Integrating Mental-RoBERTa and Generative LLMs for Phenotyping Secondary Insomnia: A Multi-Dimensional Framework for Etiology Discovery

by

Sadman Sakib
21201605
Nahmar Hossain
21201639
Kumar Prothom Pranto Sarma Chowdhury
22201736
Shehzad Hakim
22341022
A Q M Mujahidul Islam
22101014

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sadman Sakib
21201605

Nahmar Hossain
21201639

Kumar Prothom Pranto Sarma Chowdhury
22201736

Shehzad Hakim
22341022

A Q M Mujahidul Islam
22101014

Approval

The thesis/project titled "Natural Language Processing for Insomnia Symptom Classification: Developing Automated Screening Tools from Reddit Data" submitted by

1. Sadman Sakib (21201605)
2. Nahmar Hossain (21201639)
3. Kumar Prothom Pranto Sarma Chowdhury (22201736)
4. Shehzad Hakim(22341022)
5. A Q M Mujahidul Islam (22101014)

of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 2026 Year.

Examining Committee:

Supervisor:
(Member)

Md.Tawhid Anwar

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Md. Sabbir Ahmed

Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor 2:
(Member)

Dr. Md. Ashraful Alam

Associate Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Insomnia affects a significant portion of the global population, yet distinguishing between primary and secondary insomnia—where sleep disturbances stem from underlying medical, psychiatric, or environmental factors—remains critically underexplored in computational health research. This study presents a multi-stage computational framework for identifying and phenotyping secondary insomnia from naturalistic Reddit discussions. A clinically validated dataset of 600 posts was developed, with linguistic validation confirming that secondary insomnia posts contain significantly more biomedical terminology than primary insomnia posts. Mental-RoBERTa, a domain-adapted transformer, was employed (after the initial integration of RoBERTa-Base) for identifying cases, which outperformed both traditional baselines and large language models. The trained model identified over 3,400 high-confidence secondary insomnia cases from 5,000 unlabeled posts. A hybrid pipeline combining Llama-3 clinical summarization with BERTopic unsupervised clustering discovered 10 distinct etiological categories, revealing gastroesophageal reflux disease, menopausal factors, and benzodiazepine withdrawal as predominant drivers. Zero-shot emotion profiling revealed distinct psychological phenotypes across different triggers, with benzodiazepine withdrawal exhibiting the highest perplexity and environmental factors showing the greatest frustration. This framework enables the first large-scale computational characterization of secondary insomnia drivers and their psychological burdens, providing actionable insights for precision intervention design and population health surveillance in digital mental health systems.

Keywords: Insomnia, sleep disorder, machine learning, text classification, supervised learning, symptom detection, Reddit

Table of Contents

Chapter 1

Introduction

Around the globe, approximately 852 million adults struggle with insomnia, and a fraction of those people meet the criteria for chronic insomnia disorder [22]. It is already established that insomnia disrupts sleep quality, this also carries serious health implications: increasing the risk of cardiovascular diseases, causing depression and anxiety disorders and creating work-life imbalance. Despite having such major implications, clinicians continue to face a core diagnostic challenge: determining whether insomnia arised independently as in without any external cause or resulted from identifiable external factors. Although in recent times official DSM-5 and ICSD-3 guidelines for diagnosing insomnia, use a more unified term, "Chronic Insomnia Disorder," however, understanding the root cause remains very important for selecting the right approach for appropriate treatment. Traditionally, it has been a standard practice to differentiate between primary insomnia, which occurs without any clear cause and secondary insomnia, which develops in response to underlying medical conditions, psychiatric disorders, or environmental factors. This differentiation is actually very important because it has great significance when it comes to clinical treatment. For example, it is best to treat primary insomnia with sleep-related remedies such as cognitive-behavioral therapy for insomnia (CBT-I), on the contrary, secondary insomnia requires doctors to address the root cause of the sleep disturbance. Population research shows how important this differentiation is, revealing that 40% of insomnia patients have comorbid mental health conditions, compared to just 16.4%, who do not have any sleep related issues [1]. This finding shows why accurately identifying the cause of insomnia in individual patients matters so much for effective care.

Social media platforms usually generate large scale patient-generated health data, creating opportunities for mental health monitoring from a holistic perspective. Reddit is irrefutably a very important platform in this regard because it provides its users pseudo-anonymity and its specialized communities encourage users to share very detailed information regarding their health related issues. Luckily, vast amounts of sleep related data is available on two of its subreddits, r/insomnia and r/sleep. Recent studies facilitating NLP, especially through transformer-based architectures and domain-specific models like Mental-RoBERTa, have allowed researchers to automatically detect psychiatric disorders from social media text with accuracy approaching that of conventional clinical screening tools. However, when it comes to proper application and medical execution, these tools remain largely unexplored when applied to sleep disorders. The only substantial large-scale study of the r/insomnia community relied on simple keyword frequency counts to track discussions related to treatment, providing no proper tools to differentiate between insomnia subtypes and tackling underlying issues [15].

This gap in traditional research has significant clinical consequences. Usual sleep related studies depend on collecting data covering a period of years, which naturally limits researchers' ability to detect, for what reasons people are being affected by insomnia in real time. This approach of computational phenotyping using social media data has the potential to address this issue by allowing continuous real-time monitoring, therefore, supporting medicines that cater to people's individual needs.

1.1 Research Problem

Insomnia, a widespread medical condition, affects 10-30% of the global population [10], yet it is still one of most underdiagnosed and insuffciently treated medical conditions, making people seek help in specialized communities like Reddit. There are existing papers that utilzed NLP to monitor discussions about insomnia on social media [15], the current computational techniques are limited to surface-level calculations monitoring mentions of treatments (melatonin, CBT-I, Ambien) and sentiment analysis without addressing the underlying clinical medical and psychological reasons.

This gap has real-life implications as secondary insomnia—sleep disturbances that arise from comorbid conditions, most commonly found in real-world populations but is not characterized in a systematic way in digital health research. Clinical sleep medicine recognizes the importance of identifying underlying etiologies for treatment planning. There has been no published research aimed at classifying primary and secondary insomnia using automated methods employing modern NLP techniques.

The core problem is the lack of computational methods that can automatically identify, categorize, and phenotype the heterogeneous causes of secondary insomnia from patient-generated text. Existing work treats insomnia as a single condition, focusing mainly on binary detection or treatment tracking. As a result, three gaps remain unaddressed: systematic discovery of underlying etiologies, characterization of multi-dimensional emotional burden tied to those etiologies, and clinically actionable phenotyping that links triggers to prevalence, emotional signatures, and linguistic markers. These gaps limit targeted intervention design and real-world surveillance of emerging insomnia patterns.

1.2 Research Objectives

The following specific objectives are formulated for this study:

- **Dataset Construction and Linguistic Validation** : Develop a clinically validated dataset of 600 Reddit posts from r/insomnia and r/sleep, expertly annotated each post to distinguish primary insomnia (n=300) from secondary insomnia (n=300). Validate the annotation schema through medical entity density analysis using scispaCy's biomedical NER, demonstrating that secondary insomnia posts contain significantly more medical terminology than primary insomnia posts via Mann-Whitney U tests and Cohen's d effect size calculations.

- **Domain-Adapted Classification for High-Confidence Case Identification:** Train and evaluate Mental-RoBERTa (after initial integration of RoBERTa-Base) using 5-fold stratified cross-validation in order to classify presentations of insomnia, with a particular emphasis on the identification of cases of secondary insomnia. Following that, apply the trained model to 5,000 unlabeled Reddit posts with confidence-based filtering (softmax probability $\geq$ 0.70) to extract high-quality cases of secondary insomnia, thus establishing classification as a pathway to clinical phenotyping instead of merely an analytical conclusion.

- **Computational Etiology Discovery via Hybrid LLM-Topic Modeling Architecture** : Identify the various root causes that lead secondary insomnia through a multi-stage pipeline that merges Llama-3 8B clinical summarization with BERTopic's unsupervised clustering approach. Implement few-shot prompt engineering to produce clinically relevant diagnostic labels (such as benzodiazepine withdrawal, GERD, chronic pain) and assess prevalence distributions, thus offering the first extensive computational characterization of secondary insomnia factors in real-world patient narratives.

- **Emotional Burden Phenotyping and Model Interpretability** : Individual psychological experiences, resulting from different insomnia subtypes to be mapped using BART-large-MNLI zero-shot classification across six emotional dimensions (Stress, Frustration, Desperation, Helplessness, Perplexity, Dread). Construct emotion-etiology heatmaps to show emotional markers related to proper clinical approach selection. Implement Layer Integrated Gradients attribution analysis to validate that classification decisions rely on clinically meaningful features (medical terminology, comorbidity mentions, causal language) rather than spurious correlations.

1.3 Workflow

The workflow begins with data collection and the training of a Mental-RoBERTa classifier to accurately distinguish secondary insomnia cases from social media text (an initial integration of RoBERTa-Base was also carried out), validated against expert annotations. High-confidence predictions are then processed by Llama-3, which acts as a clinical reasoning agent to transform raw posts into structured medical narratives highlighting specific triggers. These narratives feed into two parallel downstream pathways: an unsupervised BERTopic framework to cluster and label etiological factors, and a zero-shot emotion profiling module to characterize the specific emotional burdens associated with each condition.

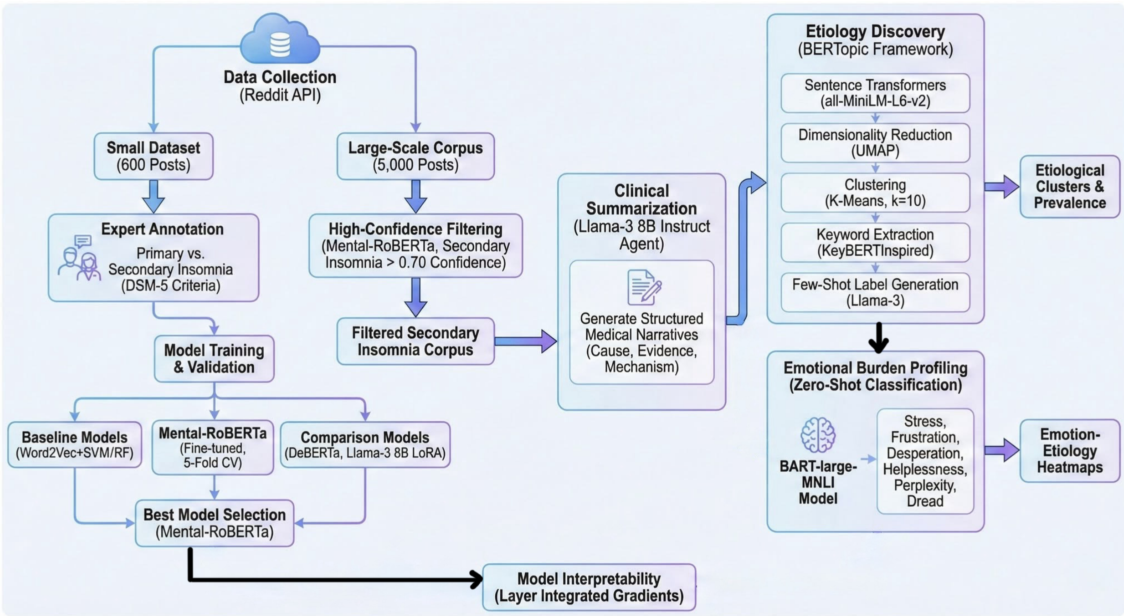


Figure 1.1: Workflow

Chapter 2

Literature Review

2.1 Traditional and Transformer-Based Foundations

2.1.1 Evolution from Classical Machine Learning to Transformers

Previously, detecting mental health conditions by analysing user's posts from social media was done by classical "word-checklist" models but now it can be done by models involving neural networks which are more contextualized. Early work established that traditional supervised learning approaches can achieve basic baseline performance. The authors of [9] demonstrated that Support Vector Machines (SVM) and Logistic Regression achieved 79-75% accuracy in detecting mental illness from with Logistic Regression and Random Forest both reaching 77.29% accuracy and 0.77 F1-score on a 60,000-post corpus from r/depression and r/SuicideWatch. These findings show that these processes had a constraint of not being able to process the text more precisely for better classification accuracy, which could have caused misclassifications. Early neural networks, such as LSTMs, introduced sequential processing to better process chain of events in text properly, which was not enough because of limitation in parallelization. This paved the way for transformer architectures, which revolutionized the field through self-attention mechanisms and bidirectional context understanding, enabling superior performance in capturing intricate emotional and psychological indicators.

Transformer architectures have markedly elevated the performance potential of mental health text classification. The authors of [8] showed that fine-tuning RoBERTabase on more than 17,000 Reddit posts yielded a macro F1 score of 0.89 across five mental disorder categories (depression, anxiety, bipolar disorder, ADHD, PTSD), substantially surpassing LSTM and standard BERT baselines. The study showed that RoBERTa's dynamic masking strategy and extensive pretraining allowed the model with enhanced capacity to identify complex semantic dependencies beyond that of static embeddings. Notably, the findings revealed persistent ambiguity between depression and anxiety due to overlapping symptoms, with the model occasionally misclassifying semantically similar conditions despite exhibiting overall robust performance. The study demonstrated that RoBERTa's dynamic masking approach and its comprehensive pre-training allowed the model with enhanced capacity to identify complex semantic dependencies compared with static embeddings. Importantly, the findings revealed enduring ambiguity between depression and anxiety arising from overlapping symptoms, with the model occasionally misclassifying semantically similar conditions despite a generally robust performance.

2.1.2 The Domain Adaptation Paradigm

The shift towards domain-specific pre-training signifies a significant methodological transformation in the field of mental health natural language processing. The authors in [7] pre-trained MentalRoBERTa, a RoBERTa variant on mental health subreddit corpora, demonstrating that exposure to domain-specific linguistic patterns—informal language, symptom descriptions, and community discourse norms—helped achieve substantial performance advantages over general-purpose models. Building upon this foundation, the authors of [17] utilized multiMentalRoBERTa to address multiclass detection across six categories (Stress, Anxiety, Depression, PTSD, Suicidal Ideation, Neutral), achieving a macro F1-score of 0.839 and significantly outperforming traditional baselines such as SVM (0.73 F1). Domain-adapted models excel at reducing ambiguity conceptually overlapping conditions like depression versus suicidal ideation by learning subtle linguistic markers specific to each condition.

Cross-condition transfer learning has surfaced as an effective approach to overcome data shortages in less-studied areas of mental health. The authors of [13] integrated StressRoBERTa, which was pre-trained on a corpus of 108 million words sourced from communities focused on depression, anxiety, and PTSD. The model reached an F1 score of 0.82 for stress detection, exceeding the original SMM4H 2022 Task 8 champion by 3 percentage points and going beyond the standard MentalRoBERTa baseline. This confirms that the semantic indicators of comorbid conditions demonstrate cross-domain relevance—a concept that is significant when utilizing mental health-adapted frameworks for the classification of insomnia even though insomnia is not a central target during pre-training.

2.2 Comparative Model Architectures: Discriminative Transformers, Generative LLMs, and Hybrid Systems

2.2.1 Performance Characteristics Across Model Classes

Recent benchmarking studies show that different models perform better than others depending on tasks, challenging the assumption that larger models universally outperform smaller specialized architectures. In a comprehensive evaluation across 51,000+ social media posts, the authors of [18] utilized TF-IDF + SVM, prompt-engineered GPT-4o-mini, and fine-tuned GPT-4o-mini for seven-class mental health classification. Surprisingly, the standard SVM approach achieved the highest accuracy of 95%, significantly exceeding the performance of fine-tuned LLMs, which scored 91% at 3 epochs, and prompt-engineered techniques (65%). The prompt-engineered model had particular challenges with 'Stress' (F1: 0.33) and 'Personality Disorder' (F1: 0.31), frequently confusing depression with suicidal ideation. The study's conclusion emphasized that specialized classical machine learning with careful feature engineering can perform better than a generalized LLM.

This observation— smaller, specialized models exhibit superior performance compared to general-purpose LLMs in nuanced discrimination tasks—has been extensively documented over time. The researchers in [14] investigated Llama 3 8B, GPT-3.5, and discriminative transformers focused on detecting anxiety, depression, and stress. The findings indicated that fine-tuned smaller models frequently outperformed their larger generative counterparts. For example, Distil-RoBERTa achieved an F1 score of 0.883 in anxiety detection, while XLNet reached an F1 score of 0.891 in depression detection. The basis of the architectural explanation lies in the training objectives: discriminative models concentrate on optimizing categorical separation by employing cross-entropy loss on labeled examples, enabling them to learn precise decision boundaries. Conversely, generative models are trained on next-token prediction from a wide range of corpora, prioritizing a comprehensive understanding of language rather than the precision of specific classifications. For tasks that require a detailed distinction between categories that semantically overlap, the application of discriminative architectures with customized pre-training results in exceptional performance.

2.2.2 Hybrid Architectures: Synergistic Combinations of Discriminative and Generative Models

Recent studies indicate that combining generative LLMs' reasoning capabilities with the classification efficiency of discriminative transformers results in systems that utilize their respective strengths effectively. The authors of [20] presented the 'Lotus' system designed for emotion classification, which utilizes a two-stage pipeline in this process, Llama-3-8B generated contextual explanations that clarify emotional ambiguities, which are then combined with the original text to fine-tune RoBERTa-large. This augmentation technique has proven to be effective in addressing issues of class imbalance and data scarcity, achieving a macro F1 score of 0.7396 compared to 0.7112 when done without the explanation. The findings illustrate that generative context disambiguation significantly improves discriminative performance by offering clear reasoning chains that can assist in resolving interpretive ambiguities.

The researchers in [12] have proposed a mental health framework that is guided by large language models (LLMs) employing advanced large language models for initial reasoning and insight extraction. This process subsequently directs the fine-tuning of domain-specific models such as MentalRoBERTa, which are utilized for the classification of disorders and the assessment of their severity. Although this approach of 'LLM-guided fine-tuning' is temporally distinct from post-classification analysis, both methodologies acknowledge an optimal division of labor: LLMs are proficient in complex medical reasoning and contextual interpretation, whereas specialized transformers are proficient in categorical discrimination.

The authors' [19] success in attaining first place in the *eRisk@CLEF* 2025 Pilot Task serves as further validation of the reliability of modular LLM systems. The dual-agent architecture they employed—featuring one Llama-3.1-8B-Instruct agent engaged in empathetic dialogue and a second agent evaluated symptom severity resulted in the highest Depression Category Hit Rate (0.66) and the lowest Average Difference in Overall Depression Levels (0.93), all while ensuring conversational efficiency (6.54 messages per run). This indicates that the implementation of task-specific LLM deployment, as opposed to a monolithic all-in-one application, boosts both performance and interpretability.

2.2.3 Zero-Shot and Few-Shot LLM Applications

The deployment of zero-shot large language models (LLMs) in clinical screening has unveiled both potential and variability in performance that are model-specific. The authors of [21] performed an evaluation of seven large language models—GPT-4o, Llama 3 8B, Cohere, and Gemini—focusing on their capacity to predict PHQ-8 depression scores derived from clinical interview transcripts, employing the RISEN prompt engineering framework. GPT-4o exhibited the highest overall performance, achieving a binary accuracy of 75.9% (F1: 0.74), with significant strengths in cognitive and emotional symptom evaluation. However, model-specific strengths became evident: Llama 3 8B excelled in identifying anhedonia, achieving an accuracy of 77.0%, while Cohere reached an accuracy of 94.0% in evaluating psychomotor symptoms. The differences in the capabilities of large language models (LLMs) across various symptom categories highlight the necessity for strategic model selection—advocating for the use of specific LLMs for particular clinical reasoning tasks rather than assuming a uniform level of superiority.

2.3 Unsupervised Neural Topic Modeling for Clinical Phenotyping

The utilization of neural network-based topic modeling in the analysis of mental health texts facilitates the automated identification of underlying patterns without the necessity for labeled training data for each distinct category. The authors of [16] focused on suicide risk detection by analyzing posts from the Reddit communities r/SuicideWatch and r/Teenagers, utilizing fine-tuned Llama 3 8B and Mistral 7B for binary classification (suicidal versus non-suicidal), achieving a test accuracy of 93.71% with Llama 3. After classification, they implemented BERTopic enhanced with KeyBERT and GPT-4o, for label generation to uncover underlying themes. The analysis highlighted significant thematic differences: suicidal posts were primarily focused on long-term, profound distress (relationship struggles: 3,285 posts; academic failure: 1,791 posts; childhood/family trauma: 1,421 posts) while nonsuicidal posts were more oriented towards temporary stressors. This approach—merging sentence transformers, UMAP dimensionality reduction, K-Means clustering, and LLM-based labeling—offered a robust framework for identifying clinically relevant categories from unstructured text.

A notable progress in the approach to clinical topic modeling involves the engineering of few-shot prompts that are utilized for the creation of category labels. It is essential to have expert analysis to convert raw keyword clusters like [“pill,” “doctor,” “stopping”] into clinically relevant diagnoses. BERTopic by default extracts keywords that are often difficult to interpret and more like technical terms instead of diagnostic labels, which requires post-processing with LLMs to produce understandable names for the themes. By applying custom stopword filtering, which removes general terms such as ”insomnia,” ”sleep,” and ”symptoms,” it compels differentiation based on unique triggers instead of common symptoms. This guarantees that any topics found will truly reflect differences in causes rather than mere similarities in vocabulary at the surface level.

2.4 Layer Integrated Gradients for Model Interpretability and Clinical Validation

The integration of explainability methods addresses the hesitation to rely on AI and validation requirements unlike black box situations in high-stakes healthcare applications. The authors of [17] employed Layer Integrated Gradients (LIG) in their multiMentalRoBERTa study to visualize token-level attributions, demonstrating that the model highlighted symptom-specific lexical triggers (e.g., “hopeless,” “end it”) rather than unrelated patterns. LIG computes attribution scores to each token by summing gradients along a path from baseline (zero vectors) to the real input embeddings, showing how much a token has effect on the model’s output.

LIG has some theoretical benefits compared to other techniques such as SHAP when it comes to interpreting transformers. Calculating SHAP’s Shapley values works great for static features. But there are some issues when applying it to dynamic contextualized embeddings, where the meaning of a token can change depending on the context. The path integration of gradients in LIG integrates effortlessly with the attention mechanism of transformers giving attributions that match the representations learned by the models. Standard LIG implementations generate subword-level attributions that cannot be clinically interpreted, which requires aggregation algorithms to reconstruct full-word importance scores.

2.5 The Insomnia Gap: Underexplored Territory in Sleep Disorder NLP

2.5.1 Current State of Sleep Disorder Research

While works on mental health involving NLP have matured substantially, computational sleep disorder research remains notably underdeveloped, relying predominantly on primitive text analysis methods. The authors of [15] conducted the only large-scale computational analysis of r/insomnia, examining 340,130 comments spanning 2008-2022 using bag-of-words modeling for 35 treatment-related keywords, regular expression analysis, and VADER sentiment analysis. Their findings revealed treatment discussion trends: melatonin emerged as most frequently mentioned (15,005 mentions), followed by CBT-I (13,461 mentions) and Ambien (11,256 mentions), with CBT-I discussions increasing from 2% (2013-2014) to over 5% (2018). The temporal analysis demonstrated growing awareness of evidence-based non-pharmacological interventions within the r/insomnia community.

On the other hand, this method identified references to treatments without making distinctions in diagnostic categories. The bag-of-words model is insufficient for recognizing the semantic differences that separate primary insomnia, defined as an independent sleep disorder, from secondary insomnia, which refers to sleep disturbances caused by coexisting medical or psychiatric conditions. Both regular expressions and sentiment analysis exhibit a deficiency in contextual comprehension, which is necessary to determine whether a post that mentions "can't sleep due to anxiety" refers to an anxiety disorder that causes insomnia (secondary) or to anxiety concerning insomnia itself (primary with secondary psychological distress). No research has employed transformer-based approaches to recognize insomnia subtypes or the root causes behind them.

2.5.2 Diagnostic Differentiation as an Unsolved Problem

The absence of prior work computationally differentiating primary versus secondary insomnia represents a significant research gap with direct clinical implications. Clinical sleep medicine distinguishes these subtypes because they require different therapeutic approaches: primary insomnia typically responds to sleep-specific interventions like CBT-I, whereas secondary insomnia necessitates treating underlying medical or psychiatric conditions alongside sleep management. At present, there is no existing paper that performs this automatic classificaion of insomnia subtypes.

In making comparisons to the studies aimed at differentiating between depression and anxiety—which have revealed significant classification ambiguity due to overlapping symptoms, the density of medical entities could serve as a measure for linguistic validation. Discussions regarding secondary insomnia should inherently include more medical context, such as the names of diseases, references to medications, and relevant comorbidities. This approach should be consistent with clinical diagnostic criteria, as the diagnosis of secondary insomnia necessitates the identification of an underlying medical cause. This assumption continues to lack empirical testing in the field of sleep disorder NLP.

2.5.3 Emotional Burden as an Unexplored Dimension

Existing sleep disorder research relies on binary sentiment analysis (positive or negative), which fails to capture clinically relevant emotional states. Depression research has demonstrated that multi-dimensional emotion profiling (e.g., hopelessness, anhedonia, psychomotor agitation) predicts treatment response better than overall mood scores. No prior work has applied emotion phenotyping to causes of insomnia, despite evidence that withdrawal-related insomnia will present with a different distress profile compared to idiopathic insomnia cases.

The characterization of emotions across multiple dimensions has the potential to deliver useful insights for clinical practice. As an illustration, significant Perplexity scores which denote confusion about the condition encourage the implementation of psychoeducational interventions. Conversely, high Helplessness scores might reveal a need for motivational interviewing. Identifying particular triggers, such as benzodiazepine withdrawal or chronic pain, and correlating them with psychological states like desperation, dread, and perplexity would align with the symptom-specific diversity observed in various mental health research forming clinically relevant phenotypes.

2.5.4 Validation Gaps and Translational Challenges

The present state of mental health research in the domain of natural language processing (NLP) is encountering a crisis of reproducibility, models that have been trained on data from Reddit demonstrate an accuracy exceeding 90%+ [18], yet they encounter challenges in clinical implementation as a result of of discrepancies in population characteristics. Models of sleep disorders are subject to the same risks: users of r/insomnia tend to be younger, more familiar with technology, and less responsive to treatment than the general insomnia demographic. Currently no existing research has confirmed the accuracy of sleep disorder classifiers derived from social media when compared to polysomnography or structured clinical interviews (SCID) which restricts their translational applicability. This scarcity of evidence shows the obstacles when it comes to clinical adoptation and pinpoints the importance of careful interpretation of Reddit findings.

Chapter 3

Methodology

This study focuses on a comprehensive computational framework for detecting and analyzing secondary insomnia cases from day-to-day social media posts. Rather than focusing on simple identification and classification accuracy this study prioritized the discovery and phenotyping of underlying conditions that drive the insomnia symptoms. The pipeline uses a supervised deep learning methodology for initial case detection, after the identification stage an LLM produces a clinical summarization, following that an unsupervised topic modelling is used to extract the etiological factor, and finally a zero-shot emotion profiling to characterize the emotional burden that often stems from various insomnia triggers.

3.1 Dataset Construction and Annotation

600 posts were meticulously chosen from r/insomnia and r/sleep subreddits through a set of keywords that mention sleep difficulties, insomnia symptoms and related comorbidities. To ensure data quality, posts shorter than 50 characters and non-English content were discarded. Two trained annotators with backgrounds in general medicine who are experts in diagnosing sleep related issues, independently performed binary labelling of each post. Following annotation criteria were adhered:

Required Criteria for Primary Insomnia:

- Difficulty falling asleep at night
- Difficulty staying asleep / frequent awakenings during the night
- Early morning awakening and being unable to return to sleep
- Non-restorative / poor-quality sleep
- Daytime impairment (fatigue, mood disturbances, concentration issues)

Required Criteria for Secondary Insomnia:

- Difficulty falling asleep at night
- Medical conditions (e.g., chronic pain, asthma, hyperthyroidism)

- Psychiatric conditions (e.g., depression, anxiety)
- Substance use (e.g., caffeine, alcohol, medications)
- Environmental factors (e.g., noise, light, uncomfortable sleep environment)

Inter-annotator reliability for the 600-post dataset was measured using Cohen's kappa. The two medical experts achieved an observed agreement of 86%, resulting in a kappa score of 0.72 ($\kappa = 0.72$). Discrepancies ($n = 84$) were resolved through a collaborative approach to ensure the quality of the dataset.

Primary Insomnia Post	**Secondary Insomnia Post**
I've been trying to sleep at night from the past 2 months and failing. When I try to sleep my head is filled with racing thoughts. I tried everything that Google advised me to do — read a book, breathing exercises, meditation, no blue screen before bed for 2 hours, drank plenty of water. What not but still I can't sleep till 4:30 am. Any tips would be greatly appreciated. Thank you.	Lately I've been having trouble sleeping. My anxiety gets worse when I'm getting ready for bed. I've had two panic attacks right before I'm able to sleep. Been having trouble breathing, I overthink everything, and I've been stressing a lot lately. Can that be the problem? Stress? Anyone else been having trouble with this? I'm taking the drivers ed, and I'm in the middle of moving. Also I recently went on an interview and I haven't heard anything yet. I'm guessing it's all just stress, but I keep overthinking that it might be something else. I've just been having really bad anxiety lately. I'm sorry for rambling, my mind has just been running so much.

Table 3.1: Examples of Primary Insomnia Posts and Secondary Insomnia Posts on Reddit Social Media

3.2 Biomedical Entity Density and Statistical Validation

To assess linguistic genuineness and semantic differences, a medical entity density analysis was conducted using scispaCy's `en_core_sci_sm` model, a biomedical named entity recognition (NER) system trained on PubMed abstracts and clinical notes. Each density was computed as the count of recognized medical terms (DISEASE, CHEMICAL, ENTITY tags) per 100 words to normalize for document length. Mann-Whitney U test, a non-parametric alternative to the independent sample t-test, was used to compare the distributions of entity densities. To assess the magnitude of difference between groups, Cohen's d was used to quantify the effect size.

3.3 Classification Models and Training

3.3.1 Baseline Models

The study utilized the pretrained Google News Word2Vec model, which was trained on a massive corpus of approximately 100 billion tokens to generate 300-dimensional dense vector representations. Pretrained embeddings were loaded via the Gensim library to convert the textual data into structured numerical data for proper classification. Support Vector Machines (SVM) with RBF kernel and Random Forest ensembles with default hyperparameters were also trained.

3.3.2 RoBERTa-base

In order to compare against domain-adapted architectures, RoBERTa-Base was used, which is a general-domain model. It was fine-tuned using a maximum sequence length of 256 tokens and a batch size of 8. 3 epochs were used to train the model. The first epoch used frozen transformer layers to make the classification head stable before going into full fine-tuning. In order to address imbalanced class, oversampling was used to the minority class while training. Five-fold cross-validation was used to measure the model's performance.

3.3.3 Mental-RoBERTa

Mental-RoBERTa, a domain-adapted variant of RoBERTa-base specifically fine-tuned on mental health subreddits, was integrated as the initial classification model. Mental-RoBERTa processes input through 12 transformer layers with 768-dimensional hidden states. The model was fine-tuned using 5-fold stratified cross-validation. The following hyperparameter configuration was implemented for each fold's training purpose:

1. Batch size of 8 with gradient accumulation over 2 steps (effective batch size = 16),
2. Maximum sequence length of 256 tokens,
3. Learning rate of 2×10^{-5} with linear warmup over 10% of training steps,
4. Weight decay of 0.01, and
5. Maximum 8 epochs with early stopping based on validation macro F1-score (patience = 2 epochs).

The following training strategy was applied: phased learning with initial encoder freezing, unfrozen for full model fine-tuning. Best performing fold was selected based on validation F1-score for subsequent downstream analysis.

3.3.4 DeBERTa

Microsoft's DeBERTa-v3-base was fine-tuned for comparative analysis. Identical cross-validation splits with a lower learning rate (6×10^{-6}) and higher weight decay (0.15) was incorporated to train DeBERTa and account for architectural differences.

3.3.5 Llama-3 8B

Additionally, Llama-3 8B was fine-tuned using a 5-Fold Stratified Cross-Validation strategy to ensure robust performance estimation. The model utilized Low-Rank Adaptation (LoRA) ($r = 8$, $\alpha = 16$, dropout 0.1) and integrated NEFTune ($\alpha = 5$) to enhance generalization. Training was conducted over 3 epochs at a learning rate of 5×10^{-5}.

3.4 Model Interpretability Analysis

Layer Integrated Gradients (LIG) from the Captum library was applied to make sure classifications were based on clinically relevant features. LIG calculates the contribution of each word when it comes to model's prediction, by measuring gradients along a path from a baseline state (zero vectors) to the actual word embeddings. This approach has theoretical advantages over SHAP (SHapley Additive exPlanations) for transformer models because LIG directly checks the embedding layer's computational process, aligning with the model's decision making process. The analysis used Primary Insomnia (class 1) as the reference category, and scores showed the words that directed the predictions toward Primary versus Secondary Insomnia (class 0). Since RoBERTa uses byte-pair encoding that divides words into pieces of subword, a custom made aggregation algorithm was used to combine these into complete words with importance scores normalized to a -1 to 1 range. The HTML visualizations utilize color coding on a green-to-red scale, where green shows words pushing classification toward Primary Insomnia (such as "can't sleep") and red indicates words linked with Secondary Insomnia (such as "medication" and "anxiety").

3.5 Large-Scale Inference and Confidence-Based Filtering

5,000 posts were curated from the target subreddits using an expanded keyword set and processed using the best-performing Mental-RoBERTa classifier, to measure real-world implications. Large-scale inference was conducted by filtering posts predicted as Secondary Insomnia (label = 0) using a confidence-based threshold on the model's softmax output.

This confidence-based filtering strategy offsets the precision–recall imbalance present in disproportionate classification and ensures that large-scale inference remains both reliable and sufficiently comprehensive for downstream clinical analysis.

3.6 Clinical Summarization via Large Language Models

In order to extract structured clinical narratives suitable for etiological discoveries, Llama-3 8B instruct was deployed as a medical reasoning agent. A prompt, which was engineered to process each post thoroughly and construct concise clinical summaries, instructed Llama-3 to:

1. Detect the underlying cause or comorbid issues related to insomnia mentioned in the post,
2. Provide direct textual evidence, and
3. Describe the mechanistic pathway disrupting sleep.

Outputs contained coherent paragraphs of 3-4 sentences avoiding bullet points or responses that are fragmented to support a downstream NLP structure. The clinical analysis summaries generated by the Large Language Model (LLM) were validated by subject matter experts. Inference was used with 4-bit quantization via bitsandbytes at batch size 4.

Post	**Clinical Analysis**
This just started in the last month or two. Each night I'm collapsing into my mattress, finally starting to relax after the worst of the nighty night hot flashes have passed, when I get a persistent creepy, achy, wormy feeling under my skin on each leg above my knees. Can't sleep. Can't stay still. Can't stand it. If this isn't menopause-related I apologize for posting here, but EVERYTHING seems to be hormone-related. I'm 5 years post-menopausal by the way. Any suggestions?	The patient's sleep disruption is driven by Restless Legs Syndrome, which creates a strong urge to move the limbs due to uncomfortable "wormy" sensations that worsen during evening rest. Despite the user's attempts to relax after their hot flashes subside, the physical discomfort in their legs creates an unavoidable sensory barrier to falling asleep. Because the inability to rest is caused by a specific neurological sensory disorder and menopausal hormonal changes rather than an idiopathic sleep defect, this is classified as secondary insomnia.

Table 3.2: Example of Clinical Summarization of a Post

3.7 Etiology Discovery via BERTopic

The input corpus for etiological discoveries of BERTopic was formed by the Llama-generated clinical summaries. Following topic modeling pipeline was used:

- **Semantic Embedding:** Clinical summary converted into 384-dimensional dense vector representation by the all-MiniLM-L6-v2 sentence transformer to capture semantic meaning.
- **Dimensionality Reduction:** UMAP (Uniform Manifold Approximation and Projection) reduced embeddings to 5 dimensions (`n_neighbors=15, min_dist=0.0, metric='cosine'`) while preserving local neighborhood structure.
- **Clustering:** K-Means clustering was applied to the reduced embeddings, guided by Silhouette analysis to balance cluster stability with clinical and semantic granularity.
- **Keyword Extraction:** Statistically definitive terms identified through KeyBERTInspired for each cluster using semantically-aware TF-IDF weighting (`ngram_range=(1,2)`).
- **Label Generation:** Llama-3 8B synthesized using few-shot prompting that clusterd keywords into clinically relevant diagnostic labels.

3.8 Emotional Burden Profiling

BART-large-MNLI, a 400-million parameter model trained on natural language inference tasks, repurposed for zero-shot classification, assessed whether input text semantically involves specific emotional states without requiring emotion-labeled training data. Based on their relevance to chronic illness, each clinical summary was scored against six candidate emotions selected: Stress, Frustration, Desperation, Helplessness, Perplexity, and Dread.

The selection of these specific psychological phenotypes—Dread, Helplessness, Frustration, Stress, Perplexity, and Desperation—is grounded in established clinical models and empirical research that characterize the complex emotional and physiological landscape of insomnia. Dread represents the pre-sleep anxiety and excessive negatively toned cognitive activity described by [3], which triggers autonomic arousal and selective attention toward sleep-related threat cues. This state often transitions into a pervasive sense of Helplessness, where the repeated inability to initiate sleep reinforces negative schemas regarding a lack of control, serving as a primary risk factor for the development of clinical depression [5]. This cognitive burden is compounded by Frustration, as patients often feel misunderstood due to a mismatch between their lived experience of daily functional impairment and the responses they receive from healthcare professionals [6]. Physiologically, these states are underpinned by Stress and the hyperactivity of the hypothalamic–pituitary–adrenal

(HPA) axis, with elevated cortisol levels serving as a marker for the systemic hyperarousal that maintains insomnia [11]. For individuals specifically experiencing secondary insomnia, Perplexity captures the diagnostic challenge of causal attribution—the difficulty patients and clinicians face in determining whether an accompanying disorder is the cause or a comorbidity of the sleep disturbance [2]. Finally, Desperation characterizes the high level of psychological distress and daytime fatigue that reaches a critical threshold, serving as the primary determinant that prompts individuals to seek clinical treatment and intervention [4].

The complex affective experiences of insomnia sufferers were reflected by the multi-label classification setting (`multi_label=True`), which allowed posts to score highly on multiple emotions simultaneously. A custom made stopword filtering process removed words that are too common such as "insomnia," "sleep," "symptoms," and "disorder" and made the model to differentiate clusters based on specific cause rather than shared symptomatology. Multi-word medical concepts like "REM behavior disorder" were captured by Bigram extraction (`ngram_range=(1,2)`).

Lastly, An Emotion-Etiology Heatmap was constructed using emotion scores that were averaged across different underlying causes, with intensity measured on a scale from 0 to 1.

Chapter 4

Experimental Results

4.1 Evaluation Metrics

In order to evaluate classifier performance, standard metrics were used as it is the industry-standard practice. Metrics used: accuracy, precision, recall, and F1-score, collected from the confusion matrix containing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics collectively assess how effectively the models detect insomnia-related posts.

1. **Accuracy** measures overall correctness by calculating the ratio of correctly classified instances to total instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Precision** evaluates the model's accuracy in positive class prediction by calculating the ratio of correctly predicted positive samples to all predicted positives:

$$Precision = \frac{TP}{TP + FP}$$

3. **Recall** is a metric that is also known as sensitivity. It quantifies the percentage of positive samples correctly predicted out of all the samples in the actual class:

$$Recall = \frac{TP}{TP + FN}$$

4. **F1-Score** is a weighted average of precision and recall. It considers both false negative and false positive results:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.2 Empirical Findings

To measure the density of medical entities (`DISEASE`, `CHEMICAL`, and `ENTITY` tags) per 100 words in the 600-sample annotated dataset, the *encoresci_sm* biomedical NER model from scispaCy, was employed.

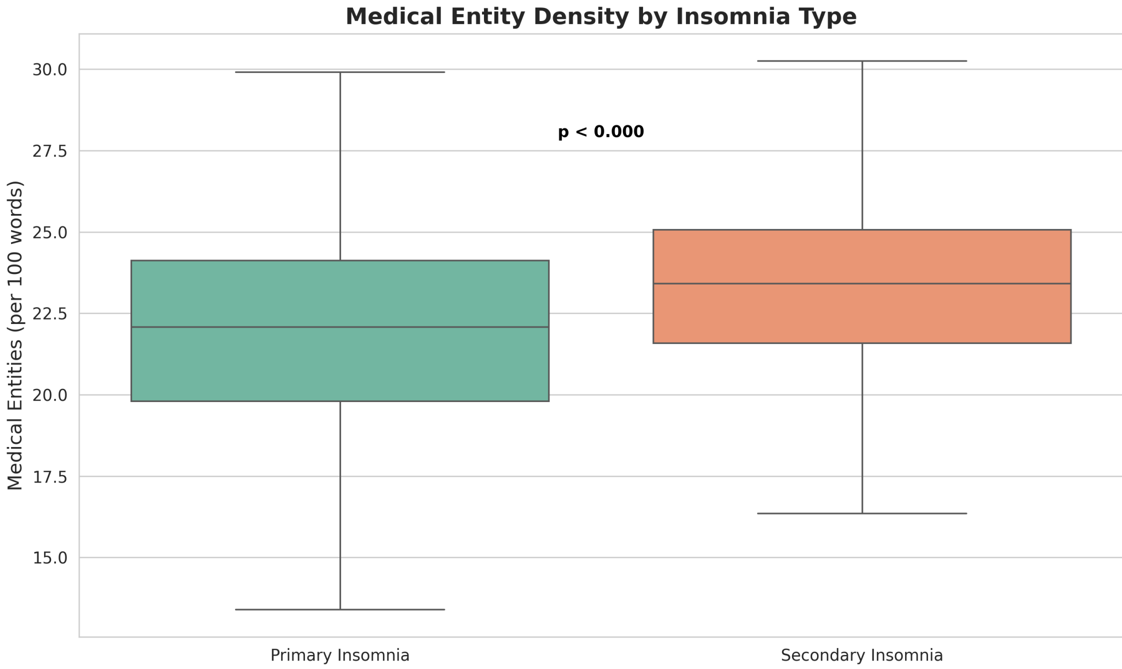


Figure 4.1: Medical Entity Density

Posts on secondary insomnia had a high number of medical terms as compared to primary insomnia posts (median: 23.41 vs. 22.08 entities per 100 words). The statistical significance (U = 55,629.5, $p < 0.00001$) and medium effect size (Cohen's d = 0.433) was proved by the Mann-Whitney U test. People who characterised secondary insomnia intuitively inserted more disease names, drug names and comorbidity references in their account and primary insomnia accounts described more of sleep manifestations without reference to external medicine.

4.3 Classification Performance

4.3.1 Baseline Models

Traditional machine learning classifiers established performance benchmarks using Google News' pre-trained 300-dimensional Word2Vec embeddings on the 600-post dataset. Random Forest achieved optimal baseline performance: 74.17% accuracy with precision 0.75, recall 0.74, and macro F1 0.74. Support Vector Machine with RBF kernel reached 72.50% accuracy (precision 0.74, recall 0.72, macro F1 0.72).

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	0.742	0.75	0.74	0.74
SVM	0.725	0.74	0.72	0.72

Table 4.1: Performance comparison of Random Forest and SVM

4.3.2 RoBERTa-base

RoBERTa-base fine-tuning using 5-fold stratified cross-validation demonstrated consistent and robust performance across folds. The model achieved a mean cross-validated accuracy of 80.50% (SD ± 1.73%) with a macro F1-score of 81.36% (SD ± 1.24%), indicating stable generalization despite limited training data. The best-performing fold attained 82.50% accuracy, with balanced class-wise discrimination reflected in strong macro-averaged precision (0.82) and recall (0.88), supporting the effectiveness of RoBERTa-base for insomnia subtype classification.

Fold	**Accuracy**	**Precision**	**Recall**	**F1-score**
Fold 1	0.7917	0.7500	0.8500	0.8092
Fold 2	0.8083	0.7879	0.8500	0.8099
Fold 3	0.7833	0.7465	0.8333	0.7969
Fold 4	0.8167	0.8033	0.8667	0.8254
Fold 5	0.8250	0.8197	0.8833	0.8264
Mean	**0.8050**	**0.7815**	**0.8500**	**0.8136**
Std.	**0.0173**	**0.0324**	**0.0264**	**0.0124**

Table 4.2: 5-Fold Cross-Validation Performance of RoBERTa Classifier

4.3.3 Mental-RoBERTa

Mental-RoBERTa fine-tuning via 5-fold stratified cross-validation substantially improved over baseline methods. Mean cross-validated accuracy reached 84.00% (SD ± 1.81%) with macro F1-score 84.23% (SD ± 1.99%). The best performing fold achieved 86.67% accuracy with balanced class performance: Secondary Insomnia (precision 0.87, recall 0.90, F1 0.88) and Primary Insomnia (precision 0.87, recall 0.83, F1 0.85).

Fold	**Accuracy**	**Precision**	**Recall**	**F1-score**
1	0.8250	0.8125	0.8667	0.8372
2	0.8250	0.7826	0.7667	0.8142
3	0.8500	0.8281	0.9000	0.8548
4	0.8333	0.8667	0.8833	0.8387
5	0.8667	0.8679	0.8667	0.8667
Mean	0.8400	0.8316	0.8567	0.8423
Std. Dev.	0.0181	0.0365	0.0522	0.0199

Table 4.3: 5-fold cross-validation performance metrics

4.3.4 Llama-3 8B

Llama-3 8B using a 5-Fold Stratified achieved approximately 77% accuracy and F1-score.

Fold	**Accuracy**
Fold 1	82.50%
Fold 2	74.17%
Fold 3	76.67%
Fold 4	77.50%
Fold 5	73.33%
Mean Accuracy	**76.83%**
Standard Deviation	**± 3.22%**

Table 4.4: 5-Fold Cross-Validation Accuracy Results

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Primary Insomnia	0.75	0.81	0.78	300
Secondary Insomnia	0.79	0.73	0.76	300
Overall Accuracy	-		**0.77**	**600**
Macro Average	0.77	0.77	0.77	600
Weighted Average	0.77	0.77	0.77	600

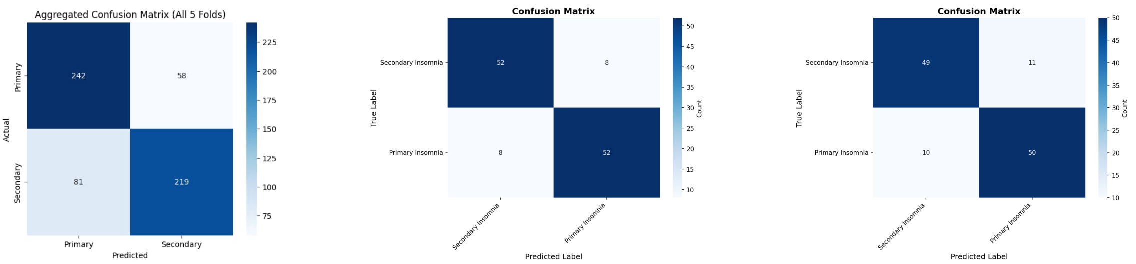
Table 4.5: Aggregated Classification Report

4.4 Comparative Analysis

Mental-RoBERTa outperformed all other models on the dataset, achieving the highest accuracy (0.8400) and F1-score (0.8423), demonstrating the benefit of domain-specific pretraining. RoBERTa-base showed competitive performance as a strong general-domain baseline, while DeBERTa-v3 and LLaMA-3 8B yielded lower and more uniform scores, indicating reduced discriminative capability for this task. Overall, the results favor domain-adapted transformer models for clinical text classification.

Model	Dataset Size	Accuracy	Precision	Recall	F1-score
RoBERTa-base	(600)	0.8050	0.7815	0.8500	0.8136
Mental-RoBERTa	(600)	0.8400	0.8316	0.8567	0.8423
DeBERTa-v3	(600)	0.7483	0.7570	0.7483	0.7462
Llama-3 8B	(600)	0.7683	0.7700	0.7700	0.7700

Table 4.6: Performance Comparison of Models



(a) Llama-3 8B (b) Mental-RoBERTa (c) RoBERTa-base

Figure 4.2: Confusion matrices

4.5 Large-Scale Filtering for High-Confidence Cases

The trained Mental-RoBERTa model was applied to 5,000 unlabeled Reddit posts from r/insomnia and r/sleep. Confidence-based filtering retained posts predicted as Secondary Insomnia (label = 0) with a softmax probability of at least 0.70. This threshold was selected as a practical operating point, balancing precision and retention based on threshold sensitivity analysis.

Applying this criterion yielded 3,423 high-confidence secondary insomnia posts (68.5% retention), while excluding 1,577 lower-confidence samples. The 0.70 cutoff provided a substantial improvement in output quality over lower thresholds while avoiding the sharp reduction in recall observed at more restrictive confidence levels, supporting its suitability for large-scale inference.

Threshold	Precision	Recall	F1-Score	Posts Kept	% Target Kept
0.50	0.8667	0.8667	0.8667	60	100.0%
0.60	0.8793	0.8500	0.8644	58	96.7%
0.70	0.8889	0.8000	0.8421	54	90.0%
0.80	0.8868	0.7833	0.8319	53	88.3%
0.90	0.9111	0.6833	0.7810	45	75.0%

Table 4.7: Confidence Threshold Analysis on Validation Set (n=60 secondary insomnia cases from best fold). Metrics reflect performance on labeled data used to select optimal threshold for large-scale inference.

4.6 Etiology Discovery via Topic Modeling

4.6.1 Prevalence Distribution

BERTopic pipeline augmented with Llama-3 8B for clinical label synthesis identified 10 distinct etiological categories from 3,423 Llama-generated clinical summaries. marked concentration can be observed using prevalence distribution : the top three categories are responsible for 56.8% of all cases.

Gastroesophageal Reflux Disease (GERD) is primarily responsible for sleep disturbances (946 posts, 27.6%), it aligns with gastroenterology literature that shows nocturnal acid reflux prompts a person to wake up abruptly. All the menopausal-related categories combined, represent 26.9%, with Menopausal Hormonal Imbalance (516 posts, 15.1%) and Menopausal Hot Flashes Trigger (405 posts, 11.8%).

Benzodiazepine Withdrawal Symptoms is the third main reason for sleep disturbances (482 posts, 14.1%), shows patients experience rebound insomnia after they stop taking GABAergic medication. This indicates substantial iatrogenic insomnia caused by prior treatment rather than independent disorder.

Additional categories include Chronic Stress Management Failure (334 posts, 9.8%), Caffeine Use Disorder (225 posts, 6.6%), Chronic Pain Syndrome (172 posts, 5.0%), Environmental Factor (133 posts, 3.9%), Nicotine Withdrawal Symptoms (107 posts, 3.1%), and Magnesium Deficiency Syndrome (103 posts, 3.0%).

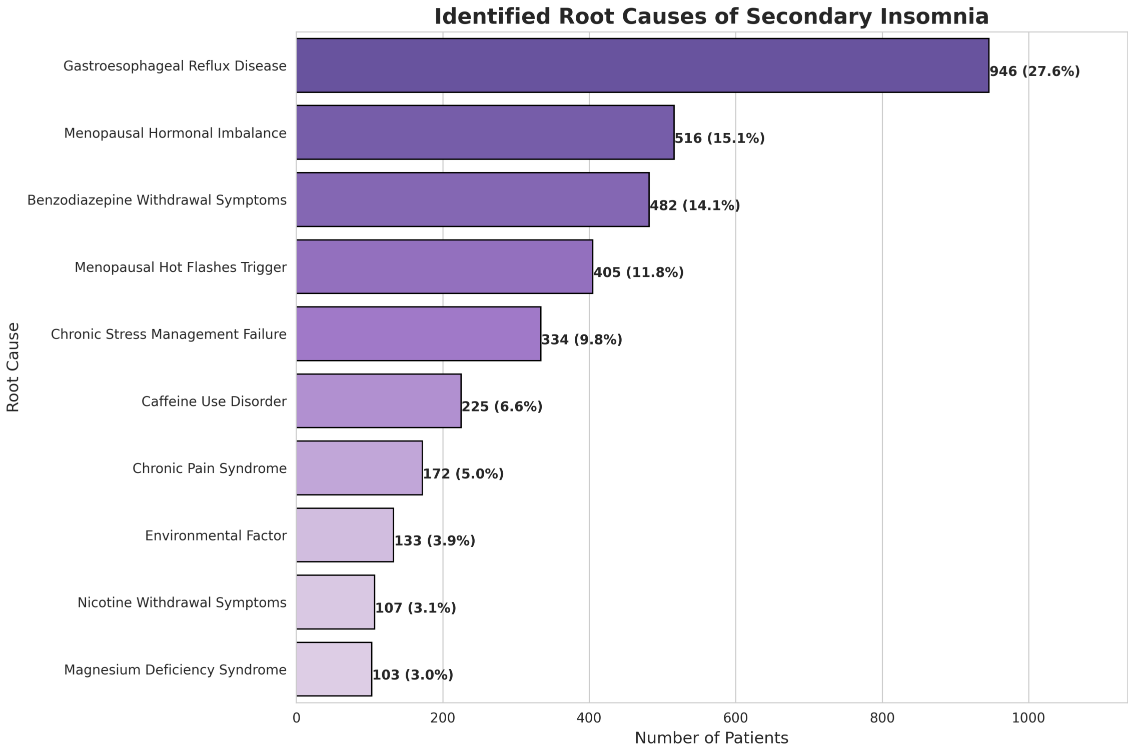


Figure 4.3: Identified Root Causes of Secondary Insomnia

4.6.2 Cluster Validation

BERTopic's multi-stage pipeline combined sentence-transformer embeddings (`all-MiniLM-L6-v2`), UMAP dimensionality reduction (5 dimensions), and KMeans clustering. Cluster quality was quantitatively assessed using Silhouette analysis across $k = 2$–20. Silhouette analysis favored k=6 (0.14), but this merged clinically distinct etiologies. We selected k=10 (Silhouette=0.08) to balance statistical separation with diagnostic granularity, accepting lower geometric cohesion for improved semantic interpretability. Manual inspection confirmed k=10 produced actionable categories (GERD, benzodiazepine withdrawal, menopausal factors) whereas k=6 conflated medication side effects with withdrawal syndromes.

4.7 Emotion Profiling of Clinical Insomnia Narratives Using Zero-Shot Classification

BART-large-MNLI zero-shot classifier applied to all 3,423 clinical summaries generated multi-dimensional emotional profiles across six candidate emotions (Stress, Frustration, Desperation, Helplessness, Perplexity, Dread). Mean intensity scores (0-1 scale) were aggregated within each etiology cluster.

Distinct emotional phenotypes emerged across etiologies. Benzodiazepine Withdrawal Symptoms exhibited highest overall emotional intensity, scoring particularly

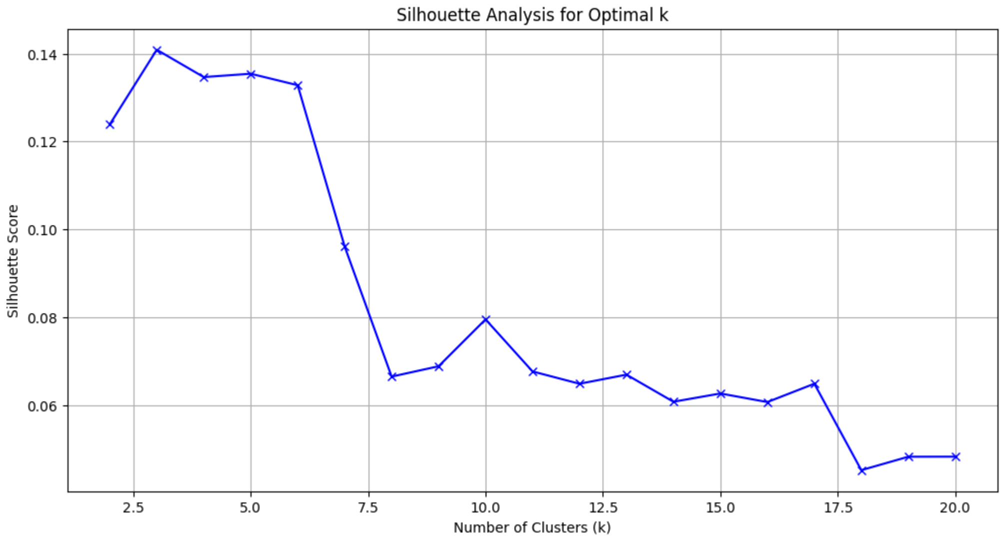


Figure 4.4: Silhouette analysis across k = 2–20

high on Perplexity (0.55) and Frustration (0.46), reflecting patient confusion about withdrawal timelines and frustration with protracted symptoms. Relatively lower Stress (0.07) and Desperation (0.23) suggest patients understand their condition but struggle with symptom duration.

Environmental Factor showed highest Frustration score (0.55) across all categories, that align with the fact that patients' think insomnia occurs from controllable but unresolved external factors. Chronic Pain Syndrome showed high level of Helplessness (0.30) and Perplexity (0.51), showing dual problems: physical suffering combined as well as uncertainty about pain-sleep interactions.

Magnesium Deficiency Syndrome showed lowest overall emotional burden, with Stress (0.08), Frustration (0.34), and Desperation (0.16) all below median values. This likely shows the fact that people think of nutritional deficiencies as straightforward, treatable conditions. On the contrary, Chronic Stress Management Failure showed elevated Stress (0.37) and Helplessness (0.31), meaning patients see stress as the primary reason but feel they are incapable of implementing effective coping strategies.

Menopausal categories show very similar profiles with moderate scores across all emotions and slightly high Perplexity (0.51 and 0.49 respectively), potentially showing uncertainty about permanence of symptom rather than temporality.

Multi-label classification was able to capture emotional complexity: posts frequently scored high on multiple dimensions continuously, showing nuanced affective experiences of chronic illness that binary sentiment analysis would be unble to capture..

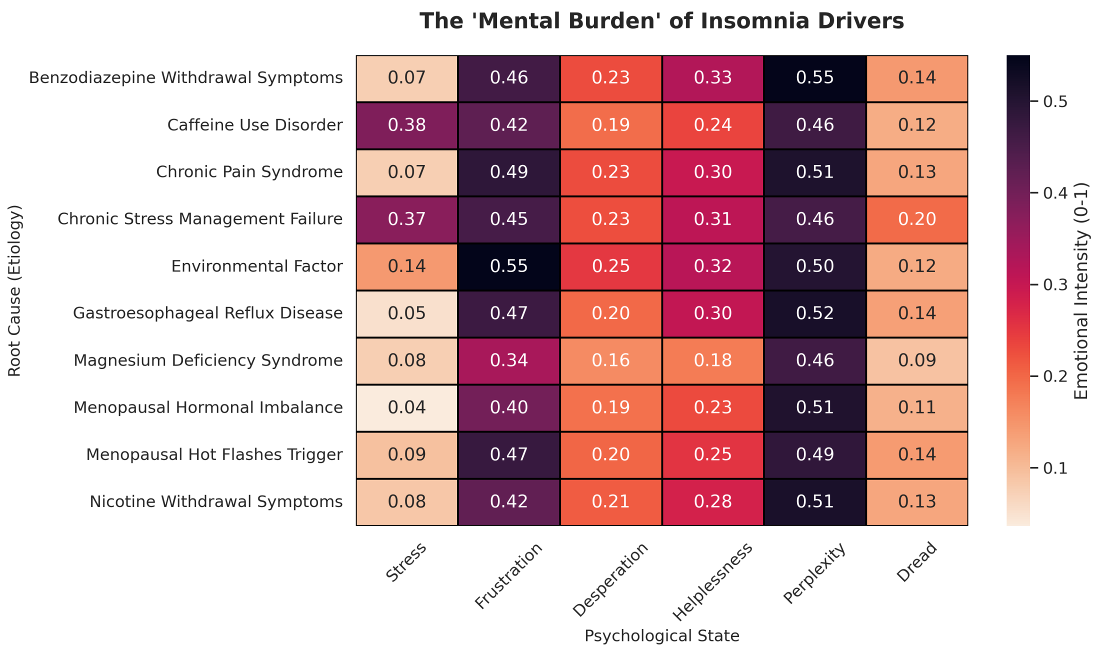


Figure 4.5: The 'Mental Burden' of Insomnia Drivers

4.8 Model Interpretability

Layer Integrated Gradients attribution analysis on Mental-RoBERTa's classification decisions showed clinically relevant decision boundaries conducive to diagnostic criteria. Token-level attributions were visualized for representative posts from both classes, with green highlighting indicating tokens driving Secondary Insomnia classification (positive attribution) and red indicating Primary Insomnia association (negative attribution).

Primary insomnia post analysis showed model focus on pure sleep symptom descriptors: temporal references ("2 hours," "minutes," "months," "night," "evening"), sleep-specific verbs ("sleep," "wake up," "repeats"), and subjective distress expressions ("sick of it," "unable to focus," "hallucinating," "headache," "fatigue"). When posts mentioned medical interventions ("tricyclic antidepressant," "neurologist"), these terms received red highlighting, indicating the model learned that mere mention of past treatment does not constitute current secondary insomnia without active comorbid condition description.

Secondary insomnia post analysis revealed strong green highlighting on medical

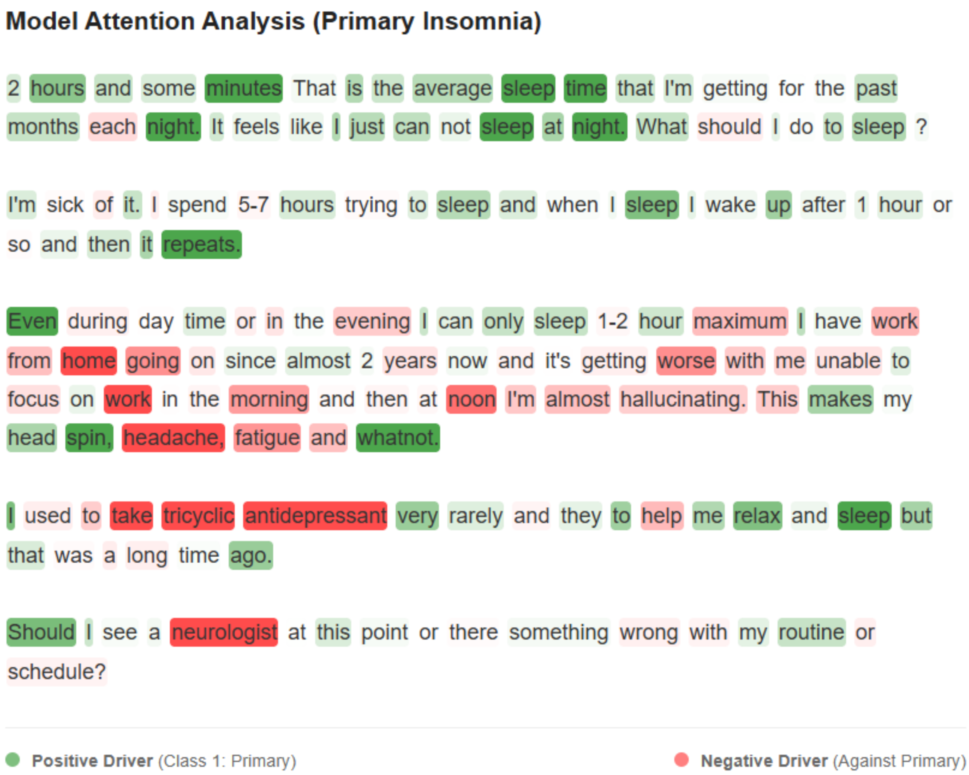


Figure 4.6: LIG attribution visualization for Primary Insomnia post

terminology and comorbidity indicators: "sleep disorder," "depression," "anxiety," "doctor," "medication," and "college" (environmental stressor). The model exhibited particular sensitivity to causal language structures. The phrase "sleep disorder appeared in October last year when I began my first year of college" demonstrated learned identification of temporal relationships between life events and symptom onset. Co-occurrence of "medication for depression and anxiety" received strong positive attribution, which indicates recognized psychiatric comorbidities as secondary insomnia drivers.

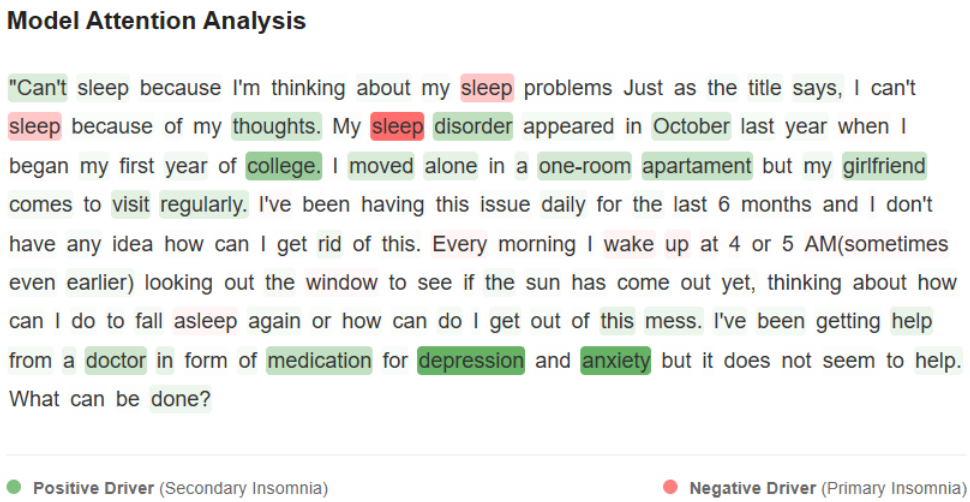


Figure 4.7: LIG attribution visualization for Secondary Insomnia

Chapter 5

Discussion

5.1 Principal Findings

This analysis provides the very first computational framework for the automatic recognition of secondary insomnia from social media discussions, additionally revealing underlying causes and emotional implications. RoBERTa-base achieved an accuracy of 80.5% and Mental-RoBERTa recorded an accuracy of 84.00% for the differentiation of primary and secondary insomnia, exceeding the performance of traditional baselines (74.17%) and large language models (Llama-3 8B: 77%). The combination of the hybrid architecture, which merges discriminative classification with LLM-based summarization, led to the discovery of 10 etiological categories. This analysis highlighted GERD as the primary contributor at 27.6%, followed by menopausal factors at 26.9% and benzodiazepine withdrawal at 14.1%. Zero-shot emotion profiling revealed various psychological phenotypes, where benzodiazepine withdrawal presented the highest perplexity (0.55) and environmental factors led to the greatest frustration (0.55).

5.2 Domain Adaptation and Model Performance

The superior performance of Mental-RoBERTa confirms that pre-training tailored to the domain of mental health discourse provides significant benefits for the classification of clinical texts. This further develops the conclusions drawn by the authors of [13] illustrated the advantages of domain adaptation in the detection of depression and stress, and now extends these findings to include sleep disorders, even though insomnia was not a primary focus during the pre-training phase. The model shows a 9.83 percentage point better result compared to general baselines, showing that engagement with mental health community discussions—featuring symptom descriptions, treatment narratives, and emotional expressions found in Reddit support forums—facilitates the model's ability to learn linguistic patterns that general-purpose embeddings cannot capture.

The observed performance gap between Mental-RoBERTa and Llama-3 8B, even with Llama-3's substantial 64-fold parameter superiority, is consistent with new evidence suggesting that smaller, task-specific models can excel beyond general-purpose large language models (LLMs) in tasks requiring nuanced discrimination. According to the authors of [14], similar observations were noted, where fine-tuned discriminative transformers showed better performance compared to Llama 3 and GPT-3.5 for anxiety and depression detection. Discriminative architectures are made to optimize categorical separation by supervised loss on labeled examples. On the other hand, generative models that are trained with next-token prediction tend to prioritize a broader understanding of language, sometimes at the expense of specific classification accuracy. This distinction becomes particularly challenging in cases of semantically overlapping categories, such as primary and secondary insomnia, which share common sleep complaints but have different underlying causes. For clinical applications that face resource constraints, the outstanding performance of Mental-RoBERTa, and its lower variance (SD $\pm$ 1.81%) and less demanding computational needs, makes it a practical option for deployment.

Although Llama-3 8B achieved respectable 79% accuracy, Mental-RoBERTa was selected (84%) for large-scale deployment based on four empirical factors beyond raw accuracy:

1. Statistical robustness: Mental-RoBERTa's 5-fold cross-validation showed high stability with low variance (SD=1.81%) and a mean accuracy of 84.00%. These results outperform Llama-3, which recorded a lower mean accuracy of 80.50% and a standard deviation of 1.73%, confirming Mental-RoBERTa's superior and more consistent performance across the different data folds.

2. Computational efficiency: Mental-RoBERTa demonstrated substantially faster inference than the comparison model, making it well suited for population-scale screening. Large-scale corpus processing was completed in a fraction of the time required by the baseline, resulting in markedly lower computational overhead and proportionally reduced cloud deployment costs for sustained, high-volume inference.

5.3 Linguistic Validation and Interpretability

Medical entity density analysis confirmed that the binary schema is capable of capturing accurate diagnostic distinctions. The greater frequency of medical terminology in secondary insomnia posts (median 23.41 vs. 22.08 entities per 100 words, $p < 0.00001$, $d = 0.433$) shows that people naturally talk about more disease names and medication intake when describing sleep related issues with identifiable external causes. This extraordinary pattern aligns with the criteria of clinical diagnosis of secondary insomnia, where clear identification of underlying medical or psychiatric conditions is required. The effect size confirms the fact that this is not just a mere difference but shows a clear shift in how people talk about and make sense of their problems.

Layer Integrated Gradients (LIG) analysis ensured Mental-RoBERTa successfully aquired the knowledge of clinically relevant decision boundaries. The model properly weighted terminoliges that are about cause or time related information ("insomnia appeared when I started college") and active condition details("currently taking medication for depression") with more importance than irrelevant medical mentions with no real connection to insomnia("saw neurologist last year"), considering context as events and time, beyond just checking some keywords. This semantic analysis, including negative attribution to past medical history without current relevance, validates that the model learned diagnostic details rather than only detecting medical vocabulary.

5.4 Hybrid Architecture Synergies

The dual-phased approach: MentalRoBERTa for binary classification followed by Llama-3 for clinical summarization utilizes the benefits of both discriminative and generative models. Mental-RoBERTa provided classifications, while Llama-3's generative capabilities explained the clinical details. First, Mental-RoBERTa was used to sort posts into primary or secondary insomnia categories. Then, Llama-3 was used to generate straightforward summaries including the medical details behind each classification. This combination worked well because each model contributed its own strength. Mental-RoBERTa determined which category a post belonged to, while Llama-3 explained the clinical reasoning. This 40-fold computational difference showed the benefit of binary classification before applying resource-heavy LLM processing on a bigger scale. The clinical summaries created by Llama-3 served two useful functions. First, the lengthy posts were made into short, focused medical descriptions that were much easier to work with in the next analysis phase. Second, the varied way in which patients described their problems was changed to consistent clinical language. This standardization proved useful when moved to the BERTopic clustering phase, as the cleaned-up Llama-3 summaries grouped together far more cleaner than the original Reddit posts, which often contained off-topic complaints and unclear descriptions.

The hybrid architecture leverages architectural complementarity: Mental-RoBERTa excels at discriminative classification through supervised fine-tuning on labeled examples, while Llama-3 8B, as a generative model, produces structured clinical summaries that Mental-RoBERTa's encoder-only design cannot generate. This division of labor—classification for case identification, generation for narrative synthesis—mirrors successful hybrid systems in mental health NLP [12][20] where discriminative precision and generative reasoning are combined rather than forcing a single model to handle both tasks. While Mental-RoBERTa excels at classification through discriminative training, LLMs trained on next-token prediction demonstrate superior performance in extracting causal relationships, synthesizing multi-sentence narratives, and producing clinically coherent explanations [8]. This architectural complementarity motivated our hybrid pipeline: Mental-RoBERTa for case identification, Llama-3 for clinical reasoning.

5.5 Etiological Findings and Clinical Implications

The analysis of over 3,400 Reddit posts gave an idea of the major problems that keep people awake at night. GERD, chronic acid reflux, was found as the biggest reason for secondary insomnia at 27.6%. This is important because people who are suffering from insomnia usually do not get asked about heartburn, making them seek consultations from separate doctors. The findings proves the presense of greater interconnection that current practice fail to recognize. Multiple posts had a common pattern: nighttime reflux wakes them up, disrupts expected sleep, and creates chronic sleep problems making them suffer even on reflux-free nights. Medications such as proton pump inhibitors may produce better results than sleeping pills in the aforementioned cases.

Sleep disturbances that are caused by Menopause-related problems were the second-largest reason at 26.9%, which is divided into two classes, hormonal effects on sleep regulation (15.1%) and hot flashes physically waking them up (11.8%). This gives a better understanding of how insomnia should be treated in middle-aged women. Many women who experience insomnia as integral menopause induced problem, like hot flashes or mood changes, should refrain from treating insomnia as a separate disorder requiring sleeping pills. This has significant implications: rather than receiving treatment for primary insomnia from generic sleep clinics, hormone replacement therapy consultations would produce better outcomes.

Emotional profiling revealed distinct psychological experiences across triggers. Benzodiazepine withdrawal patients scored highest on Perplexity (0.55), showing profound confusion about symptom duration, potential permanent damage, and appropriate responses. Environmental factor cases showed highest Frustration (0.55), suggesting benefit from practical problem-solving, sound machines, schedule adjustments, rather than cognitive therapy. Magnesium deficiency cases showed minimal emotional burden (Stress: 0.08), requiring primarily medical education about supplementation. This emotional mapping enables realistic modeling of the situation: matching intervention intensity and type to both underlying cause and psychological characteristics rather than treating all insomnia in the same manner.

5.6 Resource Allocation in Underfunded Health Systems

The reason discovery module has uses for planning resources in budget-constrained healthcare systems. The pipeline could be applied to patient portal messages or telehealth transcripts to identify emerging problems, for instance, if benzodiazepine withdrawal cases suddenly spike regionally, it signals the need for better dose reduction steps. The emotional burden heatmap enables efficient prioritization: benzodiazepine withdrawal cases scoring highest on "Perplexity" (0.55) need psychoeducational resources explaining withdrawal neurophysiology and timelines, while high "Helplessness" cases (chronic stress: 0.31, chronic pain: 0.30) require motivational interviewing to rebuild treatment engagement. For underfunded community health centers with limited psychological assessment capacity, this automated phenotyping directs limited counseling toward patients that need more help while providing self-directed materials to lower-distress cases. The computational efficiency (5,000 posts classified in under 15 minutes on standard hardware) makes continuous population-level surveillance possible even in resource-constrained settings, transforming insomnia monitoring from periodic surveys to real-time detection of emerging medication side effects, cause-specific variations requiring different intervention strategies, and evolving population health needs.

5.7 Limitations

Several factors limit immediate clinical and holistic application of the findings. Although the annotated dataset containing 600 posts can be considered sufficient for feasibility, it represents a relatively small sample for modern NLP related research. A larger annotated dataset would be required in order to get more accurate confidence intervals, and the differences in performances between different models, would also benefit from it. It would also measure whether Mental-RoBERTa still outperforms other models while dealing with a larger dataset. Future work should focus on expanding the annotated dataset to 2,000-5,000 posts, it would enable better statistical comparisons and hopefully reveal reasons of insomnia that are not commonly observed in the current sample.

The construction of dataset using Reddit data undoubtedly leads to major selection bias, which reduces the possibility of making the dataset more generalized beyond this particular online community. Reddit users represent mostly a specific age group (around 15-45), more educated, more technologically literate, and predominantly English-speaking compared to the general insomniac population seeking clinical care. The r/insomnia community likely overrepresents treatment-resistant cases as individuals with easily resolved sleep disturbances may not seek online support forums. Demographic underrepresentation of elderly populations (who experience high insomnia rates) and non-English speakers means the discoveries about causes may not cover all epidemiological patterns. The observed 27.6% GERD prevalence and 26.9% menopausal contribution may be because of Reddit users demographic composition rather than correct modeling of the population.

Clinical validation against gold-standard assessments is absent e.g., Computational classifications against structured clinical interviews (SCID), polysomnography, or physician diagnoses have not been validated. The 68.5% secondary insomnia rate in the unlabeled corpus reflects classification model outputs applied to r/insomnia and r/sleep discussions, not confirmed diagnostic prevalence. The implementation of clinical applications involves future validation studies that compare the predictions made by the model with evaluations conducted by clinical experts across a variety of patient populations.

5.8 Future Directions

Increasing the size of the annotated dataset to between 2,000 and 5,000 posts is the leading priority for enhancing the model's dataset and its generalizability. A massive collection of posts or similar data would produce superior confidence intervals for model comparisons, this would minimize variance in results of cross-validation, and capture rare cases of sleep disturbances that are not present in the existing sample of 600 posts. Moreover, future research should focus on creating a corpus of diverse population consisting people of different ages, people who speak languages other than English and most importantly elderly people who are severly under-represented in Reddit data. Expanded data would also enable supporting reserved holdout test sets for final model validation, preventing potential overfitting to cross-validation splits observed in the current experimental design. Systematic annotation guidelines with explicit decision trees for ambiguous cases would improve inter-rater reliability, while annotator confidence ratings could identify inherently difficult cases requiring expert involvement or exclusion from training data.

Cross-platform generalization testing represents an important validation step before real-world deployment. The models should be evaluated on Facebook chronic insomnia support groups, specialized sleep disorder forums (sleepio.com community boards), patient portal messages from electronic health records, and telehealth intake questionnaires to establish whether learned linguistic patterns reflect universal features of insomnia discourse or Reddit-specific conventions. Platform-specific fine-tuning strategies could be developed for contexts with different description patterns due to how the text bodies are written due to inherent differences of how the portals allow the users to describe their problems. Patient portal messages likely require different feature weightings than informal Reddit posts, while character-limited Twitter discussions may emphasize different lexical markers. Multilingual extension to non-English communities would assess whether the emotion phenotyping and cause categorization process applies across cultural contexts where symptom expression and help-seeking behaviors may differ from English-speaking Reddit users.

Chapter 6

Conclusion

This research establishes foundational methods for computational secondary insomnia detection and etiology characterization from social media discourse. The better accuracy of Mental-RoBERTa confirms the importance of domain-specific pre-training for clinical differentiation of insomnia patterns. Furthermore, hybrid architectures that use efficient classification with LLM-based summarization and topic modeling enable scalable phenotyping. Sleep issues occuring from GERD, menopausal factors, and treatment related issues, such as the challenges linked with benzodiazepine withdrawal, provides hope for people who seek help for insomnia related issues on online communities. On the other hand, unique emotional phenotypes can be used to develop case-specific intervention strategies. Despite limitations requiring clinical validation, expanded datasets, and cross-platform generalization testing, this work demonstrates that computational approaches can extract clinically relevant signals from patient narratives at scale, offering complementary methods to traditional epidemiology for continuous monitoring of evolving health needs and precision matching between patient phenotypes and evidence-based interventions.

This study used publicly available data from Reddit, collected via the official Python Reddit API Wrapper (PRAW). Data collection and processing were conducted in strict accordance with Reddit's Developer Terms of Service and API Guidelines. Since the research analyzed publicly accessible posts and did not involve any direct interaction with users, therefore it did not violate any industry-standard guidelines.

Acknowledgements

Clinical Validation and Annotation

The dataset used in this research went through rigorous clinical validation to make sure it is diagnostically accurate. We earnestly thank the following medical practitioners for their role as expert annotators:

- **Dr. Alexsandro Famous Plabon Halder, MBBS**
  Bangladesh Medical and Dental Council (BM&DC) Registration No. 164516.
- **Dr. Ferdous Al Mahabub, MBBS**
  Bangladesh Medical and Dental Council (BM&DC) Registration No. 145145.

Both practitioners gave their expert review and classification of the insomnia-related case descriptions in adherence with recognized medical guidelines and diagnostic criteria. The legality of their contributions were validated through a formal affidavit to establish the medical standing and legitimacy of the annotated data.

Bibliography

[1] D. E. Ford and D. B. Kamerow, "Epidemiological study of sleep disturbances and psychiatric disorders: An opportunity for prevention?" *JAMA*, vol. 262, no. 11, pp. 1479–1484, 1989. DOI: 10.1001/jama.1989.03430110069030

[2] C. S. McCrae and K. L. Lichstein, "Secondary insomnia: Diagnostic challenges and intervention opportunities," *Sleep Medicine Reviews*, vol. 5, no. 1, pp. 47–61, 2001. DOI: 10.1053/smrv.2000.0146 [Online]. Available: https://doi.org/10.1053/smrv.2000.0146

[3] A. G. Harvey, "A cognitive model of insomnia," *Behaviour Research and Therapy*, vol. 40, no. 8, pp. 869–893, 2002. DOI: 10.1016/S0005-7967(01)00061-4 [Online]. Available: https://doi.org/10.1016/S0005-7967(01)00061-4

[4] C. M. Morin, M. LeBlanc, M. Daley, J. P. Grégoire, and C. Mérette, "Epidemiology of insomnia: Prevalence, self-help treatments, consultations, and determinants of help-seeking behaviors," *Sleep Medicine*, vol. 7, no. 2, pp. 123–130, 2006. DOI: 10.1016/j.sleep.2005.08.008 [Online]. Available: https://doi.org/10.1016/j.sleep.2005.08.008

[5] M. L. Perlis et al., "Insomnia as a risk factor for onset of depression in the elderly," *Behavioral Sleep Medicine*, vol. 4, no. 2, pp. 104–113, 2006. DOI: 10.1207/s15402010bsm0402_3 [Online]. Available: https://doi.org/10.1207/s15402010bsm0402_3

[6] S. D. Kyle, C. A. Espie, and K. Morgan, "...not just a minor thing, it is something major, which stops you from functioning daily: Quality of life and daytime functioning in insomnia," *Behavioral Sleep Medicine*, vol. 8, no. 3, pp. 123–140, 2010. DOI: 10.1080/15402002.2010.487450 [Online]. Available: https://doi.org/10.1080/15402002.2010.487450

[7] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, *MentalBERT: Publicly available pretrained language models for mental healthcare*, 2021. arXiv: 2110.15621 `[cs.CL]`. [Online]. Available: https://doi.org/10.48550/arXiv.2110.15621

[8] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using RoBERTa," in *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, online: Association for Computational Linguistics, Apr. 2021, pp. 59–68. [Online]. Available: https://aclanthology.org/2021.louhi-1.7/

[9] P. Jain, K. R. Srinivas, and A. Vichare, "Depression and suicide analysis using machine learning and nlp," *J. Phys.: Conf. Ser.*, vol. 2161, no. 1, p. 012034, 2022. DOI: 10.1088/1742-6596/2161/1/012034 [Online]. Available: https://doi.org/10.1088/1742-6596/2161/1/012034

[10] C. M. Morin and D. C. Jarrin, "Epidemiology of insomnia," *Sleep Medicine Clinics*, vol. 17, no. 2, pp. 173–191, 2022. DOI: 10.1016/j.jsmc.2022.03.003 [Online]. Available: https://doi.org/10.1016/j.jsmc.2022.03.003

[11] G. J. Elder, E. Altena, L. Palagini, and J. G. Ellis, "Stress and the hypothalamic–pituitary–adrenal axis: How can the covid-19 pandemic inform our understanding and treatment of acute insomnia?" *Journal of Sleep Research*, vol. 32, no. e13842, 2023. DOI: 10.1111/jsr.13842 [Online]. Available: https://doi.org/10.1111/jsr.13842

[12] M. M. T. Ayyalasomayajula, A. Agarwal, and S. Khan, "Reddit social media text analysis for depression prediction: Using logistic regression with enhanced term frequency-inverse document frequency features," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 5, pp. 5998–6005, 2024. DOI: 10.11591/ijece.v14i5.pp5998-6005 [Online]. Available: https://doi.org/10.11591/ijece.v14i5.pp5998-6005

[13] A. Alqahtani, E. Kayi, and M. Diab, *Stressroberta: Cross-condition transfer learning from depression, anxiety, and ptsd to stress detection*, 2025. arXiv: 2512.23813 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2512.23813

[14] M. Arcan and D.-P. Niland, *Evaluating large language models for anxiety, depression, and stress detection: Insights into prompting strategies and synthetic data*, 2025. arXiv: 2511.07044 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2511.07044

[15] J. A. Cummins, D. J. Gottlieb, T. Sofer, and D. A. Wallace, "Applying natural language processing techniques to map trends in insomnia treatment terms on the r/insomnia subreddit: Infodemiology study," *J. Med. Internet Res.*, vol. 27, e58902, 2025. DOI: 10.2196/58902 [Online]. Available: https://doi.org/10.2196/58902

[16] E. Dan, J. Zhu, and R. Jin, "Detecting suicide risk and exploring contributing factors: Classification and topic modeling of social media data," in *Proceedings of the International Conference on Advances in Emerging Trends and Technologies*, Lisbon, Portugal: IARIA, Mar. 2025, pp. 23–28, ISBN: 978-1-68558-247-0.

[17] K. M. S. Islam, J. Fields, and P. Madiraju, *Multimentalroberta: A fine-tuned multiclass classifier for mental health disorder*, 2025. arXiv: 2511.04698 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2511.04698

[18] T. Kallstenius, A. J. Capusan, G. Andersson, et al., "Comparing traditional natural language processing and large language models for mental health status classification: A multi-model evaluation," *Scientific Reports*, vol. 15, p. 24 102, 2025. DOI: 10.1038/s41598-025-08031-0 [Online]. Available: https://doi.org/10.1038/s41598-025-08031-0

[19] A. M. Marmol-Romero, M. Garcia-Vega, M. A. Garcia-Cumbreras, and A. Montejo-Raez, *Sinai at erisk@clef 2025: Transformer-based and conversational strategies for depression detection*, 2025. arXiv: 2509.19861 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2509.19861

[20] N. Ranjbar and H. Baghbani, "Lotus at SemEval-2025 task 11: RoBERTa with llama-3 generated explanations for multi-label emotion classification," in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 431–439, ISBN: 979-8-89176-273-2. [Online]. Available: https://aclanthology.org/2025.semeval-1.60/

[21] B. G. Teferra et al., "Leveraging large language models for automated depression screening," *PLOS Digital Health*, vol. 4, pp. 1–18, Jul. 2025. DOI: 10.1371/journal.pdig.0000943 [Online]. Available: https://doi.org/10.1371/journal.pdig.0000943

[22] E. M. Wickwire et al., "Estimation of the global prevalence and burden of insomnia: A systematic literature review-based analysis," *Sleep Medicine Reviews*, vol. 82, p. 102 121, Aug. 2025, ISSN: 1087-0792. DOI: 10.1016/j.smrv.2025.102121