inverse Document Frequency (TF-IDF) to create compact, interpretable clusters, preserving the most relevant terms in topic descriptions. This approach leverages deep learning and is mostly used with the sentence-transformers embedding model, which supports document embedding extraction in more than 50 languages [148]. Figure 5.2 presents the main stages of BERTopic's topic modeling framework including: document embeddings, document clustering, and topic representation.

Text Embedding: Text embedding refers to the vector representation of text within a multidimensional space where textual contents conveying similar meanings exhibit similar embeddings. In this project, we used "*SentenceTransformers*" Python framework for state-of-the-art sentence and text embeddings [149]. Although there are many models for text embeddings, we use the sentence-transformers "*all-MiniLM-L6-v2*" model as it has been shown as one of the best-performing models within the BERTopic framework[2].

Dimension Reduction and Clustering: In this step, similar documents or sentences are grouped based on their content. It is a method of organizing large amounts of textual information into meaningful categories or clusters, providing a high-level overview of the information contained within. Before clustering the embeddings, a dimensionality reduction

[1] https://maartengr.github.io/BERTopic/index.html
[2] Comparison of *SentenceTransformers* pretrained models

Figure 5.2: BERTopic framework diagram



is implemented, as embeddings are often high in dimensionality. In this work, we use the UMAP algorithm to reduce the dimensionality of the embeddings because it can capture both the local and global structures of high-dimensional data in lower-dimensional space [150]. It has also been proven that for short text clustering, UMAP demonstrates superior results [151][152]. The hyper-parameter space involved in UMAP is manually inspected, and based on the performance of the model and presented topics, the best parameters are selected. The number of neighbors, the number of components, and the minimum distance of each component are selected as 15, 5, and 0, respectively.

Following the dimensionality reduction of our input embeddings, we need to cluster them into groups of similar embeddings to extract our topics. The method used in this chapter is HDBSCAN, introduced by Campello et al. [153]. This method is based on the density clustering method that finds clusters of different shapes and identifies outliers where

possible. Similar to UMAP, the parameters of this model are manually inspected, and the proper parameters are chosen. There is no automated method for determining values for HDBSCAN. The parameters should be set manually based on domain knowledge and understanding of the dataset. Hence, we select 10 and 5 as the minimum cluster size and sample number in each cluster, respectively.

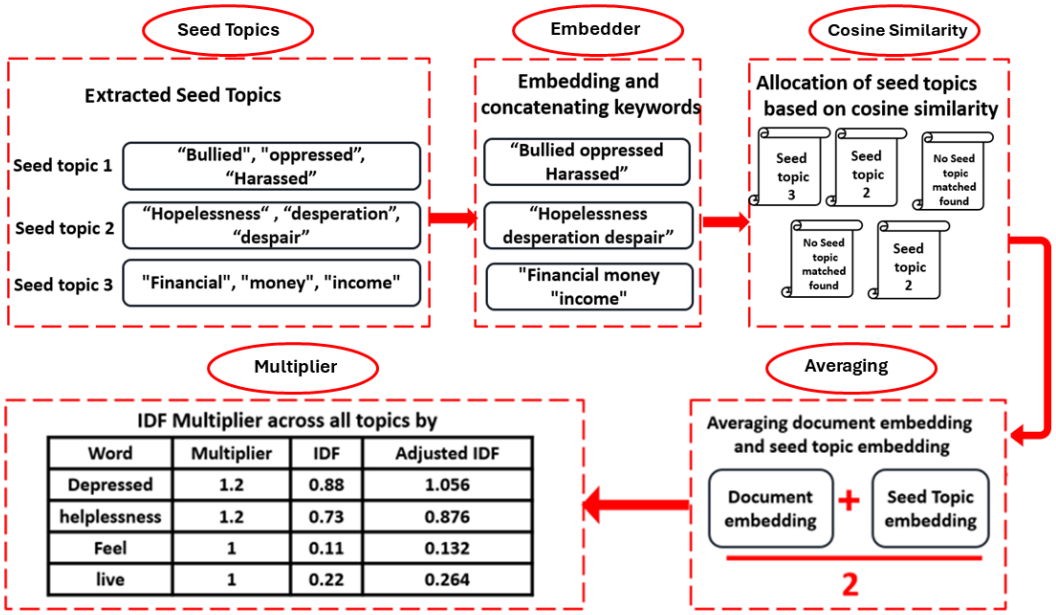
Topic Representation: In this step, Class-based TF-IDF scores are used to identify a set of keywords that represent the topic for a better interpretation of the topics' content. TF-IDF is a technique used to extract features from text documents, which is achieved by combining two components: Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency (TF) represents the simple word count within a document, treating each word count as a feature and it calculated by dividing the number of times the term occurs in the document by the total number of terms in the document. Inverse Document Frequency (IDF) gauges the informativeness of specific words by measuring their frequency within a document relative to their frequency across all other documents. The class-based TF-IDF is similar to TF-IDF but is adopted for multiple classes by joining all documents per class. Thus, each class is converted to a single document instead of a set of documents. After calculating the class-based TF-IDF scores for all the words in each topic, words with the highest scores are chosen as keywords associated with that topic.

5.3 Guided Topic Modeling

In topic modeling, dealing with data that employs specialized language can pose challenges. In such instances, the effectiveness of topic modeling techniques in capturing and expressing the semantic nuances of domain-specific terms may be hindered. To make sure that certain

domain-specific words are detected and used in topic representations, we employed *Guided Topic Modeling*. Guided topic modeling is an extension of traditional topic modeling that incorporates external information or guidance to influence the topic discovery process. In this project, we employed Guided BERTopic [3] which introduces external information in the form of seed words to guide the algorithm in discovering topics that align with the specified guidance. This guidance helps to improve the relevance and coherence of the identified topics. Here, we utilize discovered suicidal risk factors from the psychology literature as seeds to guide the topic modeling process. This deliberate approach allows us to align the resulting topics with the known psychological factors that contribute to suicide risk, creating a more focused and domain-informed representation.

Figure 5.3: Incorporation of Seed Word in the Guided BERTopic approach



[3] Guided BERTopic

Guided BERTopic involves two primary steps: Firstly, the BERTopic algorithm generates embeddings for each seeded topic by concatenating them and passing them through the document embedder. As shown in Figure 5.3, these embeddings are compared to the existing document embeddings using cosine similarity and are then assigned a label. If a document is most similar to a seeded topic, it receives that topic's label; otherwise, if it is most similar to the average document embedding, it will be categorized as an outlier (the -1 label). UMAP then applies these labels in a semi-supervised manner, guiding the dimensionality reduction process to emphasize the distinctions between seeded topics and potentially identify outliers, thereby steering the topic creation more effectively towards the seeded topics. Secondly, all words in the seed topic list are assigned to a multiplier with a value greater than 1. These multipliers are applied to increase the IDF values of the words across all topics by boosting the likelihood of a seeded topic word appearing in a topic. After having generated our topics using class-based TF-IDF , we employed *KeyBERTInspired* extraction technique that leverages BERT embeddings and simple cosine similarity to find keywords and key phrases that are most similar to a document. Then we chose the keywords with highest class-based TF-IDF score as representative seed topics for the guided topic modeling.

5.4 Results

In our scoping review of the psychology literature, we initially identified 1,042 articles related to suicide. After applying the inclusion and exclusion criteria outlined in Figure 5.1, a total of 51 studies were retained for analysis. Table 5.1 provides a list of references we investigated to extract underlying risk factors of suicide ideation. This table organizes all the studies reporting the identified risk factors. Some of these risk factors fall under broader

categories, allowing for a more structured understanding. In Section 5.1 we provided the description of each extracted risk factor and how they play a role in suicidal ideation.

In unsupervised topic modeling, we investigated the underlying topics within each social media dataset. A comparison between the extracted topics in the Unsupervised Topic Modeling and Scoping Review Knowledge Extraction sections revealed that many topics reported in psychology literature were not discovered in social media using unsupervised BERTopic. The absence of these topics was evident only when compared to the risk factors in psychology literature, and without this comparison, these important gaps in the discussion might have been overlooked.

To address this, we employed guided topic modeling, which focuses on specific risk factors of interest. This approach ensures that important but less frequently mentioned topics are identified and included in the analysis. Hence, the topics from psychology were used as seed topics for the discovery of the topics in three datasets collected from social media, including the UMD dataset, the Knowledge Aware Assessment dataset and the 2021 Reddit dataset using guided-BERTopic. The suicidal risk factors from psychology are used as topic categories, and the keywords with the highest class-based TF-IDF found with unsupervised BERTopic are manually assigned to each risk factor as seed topics. Table 5.2 presents the list of seed words used in guided topic modeling of the UMD dataset to extract topics from the datasets. We conducted seed topic extractions separately for each of the three social media datasets. Note that this list of words does not need to be exhaustive and only serves as a hint about the topic for the topic modeling algorithm.

Here, we aim to understand the distribution of topics in datasets using guided topic modeling. Table 5.3 presents the suicidal topics in each dataset along with the number of posts for each topic. Since we are interested in the distribution of these topics in relation to suicidal thoughts in labelled datasets (UMD and Knowledge-Aware suicidality), we only

Table 5.1: List of 23 risk factors for suicidal ideation extracted through a scoping review of 51 psychology and mental health published studies. References are organized by risk factor categories.

Categories	Studies
Mental Health Disorders	
Depression	[102], [103], [104], [154], [155], [156], [157], [158], [159], [12], [160], [161], [162]
Anxiety	[105],[163],[161] [164],[165], [157],[158],[159]
Bipolar	[154],[160],[12],[161], [162],[16]
Schizophernia	[164],[154],[156], [158],[12],[160],[161]
Borderline	[157], [166], [167],[12],[161]
PTSD	[106],[107], [158]
Sociodemographic Status	
Unemployment	[116].[158],[159] ,[12],[160]
Education Pressure	[122], [123],[160], [162]
Financial Crisis	[105], [121],[165],[158],[159] , [162]
Sexual Minority Stigmas	[125],[158], [126],[127],[124] [128] , [162]
Abuse	
Being Bullied	[105], [129], [130], [131]
Sexual Abuse	[168],[158] ,[160], [162],[169]
Family domain	
Death of Loved Ones	[135],[159],[160]
Family conflicts	[116],[104] ,[161]
Relationship problem	[116],[160],[161]
Personality and Psychological Traits	
Hopelessness	[102],[103],[116],[159] ,[12], [170]
Anger	[102],[158],[159] ,[115]
Perfectionism	[102],[158],[160]
Chronic Physical Pain	[144],[145], [146], [146]
Dementia	[141],[142], [143]
Racism	[136],[137], [138]
Immigration	[139],[140],[12],[160] , [162]
Substance Abuse	[117][118], [119], [120],[165],[161] , [162]

Table 5.2: List of seed words for guided topic modeling in UMD dataset

Categories	Seed Words
Depression	Depressed, Sadness, Mentally ill
Anxiety	Anxious, Phobia, Stress
Unemployment	Job, Work, Poverty
Family Issues	Family, Parents, Mom, Dad
Relationship Problems	Wife, Husband, Partner, Girlfriend, Boyfriend
Hopelessness	Despair
Anger	Outrage, Annoyed
Perfectionism	Perfection, Expectations
Financial Crisis	Money, Income
Education Pressure	College, School, Overwhelmed
Being Bullied	Bullying, Oppress, alone
Death of Loved Ones	Loss, Grief, mourn
Immigration	moving, loneliness, Culture
Racism	Discrimination, Justice, Bias, Hate
PTSD	War, Memory, Accident
Substance Abuse	Alcohol, Drug, Opioid
Chronic Physical Pain	Constant, Hurt, Escape
Sexual Minority Stigmas	LGBTQ, Identity
Dementia	Alzheimer

considered the suicidal classes for topic modelling and evaluation.

The results in Table 5.3 show that while social media provides a vast and dynamic platform for individuals to express their thoughts and experiences, it may not always comprehensively reflect the nuanced and scientifically established suicide-related topics discussed in the academic psychology literature. Specifically, we observe that suicidal narratives related to dementia, sexual minority stigmas, immigration, death of loved ones, perfectionism and anger are never discussed in these datasets and topics of being bullied, PTSD, substance abuse, and chronic physical problems are rarely mentioned. Financial crises and racism, which are two common and important risk factors of suicide, are only discussed

Table 5.3: Count of instances related to each of suicidal risk factors, extracted using guided-BERTopic in datasets collected from social media.

Topics and Risk Factors Extracted from Psychology	Datasets		
	UMD (SW Subreddit)	Knowledge-Aware Assessment	2021 SW Dataset
Total # of Posts in Each Dataset	490	500	2050
Mental Health Disorders			
Depression	115	79	806
Anxiety	52	80	297
Bipolar	5	2	2
Schizophrenia	-	-	-
Borderline	-	2	-
PTSD	5	-	4
Sociodemographic Status			
Unemployment	30	25	-
Financial Crisis	-	30	-
Education Pressure	30	-	33
Sexual Minority Stigmas	-	-	-
Abuse			
Being Bullied	8	-	13
Sexual Abuse	2	-	7
Family Domain			
Death of Loved Ones	-	-	-
Family Issues	55	14	55
Relationship Problems	25	26	38
Personality and Psychological Traits			
Anger	-	-	-
Perfectionism	-	-	-
Hopelessness	9	-	17
Racism	-	-	36
Substance Abuse	7	-	-
Immigration	-	-	-
Chronic Physical Problems	-	2	3
Dementia	-	-	-

in one of the datasets with a relatively low number of examples. The most represented topics in these datasets are depression and anxiety. This observation is not surprising since many other risk factors lead to mental pressure. In social media conversations, users only talk about the depression and anxiety that they are experiencing and not the root causes that led to these feelings.

The conversational nature of social media often includes a wide range of personal narratives, opinions, and language that may or may not align with the structured and research-driven topics found in psychological literature. Although social media can offer valuable insights into real-world expressions of mental health concerns, researchers need to carefully interpret and validate social media data to ensure its reliability and relevance to the broader body of psychological literature on suicide. Underrepresentation of certain topics in social media data, specifically stigmatized topics such as conversations around sexual minorities and racism, can lead to models that underperform in cases where users break the stigma and talk about these issues.

5.5 Conclusion

This chapter focused on the assessment of the datasets collected from social media using suicidal ideation topics and risk factors extracted from the psychology literature. Analyzing psychological literature helps NLP researchers gain insights into the specific language patterns and terminology commonly used when discussing suicidal ideation. This knowledge can improve the design of NLP models to identify and understand such content in social media. Moreover, leveraging psychology literature helps to build a responsible AI system for addressing mental health issues on social media by being aware of the contexts of suicidal posts.

We performed Guided Topic Modeling on three social media datasets (Table 5.3) to provide a comprehensive overview of the suicidal topics and risk factors existing in those datasets. These datasets offer valuable insights into the context of suicidal ideation in online platforms. However, a notable limitation within these datasets is the absence of relevant topics that are known to be critical in addressing suicidal ideation. This deficiency could create a barrier to developing a robust NLP model for suicide-related content analysis.

Furthermore, we discovered that there is a discrepancy in the distribution of topic representations across the datasets. As an example, let's consider the UMD dataset, where the topic "Anxiety" appears 10.4 times more than "PTSD" topic. Such disparities in topic representations could lead to an inherent bias within the models. Biases can occur when certain topics are over-represented or under-represented. These biases will potentially impact the effectiveness and fairness of NLP models, as they may disproportionately emphasize or neglect certain aspects of suicidal ideation. Therefore, addressing these data limitations and imbalances is crucial to ensure the development of NLP models that provide accurate and equitable insights into this critical issue.

Chapter 6

Synthetic Data Generation

The sensitivity surrounding suicide-related data poses challenges in accessing large-scale, annotated datasets necessary for training effective machine learning models. To address this limitation, we introduce an innovative strategy that leverages the capabilities of GLLMs , such as GPT-3.5 , Flan-T5, and LLaMA, to create synthetic data for suicidal ideation detection. Our data generation approach is grounded in social factors extracted from Chapter 5 and aims to ensure coverage of essential information related to suicidal ideation. Then, we benchmarked against state-of-the-art NLP classification models, specifically those centred around the BERT family structures, by fine-tuning the ALBERT model using the UMD dataset. In our next step, we combined a mere 30% of the UMD dataset with our synthetic data and then fine-tuned the ALBERT model using this augmented dataset. Models trained with this augmented dataset achieve similar performance to models trained with the full UMD dataset when tested on the UMD test set. Such results underscore the cost-effectiveness and potential of our approach in confronting major challenges in the field, such as data scarcity and the quest for diversity in data representation. Moreover, we

analyze the generated datasets, including synthetic, augmented and real datasets based on sentence structures and diversity of semantics. Finally, we perform a guided topic modelling on Synthetic and augmented datasets to verify the generated datasets' coverage in terms of the topic and risk factors.

6.1 Synthetic Data Generation Using GLLMs

We utilized three generative language models to generate a synthetic dataset related to suicidal ideation. GLLM's foundation is constructed with transformers. Transformers are a class of deep learning models, first introduced by Vaswani et al.[171] in 2017. Researchers build state-of-the-art NLP models using transformer-based architectures because they can be quickly trained on large datasets, and studies have shown that they are better at modeling long-term dependencies in natural language text [172]. GLLMs s, including GPT-3.5 , FlanT5, and LLaMA2 are designed with the primary purpose of generating coherent and contextually relevant text. They excel at tasks such as text generation [173], completion [174], and dialogue generation[175]. These models are typically based on decoder transformer architectures and focus on the generative aspect of language which involves the auto-regressive generation, where the models predict the next word based on the preceding context. Generative models are trained on a vast corpus of text, however, their main strength lies in their ability to generate text that flows naturally and contextually appropriate.

We aim to build a diverse dataset in order to train a generalizable and robust model in suicidal ideation detection. In total, nine different datasets are generated with different specifications and models.

GPT-3.5 Turbo

we used *gpt-3.5-turbo*[1] which is one of the most advanced language models developed by OpenAI at the time of the experiment. In this project, we evaluate the capability of GPT-3.5 in Zero-shot Learning and Few-shot Learning settings to generate a diverse suicidality dataset. However, we are primarily focused on Zero-shot Learning methods as GPT-3.5 has exhibited superior performance in this setting compared to Few-shot Learning for a suicidal ideation detection task. In Chapter 4, we conducted an extensive comparison of the Zero-shot and Few-shot approaches using GPT-3.5 . According to our findings, using this model in a Few-shot setting might yield poorer performance compared to Zero-shot in various scenarios. In Few-shot, with few examples available for in-context learning, the risk of overfitting increases. The model might learn specific nuances or noise within the limited Few-shot data, leading to poor generalization on unseen examples. Moreover, Few-shot learning relies on a small subset of labeled examples, which might not adequately represent the entire diversity of the dataset. The model might fail to capture the complexity and variability present in the broader dataset during fine-tuning.

The temperature hyperparameter in GPT-3.5 is a crucial parameter that influences the generated output. A higher temperature value, such as 1.0, increases the randomness and produces more varied responses. Conversely, a lower temperature value, such as 0.1, reduces randomness and generates more focused and deterministic responses. According to our findings in Chapter 4, we configured the temperature parameter of GPT-3.5 to 1 in order to reduce the inconclusiveness rate of the model's responses. This parameter decreases as the temperature parameter increases. As such, for this project, we have configured the temperature parameter of GPT-3.5 to be 1.

[1] https://platform.openai.com/docs/models/gpt-3-5

We generated five datasets using GPT-3.5 . Four of these datasets are informed by extracted suicide-related topics from the psychology literature, while one dataset is generated without providing any specific topics. For incorporating suicide-related topics in data generation, we utilized prompt engineering techniques explained in Chapter 4 to generate suicidal ideation datasets.

In both Zero-shot and Few-shot Learning, we initiated the prompt construction process with a simple initial prompt and iteratively refined it through multiple rounds of trial and error. This iterative approach allowed us to gradually evolve the prompt, making necessary adjustments based on the observed outcomes. Similar to Chapter 4, in Few-shot Learning, the prompt is structured to include two examples for each category (8 in total) from the training set of the UMD Dataset, followed by a text generation question. This approach enables the model to learn from a limited set of labeled examples before generating a dataset. Moreover, by combining the Few-shot Learning methodology with the inclusion of the psychology topics in the prompt, we aim to enhance the model's ability to generate meaningful and contextually relevant responses when dealing with suicide-related discussions. The following example shows a prompt we used to generate a binary dataset (*Suicide/non-suicide*), guided by the topics introduced in Chapter 5, in the Zero-shot Learning setting.

```
Criteria="""These are the criteria of different suicide risk level:
    Risk Level=Non Suicidal: I do not see evidence that this person is
        at risk for suicide
    Risk Level=Suicidal: I believe this person is at high risk of
        attempting suicide in the near future. """

prompt = f"""
Your task is to generate a suicidal text for each of the following "
    topics" with different Risk levels.

    1-Depression
    2-Anxiety
```

```
    3-Hopelessness
    4-Anger
    5-Perfectionism
    6-Family issues
    7-Relationship problems
    8-Unemployment
    9-Financial Crisis
    10-Education
    11-Being Bullied
    12-Death of closed one
    13-Immigration
    14-Racism
    15-PTSD
    16-Substance Abuse
    17-Sexual Minority Stigmas
    18-Dementia
    19-Chronic pain

Provide the answers in JSON format with the following columns: text,
    topic, risk level.

Risk level criteria: ```{Criteria}```

"""
```

Later in Section 6.2, we reported that the results of the models fine-tuned on Few-shot and multi-class datasets were significantly inferior to those of Zero-shot and binary datasets. Consequently, we opted to exclusively generate binary datasets in Zero-shot settings for other LLMs.

Flan-T5

FLAN-T5 models are instruction-based, fine-tuned across a diverse set of tasks, aiming to enhance their Zero-shot performance on various tasks [176]. During instruction tuning, pretrained models undergo fine-tuning using drafts of instructions that guide them on how to perform a specific task. These instructions can include real-time feedback to assist the model in learning from its mistakes and improving at a faster rate. By providing explicit

guidance and incorporating feedback mechanisms, the instruction-tuning process enables the model to refine its performance and enhance its ability to execute the given task accurately. This iterative approach of incorporating instructions and feedback facilitates the model's learning process, allowing it to adapt and improve its performance based on the provided guidance.

In this work, we utilized *Flan-T5-XXL*[2] presented by Google Research [177] in a Zero-shot setting. Two datasets are generated using Flan-T5, one with topics and another without topics. Moreover, similar to GPT-3.5 , the temperature value is set to 1, and the same prompt structure is utilized.

LLaMA 2

LLaMA is an auto-regressive language model based on the transformer architecture. Similar to other generative models, LLaMA operates by taking a sequence of words as its input and making predictions about the subsequent word, iteratively producing text in a recursive manner. It is a collection of state-of-the-art foundational language models, with parameter counts ranging from 7 billion to 65 billion. The foundation models were trained on large unlabeled datasets, making them ideal for fine-tuning on a variety of tasks. The newest version of this model at the time of the experiment, LLaMA 2, expanded its pre-training corpus size, allowing the model to learn from a more extensive and diverse set of publicly available data. Additionally, the context length of LLaMA 2 has been doubled, enabling the model to consider a more extensive context when generating responses, leading to improved output quality and accuracy [178]. In this chapter, we used LLaMA 2-13B[3], presented by Meta in the Zero-shot setting. In total, we generated two datasets with LLaMA2, one with

[2] https://huggingface.co/google/flan-t5-xxl
[3] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

Table 6.1: Detailed description of generated synthetic datasets

Dataset #	Model	Learning Method	Topic-Oriented	# of Class	# of Instances
1	Chat GPT	Zero-shot	Yes	2	748
2	Chat GPT	Zero-shot	No	2	646
3	Chat GPT	Few-shot	Yes	2	545
4	Chat GPT	Zero-shot	Yes	4	492
5	Chat GPT	Few-shot	Yes	4	594
6	Flan-T5	Zero-shot	Yes	2	561
7	Flan-T5	Zero-shot	No	2	502
8	LLaMA 2	Zero-shot	Yes	2	395
9	LLaMA 2	Zero-shot	No	2	613
10	Mix Dataset	Zero-shot	Yes	2	1352
11	Synthetic Testing Set	Zero-shot	N/A	2	318

topics and another without topics. These datasets were created using the temperature of 1 and maintained the same prompt structure as GPT-3.5 .

Generated Datasets

A total of nine datasets are generated. An extensive description of these datasets, as well as a mixed set and a test subset, is presented in Table 6.1. As shown in Table 6.1, we created binary datasets and four-class datasets, each with the option of including or not including the topics.

- **Binary Datasets**: The binary datasets contain two classes, which allows us to evaluate the model's ability to distinguish between suicidal ideation and non-suicidal instances.
- **Four-Class Datasets**: The four-class datasets involve multiple categories, enabling us to explore more nuanced predictions of suicidal ideation levels, including "*No Risk*", "*Low Risk*", "*Moderate Risk*" and "*High Risk*" classes.

- **Topic-Oriented Datasets**: The option to include or not include the topics in these datasets allows us to investigate the impact of information provided by topics on the model's performance.

- **Synthetic Test Set**: A synthetic test set is created to support the evaluation process. The synthetic test set is composed of 10% of each synthetic dataset generated in this study. The test set was annotated independently by two human annotators with background in computer science and natural language processing who were tasked with identifying posts as either suicidal or non-suicidal. The annotators were provided with detailed guidelines describing each class definition, along with representative examples, to ensure consistent interpretation. To assess annotation reliability, inter-annotator agreement was measured using percent agreement. For 89% of the samples, the suicidality labels generated by the LLM were confirmed by both annotators. However, for the remaining 11% of the data, the assigned label were revised based on human judgment. In cases where both annotators agreed, that label was retained. In cases of disagreement, the annotators engaged in discussion to reach a consensus.

By comparing the results from datasets with and without topics, we can gain insights into how incorporating topic-related data enhances or influences the model's effectiveness in suicidal ideation detection. Furthermore, as explained in Section 6.1, we created a synthetic testing dataset comprising 10% of each dataset, which is annotated by human experts. Some example of the generated binary dataset is provided in Appendix .1.

6.2 Fine-Tuned Classifiers on Synthetic and Real Datasets

In this section, we use three testing sets to evaluate the fine-tuned models:

Multi-class UMD Testing Subset:

This test dataset is a subset of the UMD datasets, which is annotated as a 4-class dataset. We employed a 10-20-70 split for validation, test, and training sets, respectively. Out of the entire dataset, 10% was allocated for validation purposes, ensuring the model's hyperparameters and configurations were appropriately set. 20% of the data was set aside as a test set to evaluate the model's performance on unseen data and ensure its generalizability. The remaining 70% formed the training set, where the bulk of the data was utilized to train the model and learn the underlying patterns. This distribution was chosen to provide substantial data for training while reserving enough distinct data for validation and robust performance testing.

Two models, ALBERT and DistilBERT are fine-tuned with a real and two generated synthetic datasets. The real dataset we used is the UMD training dataset and the synthetic datasets we used were the GPT-3.5 multi-class, topic-oriented, Zero-shot and Few-shot datasets. Table 6.2 presents the results of the performance evaluation of models fine-tuned with multi-class synthetic datasets generated by GPT-3.5 in Zero-shot and Few-shot settings, tested on the multi-class UMD test set.

In our experiments, we observed that GPT-3.5 is not able to produce high-quality multi-class datasets in either the Zero-shot or Few-shot settings. Generating multi-class datasets using LLMs such as GPT-3.5 is a challenging task due to the inherent complexities involved

Table 6.2: Performance evaluation of fine-tuned ALBERT and DistilBERT models trained with real and synthetic datasets and tested on the Multi-class UMD test dataset

Models	Metrics	UMD Training Dataset	GPT-3.5 Zero-shot	GPT-3.5 Few-shot
ALBERT	Accuracy	0.865	0.41	0.36
	F1-Score	0.87	0.43	0.27
DistilBERT	Accuracy	0.77	0.06	0.06
	F1-Score	0.75	0.1	0.12

in distinguishing between multiple and fine-grained classes. Even with the availability of a high-quality dataset, one should anticipate lower accuracies in multi-class scenarios. This is largely attributed to the ambiguous boundaries that exist between these classes. Moreover, the creation of such datasets necessitates not only a detailed prompt but also specific instructions that outline the multi-class scenarios. This process demands a nuanced understanding and a level of specificity that often poses a considerable challenge to GPT-3.5 [55]. As a result, we opted to exclusively create binary datasets and focus our investigation on how topics impact the overall generalizability of the fine-tuned models.

Binary UMD Testing Subset:

In order to perform binary classification, we binarize the UMD Dataset. Based on the definition of each class, "*No Risk*" and "*Low Risk*" classes are considered as Non-Suicidal and "*Moderate Risk*" and"*High Risk*" as Suicidal.s

Table 6.3 provides the performance of the models trained on the binary synthetic datasets generated by GPT-3.5 , Flan-T5 and LLaMA models and tested on the binary UMD testing subset. We compare the results for the synthetic datasets with those of the UMD training set. Table 6.3 shows that incorporating topic in generating the datasets significantly improves the performance of the models. For instance, for Flan-T5, the

Table 6.3: Performance evaluation of the ALBERT and DistilBERT models fine-tuned with binary datasets and tested on UMD testing subset

Models	Metrics	Non-Synthetic	GPT-3.5			Flan-T5		LLaMA 2		Mix Dataset
		UMD Training Dataset	With Topic Few-shot	Without Topic Zero-shot	With Topic Zero-shot	Without Topic	With Topic	Without Topic	With Topic	With Topic
ALBERT	Accuracy	0.87	0.67	0.70	0.71	0.48	0.62	0.33	0.75	0.77
	F1-Score	0.87	0.66	0.79	0.79	0.54	0.64	0.49	0.78	0.82
DistilBERT	Accuracy	0.77	0.61	0.63	0.64	0.59	0.77	0.32	0.75	0.76
	F1-Score	0.75	0.59	0.69	0.71	0.61	0.84	0.15	0.77	0.82

topic-oriented dataset increased the F1-score and accuracy of the ALBERT model by 10% and 14% points, respectively. We also created a mixed dataset, including all topic-oriented datasets, to further evaluate the effects of topics on the performance of the models. With both ALBERT and DistilBERT, an F1-score of 0.82 is achieved by the mixed dataset, which is significantly higher than the DistilBERT model trained on the UMD dataset and comparable with the performance of the ALBERT model fine-tuned on the UMD dataset with an F1-score of 0.87. Surprisingly, DistilBERT fined tuned with the topic-oriented dataset generated by Flan-T5 results in F1-score of 0.84. This is an exception and does not align with the general pattern that training with synthetic data results in an acceptable level of accuracy but under-performs training with real datasets.

Synthetic Testing Subset:

We created and annotated this dataset comprising 10% of each synthetic dataset generated in this study. In Section 6.1 we explained the creation and annotation process of this test set. Table 6.4 presents the results of models included in Table 6.3 but tested on synthetic testing datasets. Similar to the results of Table 6.3, all of the topic-oriented datasets show significant improvement compared to the datasets without any topics. GPT-generated training data, with an F1-score of 0.82, exhibits the best performance, while the performances of Flan-T5 and LLaMA2-generated datasets are moderate. An important

Table 6.4: Performance evaluation of the ALBERT and DistilBERT models fine-tuned with binary datasets and tested on synthetic testing subset

Models	Metrics	Non-Synthetic	GPT-3.5			Flan-T5		LLaMA 2		Mix Dataset
		UMD Training Dataset	With Topic Few-shot	Without Topic Zero-shot	With Topic Zero-shot	Without Topic	With Topic	Without Topic	With Topic	With Topic
ALBERT	Accuracy	0.67	0.71	0.81	0.81	0.34	0.63	0.48	0.70	0.83
	F1-Score	0.70	0.69	0.78	0.82	0.41	0.69	0.24	0.73	0.81
DistilBERT	Accuracy	0.40	0.65	0.83	0.85	0.63	0.86	0.49	0.63	0.78
	F1-Score	0.61	0.61	0.81	0.81	0.69	0.84	0.12	0.69	0.73

observation is that the models trained on real data perform moderately when tested on synthetic data. We attribute this observation to the presence of topics in synthetic data that are underrepresented or absent in real data. The mixed dataset shows a 0.81 F1-score, which is an 11% improvement compared to the model trained with the UMD dataset.

6.3 Data Augmentations

In the previous section, we observed that synthetic data shows promise to be used for training specialized classifiers. Given that synthetic data generation is significantly more cost-effective than collecting and labelling real-world data, we suggest using this data for training and enriching it with relatively small sets of real-world data. Data augmentation involves enriching a dataset by introducing variations to its existing instances or generating entirely new instances. This process is designed to enhance the diversity and quality of the dataset, which, in turn, can lead to improved model performance and generalization. Hence, in this study, we augment the best-performing synthetic dataset generated by GLLMs with different subset sizes of the UMD dataset. Based on the results presented in Table 6.3 and Table 6.4, the datasets generated by GPT-3.5 in the Zero-shot setting show the best results compared to the other datasets. To generate the augmented dataset, we start by combining the 10% of the UMD training subset, with the selected synthetic dataset. In each

iteration, three folds, each comprising 10% of non-overlapping random samples from the UMD dataset, are added to the synthetic data. Subsequently, the average[4] of the accuracy and F1-score are calculated and reported in Table 6.5. The augmented dataset now contains a mix of synthetic and real data instances. The augmented dataset is used to fine-tune the pretrained models and then evaluated on two separate testing sets.

If the model's performance with the augmented dataset is less than the model trained with the UMD dataset, additional real-world data is gradually incorporated. For instance, the percentage of real data can be increased to 20% in the next iteration, and the training and evaluation process is repeated.

Throughout the iterations, the model's performance is closely monitored and compared to the baseline model trained solely on the UMD dataset. The aim is to identify the point at which the augmented dataset starts producing results comparable to or even surpassing those of the baseline model. The process continues until an optimal percentage of real data is found, where the model achieves similar results as the baseline. This ratio indicates the ideal balance between synthetic and real data for achieving high model performance and generalization. We observe that when synthetic data is augmented with 30% of the UMD train set, the fine tuned model outperforms the model trained with the full UMD dataset. This ratio represents the minimum amount of augmentation required for the ALBERT model to surpass the performance results of the model trained solely on the UMD dataset. Table 6.5 shows the results of the augmentation process.

[4]we also calculated the standard deviation of the metrics which were always < 0.02.

Table 6.5: Performance evaluation of the ALBERT model fine-tuned with the augmented dataset (synthetic data + a subset of the UMD train set) and tested on UMD and synthetic testing subsets

Test Set	Metric	UMD Training Dataset	10% (Avg. of 3 Folds)*	20% (Avg. of 3 Folds)*	30% (Avg. of 3 Folds)*
UMD Testing Set	Accuracy	0.87	0.75	0.81	0.83
	F1-Score	0.87	0.79	0.84	0.88
Synthetic Testing Set	Accuracy	0.67	0.87	0.87	0.90
	F1-Score	0.70	0.83	0.86	0.88

* Standard Deviation< 2%

6.4 Dataset Analysis

For textual datasets, assessing the diversity of semantics and sentence structure of textual content is crucial, even more so when dealing with synthetic datasets. To measure these qualities, we use three sets of metrics:

Complexity: refers to the intricacy and sophistication of the language used in a text, encompassing factors such as sentence structure, vocabulary richness, and syntactic intricacies. We utilize the *Type-Token Ratio* TTR[179] to assess the complexity of a text using lexical diversity measures. The basic idea behind that measure is that if the text is more complex, the author uses a more varied vocabulary, so there is a larger number of unique words[180].

Readability: pertains to the ease with which a text can be comprehended by its intended audience, considering elements such as sentence length, word difficulty, and overall coherence. To measure the readability of text, we use the *Flesch's Reading Ease Test*[181], which quantifies readability based on sentence length and the number of syllables per word. As a benchmark, the highest score identify a text that is easily understood by an average 11-year-old, while a lowest score is best understood by university graduates. Table 6.6 presents the definition criteria for *Flesch's Reading Ease Test* score.

Table 6.6: Flesch's Reading Ease Test score definition

Score	Difficulty
90-100	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
0-29	Very Confusing

Entropy: is a metric that measures the unpredictability of a text. It assesses the information content or disorder present in the dataset. In simpler terms, entropy in text datasets gauges how diverse or varied the words or characters are within the dataset. High entropy indicates a higher degree of unpredictability, suggesting a wider range of different words or characters used in the text. Conversely, low entropy implies a more predictable or ordered text with fewer variations in the words or characters used. Shannon entropy[182], is calculated based on the frequency of occurrence of different characters, words, or other linguistic units within the text. In word-based analysis, higher entropy suggests a wider range of vocabulary, showing greater linguistic diversity [183]. Calculation of Shannon entropy involves summing the probabilities of each word occurrence in the text, weighted by the logarithm of the inverse of these probabilities. In Equation 6.1 the Shannon entropy H for a set of words with probabilities $p_1, p_2, ..., p_n$ is calculated as:

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i) \tag{6.1}$$

In this equation, H represents the calculated entropy value for the given set of words, and p_i represents the probability of the i^{th} word occurring in the text. The sum extends over all unique words. By computing Shannon entropy in text analysis, one can gain insights

into the richness, diversity, and complexity of the language used within the text dataset.

Understanding complexity and readability in synthetic datasets helps in ensuring the generated text aligns with linguistic patterns observed in real-world data. Moreover, these parameters facilitate an assessment of the synthetic dataset, specifically regarding the incorporation of suitable language complexities. This evaluation allows us to examine whether synthetic data replicates language patterns akin to those found in genuine, human-generated content. In Table 6.7, we report that synthetic datasets feature more complex sentences, evident from the higher complexity score and lower readability score compared to real datasets (e.g., UMD Dataset). However, the low standard deviation of the Shannon entropy score in the synthetic dataset suggests that, despite its increased complexity, the synthetic dataset exhibits a narrower range of vocabulary and less diversity.

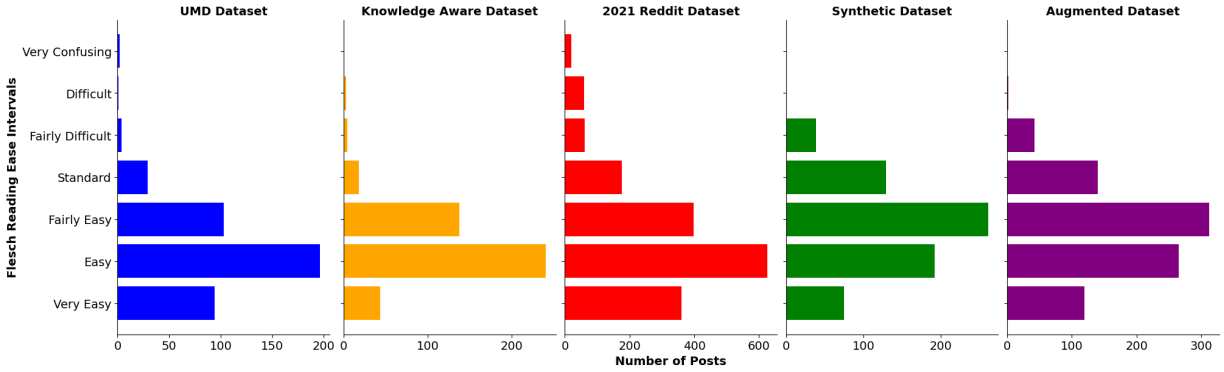
Moreover, Figure 6.1 indicates that the readability of the synthetic and augmented datasets is more challenging than that of real datasets. Specifically, 36% of the synthetic dataset falls into the 'Easy' and "Very easy" categories, while 67% of the UMD dataset is within these categories. In contrast, 33% of the UMD dataset and 64% of the synthetic dataset fall into the "Fairly easy", "Standard", "Fairly difficult", and "Difficult" categories. Furthermore, we conducted a statistical t-test with a significance level of 0.05 on the readability, complexity, and entropy metrics, comparing the synthetic dataset with each of the real datasets. Our findings indicate a significant difference in Shannon entropy between the synthetic and real datasets. However, no significant differences were observed in terms of readability and complexity between the real and synthetic datasets.

Shannon Entropy serves as a quantitative measure of diversity, reflecting the range of vocabulary in a dataset. The results of our study, which showed a lower Shannon Entropy in synthetic datasets despite higher complexity, prompt a nuanced discussion. This discrepancy indicates that, despite intricate language patterns, synthetic datasets might lack

Table 6.7: Mean and Standard deviation of Complexity, Readability, and Shannon entropy of the Datasets

Datasets	Complexity		Readability		Shannon Entropy*	
	μ	σ	μ	σ	μ	σ
UMD Dataset	61.18	16.15	82.49	10.87	4.31	0.17
Knowledge-Aware Dataset	49.04	19.53	82.51	15.27	4.17	0.75
2021 Reddit Dataset	72.87	20.13	75.76	48.02	4.27	0.30
Synthetic Dataset	72.77	18.40	75.84	10.34	4.27	0.05
Augmented Dataset	70.27	17.18	77.38	10.80	4.28	0.05

Figure 6.1: Flesch's Reading Ease Test score and its distribution for each dataset



the diverse lexical richness found in real datasets. Thus, the relationship between diversity and Shannon Entropy suggests that achieving linguistic complexity does not guarantee a broad vocabulary range. Hence, augmenting a real dataset with synthetic data can leverage the advantages of both datasets, incorporating the linguistic complexity characteristic of synthetic datasets and the broad range of diversity inherent in real datasets. Moreover, the relationship between diversity and complexity in datasets reveals a fascinating interplay. While complexity often indicates intricate language structures, the presence of a diverse range of expressions and ideas enhances the overall diversity of the dataset. However, as seen in Table 6.7, the higher complexity in synthetic datasets does not necessarily translate to the higher diversity, as reflected by the lower Shannon Entropy. This suggests that

complexity might be influenced more by the intricacy of language patterns than by a broad lexical spectrum.

6.5 Topic Verification on Synthetic and Augmented Datasets

In this study, we harnessed the innovative approach of synthetic dataset creation as a means to enhance the fairness and accuracy of NLP models in the context of detecting suicidal ideation. Using the extracted risk factors from psychology, we construct a synthetic dataset that comprehensively represents the entire spectrum of risk factors associated with suicidal ideation. We employ guided topic modeling to verify if this method reveals the suicidal topics and risk factors in the synthetic and augmented dataset. Note that these topics are known and that the synthetic dataset has been generated using them. We generate approximately 40 posts for each risk factor. However, the number of topics discovered by the guided topic modeling does not sum up to the total number of posts, which is 748 for the synthetic dataset and 957 for the augmented dataset. This discrepancy arises because some posts are associated with multiple topics and represent more than one risk factor. Table 6.8 displays the suicidal topics within each synthetic and augmented dataset, along with the respective number of posts for each topic. As expected, we observe that topic modeling shows a relatively equal distribution of topics in the synthetic dataset. Table 6.8 illustrates how effectively guided BERTopic extracts the topics within the documents. Furthermore, these results demonstrate the significance of the generation of topic-diverse synthetic and augmented datasets for various research and applications within the domain of suicide-related studies.

Table 6.8: Distribution of suicidal risk factors in synthetic and augmented datasets extracted using guided-BERTopic.

Topics and Risk Factors Extracted from Psychology	Datasets	
	Synthetic Dataset	Augmented Dataset
Total # of Posts in Each Dataset	908	1055
Mental Health Disorders		
Depression	62	173
Anxiety	46	68
Bipolar	19	20
Schizophrenia	23	23
Borderline	23	23
PTSD	22	21
Sociodemographic Status		
Unemployment	20	35
Financial Crisis	21	39
Education Pressure	21	20
Sexual Minority Stigmas	41	42
Abuse		
Being Bullied	39	39
Sexual Abuse	18	20
Family Domain		
Death of Loved Ones	40	41
Family Issues	19	20
Relationship Problems	40	43
Personality and Psychological Traits		
Anger	39	38
Perfectionism	48	48
Hoplessness	36	44
Racism	39	39
Immigration	20	18
Substance Abuse	40	46
Chronic Physical Problems	56	61
Dementia	39	48

To leverage the strengths of both real and synthetic data, we augmented 30% of the UMD dataset by incorporating our synthetic dataset. Table 6.8 provides an insightful breakdown, demonstrating the distribution of each topic within our synthetic dataset and the augmented dataset. This detailed analysis allows us to discern the prevalence of specific topics in these datasets, which, in turn, is invaluable for the fine-tuning and optimization of our NLP models. By combining real and synthetic data while ensuring topic representation is well-balanced, our approach seeks to fortify the accuracy, fairness, and overall efficacy of these models in the challenging task of detecting and understanding suicidal ideation.

6.6 Conclusion

In this chapter, we focus on the generation of synthetic datasets using generative models and subsequently assessing the performance of models fine-tuned with these datasets. Our synthetic data generation framework addresses two limitations of real-world data collection and annotation. First, we address the data scarcity and annotation cost by generating micropost-like suicidal/non-suicidal text. Second, we address the lack of diversity in real-world data by prompting the generative models to create a balanced number of examples related to each of the psychological and social factors impacting suicidality. Integrating insights from psychology into the NLP pipeline in this context can illuminate previously unexplored facets of suicide and mental health detection in social media.

Our results show the critical role of incorporating domain knowledge in synthetic data generation. We extracted the relevant social topics from the psychology literature and used that to create more focused prompts for data generation.

Moreover, our results of fine-tuned models on topic-oriented datasets and non-topic-oriented datasets show the significant effects of including the topics on the performance

of models trained with the generated datasets. Informing the data generation with topics in Flan-T5 and LLaMA2 increased the F1-Score of the ALBERT model by 10% and 29% points, respectively. Fine-tuning models on topic-oriented synthetic datasets allows them to gain diverse domain-specific knowledge and patterns. Furthermore, non-topic-oriented synthetic datasets might lack specificity, leading to noise and irrelevant content. In contrast, topic-oriented datasets are curated to focus on a specific domain, reducing the chances of generating irrelevant or out-of-context text.

We showed that the BERT family classifiers fine-tuned with real-world data can achieve an F1 score ranging from 0.75 to 0.87, depending on the complexity of their structure. Specifically, DistilBERT, a less efficient model from the BERT Family, achieves an F1-score of 0.75, while ALBERT, a more optimized model designed for speed and accuracy, attains an F1-score of 0.87. In contrast, both DistilBERT and ALBERT achieve a consistent F1-score of 0.82 when trained on purely synthetic data and tested on real-world data. With this, we demonstrate that the diversity of synthetic data, which originates from incorporating topics and mixing outputs of three distinct models, compensates for model complexity irrespective of its architecture. This not only underscores the considerable potential of synthetic data but also suggests that it can mitigate the limitations of real-world data in capturing diverse topics. Most notably, our results emphasize an optimal strategy that involves augmenting synthetic data with real data. This method achieves performances comparable to the ALBERT model trained with the full real dataset when relying on synthetic data augmented with merely 30% of the manually annotated dataset.

Our analysis of the characteristics of synthetic datasets shows that they often exhibit a distributional shift from real-world data. This shift arises due to the inherent differences in the data generation processes between synthetic and real domains. As a result, models trained solely on synthetic datasets may not be applicable in real-world situations, leading

to a lack of robustness and adaptability. Therefore, exploring hybrid approaches that combine synthetic and real-world data for training can offer a more comprehensive solution. Leveraging both sources allows models to learn from the strengths of synthetic data while adapting to the intricacies of real-world environments.

Chapter 7

Interpretability of LLMs

Despite the significant effort NLP research has put into facilitating mental health analysis, the best-performing models lack interpretability. In sensitive AI systems such as mental-health support tools, interpretability is necessary to enable debugging, safety assurance, and meaningful control over model behavior, particularly when harms arise from rare, context-dependent failure modes that are not detectable through aggregate performance metrics. Without interpretable models, developers cannot reliably identify spurious correlations, audit learned representations, or apply targeted safety interventions, undermining both accountability and responsible deployment [184].

Based on our findings in this study, GPT-3.5 model demonstrates superior performance in classifying suicidality of the text data compared to other examined GLLMs. However, it still falls short when compared to classifiers tailored for this task through supervised training. Therefore, for the rest of our research, we focus on improving the interpretability of fine-tuned classifiers.

In previous chapters, we showed that existing datasets lack topic diversity, and some

topics are over-represented in datasets while others are absent or underrepresented. Previous work shows that classifiers may develop a false causal relationship between an over-represented concept and a given label, leading to excessive dependence on that concept and potentially compromising classification accuracy [185]. Also, some classifiers might become under-sensitive to topics that are not well-represented in training sets [82]. Therefore, it is crucial to develop methods that can facilitate the study of how topics are formed and used in trained classifiers.

In suicidal ideation detection, certain risk factors such as depression, anxiety, relationship problems, and hopelessness are strong predictors of suicidal ideation and appear frequently in training corpora. These concepts are often over-represented in the suicidal class, making them susceptible to being learned as sufficient causes for suicidality. Consequently, the classifier may over-rely on these concepts and ignore the broader context, leading to reduced generalizability. Conversely, topics that appear rarely in the dataset such as racism and immigration may be under-represented in the model's internal features. Therefore, it is necessary to understand how under-representation of topics translates into feature coherence in the latent space. This insight is necessary for diagnosing gaps in topic coverage.

These observations suggest that suicidal-ideation classifiers may not always form clear internal representations of psychological risk factors. Some concepts might appear as coherent and separable directions, while others may remain entangled with unrelated features. The structure of these representations matters for safety, transparency, and targeted interventions. Topic-aware augmentation may also influence how these concepts are encoded.

These considerations motivate a closer examination of the geometry of the model's latent space. In the remainder of this chapter, we outline the theoretical foundations for understanding internal features, introduce dictionary learning as a tool for revealing them,

and analyze how augmentation changes their clarity and separability.

7.1 Theoretical Background

Mechanistic interpretability seeks not only to correlate inputs with outputs, but to explain a model's internal computation by identifying which intermediate structures are represented, how they compose, and how they causally influence behavior. In this line of work, analysis focuses on the geometry of a model's internal representations, meaning the intermediate activations produced within the network during processing, rather than solely on the final outputs. A common working assumption in mechanistic interpretability is the linear representation hypothesis, which states that meaningful features often approximate linear directions in a model's activation space. Here, the activation space refers to the vector space formed by the model's internal activations at a specific layer, where each vector corresponds to the model's internal state for a given token or input. Under this view, probing the geometry of activations can reveal human-describable concepts that the model uses internally [85].

Modern transformer models achieve strong performance by representing many latent features within fixed-width activation spaces. Empirical evidence suggests that models encode far more features than there are embedding dimensions [186]. An embedding is a specific type of activation, typically referring to input token representations or output vectors used for downstream tasks; embeddings therefore live within activation space, but do not exhaust it. The apparent mismatch between the number of useful features and the dimensionality of the activation space motivates the superposition hypothesis. superposition hypothesis states high-dimensional spaces contain many nearly orthogonal directions, allowing networks to store multiple features in overlapping subspaces[87]. When

the number of useful features exceeds the available dimensions, features become entangled, producing polysemantic representations that respond to multiple concepts [86].

In this context, a feature is defined as a direction in activation space that corresponds to a recurring internal pattern used by the model. A feature is called monosemantic if it consistently corresponds to a single interpretable concept across inputs; otherwise, it is polysemantic. Polysemanticity arises naturally under superposition and can obscure the internal logic behind model predictions. For high-stakes applications such as suicide ideation detection, understanding whether psychological risk factors are represented in monosemantic or polysemantic forms is crucial for safety, transparency, and targeted intervention [72].

Dictionary learning via sparse autoencoders provides a principled method for exposing these underlying feature directions. SAEs aim to recover a set of basis directions whose sparse linear combination reconstructs the model's internal activations. Each basis direction corresponds to a feature, while the scalar output of the encoder indicates the feature activation, that is, how strongly a given feature is present for a specific token or input. Because sparsity encourages only a small number of features to activate for any given input, dictionary learning is well suited for probing whether psychological risk factors such as anxiety or family issues emerge as identifiable and coherent geometric directions in activation space [187].

Concept-based explanation methods aim to explain model behavior using human-aligned concepts rather than raw activations. TCAV represents a foundational approach in this literature. TCAV defines a concept by gathering example instances, learns a linear separator in an internal activation space to obtain a concept direction, and then measures how sensitive a model's prediction is to movement along that direction. Sensitivity is computed via directional derivatives and aggregated into a global score of concept importance for a class. Originally introduced for image models, TCAV demonstrated that models rely on

meaningful concepts such as stripes for the class "zebra," and that these concepts' influences can be quantified and compared across classes and layers [80].

Subsequent work extended this idea by focusing on automatic concept discovery and principled importance attribution. ACE segment inputs, cluster coherent patterns, and treat the resulting clusters as candidate concepts that can be evaluated using TCAV, reducing manual curation while preserving global importance testing [78]. A complementary line of work develops axiomatic scoring schemes for concept importance. ConceptSHAP formulates a completeness objective over concepts and uses Shapley values to attribute contributions in a way that satisfies desirable axioms such as additivity and symmetry, enabling principled ranking of concept sets beyond single-concept tests [188]. Concept bottleneck models incorporate concepts directly into the model architecture by first predicting a vector of human-labeled concepts and then predicting the final task label from those concepts, enabling intervention and transparent error analysis with competitive performance [189]. Other architectural techniques explicitly align internal axes with concepts during training; for example, Concept Whitening introduces a whitening transformation that rotates latent axes to correspond to selected concepts, facilitating inspection and manipulation of concept activations while maintaining accuracy [190].

These approaches differ along three key dimensions. First, where the concept lives. Post hoc linear probes on fixed layers as in TCAV and ACE, embedded concept layers as in concept bottleneck and whitening models, or factorized feature spaces as in sparse autoencoders. Second, how the concept set is obtained. Manually curated examplesTCAV as in TCAV , automatic discovery as in ACE, or supervised concept labels as in concept bottleneck and whitening approaches. Third, how importance is quantified. Directional derivatives in TCAV, Shapley-style completeness in ConceptSHAP, or causal interventions enabled by bottleneck architectures.

In this chapter, we adopt a fundamental idea from the concept-based explanation literature: defining a concept through example sets and contrasting it with a non-concept baseline. Our goal is not to measure how strongly a concept influences the final prediction, as in sensitivity-based methods such as TCAV , but to assess whether a concept is geometrically well formed in the model's internal representation. Sensitivity scores conflate representational structure with decision-boundary effects and depend on task-specific gradients, which makes them difficult to compare across models, datasets, or layers. In contrast, separability directly reflects how clearly a concept occupies a distinct direction in activation space, independent of the classifier head. This makes separability a natural diagnostic for identifying entangled or underrepresented risk factors, particularly in settings where topic coverage and representational clarity are the primary concerns. Moreover, directional separability aligns directly with the dictionary-learning framework introduced earlier. If a concept is well represented, its examples should activate a coherent subset of learned feature directions whose average orientation diverges from unrelated text. This geometric criterion is therefore especially well suited for evaluating whether topic-aware augmentation improves the internal encoding of psychological risk factors, rather than merely their downstream predictive influence.

7.2 Methodology

In this methodology, we shift the focus from model outputs to the model's internal representations, namely the intermediate activation vectors produced during processing. These activations live in the model's residual-stream activation space and encode semantic and psychological structure implicitly. Rather than treating suicidal ideation detection as a black-box classification problem, we analyze these internal representations to identify

interpretable feature directions that correspond to recurring patterns in language use.

To expose such feature directions, we apply dictionary learning via sparse autoencoders to the model's residual-stream activations. The sparse autoencoder decomposes each activation vector into a sparse linear combination of learned basis directions. Each basis direction, represented by a decoder vector, defines a feature in activation space, while the corresponding scalar output of the encoder indicates the feature activation for a given token. The resulting set of learned feature directions serves as the foundational object of analysis in the remainder of the methodology, enabling both qualitative inspection through top-activating text examples and quantitative geometric analysis.

Once these feature directions are learned, we employ two complementary methods to assess their structure and interpretability. First, we use Uniform Manifold Approximation and Projection (UMAP) [150] as a non-linear dimensionality reduction technique to visualize relationships between feature directions. The input to UMAP consists of the normalized decoder vectors of the sparse autoencoder, so each point in the visualization corresponds to a single learned feature rather than to an input sample. This visualization allows us to examine whether features associated with similar psychological topics occupy nearby regions of activation space, indicating more selective and structured representations, or whether they are dispersed, suggesting stronger superposition. UMAP is used here as an exploratory tool that emphasizes preservation of local neighborhood structure rather than faithful reconstruction of global geometry. As such, it provides an intuitive view of local clustering, overlap, and separation among feature directions that would be difficult to interpret directly in the original high-dimensional activation space.

Second, we move beyond visualization and introduce a quantitative analysis based on cosine distance. Inspired by Concept-based explanation methods, for each psychological concept, we construct two sets of input texts: a concept set containing samples labeled

with that topic, and a non-concept set consisting of unrelated Reddit posts. Each input text is represented by an activation vector extracted from the same residual-stream layer of the model. We compute an average representation for each set by taking the mean of the corresponding activation vectors, yielding one mean vector for the concept set and one for the non-concept set. These mean vectors live in the same residual-stream activation space and represent the average internal state associated with concept-related and unrelated content, respectively. We then compute the cosine distance between these two mean vectors, which measures their angular separation. Larger cosine distance indicates that the model encodes the concept in a direction that is more clearly separated from unrelated text.

7.2.1 Dictionary Learning on Residual Streams

The linear representation hypothesis proposes that neural networks encode meaningful concepts as approximately linear directions in activation space. In this context, we refer to such directions as features. The superposition hypothesis further suggests that high-dimensional activation spaces contain many nearly orthogonal directions, enabling models to represent more features than the ambient dimensionality by packing them into overlapping subspaces. When many features compete for limited representational capacity, they may become entangled, leading to polysemantic representations.

Taken together, these hypotheses motivate the use of dictionary learning to recover a set of basis directions such that internal activations can be reconstructed as sparse linear combinations of feature directions. The goal is not only to reconstruct activations accurately, but also to investigate whether the recovered directions correspond to interpretable and semantically meaningful patterns that help explain model behavior [191]. Recent work has shown that this approach is effective for transformer models, where SAEs provide a

scalable approximation for recovering interpretable and steerable feature directions from residual-stream activations [87, 192, 193]. In this study, sparse autoencoders serve as the core mechanism for dictionary learning and allow us to probe concept-like directions within the model's residual-stream activation space.

Sparse Autoencoder (SAE)

Sparsity is a key constraint in sparse autoencoders. It enforces that only a small subset of latent units becomes active for any given input activation vector. This regularization prevents information from being diffusely distributed across many units and instead encourages each activation to be represented as a combination of a few distinct basis directions. In practice, sparsity improves interpretability because each latent unit tends to correspond to a more localized and meaningful direction in activation space. It also reduces interference between features and promotes disentanglement, which is critical for identifying semantically coherent feature directions. From an optimization perspective, sparsity acts as an information bottleneck that discourages redundant representations and leads to more stable and generalizable decompositions of internal activations.

Overcompleteness refers to using a latent dimensionality that exceeds the dimensionality of the input activation space. In the context of transformer residual streams, overcompleteness is essential for addressing superposition, where many independent features compete for limited representational capacity. By expanding the number of latent units beyond the dimensionality of the residual stream, the autoencoder gains sufficient capacity to separate overlapping directions into more specific and independent features. Overcomplete sparse autoencoders therefore provide a principled mechanism for uncovering monosemantic feature directions that would otherwise remain entangled within dense model activations.

Let $x \in \mathbb{R}^D$ denote a model activation vector extracted from the residual stream, normalized so that the average squared ℓ_2 norm equals the model dimension D. A sparse autoencoder learns F latent features with $F \gg D$, yielding an overcomplete dictionary. The encoder maps the input activation vector to sparse feature activations via

$$f_i(x) \;=\; \mathrm{ReLU}\big(\langle W^{\mathrm{enc}}_{i,:}, x\rangle + b^{\mathrm{enc}}_i\big), \tag{7.1}$$

where $W^{\mathrm{enc}} \in \mathbb{R}^{F\times D}$ and $b_{\mathrm{enc}} \in \mathbb{R}^F$ are learned parameters. The scalar quantity $f_i(x)$ represents the activation of feature i for input x.

The decoder reconstructs the original activation vector as a linear combination of learned feature directions:

$$\hat{x} \;=\; b_{\mathrm{dec}} + \sum_{i=1}^{F} f_i(x)\, W^{\mathrm{dec}}_{:,i}, \tag{7.2}$$

where $W^{\mathrm{dec}} \in \mathbb{R}^{D\times F}$ and $b_{\mathrm{dec}} \in \mathbb{R}^D$. Each column $W^{\mathrm{dec}}_{:,i}$ defines a feature direction in residual-stream activation space.

The model is trained to minimize a loss function that combines reconstruction error with a sparsity penalty:

$$\mathcal{L} \;=\; \mathbb{E}_x\left[\|x-\hat{x}\|_2^2 \;+\; \lambda \sum_{i=1}^{F} f_i(x)\,\|W^{\mathrm{dec}}_{:,i}\|_2\right], \tag{7.3}$$

where λ controls the strength of the sparsity constraint.

Including $\|W^{\mathrm{dec}}_{:,i}\|_2$ inside the penalty allows the decoder vectors to be interpreted as normalized feature directions,

$$\tilde{W}^{\mathrm{dec}}_{:,i} \;=\; \frac{W^{\mathrm{dec}}_{:,i}}{\|W^{\mathrm{dec}}_{:,i}\|_2}, \tag{7.4}$$

with corresponding rescaled feature activations

$$a_i(x) \;=\; f_i(x)\,\|W^{\text{dec}}_{:,i}\|_2. \tag{7.5}$$

Under this decomposition, each token-level activation vector is explained by a small number of nonzero feature activations drawn from a large dictionary of feature directions. This parts-based representation aligns with the goal of identifying interpretable and potentially monosemantic directions within the model's residual-stream activation space [87, 192].

Overcompleteness ($F > D$) provides sufficient capacity to separate overlapping directions that arise under superposition, while sparsity ensures that only a small number of features activate for any given token. Empirically, increasing F and appropriately tuning the sparsity coefficient improves feature specificity, reduces feature interference, and increases the effectiveness of causal interventions such as feature clamping [192, 193].

Together, these components define the dictionary-learning framework used in this study. Each learned feature corresponds to a candidate concept direction in residual-stream activation space, which can subsequently be visualized and quantitatively analyzed to assess how clearly psychological risk factors are represented in the model's internal geometry.

7.2.2 Feature–Topic Mapping

To relate learned feature directions to psychological risk factors, we explicitly define a mapping between sparse autoencoder features and annotated topics. This step is essential for interpreting the geometry of the learned feature space and for attributing semantic meaning to individual feature directions.

Let i index a learned feature direction, and let $a_i(x)$ denote the rescaled activation of feature i for input sample x as defined in Equation 7.5. Each sample x is associated with a psychological topic label $t \in \mathcal{T}$, where $\mathcal{T}$ denotes the set of all annotated topics.

For each feature i and each topic t, we compute the average feature activation over all samples belonging to that topic:

$$\bar{a}_{i,t} \;=\; \mathbb{E}_{x \sim \mathcal{D}_t}[a_i(x)]\,, \tag{7.6}$$

where $\mathcal{D}_t$ denotes the set of samples labeled with topic t. This quantity measures how strongly feature i responds, on average, to content associated with topic t.

We define the *dominant topic* of feature i as the topic for which this average activation is maximal:

$$t^*(i) \;=\; \arg\max_{t \in \mathcal{T}} \; \bar{a}_{i,t}. \tag{7.7}$$

This assignment yields a many-to-one mapping from feature directions to psychological topics. Importantly, this mapping is descriptive rather than exclusive: a feature may respond to multiple topics, but the dominant topic captures the strongest association and provides a principled basis for visualization and analysis.

This feature–topic association is used consistently throughout the remainder of the chapter. In visualization, features are colored according to their dominant topic to reveal geometric structure in the learned representation space.

7.2.3 UMAP for Latent Feature Visualization

While sparse autoencoders allow us to extract interpretable feature directions from the model's residual stream activation space, these directions live in a high-dimensional space

that is difficult to inspect directly. To gain intuition about the organization of the learned feature directions, we employ UMAP, a non-linear dimensionality reduction technique designed to preserve local neighborhood structure while revealing coarse geometric patterns [150].

In this work, UMAP is applied to the set of learned feature directions, represented by the normalized decoder vectors $\tilde{W}^{\text{dec}}_{:,i}$. Each feature direction is therefore treated as a point in residual-stream activation space, and UMAP provides a two-dimensional embedding that reflects similarities between feature directions. UMD serves a primarily exploratory and qualitative role in our methodology. It allows us to visually assess whether feature directions associated with similar psychological topics occupy nearby regions of activation space, indicating structured and selective representations, or whether they are dispersed, suggesting stronger superposition and polysemantic encoding. By comparing UMAP embeddings across models trained on different datasets, we can further examine how data augmentation affects the geometric organization of risk-related feature directions. Importantly, UMAP is not used to define or enforce feature structure, but rather to reveal patterns that emerge naturally from the learned representations.

7.2.4 Cosine Distance for Quantitative Separability Analysis

While UMAP provides an intuitive visualization of relationships between feature directions, it does not offer a quantitative measure of how clearly different psychological concepts are separated in the model's internal representation space. In residual-stream activation space and in sparse autoencoder–derived representations, vector norms can vary across tokens, samples, and datasets. Cosine distance removes the effect of magnitude and isolates whether a concept is encoded along a direction that differs from a non-concept baseline.

Cosine distance is well suited to linear representation settings. If internal representations correspond to approximately linear directions, then the angle between mean representations summarizes how consistently those directions distinguish one set of samples from another. Larger cosine distance indicates that the two mean vectors are oriented further apart in activation space, implying that a simple linear decision boundary could separate them with a larger angular margin. This aligns with the goals of interpretability, as stable directional differences are easier to analyze and relate to specific feature directions recovered by the sparse autoencoder.

To complement the qualitative insights from visualization, we introduce a quantitative analysis based on cosine distance computed in the sparse autoencoder feature space. For each psychological concept, we define a concept set consisting of posts associated with that concept and a corresponding non-concept set drawn from unrelated Reddit content. For each post, token-level residual-stream activations are first mapped into the SAE latent space, yielding a vector of feature activations.

Representations are extracted from the same model layer for both concept and non-concept sets. Within this space, we compute the mean activation vector for the concept set and the mean activation vector for the non-concept set. To obtain post-level feature representations, we aggregate token-level feature activations within each post using a max-pooling operation. For each feature, the post-level activation is defined as the maximum activation across all tokens in the post. This choice reflects the assumption that the presence of a psychologically meaningful feature anywhere in a post is sufficient to indicate its relevance, and avoids dilution of sparse but salient signals that would occur.

For each psychological topic, we then compute a mean post-level activation vector by averaging these post-level representations across all posts in the concept set, yielding μ_{concept}. An analogous mean vector μ_{non} is computed from the non-concept baseline. Larger

distances indicate that concept-related and non-concept content activate distinct sets of interpretable features, reflecting clearer separation in the model's internal representation. Separation between the two is quantified using cosine distance,

$$d_{\cos} \;=\; 1 - \frac{\mu_{\text{concept}} \cdot \mu_{\text{non}}}{\|\mu_{\text{concept}}\|_2 \, \|\mu_{\text{non}}\|_2}. \tag{7.8}$$

This construction follows the core idea introduced in TCAV, where concepts are defined using example sets and represented by averaged directions in activation space; however, instead of measuring sensitivity of model outputs to these directions, we use cosine distance to quantify their geometric separability within the model's internal activation space. Prior work in representation analysis and concept-based interpretability has similarly relied on cosine similarity or distance to assess alignment between learned representations and semantic concepts [80, 87].

This measure focuses on direction rather than length and is therefore insensitive to scaling and normalization differences across datasets and layers. Larger values indicate greater separation between the concept and the non-concept baseline. We compute $d_{\cos}$ for both the UMD fine-tuned model and the AUG fine-tuned model, each evaluated on the same mixed synthetic test set, and repeat this procedure for every topic to obtain a cosine distance per topic.

7.3 Experimental Setup and Data

We train an overcomplete SAE on residual-stream activation vectors extracted from our classifier backbone. This work adopts dictionary learning to probe the model's internal representations. Specifically, we train a sparse autoencoder on token-level residual-stream

activations from a fine-tuned ALBERT classifier in order to recover an overcomplete set of feature directions that can be inspected and analyzed. ALBERT is fine-tuned on the UMD suicidal ideation dataset and Augmented dataset, after which token-level residual-stream activations are extracted from the final transformer layer.

We focus on the final-layer residual stream because it captures high-level semantic representations that are most relevant to the downstream suicide ideation detection task, while remaining upstream of the classification head. This choice allows us to analyze psychologically meaningful directions that the model uses for decision-making, without conflating them with task-specific output weights. Prior work in mechanistic interpretability has similarly shown that late-layer residual streams contain rich, interpretable features aligned with abstract concepts, making them suitable targets for sparse dictionary learning [87, 88].

We adopt the encoder–decoder architecture described in Section 7.2.1. An overcomplete feature dictionary is used, with $F = 4096$ learned feature directions for input activation vectors of dimension $D = 768$. This expansion increases the representational capacity of the dictionary and improves the separation of overlapping directions that arise due to superposition, allowing more specific and interpretable feature directions to emerge. The token dimension is flattened and filtered using the attention mask so that only non-padding token activations are fed to the sparse autoencoder. The encoder consists of a linear map followed by a ReLU nonlinearity, yielding non-negative and selective feature activations. The decoder is a linear map that reconstructs the original activation vector from a sparse linear combination of feature directions. Under this formulation, the decoder columns represent candidate feature directions, while the encoder outputs indicate how strongly each feature direction is activated for a given token.

Optimization is performed using the AdamW optimizer, which provides stable convergence

and improved control over model capacity by decoupling weight decay from gradient updates. Early stopping is applied based on validation reconstruction loss to prevent overfitting once learning plateaus. After training, feature activations are computed for all tokens in the dataset and aggregated at the sample (post) level by taking the maximum activation per feature across tokens within each text. This aggregation reflects the assumption that a feature is present in a sample if it activates strongly for at least one token.

For each learned feature direction, we then select the top 100 samples with the highest feature activations. This procedure follows established practice in sparse autoencoder interpretability work, where a fixed number of top-activating examples has been shown to be sufficient for reliably characterizing feature semantics while remaining computationally tractable [88, 87].

7.4 Analysis I: Latent Feature Visualization

A central question in this analysis is whether suicide-related psychological risk factors such as hopelessness, anxiety, and social isolation form coherent geometric structure in the model's learned feature directions. We also ask whether a topic-aware data augmentation improves the coherency of these features.

For each model—ALBERT fine-tuned on the UMD dataset and ALBERT fine-tuned on the augmented dataset—we assign a dominant psychological topic to each learned feature direction using the following procedure. First, for a given feature, we collect its activation values across all posts in the dataset. Second, posts are grouped according to their annotated psychological topic. For each topic, we compute the average feature activation over all posts belonging to that topic. The dominant topic of the feature is defined as the topic for which

this average activation is highest. This assignment provides a coarse but interpretable mapping between learned feature directions and psychological risk factors.

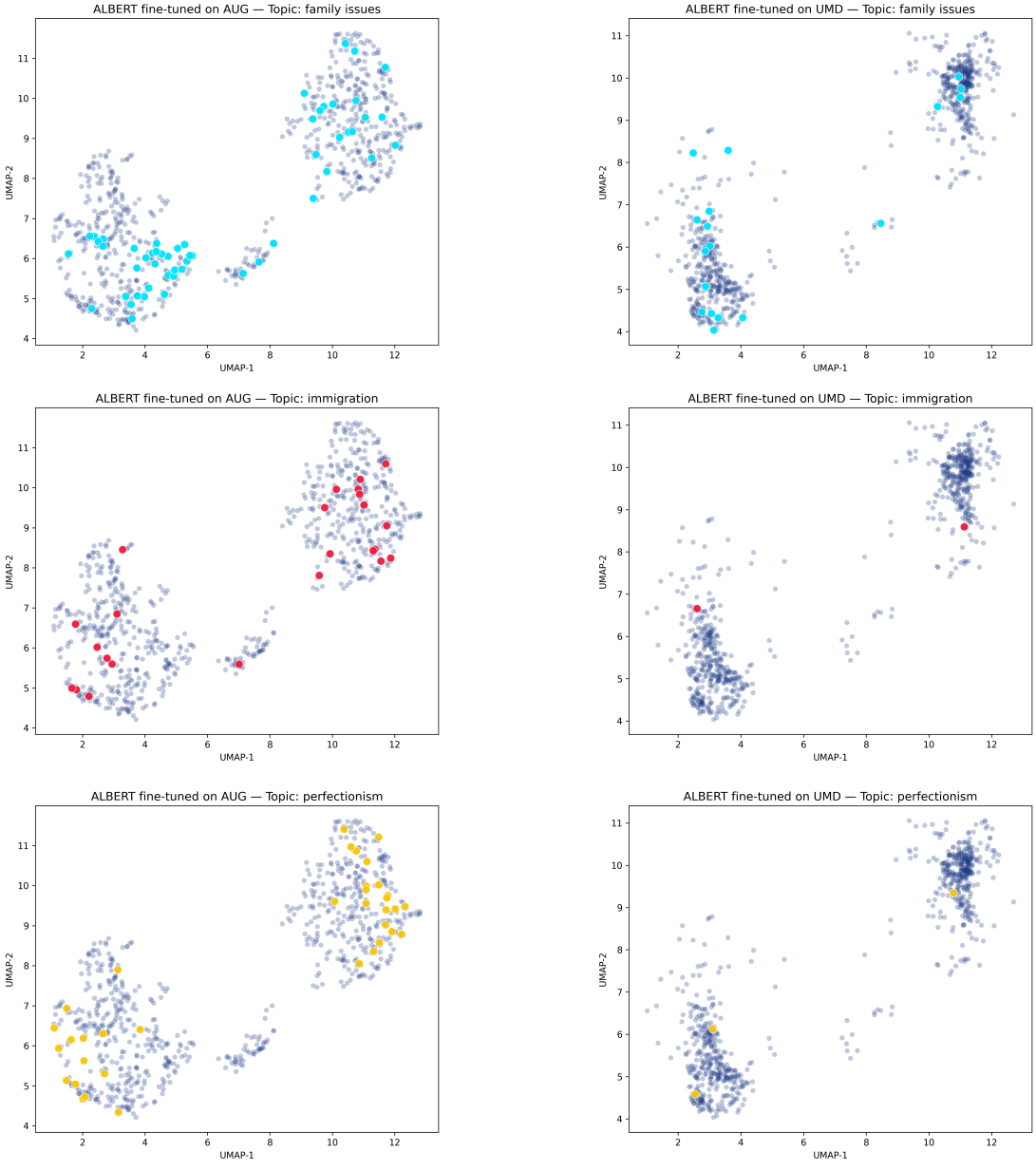
Throughout this chapter, sparse autoencoders are trained on token-level residual-stream activations in order to capture fine-grained linguistic features. However, psychological topics are defined at the post level rather than at the token level. To bridge this gap, token-level feature activations are aggregated within each post using a max-pooling operation, yielding a post-level feature activation vector. All topic assignments, visualizations, and quantitative analyses reported in this chapter are therefore performed at the post level, unless explicitly stated otherwise.

After assigning dominant topics, the full set of feature directions (decoder vectors) is projected into two dimensions using UMAP. In Figure 7.1, we visualize the geometry of the feature space one topic at a time by highlighting features whose dominant topic matches the topic of interest, while rendering all remaining feature directions as a low-opacity background. Compact highlighted regions indicate feature directions that respond selectively to the same psychological concept, suggesting movement toward monosemantic structure. In contrast, dispersed highlighted features suggest shared or overlapping representations, consistent with superposition and polysemantic encoding.

Additional topic-wise visualizations are provided in Appendix .2. For each topic, we present paired UMAP panels, with the AUG-trained model shown on the left and the UMD-trained model on the right, enabling direct comparison of how data augmentation affects the geometric organization of risk-related feature directions.

In Figure 7.1, each point corresponds to a learned feature direction, represented by its decoder vector in the model's residual-stream space. The colored points indicate features whose dominant activations are associated with the indicated psychological topic. In the

Figure 7.1: UMAP projections of features highlighting three selected topics. Left column: ALBERT fine-tuned on AUG (the augmented dataset with high topic coverage). Right column: ALBERT fine-tuned on UMD. Colored points mark features whose dominant responses align with the indicated topic; dark blue points show all other features.



UMD fine-tuned model, topics such as perfectionism and immigration activate only a small number of scattered feature directions, suggesting limited topic coverage and stronger superposition, as the same directions appear to respond to multiple phenomena.

In contrast, the augmented fine-tuned model exhibits a larger number of highlighted feature directions for the same topics, which often form compact regions within the two macro-clusters of the embedding. These regions indicate sets of directions that respond more selectively to the target topic. The family issues topic follows a similar pattern, with sparse and dispersed feature directions in the UMD model and denser, more coherent groupings in the AUG model. This shift suggests that data augmentation improves both the availability and the consistency of topic-specific directions in the learned representation space.

Monosemanticity is valuable because a compact, well-separated group of topic-selective features supports clear explanations, stable linear steering, and predictable causal edits. However, polysemanticity is not inherently undesirable. Distributed highlights can reflect useful sharing of linguistic substrates, overlapping risk factors, or genuine subtopics that co-occur in real text. Such sharing can improve generalization and parameter efficiency when concepts partially overlap. The practical concern is interference rather than polysemanticity itself; if the same features fire for unrelated content, the representation becomes noisy and separation weakens. The visualizations indicate that augmentation moved several topics toward monosemanticity by increasing the number and coherence of topic-selective directions, while still preserving beneficial sharing where concepts naturally overlap.

7.5 Analysis II: Quantifying Concept Separability

The goal of this section is to quantify how clearly each psychological topic is separated from generic, non-risk-related text in the model's internal activation space. We use cosine distance as a measure of separability because it captures directional differences between mean activation vectors associated with concept-related and non-concept-related content. If the mean vectors corresponding to two sets of points in noticeably different directions, this indicates that the model encodes them as distinct internal representations. Because cosine distance depends only on direction, it is robust to changes in vector magnitude caused by scaling, preprocessing, or dataset-specific norm shifts.

We analyze token-level activation vectors extracted from the residual stream of the final transformer block. For post-level analysis, token-level activation vectors within each post are averaged to obtain a single activation vector per post. We evaluate two models that are fine-tuned separately: one model fine-tuned on the UMD dataset and one model fine-tuned on the augmented dataset. All measurements are computed on a held-out test set consisting of a mixture of synthetic posts annotated with psychological topic labels. No training or validation data are used in these analyses.

In addition to the synthetic dataset with topic labels, we use a separate Reddit test set introduced in Section 7.3, which contains content unrelated to suicidal ideation. This Reddit set is treated as the non-concept baseline. We merge the synthetic topic-labeled posts and the Reddit posts to form the final test set used for analysis. For each psychological topic, we define the **concept set** as all test posts labeled with that topic, and the **non-concept set** as the Reddit test set.

Figure 7.2: Topic separation on the synthetic test set. Left and middle panels show cosine distance between each topic's mean activation vector and a non-concept Reddit baseline for the UMD- and AUG-fine-tuned models (shared scale; higher indicates greater separation). The right panel shows the change $\Delta = \text{AUG} - \text{UMD}$ (diverging scale centered at 0), with topics sorted by Δ.

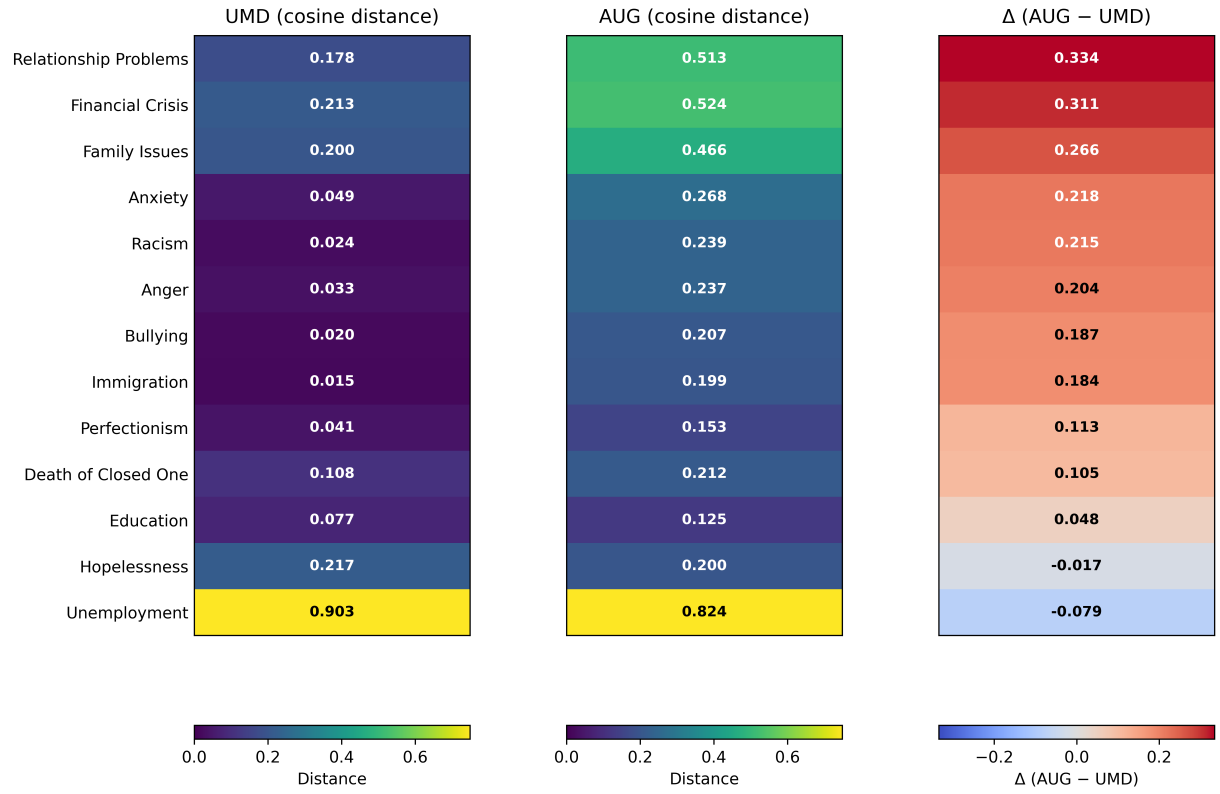


Figure 7.2 summarizes topic separation on the synthetic test set by showing the cosine distance between each topic's mean SAE feature activation vector and a non-concept Reddit baseline. Higher cosine distance indicates stronger directional separation of a topic from general Reddit content, reflecting more selective internal representations. The left and middle panels report these distances for models fine-tuned on the original UMD data and on the UMD data, respectively, using a shared color scale. The right panel visualizes the

change induced by augmentation, defined as $\Delta = \text{AUG} - \text{UMD}$, with topics sorted by their change in separation.

Across most topics, augmentation leads to a clear increase in directional separation. The largest gains are observed for relationship problems (+0.334), financial crisis (+0.311), and family issues (+0.266), indicating that augmentation substantially sharpens the representation of these psychosocial factors relative to general Reddit content. Additional consistent improvements are visible for anxiety, racism, anger, bullying, and immigration (approximately +0.18 to +0.22), suggesting that topic-aware augmentation broadly improves the model's ability to isolate diverse risk-related themes in its internal geometry.

More moderate gains are observed for perfectionism (+0.113), death of a close one (+0.105), depression (+0.092), and education (+0.048). These smaller increases indicate partial improvement in separability, consistent with topics that are either less lexically distinctive or more diffusely expressed across posts.

Two topics deviate from the overall trend. Hopelessness shows a slight decrease (−0.017), likely reflecting its strong semantic overlap with other risk factors such as depression and anxiety, which limits how distinctly it can be separated as an independent direction. Unemployment exhibits a larger decrease (−0.079); however, this topic already displays very high separation in both models (approximately 0.82–0.90). Because its baseline separation is near the upper bound of the scale, there is limited room for further improvement, and small fluctuations appear as a decrease despite the topic remaining strongly isolated from the non-concept baseline.

7.6 Conclusion

This chapter examined how models trained for suicide ideation detection represent psychologically meaningful topics and how those representations change under topic-aware data augmentation. Grounded in the linear representation and superposition hypotheses, we applied sparse dictionary learning via an overcomplete sparse autoencoder to probe internal activations, and paired this approach with representation-level analyses using UMAP projections and cosine distance between concept and non-concept means. Together, these tools allowed us to assess not only which psychological topics are encoded in the model's latent geometry, but also how selectively and distinctly they are represented.

Three findings stand out. First, the UMAP visualizations showed that models trained with topic-aware augmentation activate a larger number of feature directions that cluster more compactly for a given topic. This pattern suggests a shift toward more topic-selective structure, consistent with movement away from heavy superposition and toward increasingly monosemantic representations. Second, this qualitative change is mirrored by quantitative results: augmentation generally increased the directional separation between topic means and a general Reddit baseline in the shared test space. The largest gains were observed for relationship problems, financial crisis, and family issues, with consistent improvements across several other psychosocial topics. Third, a small number of topics behaved differently. Representations of hopelessness changed little, likely due to strong semantic overlap with other risk factors, while unemployment exhibited a decrease from an already high baseline, suggesting a ceiling effect rather than a degradation of signal.

Taken together, these results highlight an important nuance in interpretability. While monosemantic features are desirable for explanation and targeted intervention, polysemantic sharing is not inherently problematic. When psychological topics co-occur or rely on

overlapping linguistic cues, some degree of shared representation may support generalization and efficient use of model capacity. The central concern is therefore not superposition itself, but harmful interference that obscures concept boundaries. Our findings indicate that topic-aware augmentation reduces such interference for many risk factors while preserving useful representational overlap where concepts are naturally related. This balance improves both the coverage and separability of psychological topics in the model's internal geometry, strengthening the interpretability of suicide ideation detection systems without enforcing overly rigid, one-feature-per-concept representations.

Chapter 8

Conclusion and Future Works

8.1 Conclusion

The accurate identification of suicidal ideation from textual data is critical for enabling early intervention and prevention efforts. While Natural Language Processing (NLP) techniques have shown promise in this domain, the scarcity and sensitivity of suicide-related data remain major challenges. Collecting and annotating suicide-related content from social media is both resource-intensive and ethically sensitive, which motivates exploration of alternative modeling strategies.

First: We evaluated the extent to which large language models (LLMs) can mitigate data accessibility and annotation challenges in suicide ideation detection on social media. Specifically, we explored state-of-the-art LLMs in Zero-shot and Few-shot settings as predictive models, positioning them as potential alternatives to classifiers trained on manually annotated datasets. Our results show that GPT-3.5, in the Zero-shot setting, achieved an F1-score of 0.73, indicating that such models capture non-trivial signals relevant

to suicidality and may be useful as auxiliary tools, for example, in data exploration or weak supervision.

However, despite this promising performance, GPT-3.5 consistently under-performs classifiers that are fine-tuned on task-specific, high-quality annotated data. This gap highlights that, while generative LLMs can provide useful prior knowledge, fine-tuning remains necessary to achieve the level of accuracy required for reliable suicide ideation detection. These findings underscore the continued importance of curated training datasets and specialized models for this high-stakes application.

Second: We used the guided topic modeling technique to assess the quality of existing datasets collected from social media. The main goal of this assessment was to uncover the extent to which social media datasets mirror or diverge from conventional psychological understandings about suicide-related topics. We extracted risk factors linked to suicide, such as mental health challenges, relationship conflicts, and financial distress, to establish a field-related baseline grounded in psychology for our research. Our results demonstrate that while the prevalent and common risk factors, such as depression and anxiety, are over-represented in datasets collected from social media, more stigmatized topics, such as racism and immigration challenges, are entirely absent in these datasets. These results highlight the necessity of creating more diverse datasets that cover the risk factors related to under-represented social groups since such disparities in topic representations could lead to an inherent bias within the models. These biases will potentially impact the effectiveness and fairness of NLP models, as they may disproportionately emphasize or neglect certain aspects of suicidal ideation.

Third: We introduced a strategy that leverages the capabilities of generative AI models, including GPT-3.5, Flan-T5, and LLaMA2, to create synthetic data for suicidal ideation detection in order to address challenges in accessing diverse real-world examples. Our data

generation approach is grounded in social risk factors extracted from psychology literature and aims to ensure coverage of essential psychosocial contexts associated with suicidal ideation. We used the resulting synthetic datasets to fine-tune ALBERT language models for suicide detection.

Our results indicate that synthetic data generated by GPT-3.5 yielded the highest-quality training set, achieving an F1-score of 82% when evaluated on real-world test data. We further analyzed the generated datasets and compared them with real datasets. This analysis shows that the synthetic data is more balanced in terms of risk factor coverage. However, it is not significantly different from real data in terms of textual complexity and readability as measured by lexical complexity and Flesch Reading Ease readability metrics. At the same time, the diversity of vocabulary in the synthetic dataset is significantly lower than in the real dataset.

Fourth: Real-world data is often limited in size and topic coverage, leading to over-fitting and reduced model performance. We first used synthetic data as an alternative to real training data and showed that synthetic datasets created with balanced risk factors are more effective than those generated without incorporating risk factors. Then, by carefully blending synthetic data with real data, we improved the model's performance by gradually augmenting the best synthetic dataset with increasing portions of the real dataset. Our results show that models trained with synthetic data augmented with only 30% of the real data surpass the models fine-tuned with the full real-world dataset, which proved to be a cost-effective approach to training models. Our data augmentation results show that incorporating synthetic data into the training pipeline helps diversify the dataset and enhance model generalization. As a result, we improved the model's performance from 0.82 when trained with synthetic data and 0.87 when trained with real data up to 0.88 when trained with the cost-effective and topic-diverse augmented dataset.

Fifth: We investigated the internal processes of suicide detection models by applying mechanistic interpretability techniques to ALBERT models fine-tuned on both real and augmented datasets. Using sparse autoencoders and geometric analyses such as UMAP projections and cosine distance, we examined whether topic-aware augmentation reshapes the internal feature representations of the classifier. Our findings show that augmentation leads to clearer and more topic-selective latent structures. Topics that were previously weakly encoded or heavily entangled, such as immigration and family issues, form denser and more coherent feature clusters under the augmented model. Cosine distance analysis also demonstrates that for most topics, augmentation increases the directional separation between concept and non-concept representations, indicating stronger conceptual encoding and reduced interference. These results suggest that topic-aware augmentation not only improves predictive performance but also results in internal representations that better reflect meaningful psychological constructs.

8.2 Limitations

While this thesis presents a comprehensive framework that combines synthetic data generation with representation-level interpretability, several considerations provide opportunities for further refinement and future investigation.

Although synthetic data significantly improves topic coverage and supports strong model performance, it does not fully replicate the richness and variability of real-world language. As observed in our analysis, synthetic datasets tend to exhibit lower lexical diversity despite increased structural complexity. This characteristic highlights an opportunity for future work to further enhance linguistic diversity while preserving the structured coverage of psychological risk factors achieved in this study.

The generative models used in this work, including GPT-3.5, Flan-T5, and LLaMA2, represent a specific generation of large language models. With the rapid advancement of more recent models offering improved reasoning and alignment capabilities, it is likely that both classification performance and synthetic data quality can be further improved. Therefore, the findings of this thesis provide a strong foundation for evaluating and extending these approaches using newer architectures.

The annotation process for the synthetic test set was conducted by human annotators who, while experienced in text analysis, do not have formal clinical training in psychology. Annotators were provided with clear labeling guidelines and achieved a high level of agreement, which supports the overall reliability of the dataset. At the same time, incorporating domain experts in future studies could further strengthen the validity of annotations, particularly for more nuanced or ambiguous cases.

In addition, the use of a synthetic test set enables controlled evaluation of under-represented psychological topics that are difficult to assess using real-world datasets alone. While this provides valuable insights into model behavior across diverse scenarios, synthetic data may not fully capture the variability and complexity of real-world language. For this reason, results obtained on synthetic test sets are complemented with evaluations on real datasets throughout this work.

Finally, ethical considerations remain an important aspect of this research. Both real and synthetic datasets may reflect biases related to demographic, cultural, or linguistic factors, and generative models may introduce or amplify such biases during data generation. By explicitly incorporating topic diversity and structured generation strategies, this work takes initial steps toward mitigating these issues. Nevertheless, continued efforts in fairness-aware evaluation, bias detection, and human-in-the-loop validation will be essential for ensuring responsible deployment in real-world mental health applications.

8.3 Future Works

This thesis explored suicide ideation detection from multiple complementary perspectives, including the use of large language models for classification, synthetic data generation to address dataset limitations, and representation-level analysis to better understand learned decision signals. While the results demonstrate the effectiveness of topic-aware augmentation and highlight the potential of both generative and discriminative models, several important directions remain for future research.

A natural and immediate extension of this work is to evaluate more recent generations of large language models for both suicide ideation classification and synthetic data generation. Future work should also investigate the fine-tuning of more recent and powerful large language models, such as GPT-4-class or newer architectures, to assess whether improved reasoning and representation capabilities can further enhance both classification performance and synthetic data quality. This thesis primarily examined models such as GPT-3.5, Flan-T5, and LLaMA2; however, newer models with substantially improved reasoning and language understanding capabilities may yield significant gains. Future studies could assess whether state-of-the-art LLMs achieve higher accuracy in Zero-shot or Few-shot classification settings, produce higher-quality and more diverse synthetic training data, or reduce the gap between generative models and fine-tuned classifiers. Such evaluations would help determine whether advances in foundation models can meaningfully alleviate data scarcity and annotation challenges in this domain.

While the synthetic datasets generated in this thesis improved risk factor coverage and supported effective model training, limitations in lexical diversity remain. Future work could explore prompt engineering strategies, controlled generation techniques, or multi-stage generation pipelines to increase linguistic variability while preserving psychological validity.

Additionally, combining synthetic data from multiple generative models or incorporating human-in-the-loop validation may further improve realism and robustness. Evaluating synthetic data across a wider range of downstream tasks and datasets would also help establish its generalizability.

Interpretability played a supporting role in this thesis by providing insight into how topic-aware augmentation influences internal representations. Future work could build on this analysis by introducing causal interventions on learned features, such as clamping or amplifying sparse autoencoder activations, to assess their direct impact on model predictions. This would strengthen the connection between descriptive interpretability and functional model behavior, helping distinguish features that are merely correlated with suicidality from those that actively influence decisions.

The analyses in this thesis focused on individual posts treated as independent samples. However, suicidal ideation often manifests through evolving language patterns over time. Future research could extend both classification and representation analyses to longitudinal user-level data, examining how risk-related signals accumulate or change across sequences of posts. This may enable earlier detection of emerging risk and provide a more realistic modeling framework for real-world applications.

Another important direction for future work lies in translating the findings of this thesis into actionable insights for different stakeholders. For practitioners developing and deploying AI systems in mental health settings, future work could explore how topic-aware data augmentation strategies can be systematically integrated into real-world pipelines to improve model robustness and coverage of diverse user experiences. Additionally, the interpretability framework proposed in this thesis, based on sparse feature representations, could be further developed into practical tools that support model auditing and debugging in safety-critical applications.

In the context of mental health applications, future work should continue to emphasize human-centered and ethical deployment. In particular, integrating these models into expert-in-the-loop systems, where clinicians are supported rather than replaced, remains a critical consideration. Enhancing interpretability and transparency can play a central role in building trust and enabling responsible use of AI systems in sensitive domains such as suicide ideation detection.

Finally, future work should examine how both classifiers and synthetic-data-driven models perform across different populations, platforms, and linguistic styles. Assessing robustness to domain shifts and potential demographic biases is particularly important in mental health applications. Incorporating fairness-aware evaluation alongside performance and interpretability analyses would help ensure that improvements in accuracy do not come at the expense of equitable and responsible deployment.

References

[1] S. Ghosal and A. Jain, "Depression and suicide risk detection on social media using fasttext embedding and xgboost classifier," *Procedia Computer Science*, vol. 218, pp. 1631–1639, 2023.

[2] A. C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, and D. Chandran, "Identifying suicidal ideation and suicide attempts in a psychiatric clinical research database using natural language processing," *Scientific Reports*, vol. 8, no. 1, p. 7426, 2018.

[3] C. A. Bejan, M. Ripperger, D. Wilimitis, R. Ahmed, J. Kang, K. Robinson, T. J. Morley, D. M. Ruderfer, and C. G. Walsh, "Improving ascertainment of suicidal ideation and suicide attempt with natural language processing," *Scientific Reports*, vol. 12, no. 1, p. 15146, 2022.

[4] M. J. Vioules, B. Moulahi, J. Azé, and S. Bringay, "Detection of suicide-related posts in twitter data streams," *IBM Journal of Research and Development*, vol. 62, no. 1, pp. 1–7, 2018.

[5] A. Abdulsalam and A. Alhothali, "Suicidal ideation detection on social media: A review of machine learning methods," *arXiv*, vol. abs/2201.10515, 2022.

[6] R. Babbar and B. Schölkopf, "Data scarcity, robustness and extreme multi-label classification," *Machine Learning*, vol. 108, no. 8–9, pp. 1329–1351, 2019.

[7] S. I. Nikolenko, *Synthetic Data for Deep Learning*. Springer, 2021.

[8] Y. Lu, H. Wang, and W. Wei, "Machine learning for synthetic data generation: A review," *arXiv*, vol. abs/2302.04062, 2023.

[9] H. Chau, S. Balaneshin, K. Liu, and O. Linda, "Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction," in *Proceedings of the 14th Linguistic Annotation Workshop*, pp. 74–86, 2020.

[10] L. Grattidge, H. Hoang, J. Mond, D. Lees, D. Visentin, and S. Auckland, "Exploring community-based suicide prevention in the context of rural australia: A qualitative study," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, p. 2644, 2023.

[11] L. Liu, N. J. Pollock, G. Contreras, L. Tonmyr, and W. Thompson, "Pandemic-related impacts and suicidal ideation among adults in canada: A population-based cross-sectional study," *Health Promotion and Chronic Disease Prevention in Canada*, vol. 43, no. 3, p. 105, 2023.

[12] J. C. Franklin, J. D. Ribeiro, K. R. Fox, K. H. Bentley, E. M. Kleiman, X. Huang, K. M. Musacchio, A. C. Jaroszewski, B. P. Chang, and M. K. Nock, "Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research," *Psychological Bulletin*, vol. 143, no. 2, p. 187, 2017.

[13] K. H. Bentley, J. C. Franklin, J. D. Ribeiro, E. M. Kleiman, K. R. Fox, and M. K. Nock, "Anxiety and its disorders as risk factors for suicidal thoughts and behaviors: A meta-analytic review," *Clinical Psychology Review*, vol. 43, pp. 30–46, 2016.

[14] L. Orsolini, R. Latini, M. Pompili, G. Serafini, U. Volpe, F. Vellante, M. Fornaro, A. Valchera, C. Tomasetti, S. Fraticelli, *et al.*, "Understanding the complexity of suicide in depression: From research to clinical practice," *Psychiatry Investigation*, vol. 17, no. 3, p. 207, 2020.

[15] N. H. Kalin, "Insights into suicide and depression," *American Journal of Psychiatry*, pp. 877–880, 2020.

[16] L. da Silva Costa, Á. P. Alencar, P. J. N. Neto, M. d. S. V. dos Santos, C. G. L. da Silva, S. d. F. L. Pinheiro, R. T. Silveira, B. A. V. Bianco, R. F. F. P. Júnior, M. A. P. de Lima, *et al.*, "Risk factors for suicide in bipolar disorder: A systematic review," *Journal of Affective Disorders*, vol. 170, pp. 237–254, 2015.

[17] J. Paris, "Suicidality in borderline personality disorder," *Medicina*, vol. 55, no. 6, p. 223, 2019.

[18] K. H. Lee, J. S. Jun, Y. J. Kim, S. Roh, S. S. Moon, N. Bukonda, and L. Hines, "Mental health, substance abuse, and suicide among homeless adults," *Journal of Evidence-Informed Social Work*, vol. 14, no. 4, pp. 229–242, 2017.

[19] C. Okolie, M. Dennis, E. S. Thomas, and A. John, "A systematic review of interventions to prevent suicidal behaviors and reduce suicidal ideation in older people," *International Psychogeriatrics*, vol. 29, no. 11, pp. 1801–1824, 2017.

[20] E. M. Kleiman, B. J. Turner, S. Fedor, E. E. Beale, J. C. Huffman, and M. K. Nock, "Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies," *Journal of Abnormal Psychology*, vol. 126, no. 6, p. 726, 2017.

[21] N. Leigh-Hunt, D. Bagguley, K. Bash, V. Turner, S. Turnbull, N. Valtorta, and W. Caan, "An overview of systematic reviews on the public health consequences of social isolation and loneliness," *Public Health*, vol. 152, pp. 157–171, 2017.

[22] J. Holt-Lunstad, T. B. Smith, M. Baker, T. Harris, and D. Stephenson, "Loneliness and social isolation as risk factors for mortality: A meta-analytic review," *Perspectives on Psychological Science*, vol. 10, no. 2, pp. 227–237, 2015.

[23] A. Allchin, V. Chaplin, and J. Horwitz, "Limiting access to lethal means: Applying the social ecological model for firearm suicide prevention," *Injury Prevention*, vol. 25, no. Suppl 1, pp. i44–i48, 2019.

[24] K. A. Van Orden, T. K. Witte, K. C. Cukrowicz, S. R. Braithwaite, E. A. Selby, and T. E. Joiner Jr, "The interpersonal theory of suicide," *Psychological Review*, vol. 117, no. 2, p. 575, 2010.

[25] A. Wenzel and A. T. Beck, "A cognitive model of suicidal behavior: Theory and treatment," *Applied and Preventive Psychology*, vol. 12, no. 4, pp. 189–201, 2008.

[26] R. J. Cramer and N. D. Kapusta, "A social-ecological framework of theory, assessment, and prevention of suicide," *Frontiers in Psychology*, vol. 8, p. 1756, 2017.

[27] R. C. Hsiung, "A suicide in an online mental health support group: Reactions of the group members, administrative responses, and recommendations," *CyberPsychology & Behavior*, vol. 10, no. 4, pp. 495–500, 2007.

[28] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle, "Tracking suicide risk factors through twitter in the us," *Crisis*, 2014.

[29] G. B. Colombo, P. Burnap, A. Hodorog, and J. Scourfield, "Analysing the connectivity and communication of suicidal users on twitter," *Computer Communications*, vol. 73, pp. 291–300, 2016.

[30] D. Sikander, M. Arvaneh, F. Amico, G. Healy, T. Ward, D. Kearney, E. Mohedano, J. Fagan, J. Yek, A. F. Smeaton, *et al.*, "Predicting risk of suicide using resting state heart rate," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, IEEE, 2016.

[31] N. Jiang, Y. Wang, L. Sun, Y. Song, and H. Sun, "An erp study of implicit emotion processing in depressed suicide attempters," in *Proceedings of the 7th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 37–40, IEEE, 2015.

[32] W.-C. Chiang, P.-H. Cheng, M.-J. Su, H.-S. Chen, S.-W. Wu, and J.-K. Lin, "Socio-health with personal mental health records: Suicidal-tendency observation system on facebook for taiwanese adolescents and young adults," in *Proceedings of the IEEE 13th International Conference on e-Health Networking, Applications and Services*, pp. 46–51, IEEE, 2011.

[33] R. S. Skaik and D. Inkpen, "Predicting depression via automatic questionnaire filling," *IEEE Access*, vol. 10, pp. 102033–102047, 2022.

[34] D. Inkpen, R. Skaik, P. Buddhitha, D. Angelov, and M. T. Fredenburgh, "uottawa at erisk 2021: Automatic depression questionnaire filling," 2021.

[35] K. R. Chowdhary, "Natural language processing for word sense disambiguation and information extraction," *arXiv*, vol. abs/2004.02256, 2020.

[36] K. S. Babulal and B. K. Nayak, "Suicidal analysis on social networks using machine learning," in *Internet of Medical Things (IoMT) and Telemedicine Frameworks*, pp. 230–247, IGI Global, 2023.

[37] K. D. Varathan and N. F. Q. Abu Talib, "Suicide detection system based on twitter," in *Science and Information Conference*, pp. 785–788, 2014.

[38] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression detection," *CLPsych Workshop*, p. 32, 2017.

[39] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," in *Proceedings of CLPsych*, pp. 88–97, 2018.

[40] R. Skaik and D. Inkpen, "Using twitter for depression detection in canada," in *Artificial Intelligence and Cloud Computing Conference*, pp. 109–114, 2020.

[41] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, W. Alnumay, and A. P. Smith, "A lexicon-based approach to detecting suicide-related messages on twitter," *Biomedical Signal Processing and Control*, vol. 65, p. 102355, 2021.

[42] J. H. K. Seah and K. J. Shim, "Data mining approach to the detection of suicide in social media: A case study of singapore," in *IEEE International Conference on Big Data*, pp. 5442–5444, 2018.

[43] K. P. Linthicum, K. M. Schafer, and J. D. Ribeiro, "Machine learning in suicide science: Applications and ethics," *Behavioral Sciences & the Law*, vol. 37, no. 3, pp. 214–222, 2019.

[44] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury, "Detecting changes in suicide content manifested in social media following celebrity suicides," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 85–94, 2015.

[45] Y. Chishima and I.-T. H.-C. Liu, "Mental health during the covid-19 pandemic in japan: Applying topic modeling in daily life descriptions," *International Journal of Mental Health and Addiction*, pp. 1–20, 2021.

[46] Y. Hua, F. Liu, K. Yang, Z. Li, Y.-H. Sheu, P. Zhou, L. V. Moran, S. Ananiadou, and A. Beam, "Large language models in mental health care: A scoping review," *arXiv*, vol. abs/2401.02984, 2024.

[47] A. Radford, I. Sutskever, R. Child, G. Krueger, and J. W. Kim, "Chat with gpt: Improving language generation and task-oriented dialogue." https://openai.com/blog/chatgpt-plus, 2021. Online.

[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019.

[49] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[50] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, *et al.*, "So what if chatgpt wrote it? multidisciplinary perspectives on generative conversational ai," *International Journal of Information Management*, vol. 71, p. 102642, 2023.

[51] N. Lambert, L. Castricato, L. von Werra, and A. Havrilla, "Illustrating reinforcement learning from human feedback (rlhf)," *Hugging Face Blog*, 2022. https://huggingface.co/blog/rlhf.

[52] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is chatgpt to human experts? comparison, evaluation, and detection," *arXiv*, vol. abs/2301.07597, 2023.

[53] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, and J. Ricke, "Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports," *arXiv*, vol. abs/2212.14882, 2022.

[54] A. Gilson, C. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash, "How well does chatgpt perform on medical licensing exams? implications for medical education," *medRxiv*, 2022.

[55] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, *et al.*, "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," *arXiv*, vol. abs/2302.04023, 2023.

[56] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of ChatGPT and emotion-enhanced prompting for mental health analysis," *arXiv*, vol. abs/2304.03347, 2023.

[57] P. P. Sinha, R. Mishra, R. Sawhney, D. Mahata, R. R. Shah, and H. Liu, "#suicidal: A multipronged approach to identify and explore suicidal ideation in twitter," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 941–950, 2019.

[58] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak, "Knowledge-aware assessment of severity of suicide risk for early intervention," in *Proceedings of the World Wide Web Conference*, pp. 514–525, 2019.

[59] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead, "CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, (Minneapolis), June 2019.

[60] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 25–36, 2018.

[61] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Predicting risk of suicide attempts over time through machine learning," *Clinical Psychological Science*, vol. 6, no. 3, pp. 345–361, 2018.

[62] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, vol. 34, no. 13, pp. 10309–10319, 2022.

[63] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," *arXiv*, vol. abs/2110.15621, 2021.

[64] X. He, I. Nassar, J. Kiros, G. Haffari, and M. Norouzi, "Generate, annotate, and learn: NLP with synthetic text," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 826–842, 2022.

[65] A. Rashid, V. Lioutas, and M. Rezagholizadeh, "MATE-KD: Masked adversarial text, a companion to knowledge distillation," *arXiv*, vol. abs/2105.05912, 2021.

[66] L. Bonifacio, H. Abonizio, M. Fadaee, and R. Nogueira, "Inpars: Unsupervised dataset generation for information retrieval," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2387–2392, 2022.

[67] E. Balkır, S. Kiritchenko, I. Nejadgholi, and K. C. Fraser, "Challenges in applying explainability methods to improve fairness in nlp," *arXiv*, vol. abs/2206.03945, 2022.

[68] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of KDD*, pp. 1135–1144, 2016.

[69] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[70] E. Balkır, I. Nejadgholi, K. C. Fraser, and S. Kiritchenko, "Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection," *arXiv*, vol. abs/2205.03302, 2022.

[71] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning*, pp. 1885–1894, 2017.

[72] Q. Zhang, Y. Wang, J. Cui, X. Pan, Q. Lei, S. Jegelka, and Y. Wang, "Beyond interpretability: Monosemanticity and model robustness," *arXiv preprint arXiv:2410.21331*, 2024.

[73] C.-K. Yeh, J. Kim, I. E. H. Yen, and P. Ravikumar, "Representer point selection for explaining deep neural networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[74] S. O. Arik and T. Pfister, "Protoattend: Attention-based prototypical learning," *Journal of Machine Learning Research*, vol. 21, no. 210, pp. 1–35, 2020.

[75] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.

[76] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Contrastive explanations with pertinent negatives," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[77] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

[78] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[79] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20554–20565, 2020.

[80] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Testing with concept activation vectors," in *International Conference on Machine Learning*, pp. 2668–2677, 2018.

[81] I. Nejadgholi, E. Balkır, K. C. Fraser, and S. Kiritchenko, "Towards procedural fairness: Uncovering biases in toxic language classifiers," *arXiv*, vol. abs/2210.10689, 2022.

[82] I. Nejadgholi, K. C. Fraser, and S. Kiritchenko, "Improving generalizability in implicitly abusive language detection with concept activation vectors," *arXiv*, vol. abs/2204.02261, 2022.

[83] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, 2017.

[84] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, vol. 3, no. 3, p. e10, 2018.

[85] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, "Zoom in: An introduction to circuits," *Distill*, vol. 5, no. 3, p. e00024, 2020.

[86] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, and C. Chen, "Toy models of superposition," *arXiv preprint arXiv:2209.10652*, 2022.

[87] T. Bricken, L. Chan, N. Elhage, and C. Olah, "Towards monosemanticity: Decomposing language models with dictionary learning." https://transformer-circuits.pub/2023/monosemantic-features/index.html, 2023.

[88] A. Templeton, *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*. Anthropic, 2024.

[89] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, "Sparse autoencoders find interpretable features in language models," *arXiv preprint arXiv:2309.08600*, 2023.

[90] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Exploring AI ethics of ChatGPT: A diagnostic analysis," *arXiv*, vol. abs/2301.12867, 2023.

[91] R. Puri and B. Catanzaro, "Zero-shot text classification with generative language models," *arXiv*, vol. abs/1912.10165, 2019.

[92] D. J. Corbitt-Hall, J. M. Gauthier, M. T. Davis, and T. K. Witte, "College students' responses to suicidal content on social networking sites: An examination using a simulated facebook newsfeed," *Suicide and Life-Threatening Behavior*, vol. 46, no. 5, pp. 609–624, 2016.

[93] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[94] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Hugging face's transformers: State-of-the-art natural language processing," *arXiv*, vol. abs/1910.03771, 2019.

[95] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, vol. abs/1810.04805, 2018.

[96] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *arXiv*, vol. abs/1909.11942, 2019.

[97] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv*, vol. abs/1910.01108, 2019.

[98] Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? guidance for authors," *BMC Medical Research Methodology*, vol. 18, pp. 1–7, 2018.

[99] R. Armstrong, B. J. Hall, J. Doyle, and E. Waters, "Scoping the scope of a cochrane review," *Journal of Public Health*, vol. 33, no. 1, pp. 147–150, 2011.

[100] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *International Journal of Social Research Methodology*, vol. 8, no. 1, pp. 19–32, 2005.

[101] M. D. J. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares, "Guidance for conducting systematic scoping reviews," *JBI Evidence Implementation*, vol. 13, no. 3, pp. 141–146, 2015.

[102] J. Boergers, A. Spirito, and D. Donaldson, "Reasons for adolescent suicide attempts: Associations with psychological functioning," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 37, no. 12, pp. 1287–1293, 1998.

[103] E. D. Klonsky, A. M. May, and B. Y. Saffer, "Suicide, suicide attempts, and suicidal ideation," *Annual Review of Clinical Psychology*, vol. 12, pp. 307–330, 2016.

[104] R. Vilhjálmsson, E. Sveinbjarnardottir, and G. Kristjansdottir, "Factors associated with suicide ideation in adults," *Social Psychiatry and Psychiatric Epidemiology*, vol. 33, pp. 97–103, 1998.

[105] C. Lázaro-Pérez, P. Munuera Gómez, J. Á. Martínez-López, and J. Gómez-Galán, "Predictive factors of suicidal ideation in spanish university students: A health, preventive, social, and cultural approach," *Journal of Clinical Medicine*, vol. 12, no. 3, p. 1207, 2023.

[106] I. A. Florez, L. J. Allbaugh, C. E. Harris, A. C. Schwartz, and N. J. Kaslow, "Suicidal ideation and hopelessness in ptsd: Spiritual well-being mediates outcomes over time," *Anxiety, Stress, & Coping*, vol. 31, no. 1, pp. 46–58, 2018.

[107] N. Noor, C. Pao, M. Dragomir-Davis, J. Tran, and C. Arbona, "Ptsd symptoms and suicidal ideation in us female firefighters," *Occupational Medicine*, vol. 69, no. 8-9, pp. 577–585, 2019.

[108] T. R. Goldstein, W. Ha, D. A. Axelson, B. I. Goldstein, F. Liao, M. K. Gill, N. D. Ryan, S. Yen, J. Hunt, and H. Hower, "Predictors of suicide attempts among youth with bipolar disorder," *Archives of General Psychiatry*, vol. 69, no. 11, pp. 1113–1122, 2012.

[109] F. K. Goodwin and K. R. Jamison, *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression*, vol. 2. Oxford University Press, 2007.

[110] S. H. Schultz, S. W. North, and C. G. Shields, "Schizophrenia: A review," *American Family Physician*, vol. 75, no. 12, pp. 1821–1829, 2007.

[111] K. Hawton, L. Sutton, C. Haw, J. Sinclair, and J. J. Deeks, "Schizophrenia and suicide: A systematic review of risk factors," *The British Journal of Psychiatry*, vol. 187, no. 1, pp. 9–20, 2005.

[112] L. Sher and R. S. Kahn, "Suicide in schizophrenia: An educational overview," *Medicina*, vol. 55, no. 7, p. 361, 2019.

[113] J. Paris, "Chronic suicidality in borderline personality disorder," *Psychiatric Services*, vol. 53, no. 6, pp. 738–742, 2002.

[114] B. S. Brodsky, K. M. Malone, S. P. Ellis, R. A. Dulit, and J. J. Mann, "Characteristics of borderline personality disorder associated with suicidal behavior," *American Journal of Psychiatry*, vol. 154, no. 12, pp. 1715–1719, 1997.

[115] J.-M. Jang, J.-I. Park, K.-Y. Oh, K.-H. Lee, M. S. Kim, M.-S. Yoon, S.-H. Ko, H.-C. Cho, and Y.-C. Chung, "Predictors of suicidal ideation in a community sample," *Psychiatry Research*, vol. 216, no. 1, pp. 74–81, 2014.

[116] J.-I. Lee, M.-B. Lee, S.-C. Liao, C.-M. Chang, S.-C. Sung, H.-C. Chiang, and C.-W. Tai, "Prevalence of suicidal ideation and associated risk factors in the general population," *Journal of the Formosan Medical Association*, vol. 109, no. 2, pp. 138–147, 2010.

[117] J. Landberg, "Alcohol and suicide in eastern europe," *Drug and Alcohol Review*, vol. 27, no. 4, pp. 361–373, 2008.

[118] A. Mino, A. Bousquet, and B. Broers, "Substance abuse and drug-related death, suicidal ideation, and suicide: A review," *Crisis*, vol. 20, no. 1, p. 28, 1999.

[119] M. M. Rizk, H. Galfalvy, J. M. Miller, M. Milak, R. Parsey, M. Grunebaum, A. Burke, M. E. Sublette, M. A. Oquendo, and B. Stanley, "Characteristics of depressed suicide

attempters with remitted substance use disorders," *Journal of Psychiatric Research*, vol. 137, pp. 572–578, 2021.

[120] H. W. Andersson, M. P. Mosti, and T. Nordfjærn, "Suicidal ideation among inpatients with substance use disorders: Prevalence, correlates and gender differences," *Psychiatry Research*, vol. 317, p. 114848, 2022.

[121] C. J. Bryan and A. O. Bryan, "Financial strain, suicidal thoughts, and suicidal behavior among us military personnel in the national guard," *Crisis*, 2019.

[122] A. Farabaugh, S. Bitran, M. Nyer, D. J. Holt, P. Pedrelli, I. Shyu, S. D. Hollon, S. Zisook, L. Baer, W. Busse, *et al.*, "Depression and suicidal ideation in college students," *Psychopathology*, vol. 45, no. 4, pp. 228–234, 2012.

[123] F. O. Okechukwu, K. T. U. Ogba, J. I. Nwufo, M. O. Ogba, B. N. Onyekachi, C. I. Nwanosike, and A. B. Onyishi, "Academic stress and suicidal ideation: Moderating roles of coping style and resilience," *BMC Psychiatry*, vol. 22, no. 1, pp. 1–12, 2022.

[124] T. Hatchel, J. R. Polanin, and D. L. Espelage, "Suicidal thoughts and behaviors among lgbtq youth: Meta-analyses and a systematic review," *Archives of Suicide Research*, vol. 25, no. 1, pp. 1–37, 2021.

[125] E. A. Kaufman, B. Meddaoui, N. E. Seymour, and S. E. Victor, "The roles of minority stress and thwarted belongingness in suicidal ideation among cisgender and transgender/nonbinary LGBTQ+ individuals," *Archives of Suicide Research*, pp. 1–16, 2022.

[126] M. Sutter and P. B. Perrin, "Discrimination, mental health, and suicidal ideation among LGBTQ people of color," *Journal of Counseling Psychology*, vol. 63, no. 1, p. 98, 2016.

[127] H. Rhoades, J. A. Rusow, D. Bond, A. Lanteigne, A. Fulginiti, and J. T. Goldbach, "Homelessness, mental health and suicidality among lgbtq youth accessing crisis services," *Child Psychiatry & Human Development*, vol. 49, pp. 643–651, 2018.

[128] J. de Lange, L. Baams, D. D. van Bergen, H. M. W. Bos, and R. J. Bosker, "Minority stress and suicidal ideation and suicide attempts among LGBT adolescents and young adults: A meta-analysis," *LGBT Health*, vol. 9, no. 4, pp. 222–237, 2022.

[129] Z. Luo, J. Wang, Y. Zhou, Q. Mao, B. Lang, and S. Xu, "Workplace bullying and suicidal ideation and behaviour: A systematic review and meta-analysis," *Public Health*, vol. 222, pp. 166–174, 2023.

[130] M. M. Husky, A. Bitfoi, M. G. Carta, D. Goelitz, C. Koç, S. Lesinskiene, Z. Mihova, R. Otten, and V. Kovess-Masfety, "Bullying involvement and suicidal ideation in elementary school children across europe," *Journal of Affective Disorders*, vol. 299, pp. 281–286, 2022.

[131] A. Zaborskis, G. Ilionsky, R. Tesler, and A. Heinz, "The association between cyberbullying, school bullying, and suicidality among adolescents," *Crisis*, 2018.

[132] D. Finkelhor, "Sexual abuse: A sociological perspective," *Child Abuse & Neglect*, vol. 6, no. 1, pp. 95–102, 1982.

[133] S. S. Brokke, T. B. Bertelsen, N. I. Landrø, and V. Ø. Haaland, "The effect of sexual abuse and dissociation on suicide attempt," *BMC Psychiatry*, vol. 22, pp. 1–8, 2022.

[134] J. Lindert, O. S. von Ehrenstein, R. Grashow, G. Gal, E. Brähler, and M. G. Weisskopf, "Sexual and physical abuse in childhood and mental health outcomes," *International Journal of Public Health*, vol. 59, pp. 359–372, 2014.

[135] J. R. Peteet, G. Maytal, and H. Rokni, "Unimaginable loss: Contingent suicidal ideation in family members of oncology patients," *Psychosomatics*, vol. 51, no. 2, pp. 166–170, 2010.

[136] B. T. Keum, M. J. Wong, and R. Salim-Eissa, "Gendered racial microaggressions, internalized racism, and suicidal ideation among emerging adult asian american women," *International Journal of Social Psychiatry*, vol. 69, no. 2, pp. 342–350, 2023.

[137] W. L. Robinson, C. R. Whipple, L. A. Jason, and C. E. Flack, "African american adolescent suicidal ideation and behavior: The role of racism and prevention," *Journal of Community Psychology*, vol. 49, no. 5, pp. 1282–1295, 2021.

[138] B. T. Keum and M. J. Wong, "Covid-19 anti-asian racism, perceived burdensomeness, thwarted belongingness, and suicidal ideation among asian american emerging adults," *International Review of Psychiatry*, pp. 1–8, 2023.

[139] K. A. Ratkowska and D. De Leo, "Suicide in immigrants: An overview," 2013. Review article.

[140] J. D. Hovey, "Acculturative stress, depression, and suicidal ideation in mexican immigrants," *Cultural Diversity and Ethnic Minority Psychology*, vol. 6, no. 2, p. 134, 2000.

[141] H. Naismith, R. Howard, R. Stewart, A. Pitman, and C. Mueller, "Suicidal ideation in dementia: Associations with neuropsychiatric symptoms and subtype diagnosis," *International Psychogeriatrics*, vol. 34, no. 4, pp. 399–406, 2022.

[142] R. Cui, M. Maxfield, and A. Fiske, "Dementia-related anxiety and coping styles associated with suicidal ideation," *Aging & Mental Health*, vol. 24, no. 11, pp. 1912–1915, 2020.

[143] M. J. Armstrong, K. Moore, C. E. Jacobson, N. Bedenfield, B. Patel, and J. L. Sullivan, "Frequency of suicidal ideation and associated clinical features in lewy body dementia," *Parkinsonism & Related Disorders*, vol. 90, pp. 33–37, 2021.

[144] M. M. Fässberg, G. Cheung, S. S. Canetto, A. Erlangsen, S. Lapierre, R. Lindner, B. Draper, J. J. Gallo, C. Wong, and J. Wu, "A systematic review of physical illness, functional disability, and suicidal behaviour among older adults," *Aging & Mental Health*, vol. 20, no. 2, pp. 166–194, 2016.

[145] H.-M. Vasiliadis, C. D'Aiuto, C. Lamoureux-Lamarche, I. Pitrou, S. Gontijo Guerra, and D. Berbiche, "Pain, functional disability and mental disorders as potential mediators of the association between chronic physical conditions and suicidal ideation in community living older adults," *Aging & Mental Health*, vol. 26, no. 4, pp. 791–802, 2022.

[146] M. Racine, "Chronic pain and suicide risk: A review," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 87, pp. 269–280, 2018.

[147] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv*, vol. abs/2203.05794, 2022.

[148] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in Sociology*, vol. 7, p. 886498, 2022.

[149] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT networks," in *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019.

[150] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv*, vol. abs/1802.03426, 2018.

[151] M. S. Asyaky and R. Mandala, "Improving the performance of hdbscan on short text clustering by using word embedding and umap," in *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pp. 1–6, IEEE, 2021.

[152] N. Oskolkov, "Dimensionality reduction: Overview, technical details, and some applications," *Applied Data Science in Tourism*, pp. 151–167, 2022.

[153] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 160–172, Springer, 2013.

[154] G. Giupponi, M. Innamorati, R. J. Baldessarini, D. De Leo, F. de Giovannelli, R. Pycha, A. Conca, P. Girardi, and M. Pompili, "Factors associated with suicide: A case-control study," *Comprehensive Psychiatry*, vol. 80, pp. 150–154, 2018.

[155] A. M. Lasserre, H. Marti-Soler, M.-P. F. Strippoli, J. Vaucher, J. Glaus, C. L. Vandeleur, E. Castelao, P. Marques-Vidal, G. Waeber, and P. Vollenweider, "Clinical characteristics of depression and mortality risk," *Journal of Affective Disorders*, vol. 189, pp. 17–24, 2016.

[156] M. Moitra, D. Santomauro, L. Degenhardt, P. Y. Collins, H. Whiteford, T. Vos, and A. Ferrari, "Estimating suicide risk associated with mental disorders," *Journal of Psychiatric Research*, vol. 137, pp. 242–249, 2021.

[157] E. C. Harris and B. Barraclough, "Suicide as an outcome for mental disorders: A meta-analysis," *The British Journal of Psychiatry*, vol. 170, no. 3, pp. 205–228, 1997.

[158] V. M. Langford, "Risk factors for suicide in men," *Nursing Clinics*, vol. 58, no. 4, pp. 513–524, 2023.

[159] L. Favril, R. Yu, J. R. Geddes, and S. Fazel, "Individual-level risk factors for suicide mortality: An umbrella review," *The Lancet Public Health*, vol. 8, no. 11, pp. e868–e877, 2023.

[160] A. Pemau, C. Marin-Martin, M. Diaz-Marsa, A. de la Torre-Luque, W. Ayad-Ahmed, A. Gonzalez-Pinto, N. Garrido-Torres, L. Garrido-Sanchez, N. Roberto, and P. Lopez-Peña, "Risk factors for suicide reattempt: A systematic review and meta-analysis," *Psychological Medicine*, pp. 1–8, 2024.

[161] C. Grover, J. Huber, M. Brewer, A. Basu, and M. Large, "Meta-analysis of clinical risk factors for suicide in emergency settings," *Acta Psychiatrica Scandinavica*, vol. 148, no. 6, pp. 491–524, 2023.

[162] S. Jha, G. Chan, and R. Orji, "Identification of risk factors for suicide and prevention insights," *Human Behavior and Emerging Technologies*, vol. 2023, no. 1, p. 3923097, 2023.

[163] J. Anderberg, M. Bogren, C. Mattisson, and L. Brådvik, "Long-term suicide risk in anxiety disorders," *Archives of Suicide Research*, vol. 20, no. 3, pp. 463–475, 2016.

[164] A. Athey, J. Overholser, C. Bagge, L. Dieter, E. Vallender, and C. A. Stockmeier, "Risk-taking behaviors and stressors predict suicidal outcomes," *Psychiatry Research*, vol. 270, pp. 160–167, 2018.

[165] A. Page, S. Morrell, C. Hobbs, G. Carter, M. Dudley, J. Duflou, and R. Taylor, "Suicide in young adults: A case-control study," *BMC Psychiatry*, vol. 14, pp. 1–9, 2014.

[166] K. Lieb, M. C. Zanarini, C. Schmahl, M. M. Linehan, and M. Bohus, "Borderline personality disorder," *The Lancet*, vol. 364, no. 9432, pp. 453–461, 2004.

[167] M. C. Zanarini, F. R. Frankenburg, J. Hennen, D. B. Reich, and K. R. Silk, "Prediction of the 10-year course of borderline personality disorder," *American Journal of Psychiatry*, vol. 163, no. 5, pp. 827–832, 2006.

[168] M. D. Braquehais, M. A. Oquendo, E. Baca-García, and L. Sher, "Is impulsivity a link between childhood abuse and suicide?," *Comprehensive Psychiatry*, vol. 51, no. 2, pp. 121–129, 2010.

[169] K. M. Devries, J. Y. T. Mak, J. C. Child, G. Falder, L. J. Bacchus, J. Astbury, and C. H. Watts, "Childhood sexual abuse and suicidal behavior: A meta-analysis," *Pediatrics*, vol. 133, no. 5, pp. e1331–e1344, 2014.

[170] K. L. Wolfe, P. A. Nakonezny, V. J. Owen, K. V. Rial, A. P. Moorehead, B. D. Kennard, and G. J. Emslie, "Hopelessness as a predictor of suicidal ideation in adolescents," *Suicide and Life-Threatening Behavior*, vol. 49, no. 1, pp. 253–263, 2019.

[171] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5–8, 2017.

[172] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

[173] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

[174] X. Xie, N. Zhang, Z. Li, S. Deng, H. Chen, F. Xiong, M. Chen, and H. Chen, "From discrimination to generation: Knowledge graph completion with generative transformer," in *Companion Proceedings of the Web Conference 2022*, pp. 162–165, 2022.

[175] F. Mi, Y. Li, Y. Zeng, J. Zhou, Y. Wang, C. Xu, L. Shang, X. Jiang, S. Zhao, and Q. Liu, "PanGu-Bot: Efficient generative dialogue pre-training from pre-trained language model," *arXiv*, vol. abs/2203.17090, 2022.

[176] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv*, vol. abs/2109.01652, 2021.

[177] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, "Scaling instruction-finetuned language models," *arXiv*, vol. abs/2210.11416, 2022.

[178] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "LLaMA 2: Open foundation and fine-tuned chat models," *arXiv*, vol. abs/2307.09288, 2023.

[179] J. W. Chotlos, "A statistical and comparative analysis of individual written language samples," *Psychological Monographs*, vol. 56, no. 2, p. 75, 1944.

[180] B. Richards, "Type/token ratios: What do they really tell us?," *Journal of Child Language*, vol. 14, no. 2, pp. 201–209, 1987.

[181] R. F. Flesch, "Estimating the comprehension difficulty of magazine articles," *The Journal of General Psychology*, vol. 28, no. 1, pp. 63–80, 1943.

[182] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[183] E. L. Snow, L. K. Allen, M. E. Jacovina, S. A. Crossley, C. A. Perret, and D. S. McNamara, "Keys to detecting writing flexibility over time: Entropy and natural language processing," *Journal of Learning Analytics*, vol. 2, no. 3, pp. 40–54, 2015.

[184] C. Rudin, "Stop explaining black box models for high-stakes decisions," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[185] I. Nejadgholi, S. Kiritchenko, K. C. Fraser, and E. Balkır, "Concept-based explanations for detecting false causal relationships," *arXiv*, vol. abs/2307.01900, 2023.

[186] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Carter, T. Brown, and C. Olah, "Towards monosemanticity: Decomposing language models with dictionary learning." https://transformer-circuits.pub/2023/monosemanticity/, 2023.

[187] A. Mudide, J. Engels, E. J. Michaud, M. Tegmark, and C. S. de Witt, "Efficient dictionary learning with switch sparse autoencoders," *arXiv preprint arXiv:2410.08201*, 2024.

[188] C.-K. Yeh, C.-J. Hsieh, A. Suggala, D. I. Inouye, and P. Ravikumar, "On concept-based explanations in deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10278–10289, 2020.

[189] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, G. Valiant, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning (ICML)*, pp. 5338–5348, 2020.

[190] C. Chen, O. Li, C. Tao, A. Barnett, J. Su, and C. Rudin, "Concept whitening for interpretable image recognition," in *International Conference on Machine Learning (ICML)*, pp. 1135–1145, 2020.

[191] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[192] J. Templeton, J. Mu, A. Jones, A. Floreano, and S. Eyuboglu, "Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet." https://www.anthropic.com/research/scaling-monosemanticity, 2024.

[193] OpenAI, "Sparse autoencoders for interpretability." https://openai.com/index/introducing-sparse-autoencoders/, 2024.

[194] Z. Li, J. Zhou, Z. An, W. Cheng, and B. Hu, "Deep hierarchical ensemble model for suicide detection on imbalanced social media data," *Entropy*, vol. 24, no. 4, p. 442, 2022.

[195] D. Kodati and R. Tene, "Identifying suicidal emotions on social media through transformer-based deep learning," *Applied Intelligence*, vol. 53, no. 10, pp. 11885–11917, 2023.

[196] E. R. Kumar and N. Venkatram, "Predicting and analyzing suicidal risk behavior using rule-based approach in twitter data," *Soft Computing*, pp. 1–9, 2023. Early access.

[197] Q. Wei, A. Franklin, T. Cohen, and H. Xu, "Clinical text annotation: What factors are associated with the cost of time?," in *Proceedings of the AMIA Annual Symposium*, p. 1552, American Medical Informatics Association, 2018.

[198] S. Robertson, H. Zaragoza, *et al.*, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[199] H. Ghanadian, I. Nejadgholi, and H. A. Osman, "Chatgpt for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials, and limitations," *arXiv*, vol. abs/2306.09390, 2023.

[200] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[201] M. N. Team, "Introducing mpt-30b: Raising the bar for open-source foundation models." https://www.mosaicml.com/blog/mpt-30b, 2023. Online; accessed June 2023.

[202] Z. Hu, Y. Xu, W. Yu, S. Wang, Z. Yang, C. Zhu, K.-W. Chang, and Y. Sun, "Empowering language models with knowledge graph reasoning for open-domain question answering," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, (Abu Dhabi, United Arab Emirates), pp. 9562–9581, Association for Computational Linguistics, 2022.

[203] D. Hovy and D. Yang, "The importance of modeling social factors of language: Theory and practice," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Online), pp. 588–602, Association for Computational Linguistics, 2021.

[204] D. De Berardis, G. Martinotti, and M. Di Giannantonio, "Understanding the complex phenomenon of suicide: From research to clinical practice," *Frontiers in Psychiatry*, vol. 9, p. 61, 2018.

[205] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.

[206] A. Raza, F. Rustam, H. U. R. Siddiqui, I. d. l. T. Diez, B. Garcia-Zapirain, E. Lee, and I. Ashraf, "Predicting genetic disorder and types of disorder using chain classifier approach," *Genes*, vol. 14, no. 1, p. 71, 2022.

[207] M. M. Sadiq Fareed, A. Raza, N. Zhao, A. Tariq, F. Younas, G. Ahmed, S. Ullah, S. F. Jillani, I. Abbas, and M. Aslam, "Predicting divorce prospect using ensemble learning: Support vector machine, linear model, and neural network," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[208] R. Sawhney, H. Joshi, S. Gandhi, and R. R. Shah, "Towards ordinal suicide ideation detection on social media," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 22–30, 2021.

[209] S. M. Joseph, S. Citraro, V. Morini, G. Rossetti, and M. Stella, "Cognitive network neighborhoods quantify feelings expressed in suicide notes and reddit mental health communities," *Physica A*, vol. 610, p. 128336, 2023.

[210] H. Ghanadian, I. Nejadgholi, and H. Al Osman, "Socially aware synthetic data generation for suicidal ideation detection using large language models," *IEEE Access*, 2024.

[211] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv*, vol. abs/1801.06146, 2018.

[212] W. H. Organization, "Suicide." https://www.who.int/news-room/fact-sheets/detail/suicide, 2021. Accessed 2023.

[213] V. V. Ramaswamy, S. S. Y. Kim, R. Fong, and O. Russakovsky, "Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10932–10941, 2023.

[214] C. M. Bishop, "Training with noise is equivalent to tikhonov regularization," *Neural Computation*, vol. 7, no. 1, pp. 108–116, 1995.

[215] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment*, vol. 11, p. 269, 2017.

[216] L. T. Bayliss, S. Christensen, A. Lamont-Mills, and C. du Plessis, "Suicide capability in the ideation-to-action framework: A scoping review," *PLOS ONE*, vol. 17, no. 10, p. e0276070, 2022.

[217] S. Homan, M. Gabi, N. Klee, S. Bachmann, A.-M. Moser, S. Michel, A.-M. Bertram, A. Maatz, G. Seiler, and E. Stark, "Linguistic features of suicidal thoughts: A systematic review," *Clinical Psychology Review*, vol. 95, p. 102161, 2022.

[218] P. Boonyarat, D. J. Liew, and Y.-C. Chang, "Enhanced bert models for detecting suicidal ideation in thai social media," *Information Processing & Management*, vol. 61, no. 4, p. 103706, 2024.

[219] L. G. Singh, J. Mao, R. Mutalik, and S. E. Middleton, "Extraction and summarization of suicidal ideation using llms," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, 2024.

[220] "Merriam-webster dictionary." https://www.merriam-webster.com/, 2024.

[221] E. Arensman, C. Larkin, J. McCarthy, S. Leitao, P. Corcoran, E. Williamson, C. McAuliffe, I. J. Perry, E. Griffin, and E. M. Cassidy, "Psychosocial and psychiatric risk factors for suicide in ireland," *BMC Psychiatry*, vol. 19, pp. 1–11, 2019.

[222] T. Brockie, M. Kahn-John, L. Mata Lopez, E. Bell, T. Brockie, T. Brockie, E. Decker, N. Glass, H. Has Eagle, and K. Helgeson, "Factors contributing to suicide clusters among native american youth," *Frontiers in Public Health*, vol. 11, p. 1281109, 2024.

[223] R. C. O'Connor and M. K. Nock, "The psychology of suicidal behaviour," *The Lancet Psychiatry*, vol. 1, no. 1, pp. 73–85, 2014.

[224] R. J. Gallagher, K. Reing, D. Kale, and G. Ver Steeg, "Anchored correlation explanation: Topic modeling with minimal domain knowledge," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 529–542, 2017.

[225] L. Gan, T. Yang, Y. Huang, B. Yang, Y. Y. Luo, L. W. C. Richard, and D. Guo, "Comparison of topic modeling methods: Lda, top2vec, and bertopic," in *International Symposium on Artificial Intelligence and Robotics*, pp. 376–391, 2023.

[226] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[227] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, and E. Baralis, "Concept-based explainable artificial intelligence: A survey," *ACM Computing Surveys*, 2023.

APPENDICES

.1 Examples of Synthetic Data

Table 1 displays a selection of binary samples generated by ChatGPT within the synthetic dataset using social topics. The table provides an illustration of specific examples generated by this GLLM.

.2 UMAP Projections of Features

Left column: ALBERT fine tuned on AUG. Right column: ALBERT fine tuned on UMD. Colored points mark features whose dominant responses align with the indicated topic; Dark blue points show all other features.

Table 1: Generated synthetic samples using extracted social and psychological topics by ChatGPT

Topics	Labels	Sample
Depression	1	Growing up, I always felt like nobody wanted me. No one ever seemed to care about my thoughts, feelings, or emotions. As I got older, these feelings intensified, and I started to feel like I was a burden on everyone around me. I never felt like I belonged anywhere and I constantly struggled with my depression. The constant emptiness and loneliness eventually became too much to bear, and I began to contemplate ending my life.
	0	I've been struggling with mental health problems for as long as I can remember. Depression have been my constant companions. But I've never let it stop me from living my life. I've learned to manage my symptoms with the help of therapy and medication. Some days are harder than others, but I keep pushing forward and focusing on the good in my life. I've come to accept that mental illness is a part of me, but it doesn't define me.
Bullying	1	I felt like I was constantly being attacked by my classmates. They called me names, pushed me around, and even vandalized my belongings. The worst part was feeling so alone and misunderstood. I didn't want to burden my parents with my problems so I kept it all inside. The idea of ending it all seemed like the only way to escape the pain and torment.
	0	When I was in middle school, I was constantly bullied for being overweight. It was tough going to school every day and constantly being made fun of by my peers. I felt so alone and helpless. But I didn't let the bullies control my life. I joined a support group and found friends who lifted me up. I also started exercising and eating healthier, not to please anyone else, but to feel better about myself. It wasn't an easy journey, but it was worth it. I'm happy to say that today I am confident and proud of who I am.
Unemployment	1	After losing my job, I stayed unemployed for months. My savings ran out and with no source of income, my bills piled up. The constant fear of not being able to provide for myself and my family drove me to the brink of despair. I feel worthless and like a burden on everyone around me. The future seems bleak and hopeless, and I wonder if it's worth it to keep going.
	0	After graduating from college, I struggled to find a job in my field for a few months. It was frustrating and disheartening, but I kept applying and networking. Eventually, I landed a job in a related field that I enjoy. It wasn't my dream job, but it paid the bills and gave me experience. I'm still looking for my dream job, but I'm grateful for what I have and optimistic about my future prospects.
Relationship problems	1	I thought I had found the one but it seems like I was wrong because he left me for someone else. I don't know how to deal with this pain . I can't sleep, I can't eat, and I just want to disappear. Maybe everything would be easier if I just ended it all.
	0	My relationship with my partner hasn't been going well lately. We have been arguing over small things, and it's affecting our mental health. We decided to go for couples therapy, and it's been a turning point for us. We learned to communicate better and understand each other's perspective. Now we are in a better place and happier than ever before.

Figure A.1: UMAP projections of learned features. Top row: Anger; Bottom row: Anxiety. Left: AUG, Right: UMD.

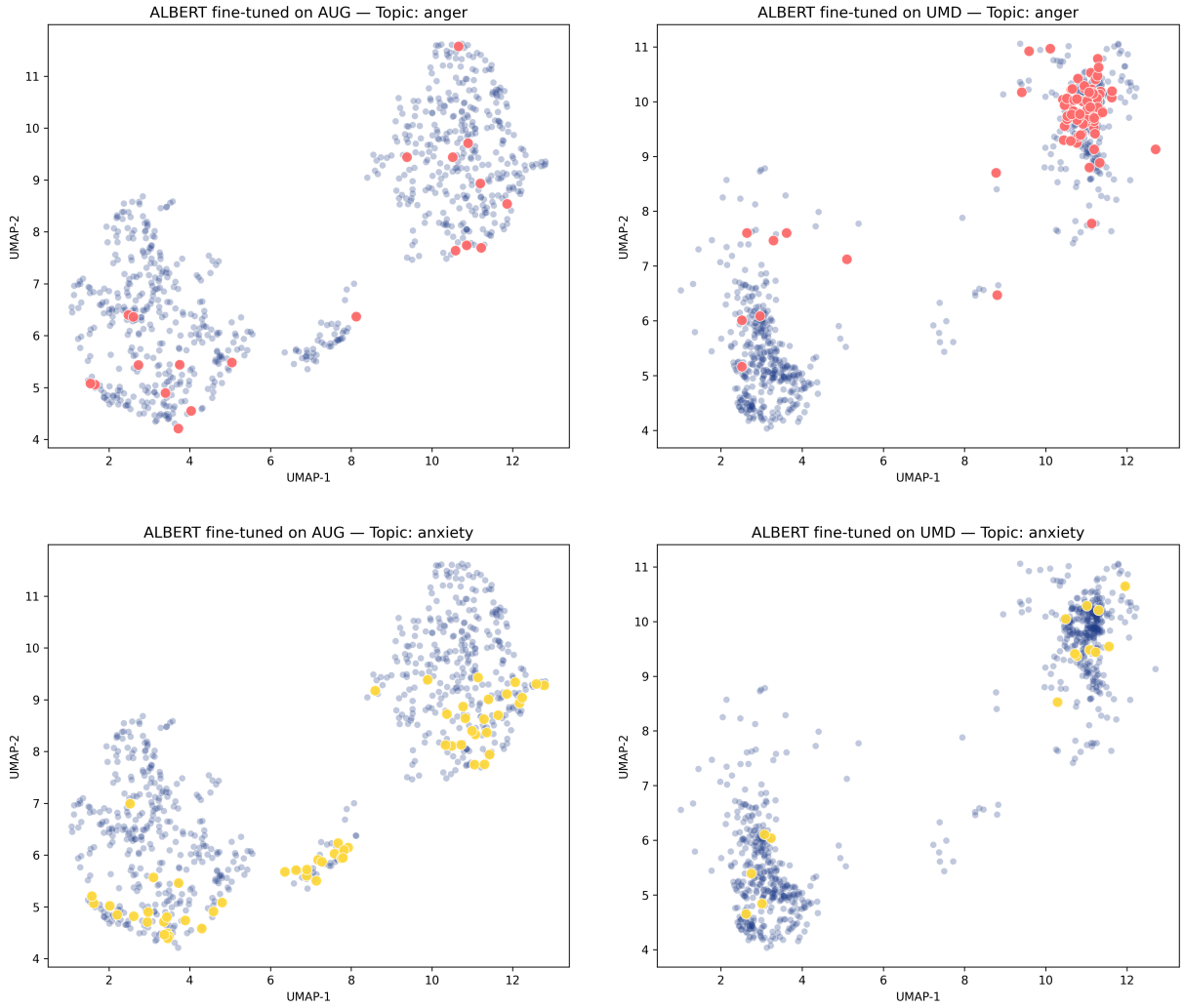


Figure A.2: UMAP projections of learned features. Top row: Bullying; Bottom row: Education. Left: AUG, Right: UMD.

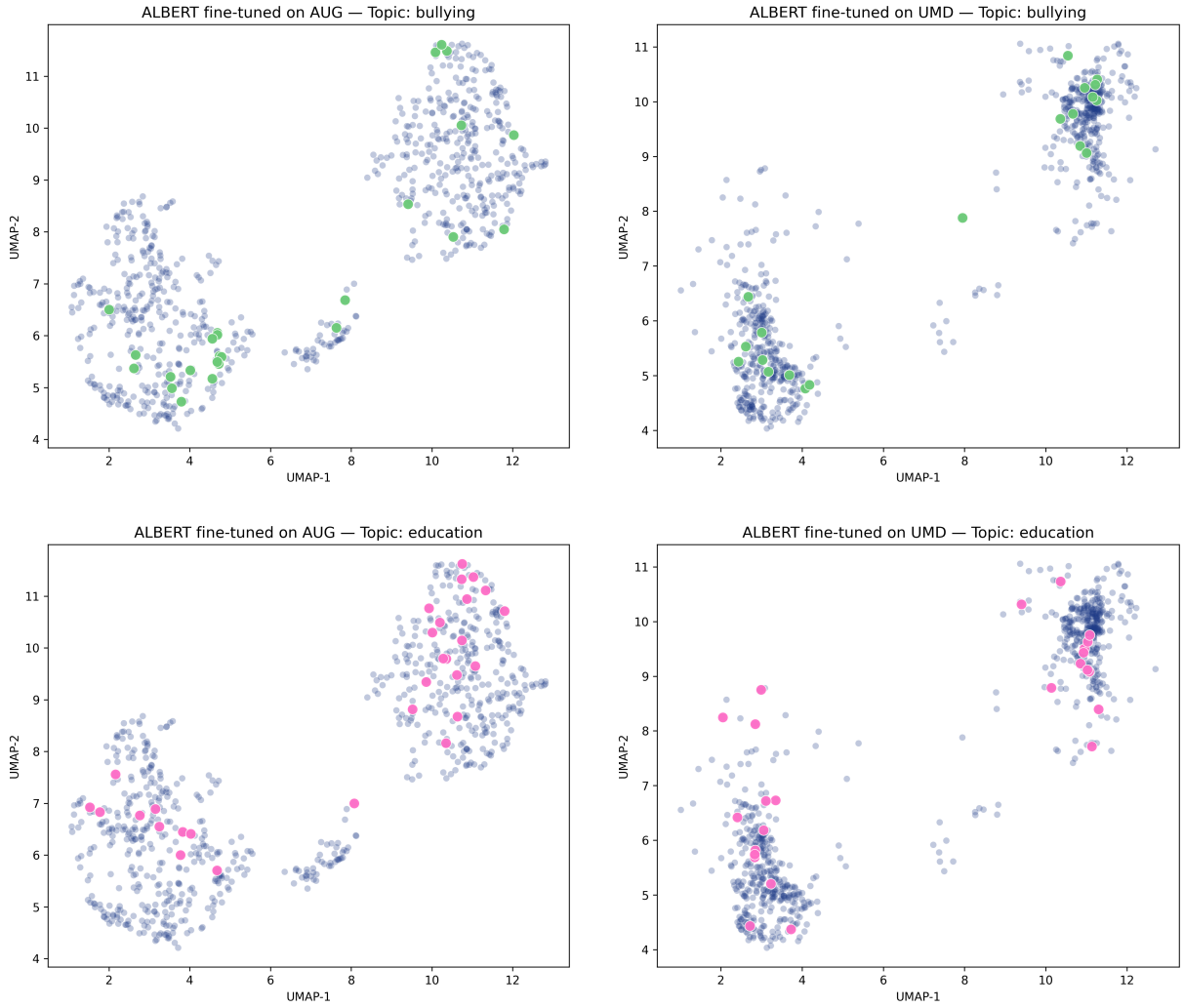


Figure A.3: UMAP projections of learned features. Top row: Death of a close one; Bottom row: Financial crisis. Left: AUG, Right: UMD.

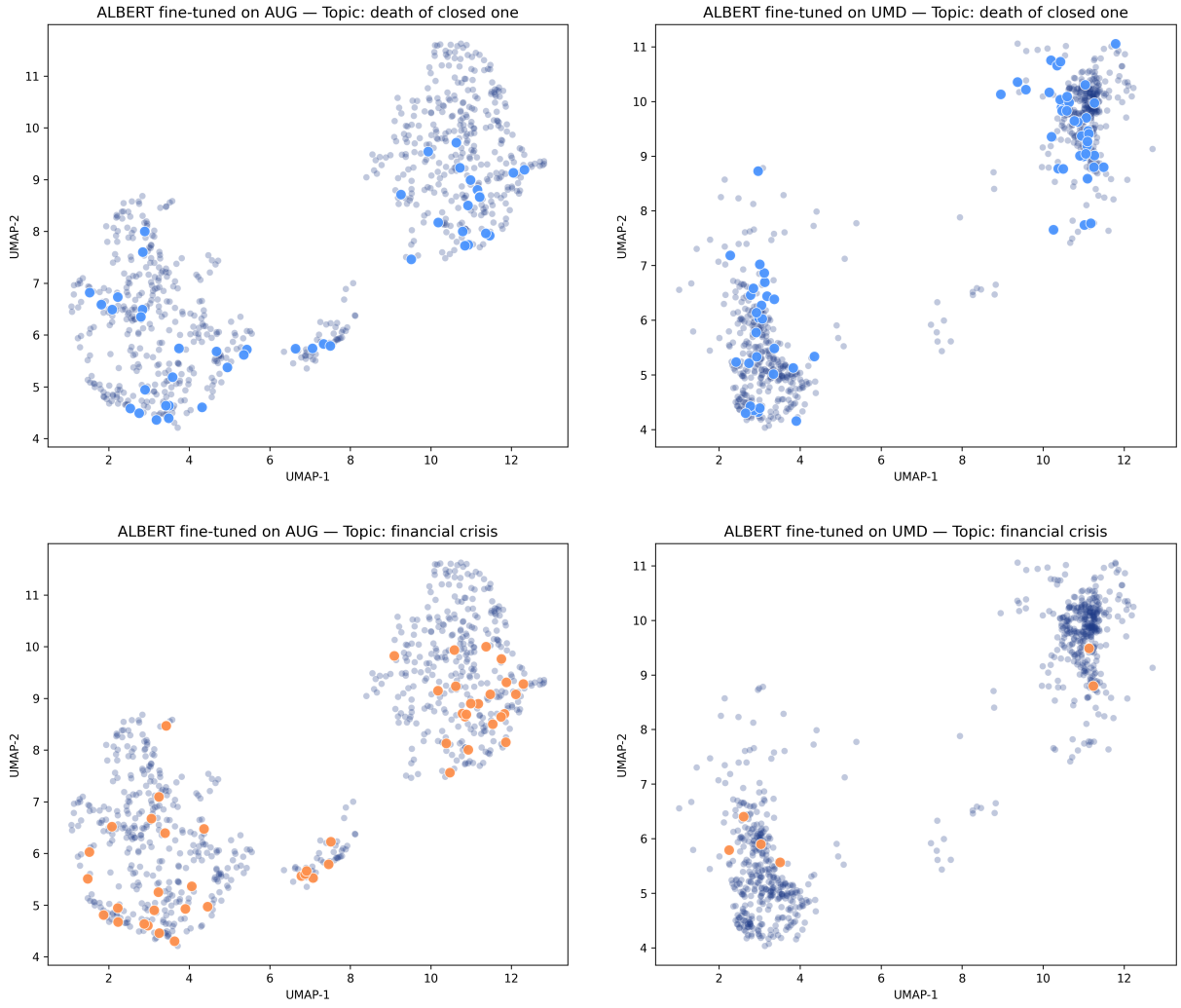


Figure A.4: UMAP projections of learned features. Top row: Hopelessness; Bottom row: Racism. Left: AUG, Right: UMD.

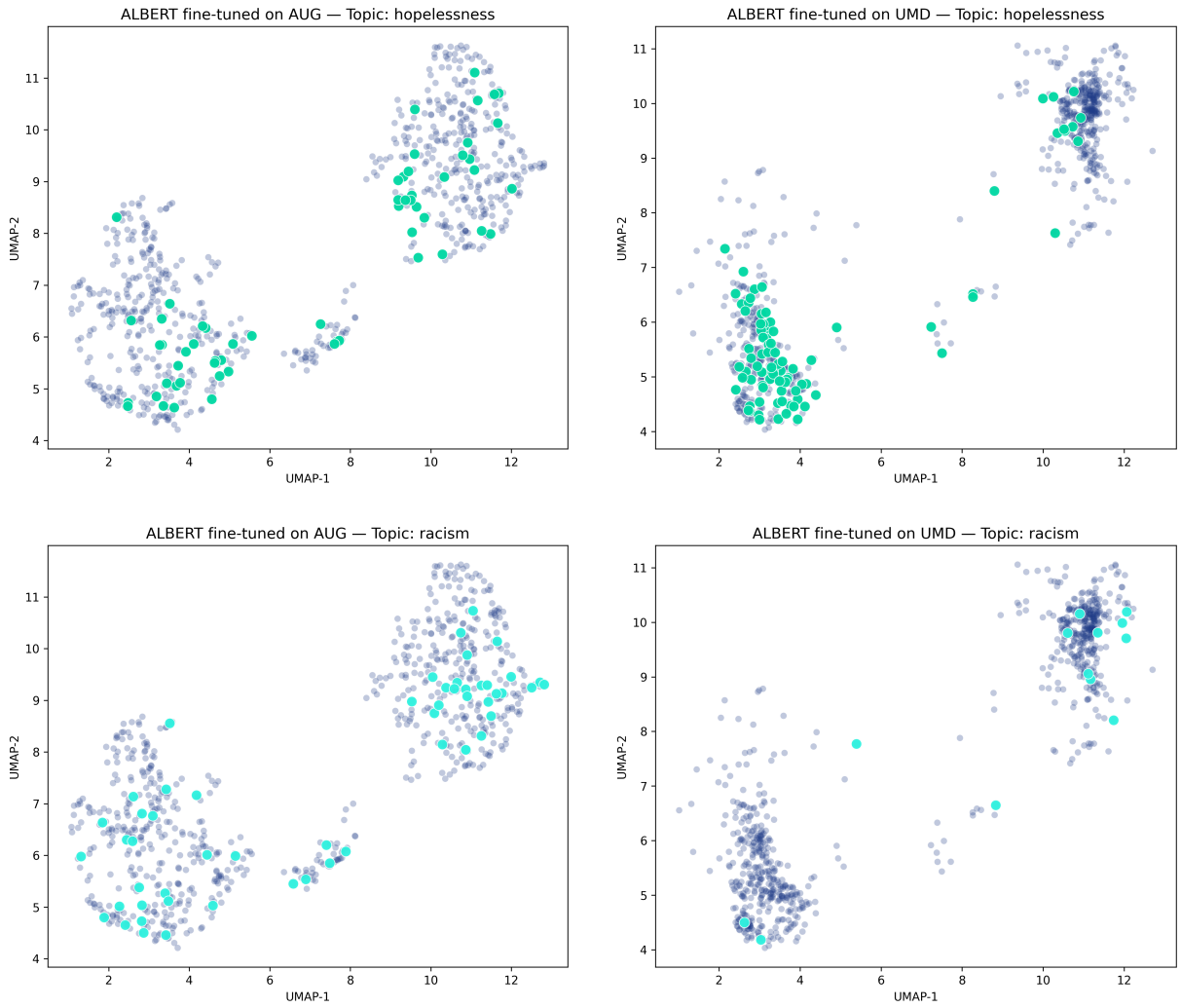


Figure A.5: UMAP projections of learned features. Top row: Relationship problems; Bottom row: Unemployment. Left: AUG, Right: UMD.