NEURO SYMBOLIC AI FOR ASSESSING EMPLOYEE MENTAL HEALTH

J.A.D.L.Wickramasinghe

239187E

MSc in Data Science and Artificial Intelligence

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

January 2025

NEURO SYMBOLIC AI FOR ASSESSING EMPLOYEE MENTAL HEALTH

J.A.D.L.Wickramasinghe

239187E

Dissertation submitted in partial fulfillment of the requirements for the degree
MSc in Data Science and Artificial Intelligence

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

January 2025

DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 24/06/2025

The supervisor should certify the Dissertation with the following declaration.

The above candidate has carried out research for the MSc in Data Science and Artificial Intelligence Dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. A.L.A.T.D.Thanuja Ambegoda

Signature of the Supervisor: Date:27/06/2025

DEDICATION

This thesis is dedicated to all individuals who strive to foster healthier workplaces by embracing mental well-being and inclusivity. It is also dedicated to researchers and innovators in the field of Artificial Intelligence, whose tireless efforts inspire transformative solutions for real-world challenges. Lastly, I dedicate this work to my family and friends, whose unwavering support and encouragement have been my greatest source of strength throughout this academic journey.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to all those who supported me throughout the Master's degree in Data Science and Artificial Intelligence. First and foremost, I extend my deepest appreciation to my supervisor, Dr. A.L.A.T.D. Thanuja Ambegoda, for his invaluable guidance, insightful feedback, and continuous encouragement. His expertise and mentorship have been instrumental in shaping the direction and quality of this research.

I am sincerely grateful to the academic staff of the Department of Computer Science & Engineering, Faculty of Engineering, University of Moratuwa, for providing a robust academic foundation and a conducive research environment. Their teachings and guidance throughout my MSc program have significantly contributed to the successful completion of this thesis.

Special thanks go to my colleagues and friends for their intellectual support, collaboration, and valuable discussions, which helped me overcome various challenges during the research process.

I would also like to thank my family for their unwavering love, patience, and moral support. Without their constant encouragement and understanding, this journey would not have been possible.

Finally, I acknowledge the countless researchers and developers whose work inspired the development of innovative solutions in this project. Their contributions to the field of Artificial Intelligence and Natural Language Processing have provided a solid foundation for my research.

ABSTRACT

In the rapidly evolving corporate landscape, employee mental well-being has become integral to productivity and organizational success. This thesis introduces a groundbreaking Neuro-Symbolic Artificial Intelligence (NSAI) framework that integrates conversational data analysis to monitor and enhance workplace mental health. At its core, the Mentalisys Health Application leverages H2O Wave to provide user-friendly dashboards equipped with real-time sentiment analysis, stress, and depression detection capabilities. A novel Commonsense-Driven Symbolic ReAct-NLI (CSR-NLI) technique, based on OpenAI's language models, combines symbolic reasoning and natural language inference to uncover causality in workplace communication. Through interactive admin and user-specific dashboards, the system fosters proactive mental health interventions and personalized support, promoting a healthier workplace environment.

The study's primary contribution lies in advancing NSAI for robust causal understanding, going beyond conventional sentiment analysis. Results demonstrate significant potential in improving employee well-being and productivity via timely interventions and precise health risk assessments. This work underscores the transformative role of AI in addressing real-world mental health challenges, driving organizational growth, and enhancing employee satisfaction, while setting a new benchmark for AI-driven solutions in corporate mental health management.

Keywords: Neuro-Symbolic AI, Sentiment and Emotion Analysis, Commonsense Reasoning, Natural Language Inference, Workplace Mental Health, Employee Productivity Enhancement, Stress and Depression Detection, Corporate Communication Platforms, Mentalisys Health Application, Real-Time Analytics, Proactive Workplace Monitoring, Explainable AI Solutions

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The modern workplace is undergoing significant transformation, driven by the increasing adoption of collaborative and employee-centric environments. While this shift fosters innovation, engagement, and agility, it also brings about unique challenges related to employee mental health. In this evolving corporate landscape, organizations are recognizing that mental well-being is crucial not only for individual employees but also for overall organizational success. Poor mental health can result in diminished productivity, increased absenteeism, and higher healthcare costs. Conversely, prioritizing mental health can enhance employee commitment, reduce turnover, and improve financial performance.

Despite growing awareness, many organizations struggle to effectively monitor and address employee mental well-being. Traditional approaches, such as periodic surveys and self-reported assessments, often fail to provide real-time insights or capture the complex emotional states employees experience during their daily interactions. Moreover, conventional sentiment analysis techniques, while useful, tend to oversimplify emotional states, limiting their ability to inform proactive interventions.

Recent global trends further highlight the urgency of addressing mental health issues in the workplace. A comprehensive report on Workplace Mental Health Trends by Spring Health indicates that seven out of ten people globally experience mental health challenges, resulting in a staggering $1 trillion in lost productivity annually [1]. Furthermore, KPMG reports that the aftermath of the COVID-19 pandemic has exacerbated mental health issues, with remote and hybrid work introducing additional stressors, such as feelings of isolation and video call fatigue [2].

A significant gap persists between the acknowledged importance of discussing mental health at work and employees' comfort levels in initiating such conversations. According to the National Alliance on Mental Illness (NAMI) workplace mental health poll conducted in January 2024 underscores this disconnect, revealing that while most employees recognize the need for open discussions on mental health, many do not feel equipped or comfortable to engage in them [3]. This disconnect emphasize the prevalence of stigma in workplaces and highlights the pressing need for proactive, technology-driven solutions that foster a supportive mental health culture.

Emerging technologies, particularly in the fields of Artificial Intelligence (AI) and Neuro-Symbolic AI (NSAI), offer transformative potential for addressing these challenges. By integrating advanced emotion analysis models with commonsense reasoning and symbolic logic, organizations can gain deeper insights into employee emotions,

detect early signs of mental health issues, and implement timely interventions.

This research focuses on developing an innovative framework combining sentiment analysis, stress detection, and depression assessment models with a novel Commonsense-Driven Symbolic ReAct-NLI (CSR-NLI) prompting technique. CSR-NLI integrates Natural Language Inference (NLI) and symbolic reasoning to infer causality from employee communications, providing an advanced, explainable AI-driven approach to mental health monitoring. The framework is implemented within the Mentalisys Health Application, a comprehensive platform designed to analyze conversational data from corporate communication tools such as Slack. The system includes interactive dashboards for both employees and administrators, enabling real-time monitoring, personalized insights, and proactive support.

By leveraging the latest advancements in NSAI and machine learning, this study aims to bridge the gaps in current methodologies, offering a sophisticated, real-time, and scalable solution for mental health assessment. Ultimately, this research endeavors to foster healthier workplace environments, and improve employee satisfaction.

1.2 Research Context

In today's business environment, the emphasis on employee well-being extends beyond physical health, with mental well-being increasingly recognized as a critical determinant of organizational success. Research highlights the tangible and intangible benefits of prioritizing mental health in the workplace, ranging from increased productivity and job satisfaction to significant reductions in healthcare costs and absenteeism. However, traditional approaches to addressing workplace mental health often rely on outdated methodologies, such as infrequent surveys or self-reported data, which fail to provide real-time insights or capture the full spectrum of emotional states. Consequently, there is a growing demand for advanced, technology-driven solutions that offer a more nuanced and proactive approach to mental health monitoring.

This research aims to bridge this gap by leveraging Neuro-Symbolic AI (NSAI) to analyze workplace communication data and provide real-time, actionable insights into employee mental well-being. Existing sentiment analysis tools, while useful, often provide a limited perspective on emotional states, failing to capture deeper psychological factors such as stress and depression. By integrating NSAI along with advanced sentiment, emotion, and stress analysis models, this study proposes a holistic approach to monitoring workplace mental health. The innovation lies in the development of the Commonsense-Driven Symbolic ReAct-NLI (CSR-NLI) prompting technique, which combines natural language inference and commonsense reasoning to infer causality in employee messages. This technique enhances the interpretability and effectiveness of the Mentalisys Health Application a comprehensive platform that utilizes communication data from tools like Slack to detect early signs of mental health issues.

Impact analysis from previous research supports the economic rationale for investing in workplace mental health assessments. For instance, a seminal study in the American Journal of Health Promotion underscored that comprehensive wellness programs have been shown to reduce absenteeism, healthcare costs, and workers' compensation claims by up to 25% [4]. Moreover, according to the World Health Organization (WHO), for every $1 invested in the treatment of common mental health disorders, there is a return of $4 in improved health and productivity [5]. Real-world examples, such as Johnson & Johnson's wellness program, demonstrate that companies can achieve substantial financial benefits through mental health initiatives. Over a ten-year period, Johnson & Johnson saved $250 million on healthcare costs, yielding a return of $2.71 for every dollar invested [6].

This research aligns with these findings by proposing a solution that integrates mental health assessment into workplace wellness programs, ultimately contributing to both employee well-being and corporate sustainability. By combining advanced AI techniques with practical applications in corporate settings, this study endeavors to redefine how organizations approach mental health, offering a scalable, real-time, and impactful solution to an ever-growing concern in the modern workplace.

1.3 Research Problem

In the contemporary corporate landscape, employee mental well-being is increasingly recognized as a cornerstone of organizational success. Despite numerous initiatives to address workplace mental health, significant gaps remain in how organizations assess and manage mental health risks effectively. Traditional methods such as subjective self-reports, periodic surveys, and rudimentary sentiment analysis tools offer limited insight into real time emotional well-being. These approaches fail to capture subtle emotional patterns or causal factors that could signal emerging issues such as stress, frustration, or burnout. As a result, organizations often struggle to implement timely interventions, leading to diminished productivity and employee satisfaction.

Existing sentiment analysis models primarily focus on simplistic emotional classifications (positive, negative, neutral) and overlook the complexity of workplace mental health indicators. These models lack the depth required to identify early signs of stress or depression and fail to infer causality from employee communication data. This limitation hampers organizations' ability to proactively address mental health challenges, resulting in missed opportunities to improve employee well-being and productivity.

1.3.1 Research Gap

Despite the growing recognition of the importance of employee mental health, existing research and tools fail to provide a comprehensive solution that integrates nuanced

sentiment analysis with contextual and causal interpretation of workplace communication. Current sentiment analysis models often oversimplify emotional states, focusing on basic polarity (positive, neutral, negative), and lack the sophistication required to identify subtle patterns of stress, anxiety, and depression. This results in missed opportunities for early intervention and fails to address the complex interplay of emotional triggers within workplace communication.

Furthermore, while neural models excel in predictive capabilities, their lack of transparency and inability to infer causality from data limit their applicability in corporate mental health monitoring. The absence of robust, explainable AI solutions that combine neural learning with symbolic reasoning highlights a significant gap in the field. Few studies leverage Neuro-Symbolic AI (NSAI) to enrich emotional and mental health analysis by integrating commonsense reasoning and natural language inference (NLI). This gap underscores the need for a novel framework that not only identifies emotional states but also provides actionable insights by uncovering the contextual and causal factors behind employee stress or burnout.

1.3.1.1 Comparison with Existing Neuro-Symbolic AI Methods

Existing Neuro-Symbolic AI approaches predominantly focus on enhancing model interpretability by combining neural networks with symbolic reasoning frameworks. The primary aim of these methods is to improve decision transparency, particularly in emotion analysis applications. Prominent frameworks such as TAM-SENTICNET [7] and Sentic PROMs [8] have employed symbolic reasoning mechanisms to provide interpretable insights, enhancing the detection of depressive language patterns and improving health-related quality of life assessments. However, these models exhibit several limitations that hinder their applicability to real-time corporate mental health monitoring.

One of the major limitations of existing NSAI methods is their limited capacity for causal reasoning. While these models are proficient at identifying emotional states, they often fail to infer the underlying causes of emotional shifts. This shortcoming reduces their effectiveness in providing actionable insights for early interventions. Additionally, many of the current approaches are domain-specific, exhibiting satisfactory performance only within narrowly defined contexts. This lack of generalization makes them unsuitable for diverse workplace environments where communication styles and stressors vary significantly.

Another challenge pertains to the scalability of existing NSAI frameworks. Integrating symbolic reasoning into neural models generally increases computational complexity, thereby compromising the efficiency required for real-time analysis of large-scale communication data. As organizations continue to produce extensive amounts of communication data daily, the need for scalable and efficient frameworks becomes

increasingly critical.

1.3.1.2 Novelty of CSR-NLI

The proposed CSR-NLI framework introduces a novel approach to addressing these limitations through a Commonsense-Driven Symbolic ReAct-NLI prompting technique. Unlike prior models that solely focus on emotional state detection, CSR-NLI integrates structured commonsense reasoning to uncover causal relationships within employee communication. This approach enhances the ability of the model to provide meaningful, contextually aware insights, thereby supporting early intervention strategies more effectively.

Furthermore, the CSR-NLI framework demonstrates improved scalability through efficient prompting techniques. By employing iterative refinement processes, the framework achieves high reasoning coherence without significantly increasing computational overhead. This ensures that real-time analysis of employee communication remains feasible even in large-scale organizational settings. Moreover, the generalization capabilities of CSR-NLI extend beyond domain-specific applications, making it adaptable to various communication platforms such as Slack, Microsoft Teams, and other corporate communication tools.

Another critical advantage of the CSR-NLI framework is its enhanced interpretability. The integration of commonsense-driven reasoning ensures that the insights derived from communication data are both accurate and interpretable. This addresses a significant limitation of traditional black-box models that offer little to no explanation of their predictions. By providing clear and understandable explanations, CSR-NLI fosters greater trust and usability among corporate stakeholders.

Overall, the novelty of the CSR-NLI framework lies in its ability to effectively combine causal reasoning, scalability, generalization, and interpretability. These advancements mark a significant departure from existing Neuro-Symbolic AI approaches, providing a more robust and practical solution for real-time corporate mental health monitoring. The proposed framework not only bridges the research gaps identified but also establishes a solid foundation for further development and application of Neuro-Symbolic AI in workplace mental health assessment.

1.3.1.3 Limitations of Traditional Sentiment Analysis

Conventional sentiment analysis tools provide limited insights by relying on basic polarity detection, which is insufficient for understanding the nuanced emotional states critical in mental health assessments. These tools lack the ability to capture subtle emotional cues or patterns that signal early indicators of stress or depression. This gap restricts organizations from implementing timely interventions to prevent mental health challenges from escalating.

1.3.1.4 Inadequate Integration of Symbolic Reasoning

While neural network-based models demonstrate strong predictive performance, they are often opaque and fail to provide interpretable insights into emotional causality. Symbolic reasoning offers the potential to address these limitations by providing a transparent framework for understanding the causes and implications of detected emotions. However, most existing methodologies neglect this approach, leaving a significant gap in explainable AI solutions for real-time mental health monitoring.

1.3.2 Significance of Addressing the Research Problem

Addressing the identified research problem is pivotal for enhancing organizational well-being and driving workplace innovation. This study aims to provide a transformative solution to proactively monitor and manage employee mental health, ultimately fostering a healthier, more productive work environment.

1.3.2.1 Enhancing Employee Well-being

By developing a Neuro-Symbolic AI-driven framework, organizations can shift from reactive mental health management to proactive, real-time intervention strategies. This framework enables the early detection of mental health challenges such as stress, anxiety, and burnout by analyzing nuanced emotional and contextual cues within workplace communication. Equipped with actionable insights, organizations can implement timely interventions to improve employee morale, reduce absenteeism, and enhance overall job satisfaction and productivity.

1.4 Research Objectives

The primary aim of this research is to develop an advanced, real-time framework that integrates Neuro-Symbolic AI (NSAI) with emotion analysis to comprehensively monitor employee mental health, provide actionable insights, and promote healthier workplace environments. The specific objectives are:

1. **Develop a Neuro-Symbolic Emotion Analysis Framework**

    - Design and implement an emotion analysis system capable of detecting nuanced emotional states such as stress, frustration, and burnout within workplace communication data.
    - Leverage advanced Natural Language Processing (NLP) models and Neuro-Symbolic AI techniques to identify emotional patterns and causal factors in real-time.

- Evaluate the framework's performance using quantitative metrics such as accuracy, precision, recall, and F1-score against benchmark datasets.

2. **Integrate Causal Reasoning Using CSR-NLI**
    - Incorporate the CSR-NLI prompting technique to infer causality and context in employee communications.
    - Demonstrate the effectiveness of CSR-NLI in uncovering the underlying reasons for emotional shifts, such as stress triggers, in workplace interactions.
3. **Monitor and Analyze Workplace Communication**
    - Implement the framework to process and analyze communication data from platforms like Slack, enabling continuous and automated tracking of employee emotional states.
    - Provide actionable insights to enable HR teams and managers to address potential mental health concerns proactively.
4. **Develop a User-Friendly Dashboard for Insights**
    - Create an interactive, intuitive dashboard using H2O Wave to visualize real-time analytics and provide insights at both organizational and individual levels.
    - Enable role-specific features, including aggregated organizational trends for administrators and personalized feedback for employees.
5. **Validate the Solution in Real-World Scenarios**
    - Conduct thorough testing of the developed framework in corporate environments to assess its usability, scalability, and effectiveness.
    - Collect feedback from corporate users, including HR teams, managers, and employees, to evaluate the practical impact of the solution on workplace well-being.

1.4.1 Rationale for Objectives

These objectives address the identified gaps in existing workplace mental health monitoring tools by focusing on explainable and actionable AI-driven solutions. The incorporation of Neuro-Symbolic AI and the CSR-NLI prompting technique ensures robust causal analysis and interpretable insights. By delivering a real-time, scalable, and ethical solution, this research contributes to enhancing employee well-being, fostering a supportive work environment, and driving organizational productivity.

1.5 Justification of the Research

This research addresses critical challenges in monitoring and managing employee mental health by leveraging Neuro-Symbolic AI (NSAI) to develop a scalable, real-time solution. Employee well-being significantly influences productivity, job satisfaction, and organizational success, yet traditional methods like periodic surveys and basic sentiment analysis fail to provide actionable, real-time insights into nuanced emotional states. This study bridges these gaps with a robust framework that integrates emotion detection, stress analysis, and commonsense reasoning.

1.5.1 Enhancing Employee Well-being

The proposed solution empowers organizations to detect early signs of stress, anxiety, and depression through real-time analysis of workplace communication. By uncovering nuanced emotional patterns and their causes, the system enables timely interventions that reduce burnout, absenteeism, and improve overall morale and productivity, fostering a supportive workplace culture.

1.5.2 Advancing Explainable AI Solutions

Integrating symbolic reasoning with neural networks ensures accurate and interpretable insights, addressing the limitations of traditional black-box AI models. The CSR-NLI technique enhances understanding of emotional triggers, positioning this research at the forefront of AI-driven mental health monitoring in corporate environments.

1.6 Scope and Limitations

This research focuses on the development and implementation of a Neuro-Symbolic AI-driven framework to monitor and assess employee mental health in real-time. The scope includes the design and integration of advanced sentiment, emotion, and stress analysis models, augmented by the CSR-NLI technique, within the practical and user-friendly Mentalisys Health Application. The system is designed for corporate communication platforms, with Slack serving as the primary testbed for analyzing textual communication data.

1.6.1 Scope of the Research

1. **Target Platforms:** The study is focused on Slack as a primary platform for analysis due to its widespread use in professional environments. This ensures the system is tailored to real-world workplace communication data.

2. **Textual Data Analysis:** The research analyzes textual communication data to detect emotional states, including stress, frustration, and burnout. It leverages the CSR-NLI technique to infer causality and identify triggers behind emotional patterns.

3. **Dashboard Development:** A comprehensive, interactive dashboard was developed using H2O Wave, designed for real-time insights. The dashboard provides aggregated organizational analytics for administrators and personalized mental health insights for employees.

4. **Framework Evaluation:** The framework was evaluated using both quantitative metrics (e.g., accuracy, precision, recall, and F1-score) for model performance and qualitative feedback from corporate users on usability and practical value.

1.6.2 Limitations of the Research

1. **Platform Dependency:** The research is specific to Slack as the testing platform. The applicability of the solution to other platforms, such as Microsoft Teams or other industries, may require additional adaptation and validation.

2. **Textual Data Focus:** This study exclusively focuses on analyzing textual communication. The inclusion of multimodal data (e.g., audio, video, or physiological metrics) could enhance the depth of mental health assessments but is outside the scope of this research.

3. **Generalizability of Findings:** The framework's effectiveness may vary depending on organizational culture, communication styles, and employee demographics. Broader validation in diverse corporate settings is necessary to ensure wider applicability.

4. **Scalability Challenges:** While the Neuro-Symbolic AI framework enhances interpretability, scaling the system for organizations with high-frequency communication or large datasets may require further optimization and computational resources.

5. **Data Privacy and Ethical Constraints:** The study relies on strict ethical guidelines to ensure data privacy, consent, and anonymization. These constraints influence the scope of data collection and may limit access to richer datasets or sensitive workplace communications.

Despite these limitations, the Mentalisys Health Application demonstrates significant potential to address critical gaps in workplace mental health monitoring. By providing real-time, explainable, and actionable insights, this research contributes to

fostering healthier workplace environments and enabling proactive mental health interventions. Future work can build on this foundation to expand the system's scope and scalability.

1.7 Structure of the Thesis

This thesis is organized into six chapters, providing a comprehensive exploration of the research problem, methodology, results, and conclusions. The opening section introduces the research, detailing its background, context, objectives, and scope. It defines the research problem and outlines the motivation and significance of addressing employee mental health, establishing a solid foundation for the study.

The next section reviews the existing body of knowledge, synthesizing research on employee mental health monitoring, sentiment and emotion analysis, and Neuro-Symbolic AI. Key concepts such as Natural Language Inference, commonsense reasoning, and emotion cause analysis are discussed, alongside identified gaps in current methodologies that justify the need for the proposed solution.

The methodology section describes the design and implementation of the Mentalisys Health Application. This includes the development of sentiment, emotion, and stress analysis models and the integration of the CSR-NLI technique. The technical architecture is detailed, covering the use of Slack APIs for data extraction, MongoDB for data storage, and H2O Wave for creating interactive dashboards. The workflow and modular system design are also outlined to demonstrate scalability and practicality.

Subsequent sections focus on experiments and results, presenting the validation of the developed models and framework. Metrics such as accuracy, precision, recall, and F1-score are used to evaluate performance, while usability testing provides practical insights into the effectiveness and adoption potential of the Mentalisys Health Application in real-world corporate settings.

The discussion section interprets the results in the context of the research objectives and compares them with existing mental health monitoring tools. The implications of the findings for workplace well-being and the innovative use of Neuro-Symbolic AI for causality detection are explored. Challenges encountered during the research, including ethical considerations and scalability issues, are discussed along with lessons learned.

The concluding section summarizes the key contributions of the research, emphasizing its potential to transform workplace mental health monitoring. It provides practical recommendations for implementation and suggests future research directions, such as expanding the framework to other communication platforms and integrating multimodal data.

The thesis concludes with references listing all cited scholarly works and an appendix for supplementary materials.

CHAPTER 2

LITERATURE REVIEW

The corporate landscape has evolved toward more agile and employee-centric models, creating both opportunities and challenges in maintaining employee mental health. Recognized as a critical determinant of organizational success, employee well-being directly influences productivity, job satisfaction, and financial outcomes. This growing importance has prompted the need for innovative, technology-driven approaches to assess and manage workplace mental health effectively.

This literature review provides a comprehensive exploration of prior studies relevant to employee well-being and the integration of Artificial Intelligence (AI) in corporate wellness strategies. It begins with a historical background of the field, tracing the evolution of mental health assessment and its integration with workplace practices. The review then examines existing studies across several key domains: the relationship between employee well-being, emotion recognition in workplace conversations, emotion cause analysis, and advancements in Natural Language Inference (NLI) and commonsense reasoning. Particular attention is given to Neuro-Symbolic AI (NSAI), an emerging approach that integrates symbolic reasoning with neural networks, offering promising solutions for nuanced and interpretable mental health detection.

By synthesizing research across these domains, the literature review contextualizes the current state of workplace mental health assessment and monitoring. It identifies critical gaps in traditional methodologies, such as the oversimplification of emotional states, limited understanding of emotional causality, and the lack of interpretability in AI models. These gaps highlight the need for a holistic and scalable framework that combines emotion recognition with causal analysis, offering actionable insights for proactive mental health management in workplace settings.

2.1 Historical Background and Development of the Field

The field of emotion recognition has evolved significantly, drawing from foundational research in psychology, neuroscience, and artificial intelligence (AI). Early studies, such as those by Paul Ekman, focused on understanding basic human emotions through observable cues like facial expressions, body language, and physiological responses [9]. These efforts laid the groundwork for the systematic study of emotions and their impact on human behavior. With the advent of computational technologies, research shifted toward developing machine learning models capable of analyzing digital data sources such as images, videos, and biometric signals, enabling more scalable and automated emotion detection.

Concurrently, the importance of employee well-being began to gain recognition,

particularly as studies highlighted the detrimental effects of work-related stress on mental health and organizational productivity [10]. Early approaches to addressing employee mental health relied on periodic surveys and self-reports, which offered limited insights into real-time emotional states. As AI technologies matured, their integration into workplace applications emerged as a novel way to monitor and enhance employee well-being, supported by research demonstrating the correlation between emotional states, productivity, and job satisfaction [11]. These advancements signaled a shift from reactive to proactive approaches in addressing mental health challenges in the workplace.

The introduction of Neuro-Symbolic AI represents a pivotal development in emotion recognition, combining the interpretability of symbolic AI with the learning capabilities of neural networks [12]. By enabling context-aware and nuanced emotional analysis, this approach overcomes several limitations of traditional AI methods, such as the lack of causal understanding and interpretability. Neuro-Symbolic AI holds significant promise for workplace applications, offering scalable and accurate tools for assessing employee well-being and facilitating timely interventions in diverse organizational settings.

2.2 Overview of Existing Studies

The exploration of artificial intelligence (AI) applications in enhancing corporate wellness and mental health assessments reveals a dynamic and rapidly evolving field. This section synthesizes significant contributions from existing research, highlighting advancements, challenges, and the potential for Neuro-Symbolic AI (NSAI) in revolutionizing mental health detection and support within organizational settings. By examining various studies across multiple domains, this review establishes a coherent understanding of how AI technologies can be applied to improve employee well-being, reduce insurance costs, and provide actionable insights through robust mental health monitoring frameworks.

2.2.1 Employee Well-Being

This section explores the relationship between employee well-being, emphasizing the role of wellness programs and compensation structures in enhancing productivity. By reviewing various studies focused on wellness interventions, and healthcare benefits, this section aims to provide a comprehensive understanding of how promoting employee well-being can positively impact organizational outcomes.

2.2.1.1 Employee Well-Being and Impact of Corporate Wellness Programs for productivity

The study by Gubler et al. [13] establish causal evidence linking wellness programs and health improvements to increased worker productivity. Their investigation into the impact of a corporate wellness program on worker productivity utilizes empirical data, although the study acknowledges limitations such as a relatively small sample size and challenges in precisely estimating parameters. The emphasis on specific program design elements and the recognition of measurement challenges highlight the complexities in evaluating the financial and health impact of wellness programs, contributing to broader discussions on the effectiveness of such initiatives.

Basi Nska-Zych and Springer [14] conducted a systematic review of workplace health promotion interventions (WHPIs) to identify outcomes for employers and employees. The study explores the diversity in outcomes measured and research methods used in WHPI assessments. The focus on qualitative and quantitative analysis of outcomes and methods employed in WHPIs contributes to understanding the effectiveness of implemented programs. However, the lack of a uniform approach to evaluation and the diversity in measurement methods and outcomes make it challenging to draw clear conclusions regarding the effectiveness of WHPIs, particularly in small- and medium-sized enterprises.

Danna and Griffin [15] offer a thorough examination of the literature on health and well-being in the workplace, highlighting its fragmented nature and the need for a cohesive understanding. Their framework addresses various factors influencing health and well-being, emphasizing the interconnectedness between work and personal life. They underscore the implications of workplace characteristics and threats on workers' health, as well as the consequences of poor health and well-being for individuals and organizations. Advocating for interdisciplinary perspectives, the authors call for the development of a unified model or theory in this domain. Their work provides valuable insights and sets a direction for future research and practice, aiming to elevate the importance of health and well-being in organizational science.

Krekel et al. [16] present a thorough investigation into the relationship between employee well-being and productivity, focusing on establishing both correlation and causation. Through a meta-analysis encompassing 339 studies with data from over 1.8 million employees and 82,000 business units across 73 countries, the authors identify a significant correlation between employee well-being and firm-level performance, particularly in areas such as customer satisfaction and staff turnover, which ultimately affect overall profitability. The paper advocates for consistent measurement and reporting of employee well-being alongside productivity metrics and recommends interventions targeting key drivers of well-being, such as social relationships and work-life balance, with a call for rigorous evaluation through randomized controlled trials.

By synthesizing empirical evidence, case studies, and meta-analysis, the paper offers compelling insights into the business case for prioritizing employee well-being in the workplace, suggesting avenues for enhancing both individual and organizational outcomes.

2.2.1.2 Legal, Ethical, and Practical Challenges

Mujtaba and Cavico [17] offer an overview of corporate wellness efforts in the American workplace, exploring challenges faced by employers in implementing wellness programs. The focus on legal, ethical, and practical aspects provides valuable insights into program design elements. However, legal challenges associated with potential discrimination, privacy concerns, and the lack of a statutory definition for "wellness program" indicate a complex landscape for employers implementing such initiatives.

2.2.1.3 Summary of Employee Well-Being

The literature indicates a strong link between corporate wellness programs and enhanced employee well-being, which significantly contributes to increased productivity and organizational success. Studies such as those by Gubler et al. [13] and Basi Nska-Zych and Springer [14] demonstrate the positive effects of wellness initiatives on health outcomes and productivity, despite challenges such as small sample sizes and diverse measurement methods. These findings underscore the value of well-implemented wellness programs in promoting employee health.

Legal, ethical, and practical challenges in implementing corporate wellness programs underscore the complexity of this endeavor. Despite these challenges, the collective body of research advocates for a holistic approach to employee well-being, incorporating strategic wellness programs to mitigate health-related expenses and ultimately influence insurance premium costs positively.

In this context, the proposed Mentalisys Health Application aims to build upon these findings by offering a comprehensive, real-time mental health monitoring solution powered by Neuro-Symbolic AI techniques. By accurately identifying stress, anxiety, and other mental health challenges through continuous analysis of workplace communication data, the Mentalisys application enables organizations to proactively address well-being issues. This proactive approach enhances employee health and productivity and presents a scalable, efficient, and effective framework for optimizing employee well-being more effectively.

2.2.2 Sentiment and Emotion Recognition in Conversation

This section examines the application of AI techniques for sentiment and emotion recognition within workplace communication. The focus is on evaluating various mod-

els and frameworks designed to detect and interpret emotional states from textual data. Additionally, the challenges and advancements in analyzing sentiments across social media platforms and corporate communication tools are discussed, highlighting their implications for workplace mental health monitoring.

2.2.2.1 Sentiment Analysis in Social Media Conversations

Lim et al. [18] address the challenges posed by inept regulations on social media through text sentiment analysis on Twitter. Using TF-IDF, Word2Vec, and ELMo, the authors explore the identification of problems caused by irresponsible users. Challenges include the reliance on the sentiment140 dataset with potential limitations in representing social media diversity and the performance of ELMo being hindered by a relatively small dataset. The study, while valuable for sentiment analysis, may not encompass all aspects of social media problems, such as privacy and misinformation regulation challenges.

Blair et al. [19] explore the use of Twitter to assess the mental well-being of essential workers during the COVID-19 pandemic. Their study applies sentiment analysis using the VADER tool to evaluate the polarity (positive, neutral, and negative) of tweets authored by essential workers compared to the general Twitter user population. The results revealed that, despite the challenges faced by essential workers, their tweets exhibited higher positive sentiment scores than those of average users, both before and during the pandemic. However, essential workers were more likely to tweet about mental health-related topics, indicating increased awareness or discussion of mental health concerns within this group. The study acknowledges methodological limitations, including reliance on publicly visible Twitter accounts and potential biases in self-identification. While the findings provide valuable insights into the potential of social media data and sentiment analysis for mental well-being assessments, the authors note the complexities of inferring causation and understanding underlying factors influencing the observed sentiment trends. This research underscores the role of sentiment analysis in highlighting mental health trends but also points to the need for more nuanced methods to capture emotional context and causality.

2.2.2.2 Advanced Methods for Emotion Recognition

The study Ghosal et al. [20], introduce the Dialogue Graph Convolutional Network (DialogueGCN) to address emotion recognition in conversations. By leveraging self and inter-speaker dependency and utilizing a graph network to overcome context propagation issues, DialogueGCN outperforms existing methods on benchmark emotion classification datasets. The use of Graph Convolutional Networks, Bidirectional Gated Recurrent Units, and Convolutional Neural Networks enables effective modeling of

conversational context. However, potential challenges may include the reliance on labeled data, scalability, and the generalization of the model to diverse conversational patterns.

The COSMIC by Ghosal et al. [21], presents a framework for emotion identification in conversations using commonsense knowledge. Incorporating mental states, events, and causal relations, COSMIC achieves new state-of-the-art results and addresses context propagation challenges, emotion shift detection, and the differentiation of related emotion classes. The integration of RoBERTa, COMET, and Bidirectional GRU cells enhances context-independent feature extraction. While the paper does not explicitly mention limitations, potential challenges may include the generalization of commonsense knowledge, scalability, and the need for high-quality annotated data.

2.2.2.3 Emotion Analysis in Corporate Communication Platforms

Feislachen et al. [22] explore communication during an online hackathon through sentiment analysis of interactions facilitated by Slack. Natural Language Processing techniques and sentiment analysis of emojis reveal positive sentiments in messages related to motivation and achievements. Challenges include the dataset's limited geographical context (South Asia), difficulty in considering cultural differences, and the potential influence of factors like politeness on emoji usage. Additionally, the study's focus on short-term virtual teams using Slack may limit generalizability to other communication platforms or contexts.

Chatterjee et al. [23] focus on software-related Q&A chat conversations, presenting a dataset from public Slack communities and a disentanglement algorithm. Challenges include the unstructured nature of chat forums, platform-specific modifications to the algorithm, and potential bias arising from public channels. The study contributes to understanding communication dynamics in open-source software communities but may require adaptation for other chat platforms.

Wang et al. [24] identify nine categories of Slack group chat channels, leveraging qualitative coding and machine learning classification. Challenges involve context-specific findings within an R&D department and potential limitations in generalizing the machine learning model to other platforms. The study contributes valuable insights into understanding group communication styles, but the context-specific success metric and potential generalization issues should be considered.

Chatterjee et al. [25] focus on the potential usefulness of mining developer Q&A conversations in Slack, comparing content with Stack Overflow. Challenges include the context-specific focus on public chat communities, difficulties in automated analysis, and the transient nature of Slack conversations. The study provides valuable insights into the challenges and potential benefits of mining developer interactions in Slack for software engineering tools.

2.2.2.4 Summary of Emotion Recognition in Conversation

Recent advancements in emotion recognition in conversation, as evidenced by studies such as Ghosal et al. [20] with DialogueGCN and COSMIC by Ghosal et al. [21], mark significant progress in accurately classifying emotions using advanced neural networks and integrating commonsense knowledge. Despite these successes, challenges such as model scalability, reliance on labeled datasets, and generalization across diverse conversational patterns persist. Furthermore, sentiment analysis in social media and corporate communication platforms like Twitter and Slack underscores the complexity of emotion recognition across different contexts. These studies highlight the need for adaptable models that can account for varied datasets and the importance of addressing potential biases. The relevance of emotion recognition to enhancing business performance, understanding consumer sentiment, and monitoring employee well-being in corporate communication is evident.

In particular, the insights gained from existing research on sentiment and emotion detection in platforms like Slack can be directly applied to the development of the Mentalisys application. By evaluating and selecting the most effective models from previous studies, the Mentalisys application aims to integrate real-time emotion and sentiment detection into workplace communication systems to improve employee well-being monitoring and intervention capabilities.

2.2.3 Emotion Cause Analysis

This section provides an overview of emotion cause analysis, a critical area of research aimed at identifying the underlying triggers of emotional states from textual data. The discussion includes key contributions related to emotion annotation, dataset creation, and modeling approaches. Furthermore, the challenges associated with accurately detecting emotion causes and potential solutions to enhance model generalization and robustness are addressed.

2.2.3.1 Emotion Annotation and Datasets

In the realm of Emotion Cause Analysis, foundational work has been laid in the annotation of emotions and the creation of datasets. Liu [26] provided a significant contribution with their work on "Sentiment Analysis and Opinion Mining," which introduced key concepts such as "aspect category" and "aspect expression." However, a limitation of this work lies in its omission of explicit discussions on methodological constraints or limitations associated with its proposed techniques.

More recently, Bostan et al. [27] introduced the "GoodNewsEveryone" corpus, a resource annotated for emotions, semantic roles, and reader perception in news headlines. Their two-phase crowdsourcing annotation approach and baseline model for

semantic role prediction enable further research in emotion classification and cause detection. However, challenges arise due to the subjective nature of emotion perception and annotation, emphasizing the complexity of predicting semantic roles in text.

Similarly, the "WRIME: A New Dataset for Emotional Intensity Estimation" by Kajiwara et al. [28] addresses emotional intensity differences between writers and readers in a Japanese SNS dataset. While employing various models like Bag-of-Words with Logistic Regression and pre-trained BERT, the study highlights challenges in predicting subjective emotional intensity accurately. This underscores the need for a deeper understanding of writers' subjective emotions compared to more apparent objective emotions.

The CAMS dataset (Causal Analysis of Mental Health Issues in Social Media Posts) introduced by Garg et al. [29] represents a significant advancement in emotion cause analysis, particularly in mental health research. This annotated corpus combines 3,155 Reddit posts and 1,896 re-annotated instances from the SDCNL dataset, providing a rich resource for understanding the underlying causes of mental health challenges. Unlike traditional datasets, CAMS is specifically designed for interpretable causal analysis, categorizing causes into six key classes, such as relationships, jobs/careers, and alienation. By offering a robust annotation schema and demonstrating its utility through machine learning models, CAMS bridges critical gaps in automated causal understanding. While its focus on social media data aligns with trends in digital mental health research, the dataset also highlights challenges in subjective annotation and imbalanced class distribution, emphasizing the need for nuanced methodologies in emotion and cause detection.

2.2.3.2 Modeling Approaches

Several modeling approaches have been proposed to address Emotion Cause Analysis. Yuan et al. [30] formulate emotion-cause pair extraction as a sequence labeling problem, achieving state-of-the-art performance. However, they do not explicitly discuss potential limitations or constraints associated with the proposed model. The "Dual-Questioning Attention Network" (DQAN) introduced by Sun et al. [31] excels in Emotion-Cause Pair Extraction (ECPE). The hierarchical structure of DQAN, incorporating word and clause-level analysis with a dual-questioning mechanism, outperforms baseline models. However, the paper does not explicitly discuss the limitations or constraints of the DQAN model or its methodology.

Teodorescu and Mohammad [32] evaluates emotion arcs, emphasizing the efficacy of lexicon-only methods in generating high-quality emotion arcs from text streams. While the study successfully demonstrates the effectiveness of simple, interpretable methods for emotion arc generation, the focus on aggregate-level emotion arcs may limit its applicability to coherent narratives.

2.2.3.3 Challenges and Limitations

Despite the advancements, various challenges and limitations persist in Emotion Cause Analysis. The subjective nature of emotion annotation, limitations in model discussions, and complexities in applicability and generalization are common themes across the literature.

For instance, Bostan and Klinger [33] focuses on aggregating diverse emotion classification datasets into a unified format, providing a standardized format but facing challenges inherent in unifying datasets with differing annotation schemes and domains, which may affect the generalizability of their findings.

Similarly, the multi-label approach for emotion cause detection presented by Chen et al. [34] shows significant performance improvement. However, the dependency on linguistic cues, the limited scope of the corpus, and potential challenges in cross-linguistic applications may affect its broader applicability.

2.2.3.4 Applications and Domains

Studies such as the scoping review by Maleki et al. [35], exploring root causes of insurance deductions in Iranian hospitals, provide valuable insights into specific application domains. However, the difficulty in finding suitable terms in English databases and potential challenges in cross-cultural generalization may impact the wider applicability of such studies.

Similarly, the text-document clustering-based cause and effect analysis methodology by Verma and Maiti [36], identifying root causes of incidents in a steel plant, provides a valuable approach. However, dependence on employee perceptions, short incident descriptions, and the static nature of the analysis may limit its effectiveness in dynamic environments.

2.2.3.5 Summary of Emotion Cause Analysis

Emotion Cause Analysis has seen considerable progress through foundational studies such as Liu [26] and Bostan et al. [27], which introduced comprehensive annotation frameworks and datasets. Despite these advancements, challenges such as subjective annotation, limited generalization across domains, and difficulty in capturing complex semantic roles persist. Approaches like Yuan et al. [30]'s sequence labeling and Sun et al. [31]'s Dual-Questioning Attention Network have demonstrated improvements in emotion-cause pair extraction, yet they primarily rely on data-driven methods lacking explainability and causal reasoning. The ongoing struggle to generalize findings across diverse linguistic and cultural contexts underscores the limitations of current approaches.

This gap highlights the potential of Neuro-Symbolic AI, which offers a promising path toward integrating symbolic reasoning with neural learning to enhance interpretability, scalability, and robustness in emotion cause analysis. Leveraging such techniques can provide deeper insights into causal relationships and improve the reliability of emotion detection systems, which is particularly relevant for developing advanced solutions like the Mentalisys application.

2.2.4 Neuro Symbolic AI-Based Mental Health Detection

Neuro-Symbolic AI combines the strengths of neural networks in learning patterns from large datasets with the interpretability and domain knowledge provided by symbolic reasoning. This integration is particularly critical in the domain of mental health detection, where interpretability, reliability, and contextual understanding are essential due to the sensitive and high-stakes nature of the applications [37].

A significant advancement in this field is the TAM-SENTICNET model, introduced in a study on early detection of depression using Neuro-Symbolic AI via social media analysis [7]. This approach leverages a symbolic reasoning layer built upon SenticNet to provide interpretable insights into emotion and sentiment analysis, addressing challenges in detecting subtle mental health cues. By combining the linguistic capabilities of neural networks with the structured knowledge of symbolic systems, TAM-SENTICNET demonstrates improved accuracy and transparency in identifying depressive language patterns.

In the broader context of healthcare applications, Neuro-Symbolic AI has been shown to address challenges in decision-making, particularly in scenarios where neural networks alone fail to provide reliable and interpretable results [38]. By integrating symbolic reasoning, systems can incorporate domain-specific knowledge and causal relationships, improving their ability to detect mental health conditions like stress and depression. This integration is critical for ensuring the trustworthiness of AI systems, especially in healthcare settings.

The Sentic PROMs framework, proposed by Antoniou et al. [8], further illustrates the potential of Neuro-Symbolic AI in mental health detection. Sentic PROMs bridge structured questionnaire data with unstructured natural language inputs, enabling more comprehensive assessments of health-related quality of life. This framework highlights the importance of explainability in mental health applications, as it allows clinicians to better understand the rationale behind AI-driven predictions.

While advancements in Neuro-Symbolic AI have improved mental health detection, challenges remain. The reliance on high-quality annotated datasets, such as those provided for stress detection and depression analysis, underscores the need for diverse and representative data sources [39]. Additionally, integrating symbolic reasoning with neural models often requires extensive computational resources, making scalability an

ongoing concern [40]. Despite these limitations, the adoption of Neuro-Symbolic AI offers a promising path forward for building reliable, interpretable, and scalable systems for mental health monitoring and early detection.

2.2.5 Natural Language Inference (NLI)

Natural Language Inference (NLI) forms the backbone of many Natural Language Processing (NLP) applications, enabling systems to deduce logical relationships between a premise and a hypothesis. It is foundational to tasks such as question answering, summarization, and dialogue systems. Benchmark datasets like SNLI and MultiNLI have played a pivotal role in advancing the field, providing large-scale annotated resources for evaluating entailment, contradiction, and neutrality [41, 42]. However, these datasets primarily focus on general-purpose language understanding, raising concerns about their applicability in specialized fields that require domain-specific reasoning or contextual awareness.

Domain-specific datasets like SCITAIL have addressed some of these limitations by introducing entailment tasks grounded in science question answering [43]. These tasks demand a deeper integration of domain knowledge and reasoning capabilities, exposing the limitations of shallow inference models. Similarly, datasets like e-SNLI extend traditional benchmarks by including human explanations for inferences, providing a step toward enhancing interpretability and causal understanding [44].

Recent innovations in NLI emphasize improving interpretability and robustness to address critical limitations of earlier models. For example, graph-constrained reasoning leverages structured knowledge graphs to enhance faithfulness and transparency, mitigating issues such as hallucinations in transformer-based models [45]. Furthermore, commonsense reasoning frameworks like ATOMIC [46] and COMET [47] have introduced causal relationships into NLI tasks, enabling models to better understand the "why" behind entailment decisions. Despite these advancements, challenges such as scalability, generalization, and dataset biases persist, limiting the applicability of NLI models in real-world domains like emotion cause analysis and mental health monitoring.

Addressing these challenges requires a shift from shallow pattern recognition toward deeper reasoning and causal analysis. The absence of robust causal inference capabilities in current NLI systems hinders their application in domains where interpretability and context are critical, such as identifying stressors or emotional triggers in workplace communications. Future research must focus on integrating symbolic reasoning, commonsense knowledge, and real-world datasets to develop holistic and scalable NLI solutions.

2.2.6 Commonsense Reasoning

Commonsense reasoning, the ability of systems to infer unstated assumptions, implicit relationships, and cause-effect dynamics, is a cornerstone of building intelligent AI systems. It extends beyond mere pattern recognition to encompass the interpretation of real-world scenarios and contextual knowledge, which is crucial for tasks like emotion analysis, causal inference, and decision-making.

Recent advancements, such as Generated Knowledge Prompting (GKP), have significantly contributed to this field. GKP enables models to generate task-specific knowledge without relying on predefined structured knowledge bases, achieving state-of-the-art performance in tasks like numerical reasoning (NumerSense), general commonsense question answering (CommonsenseQA 2.0), and scientific reasoning (QASC) [48]. However, while GKP reduces dependence on static knowledge repositories, it introduces challenges related to the quality and consistency of generated knowledge. Without validation mechanisms, GKP-based models risk producing hallucinated outputs, which limits their reliability in real-world applications.

Techniques like Chain-of-Thought (CoT) prompting have further advanced commonsense reasoning by encouraging step-by-step problem solving. CoT improves both interpretability and accuracy by decomposing complex reasoning tasks into intermediate steps [49]. Building on this approach, the Tree-of-Thought (ToT) framework introduces a generalized reasoning method that explores multiple reasoning paths and allows for backtracking, enhancing deductive and abductive reasoning [50]. While CoT and ToT frameworks demonstrate impressive performance on commonsense reasoning benchmarks, their computational cost remains a significant barrier to scalability in dynamic, real-time systems.

Datasets like CommonsenseQA, which challenge models to answer multiple-choice questions requiring commonsense knowledge [51], and ATOMIC, which focuses on causal and inferential relationships in everyday situations, have been instrumental in advancing the field. Additionally, frameworks such as COMET, built upon ATOMIC, enable models to predict causal, temporal, and social relationships. These resources highlight the importance of structured commonsense knowledge and its integration into neural architectures. However, many commonsense datasets still lack cultural diversity, which limits their ability to generalize across different languages and contexts.

Despite advancements, commonsense reasoning systems face challenges such as biases in pre-trained language models, the inherent subjectivity of commonsense knowledge, and the limitations of benchmarks like CommonsenseQA and ATOMIC, which focus on static, single-turn tasks rather than dynamic, multi-turn reasoning. Future work should prioritize integrating symbolic and neural reasoning techniques, developing diverse and comprehensive datasets, and creating scalable frameworks to enhance their applicability to real-world tasks like emotion cause analysis and mental health

monitoring, where understanding context and causality is essential.

2.2.7 Prompting in Natural Language Processing

Prompt engineering has revolutionized the application of large language models (LLMs), enabling task-specific adaptation through structured natural language instructions. This paradigm reduces the reliance on extensive fine-tuning by leveraging the pre-trained capabilities of LLMs. Techniques such as few-shot prompting and zero-shot prompting have demonstrated remarkable generalization across tasks, exemplified by models like GPT-3, which achieve strong performance with minimal labeled data [52, 53]. By presenting task instructions directly within the input, these strategies exploit the latent capabilities of LLMs, offering practical solutions for domains with limited annotated datasets.

One prominent advancement in prompting is Generated Knowledge Prompting (GKP), which combines the power of LLMs with external knowledge elicitation. GKP allows models to generate task-relevant knowledge dynamically, improving performance on commonsense reasoning tasks such as CommonsenseQA and NumerSense [48]. However, GKP faces challenges related to the quality and consistency of generated knowledge, as models are prone to hallucinations and inaccuracies when external validation mechanisms are absent. This underscores the importance of combining GKP with post-generation validation frameworks to ensure the reliability of outputs in high-stakes applications.

Frameworks like ReAct (Reasoning and Acting) have further advanced prompting by enabling models to dynamically interact with external systems, such as APIs, databases, or knowledge graphs, during inference [54]. ReAct bridges the gap between static reasoning and real-time action, allowing iterative decision-making that improves contextual understanding and performance in complex, multi-step inference tasks. For instance, ReAct has been particularly effective in scenarios requiring access to dynamic tools, such as retrieving up-to-date information or performing calculations, which traditional prompting approaches cannot achieve in isolation. However, the computational complexity and system integration requirements of ReAct present scalability challenges, particularly in real-time applications or environments with limited resources.

Despite significant advancements, prompting techniques face key limitations, including sensitivity to phrasing, which reduces robustness for large-scale applications, and reliance on immediate context, which limits their effectiveness in tasks requiring long-context reasoning or causal inference. The lack of a universal framework for evaluating prompts further complicates development, as benchmarks often fail to address real-world complexities like user errors or incomplete instructions. While resources like the Prompting Guide offer valuable insights into development and opti-

mizing prompts for tasks such as summarization and translation, practical applications still demand extensive manual experimentation, as prompt design remains an intricate and subjective process.

2.2.8 Neuro-Symbolic AI for Workplace Mental Health Monitoring

The application of Neuro-Symbolic AI in mental health monitoring represents a promising direction for developing explainable and effective models that can capture complex emotional states and their causes. Neuro-Symbolic AI integrates the pattern recognition capabilities of neural networks with the interpretability of symbolic reasoning, making it particularly suitable for high-stakes domains like mental health monitoring where interpretability and contextual understanding are essential [37].

Several studies have demonstrated the potential of Neuro-Symbolic AI for emotion and sentiment analysis. The TAM-SENTICNET model, for example, effectively combines symbolic reasoning with neural networks to provide interpretable insights into depressive language patterns from social media posts [7]. While this model achieves improved transparency and accuracy in identifying depressive states, it is primarily designed for social media data and lacks applicability in workplace communication platforms like Slack.

Similarly, the Sentic PROMs framework proposed by Antoniou et al. [8] bridges structured questionnaire data with unstructured natural language inputs, highlighting the importance of explainability in mental health applications. However, it relies heavily on pre-defined structures, which limits its adaptability to dynamic, informal communication often encountered in workplace settings. Furthermore, these existing frameworks do not effectively address the challenge of detecting complex emotional states like stress and frustration arising from contextual interactions within corporate platforms.

Despite these advancements, significant gaps remain in applying Neuro-Symbolic AI to workplace mental health monitoring. Current models face challenges such as reliance on high-quality annotated datasets, limited generalization capabilities, and the need for extensive computational resources [39, 40]. More critically, they lack robust methods for integrating causality and subjective context effectively, particularly in real-time communication environments like Slack, where textual exchanges are brief, dynamic, and context-dependent.

To address these limitations, this research proposes the development of a novel framework known as the **Commonsense-Driven Symbolic ReAct-NLI (CSR-NLI) Prompting Framework**. The CSR-NLI framework integrates Natural Language Inference (NLI), commonsense reasoning, and Neuro-Symbolic AI to provide context-aware, explainable, and scalable mental health monitoring solutions tailored to workplace communication data. Unlike previous approaches, CSR-NLI is specifically de-

signed to infer causal relationships from employee messages, thereby improving the detection of nuanced emotional states such as stress, frustration, and burnout.

Furthermore, by leveraging advanced prompting techniques and symbolic reasoning, the CSR-NLI framework enhances the interpretability of model outputs and addresses the limitations of existing models in terms of scalability and generalization. This integration offers a unique opportunity to bridge the gap between traditional sentiment analysis tools and comprehensive mental health monitoring systems, making it particularly valuable for corporate environments where employee well-being and productivity are closely intertwined.

The proposed CSR-NLI framework represents a significant step forward in the development of Neuro-Symbolic AI-based mental health monitoring tools. By addressing the limitations of existing methods and providing a holistic approach to emotion recognition, causality detection, and interpretability, this framework offers a robust solution for enhancing employee well-being and fostering a healthier workplace culture.

CHAPTER 3

METHODOLOGY

3.1 Mentalisys Health Application Development

3.1.1 Overview of the Mentalisys Health Application

The Mentalisys Health Application is an advanced mental health analysis tool designed to assess workplace well-being by analyzing communication data extracted from Slack. Developed using the H2O Wave framework, the application integrates sophisticated data visualization and user-friendly interfaces to provide actionable insights for fostering a supportive work environment. Its core purpose is to detect potential signs of stress, depression, and emotional states in employee communications, enabling organizations to implement timely interventions that promote mental health and productivity.

At the heart of the application is an AI/ML analytical framework that employs Natural Language Processing (NLP) techniques to evaluate textual communication patterns. By analyzing message tone, linguistic changes, and interaction frequencies, the application identifies early indicators of emotional distress, such as depression and stress tendencies. This proactive approach empowers organizations to prioritize mental health while improving workplace engagement and performance.

The Mentalisys Health Application is built on a Model-View-Controller (MVC) architecture, ensuring scalability, modularity, and maintainability:

- **Model:** Handles the data representation and business logic, including managing Slack messages, user information, and team metadata.
- **View:** Developed using H2O Wave, the View component provides interactive interfaces such as login pages, admin dashboards, and employee dashboards. These interfaces offer visualizations of key metrics, such as sentiment trends, daily message counts, and emotional proportions.
- **Controller:** Acts as the intermediary between the Model and View, processing user inputs, handling requests, and ensuring seamless data flow.

The application features two primary dashboards tailored to different user roles:

- **Admin Dashboard:** Offers a comprehensive overview of the organization's mental health status through interactive visualizations. Metrics include channel-wise message counts, sentiment distributions, employee depression tendencies, and active versus passive user engagement.

- **Employee Dashboard:** Provides personalized insights for employees, including their current depression tendency score, weekly sentiment trends, emotional frequencies, and the influence of their interactions on colleagues' emotional states.

Key components of the Mentalisys system include:

- A Slack data extraction bot that retrieves and preprocesses communication data.
- AI/ML modules for sentiment analysis, emotion state classification, and stress and depression detection.
- MongoDB for storing and managing processed data, ensuring efficient retrieval and analysis.
- H2O Wave-based dashboards for presenting analytics in an intuitive and actionable format.

A distinguishing feature of the Mentalisys Health Application is the integration of a Neuro-Symbolic Prompting Technique (CSR-NLI) for advanced causal reasoning and classification. This technique, powered by OpenAI APIs, augments traditional sentiment analysis by providing deeper insights into the reasoning and context behind messages.

By leveraging these capabilities, the Mentalisys Health Application delivers a robust, real-time solution for monitoring employee mental well-being. Its ability to translate communication data into meaningful insights helps organizations identify potential risks early and create healthier, more supportive work environments.

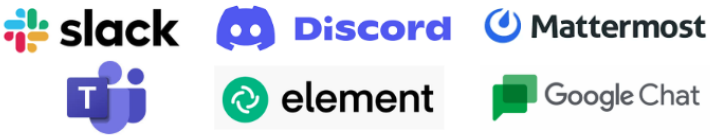
3.1.2 Justification for Choosing Slack as the Primary Communication Platform

The Mentalisys Health Application utilizes Slack as its primary communication platform for analyzing workplace well-being. The selection of Slack over other platforms was made based on several critical factors that enhance the generalizability of the proposed approach to other workplace communication tools.

Slack offers a comprehensive API that allows for seamless data extraction, processing, and integration with third-party analytical tools. Its architecture supports both public and private channels, making it well-suited for capturing diverse communication patterns within organizations. Moreover, Slack's widespread adoption across various industries, particularly within innovative and tech-forward companies, provides a robust and relevant dataset for mental health analysis.

The comparative analysis of user demography across various communication platforms is presented in Figure 3.1. The analysis highlights Slack's considerable market share, number of customers, and growth rate, which further justifies its selection for this study. Additionally, while Slack is the primary platform, the methodology can be

adapted to other platforms such as Microsoft Teams, Discord, Element, Mattermost, and Google Chat, enhancing the generalizability of the proposed approach.



	Market Share %	# of Customers (mil.)	Growth Percentage
Slack	2.9	12	25
Microsoft Teams	2.7	115	50
Discord	2.3	140	37.3
Element	0.8	0.2	100
Mattemost	0.01	N/A	N/A
Google Chat	2.4	6	20

Fig. 3.1: User Demography and Market Share Analysis of Communication Platforms

3.2 Application Architecture

The Mentalisys Health Application is built upon a robust Model-View-Controller (MVC) architectural pattern. This design ensures a clear separation of concerns, enhances scalability, and improves maintainability. The architecture is divided into three primary components: the Model, the View, and the Controller, each of which plays a distinct role in the functionality of the system. Additionally, the application integrates Slack APIs for data extraction, MongoDB for data storage, and H2O Wave for data visualization.

3.2.1 Model

The Model component is responsible for managing the core data and business logic of the application. It interacts directly with the MongoDB database, defining the structure of various entities and handling operations such as data retrieval, validation, and updates.

In the context of Mentalisys, the key models include:

- **User Model:** Stores user-related data such as usernames, passwords, email addresses, registration status, and mental health indicators such as depression tendency. Each user is assigned a *role* (e.g., admin or employee) for access control within the system.

- **Team Model:** Represents an organizational team, storing metadata such as the *team name* and *team members*. A list of associated users ensures structured access and interaction within a workplace.
- **Message Model:** Represents individual Slack messages, capturing message content, timestamps, and references to users, channels, and threads. Additionally, this model incorporates embedded predictions for *sentiment analysis*, *emotion recognition*, and *depression tendency assessment*, facilitating AI-driven mental health monitoring.
- **Channel Model:** Stores information about Slack channels, including the *channel ID*, *channel name*, *privacy status* (public or private), and its associated team. This model ensures structured data organization across workplace conversations.

The Model implements CRUD (Create, Read, Update, Delete) operations, enforces business rules, and ensures data consistency. This layer is critical for seamless integration with both the Slack APIs and the AI/ML analytical modules.

3.2.2 View

The View component is responsible for rendering the user interface and presenting data in an intuitive and interactive format. Built using the H2O Wave framework, this component ensures real-time updates, dynamic visualizations, and a seamless user experience. The Mentalisys application consists of the following key views:

- **Login and Registration Pages:** Enable secure authentication, allowing users to log in, register, and reset passwords. The UI includes validation handling and notification-based feedback to enhance security and accessibility.
- **Admin Dashboard:** Provides administrators with high-level analytics, including:
  - Channel-wise and user-wise message counts.
  - Sentiment and emotion distribution metrics.
  - Depression tendency tracking over time.
  - Interactive visualizations such as pie charts, bar graphs, and statistical overviews.
- **Employee Dashboard:** Displays personalized mental health insights for users, including:
  - Individual depression tendency scores and sentiment trends.

  - Communication activity analytics (messaging patterns and engagement).
  - Impact analysis of user interactions on colleagues' well-being.
  - Box plots, line charts, and structured tables for interactive data representation.

The View component dynamically updates based on underlying data changes and ensures a fluid, responsive interaction for users. With its structured interface and real-time analytics, the Mentalisys application effectively bridges the gap between conversational data and mental health insights.

3.2.3 Controller

The Controller serves as the intermediary between the Model and the View, orchestrating user inputs, data processing, and interface updates. It ensures that the right data is retrieved, processed, and displayed effectively. In Mentalisys, the Controller handles the following key tasks:

- **User Authentication:** Manages secure registration, login, and session handling. Ensures users are verified before accessing the system.
- **Data Retrieval:**
  - Fetches Slack communication data, user details, and AI/ML-generated predictions from the MongoDB database.
  - Retrieves user-specific analytics, including depression tendency scores and engagement patterns.
  - Extracts message statistics by sentiment, emotion, and channel activity.
- **Data Processing and Analysis:**
  - Implements AI/ML-based sentiment analysis, emotion recognition, and stress detection models.
  - Aggregates user messages to track depression tendency trends over time.
  - Analyzes channel-wise engagement to assess how conversations impact workplace well-being.
- **Channel and User Engagement Tracking:**
  - Monitors Slack channel interactions, including message frequency and user participation.
  - Identifies the most active and passive users based on messaging behavior.

– Extracts insights on the influence of user communication on team mental health.

The Controller ensures seamless integration between the backend and the user interface, dynamically updating dashboards based on real-time data. By managing data flow efficiently, it enables a robust and interactive experience for administrators and employees alike.

3.2.4 Integration of Slack APIs, MongoDB, and H2O Wave

The Mentalisys Health Application integrates various technologies to ensure seamless data extraction, processing, storage, and visualization. Slack APIs are used for data extraction, MongoDB for storing structured data, and H2O Wave for developing interactive dashboards. The overall system workflow is illustrated in Figure 3.2.

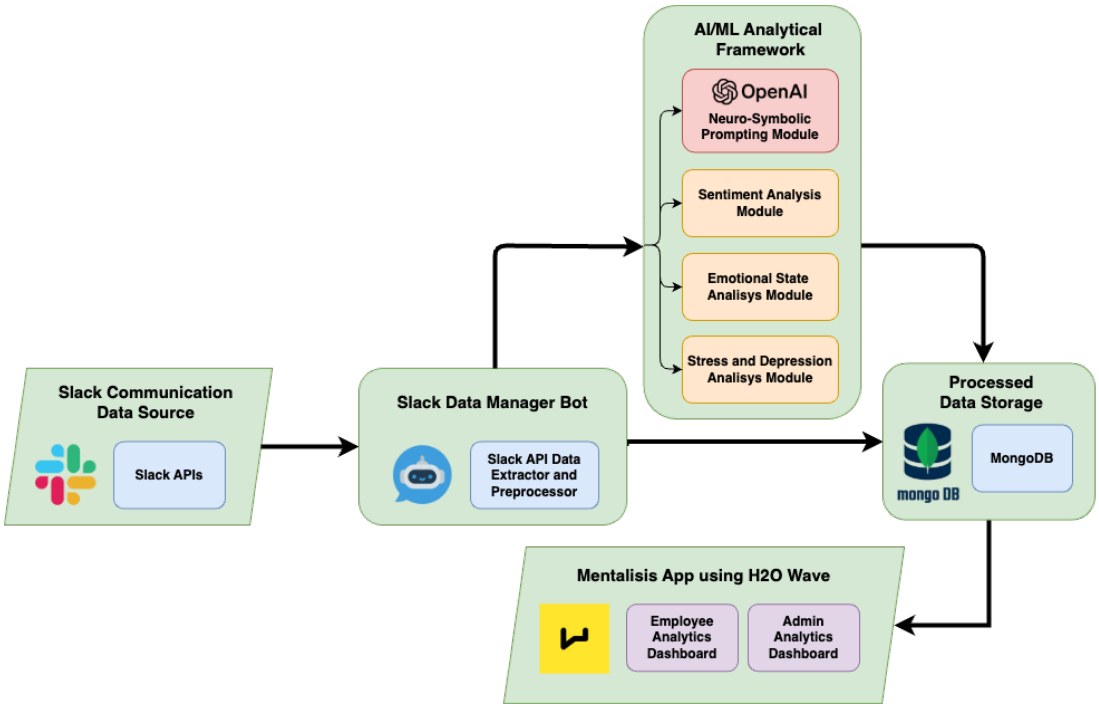


Fig. 3.2: Architecture of the Mentalisys Health Application

3.2.5 Workflow Explanation

The Mentalisys Health Application architecture, as shown in Figure 3.2, comprises the following key components:

- **Slack Communication Data Source:** Slack APIs are used to extract messages and metadata from Slack channels.

- **Slack Data Manager Bot:** This custom-built bot processes and pre-processes extracted Slack data for further analysis.
- **AI/ML Analytical Framework:** Includes modules for sentiment analysis, emotional state classification, stress and depression detection, and the Neuro-Symbolic Prompting Module powered by OpenAI APIs.
- **Processed Data Storage:** MongoDB serves as the repository for both raw and processed data.
- **Mentalisys Application:** Built on H2O Wave, this application visualizes the processed data via admin and employee dashboards.

This architecture ensures seamless data flow and integration between components, enabling real-time monitoring and actionable insights into employee well-being.

3.2.6 Benefits of the MVC Architecture

The adoption of the MVC architecture enhances the modularity of the Mentalisys Health Application. Each component can be independently developed, tested, and maintained, ensuring scalability and adaptability for future updates. This design also facilitates seamless integration of advanced AI/ML modules, such as the Neuro-Symbolic Prompting Module, while maintaining a clean and efficient system structure.

By leveraging the MVC pattern and integrating state-of-the-art technologies, the Mentalisys Health Application delivers a scalable, maintainable, and user-friendly platform for workplace mental health monitoring.

3.3 Slack Data Manager Bot

The Slack Data Manager Bot serves as the primary data extraction module within the Mentalisys Health Application. This bot interacts with Slack APIs to retrieve, preprocess, and store communication data, forming the foundation for advanced sentiment and emotion analysis. It is designed to extract Slack messages from public and private channels, ensuring comprehensive data collection for workplace well-being analysis.

3.3.1 Core Functionalities

The Slack Data Manager Bot performs the following key functions:

- **Message Extraction:** Retrieves both general messages and threaded replies from specified Slack channels using the `conversations_history` and `conversations_replies` API methods. It includes metadata such as timestamps, user IDs, and reactions.

- **Thread Handling:** Identifies threaded messages by checking the `thread_ts` field and retrieves parent messages along with their replies, ensuring a complete understanding of conversations.
- **Metadata Collection:** Captures supplementary data such as user information, team identifiers, and channel details for contextual analysis.
- **Time Filtering:** Allows for time-specific data extraction to target messages within a defined period, facilitating periodic analysis and monitoring.
- **Preprocessing:** Formats extracted data into structured records, removing duplicates and ensuring only relevant attributes are retained. The bot filters out system messages (e.g., `bot_message`, `channel_join`) while retaining user-generated content.
- **Reactions and Mentions Processing:** Extracts and stores reactions within the `messages` collection as an array of reaction details. Mentions (`<@user_id>`) are resolved using the `users` collection.
- **Database Integration:** Stores the preprocessed data in MongoDB, maintaining indexed collections for users, teams, channels, and messages. Indexes are placed on `user_id`, `thread_ts`, and `timestamp` for efficient querying.

3.3.2 Slack Data Extraction Pipeline

The Slack Data Extraction Pipeline is designed to capture and organize communication data from Slack into a structured format for further processing and analysis. Figure 3.3 illustrates the data model derived from `SlackExtractionPipelineDDL.png`.

3.3.2.1 Data Model Overview

The pipeline is designed around several key entities, each playing a critical role in organizing Slack communication data:

- **User:** Represents Slack users interacting within the workspace. Users are linked to messages they create and reactions they provide.
- **Team:** Corresponds to the Slack workspace, grouping users and channels.
- **Channel:** Denotes communication channels within the workspace where messages and threads are exchanged.
- **Thread:** Represents discussions extending from a parent message within a channel. Tracks parent-child relationships between messages.

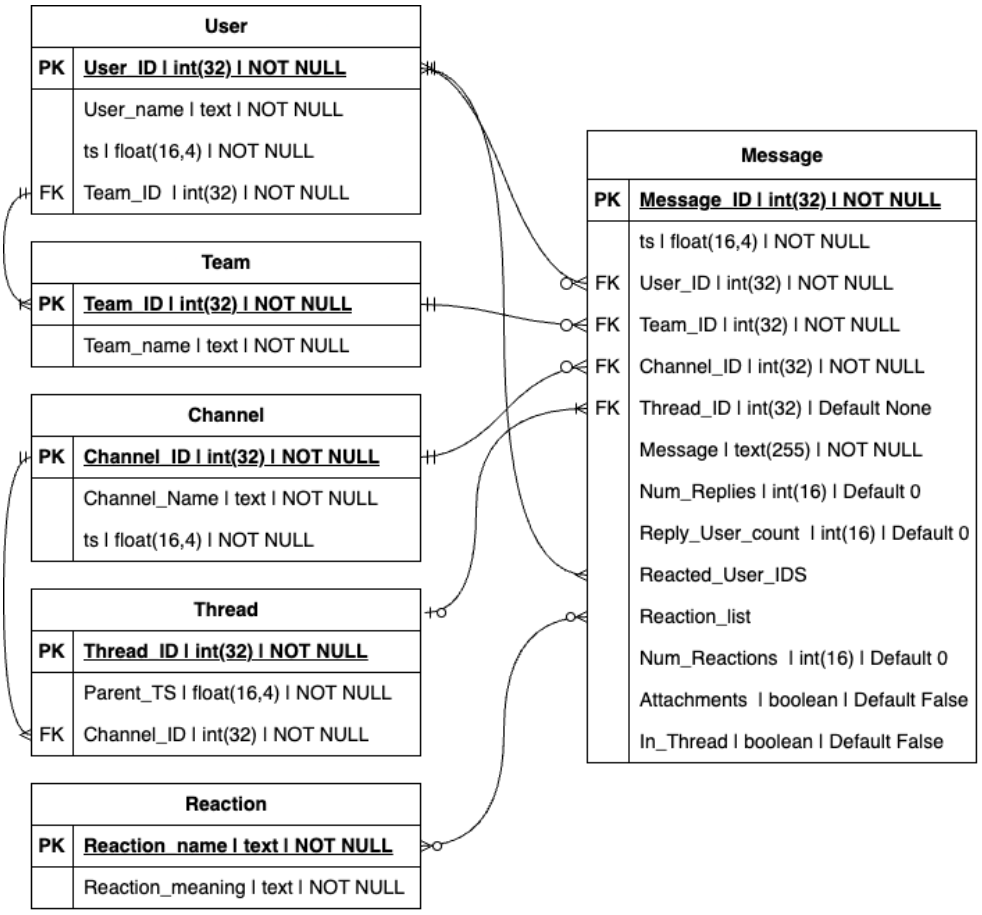


Fig. 3.3: Slack Data Extraction Pipeline Data Model

- **Message:** Represents the core communication exchanged in Slack, containing textual content, timestamps, and metadata.
- **Reaction:** Tracks user interactions with messages through emojis, providing additional context about the message's reception.

3.3.2.2 Entity Relationships

The relationships between these entities ensure comprehensive data organization and contextual understanding:

- **User → Message:** Captures messages authored by users.
- **Team → Channel:** Represents the association between Slack workspaces and their channels.
- **Channel → Thread → Message:** Shows the hierarchical flow from channels to threads and individual messages within those threads.

- **Reaction → Message:** Highlights user interactions with specific messages.

This structured data pipeline facilitates efficient storage, querying, and analysis, forming the backbone of the Mentalisys Health Application's analytical capabilities.

3.3.3 Development Steps

The development of the Slack Data Manager Bot involved several key steps:

1. **Setup and Dependencies:** The bot is implemented in Python using the following libraries:

```
slack-sdk==3.23.0
python-dotenv==0.21.0
pymongo==4.5.0
```

2. **Channel Enumeration:** The bot uses the `conversations_list` API method to enumerate all public and private channels where it is a member.

3. **Message Retrieval:** Messages are fetched using `conversations_history`, while threaded replies are captured with `conversations_replies`. The retrieved data includes reactions, mentions, and emojis.

4. **Preprocessing:** Extracted data is cleaned to remove unnecessary records, duplicates, non-user messages (e.g., bot messages), and deleted messages.

5. **Data Storage:** The cleaned data is stored in MongoDB collections optimized for performance:

    - **Users:** Stores user IDs, roles, and team affiliations.
    - **Teams:** Captures workspace-specific metadata.
    - **Channels:** Contains channel names and visibility details.
    - **Messages:** Stores message content, timestamps, metadata, reactions, and mentions.

3.3.4 Slack Bot Setup and Permissions

To extract messages and interact with Slack APIs, the Slack Data Manager Bot must be configured within the Slack Developer Console. Figure 3.4 shows the bot's setup in the Slack API console.

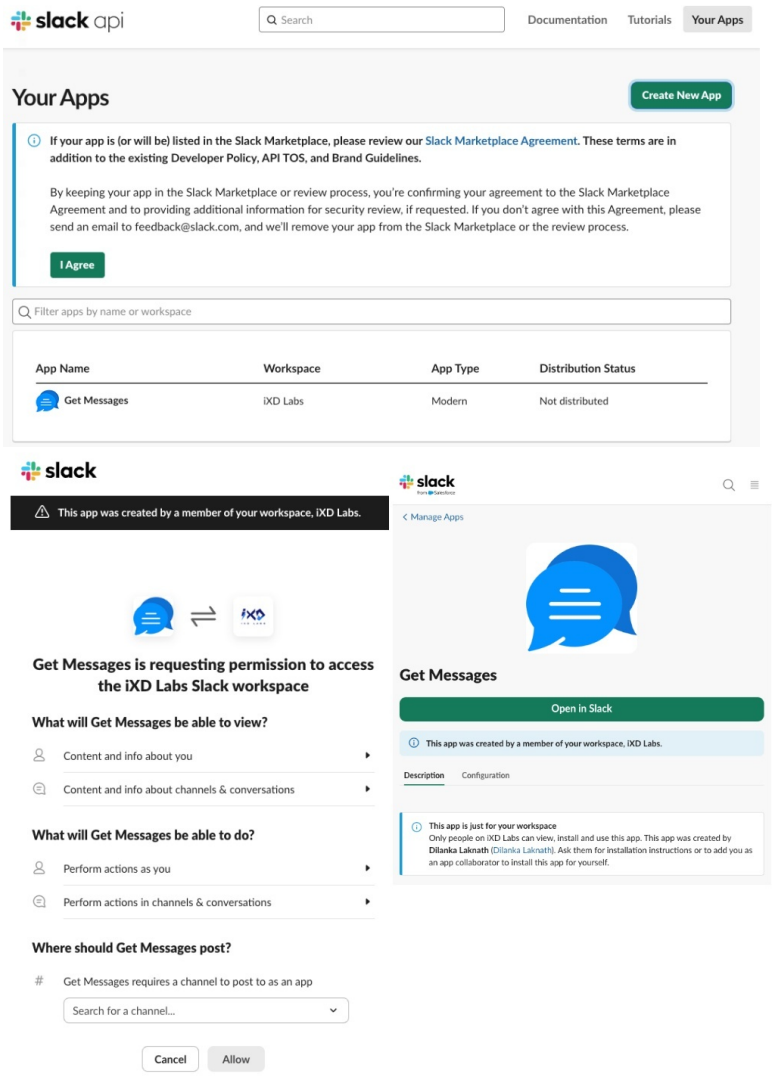


Fig. 3.4: Slack API Console: Bot Dashboard and Permissions Setup

3.3.4.1 Bot Configuration Workflow

The configuration workflow involves:

- Creating a new Slack app in the Developer Console and naming it appropriately (e.g., "Get Messages").
- Assigning scopes such as:
  - `channels:history`, `channels:read` - To fetch messages.
  - `users:read` - To retrieve user metadata.
  - `conversations:read` - Required to fetch private and direct messages.
- Deploying the app within the workspace and granting it access permissions.

Once configured, the bot can interact with public and private Slack channels to extract and process data efficiently.

3.3.5 Future Enhancements

The following features are planned for future development:

- **Shared Channels:** Extend support for Slack Connect channels across workspaces.
- **File Attachments:** Extract and categorize messages containing files or links.
- **Message Edits and Deletions:** Implement tracking of edited and deleted messages.

The Slack Data Manager Bot is a critical module in the Mentalisys Health Application, facilitating continuous and efficient data extraction. By leveraging Slack APIs and advanced preprocessing techniques, it ensures a robust foundation for real-time mental health monitoring and analysis.

3.4 AI/ML Analytical Framework

3.4.1 Emotional State Analysis Module

The Emotional State Analysis Module is a key component of the Mentalisys application, providing real-time analysis and categorization of user messages into emotional states. The module utilizes advanced machine learning techniques and is powered by well-curated datasets.

3.4.1.1 Dataset Benchmarking and Selection

Several datasets were considered for building the emotion analysis model. Table 3.1 presents a detailed benchmarking of prominent datasets for emotion recognition in conversations.

After evaluating the datasets, the CARER (Contextualized Affect Representations for Emotion Recognition) dataset was selected due to its large-scale, high-coverage emotion representation model and superior generalizability for textual emotion recognition.

3.4.1.2 CARER Dataset

The CARER dataset introduces a novel approach to textual emotion recognition by leveraging contextualized affect representations, which enrich emotion detection via pattern-based embedding techniques [55]. The dataset was created using a distant supervision approach by collecting emotion-labeled tweets based on hashtag annotations.

TABLE 3.1: Benchmarking of Emotion Recognition Datasets

Dataset Name	Description	# Samples / Utterances	# Classes (Emotions)	Evaluation Metric	SOTA Performance
CARER	Contextualized affect representations for textual emotion recognition based on enriched pattern-based embeddings [55].	10M tweets (distant supervision)	Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust	F1: 79%	CARER
GoEmotions	58k English Reddit comments labeled for 27 emotions and neutral.	58,009 utterances	27 emotions + Neutral	Average F1: 46%	BERT
Emotion Dataset	Dataset of English Twitter messages labeled with six emotions.	2,000 tweets	Anger, Fear, Joy, Love, Sadness, Surprise	F1: 93.8%	DistilBERT
EmoryNLP	Textual data from the *Friends* TV series; annotated with seven emotion categories.	12,606 utterances	Sad, Mad, Scared, Powerful, Peaceful, Joyful, Neutral	Weighted F1: 42.08%	CKERC
EmoContext	Three-turn English Tweets labeled for four emotions.	30,160 dialogues	Happy, Sad, Angry, Others	Micro F1: 77.65%	NELEC
RECCON	Dataset for emotion cause recognition in conversations.	10,600 utterances	Angry, Excited, Fear, Sad, Surprise, Frustration, Happy, Disappointed, Disgust, Neutral	Micro F1: 75.71%	SpanBERT
DailyDialog	Multi-turn English dialogues annotated with emotions.	13,118 dialogues	Anger, Disgust, Fear, Happiness, Sadness, Surprise	Micro F1: 64.07%	S+PAGE

CARER is specifically designed to capture nuanced emotion expression across diverse text sources.

3.4.1.2.1 Key Features of CARER

- **Distant Supervision:** The dataset includes over 10M tweets annotated for emotions based on 339 emotion-related hashtags, providing a vast, diverse training corpus.
- **Pattern-Based Embeddings:** Unlike traditional lexicon or rule-based approaches, CARER employs enriched emotion pattern extraction using graph-based representations, significantly improving emotion recognition performance.
- **Context-Aware Modeling:** The dataset incorporates contextual dependencies, enabling models to capture emotions beyond mere word presence and infer latent emotional cues in textual interactions.

3.4.1.2.2 Dataset Statistics Table 3.2 provides an overview of the CARER dataset distribution and its emotion categories.

3.4.1.2.3 Application in Mentalisys: The CARER dataset is highly suitable for Mentalisys, as it provides a scalable, robust, and domain-adaptive model for text-based

TABLE 3.2: CARER Dataset Statistics

Statistic	**Count**
Number of Tweets	10M+
Number of Unique Hashtags Used	339
Number of Emotion Categories	8
Emotion Distribution:	
Anger	102,289
Anticipation	3,975
Disgust	8,934
Fear	102,460
Joy	167,027
Sadness	214,454
Surprise	46,101
Trust	19,222

emotion recognition. Its contextualized approach improves the detection of subtle emotional cues in workplace communication, making it ideal for assessing employee emotion within Slack conversations.

3.4.1.3 Emotion Analysis Model Development

The Emotion Analysis Module classifies Slack messages into one of several emotional categories to gain insights into workplace communication patterns. The model was developed using the CARER dataset through the following process:

- **Preprocessing:** Slack messages were cleaned using tokenization, stopword removal, and lemmatization to standardize text data. Emotion labels were encoded for classification.
- **Feature Engineering:** Textual features were extracted using TF-IDF vectorization to capture relevant linguistic patterns.
- **Model Training:** Multiple machine learning models, including Logistic Regression, Naïve Bayes, Decision Trees, Random Forest, and SVM, were trained to classify emotions.
- **Evaluation:** The models were assessed using accuracy, precision, recall, and F1-score to determine the most effective classifier.
- **Integration:** The selected emotion classification model was deployed into Mentalisys for real-time emotion detection, enabling dynamic analysis of workplace communication.

By integrating emotion classification into Mentalisys, this module provides valuable insights into employee emotional states, supporting mental health monitoring and organizational well-being.

3.4.2 Sentiment, Stress, and Depression Analysis Modules

The Mentalisys application incorporates modules for Sentiment Analysis, Stress Analysis, and Depression Detection to assess employee communication. These modules work together to provide insights into workplace sentiment and mental well-being.

A common dataset, Dreaddit, was selected as the benchmark dataset for training and evaluation due to its rich annotations covering both sentiment and stress indicators. The dataset comprises 190K Reddit posts, with 3.5K manually labeled segments, making it well-suited for analyzing emotional and mental health trends in text-based conversations.

3.4.2.1 Dataset Benchmarking and Selection

To determine the most suitable dataset, multiple publicly available datasets were evaluated. Table 3.3 presents a summary of the datasets considered. Dreaddit was chosen for its multi-domain coverage, detailed sentiment annotations (positive, neutral, negative), and binary stress labels (stress, non-stress), making it applicable to both sentiment analysis, stress analysis, and depression detection.

TABLE 3.3: Benchmarking of Sentiment and Stress Analysis Datasets

Dataset Name	**Description**	**# Samples**	**# Classes**	**Evaluation Metric**	**SOTA Performance**
Dreaddit	A dataset from Reddit annotated for sentiment and stress analysis.	3,554 labeled segments	3 (Positive, Neutral, Negative) + 2 (Stress, Non-Stress)	Accuracy, F1-Score, Precision, Recall	N/A
Kayalvizhi	Depression-level detection dataset from social media.	16,632	3 (Not Depressed, Moderately Depressed, Severely Depressed)	F1-Score, Accuracy	Word2Vec + Random Forest: 87.7%
MDDL	Multimodal learning dataset for depression detection.	Tweets from 2009-2016	2 (Depressed, Non-Depressed)	F1-Score	MDL: 85%
RSDD	Reddit posts from self-reported depression-diagnosed users and controls.	9,000 users with depression; 107,000 controls	2 (Depressed, Non-Depressed)	F1-Score	CNN: 51%

3.4.2.2 Sentiment Analysis Model Development

The Sentiment Analysis Module classifies Slack messages into positive, neutral, or negative sentiment to monitor workplace mood. The model was developed using the Dreaddit dataset through the following process:

- **Preprocessing:** Slack messages were cleaned by removing special characters, stopwords, and redundant spaces. Sentiment labels were encoded for training.
- **Feature Engineering:** TF-IDF vectorization was applied to extract meaningful text representations.
- **Model Training:** Multiple machine learning models, including Logistic Regression, Support Vector Machines (SVM), Multinomial Naïve Bayes, Decision Trees, and Random Forest, were trained to classify sentiment.
- **Evaluation:** The models were assessed using accuracy, precision, recall, and F1-score to determine the best-performing sentiment classifier.
- **Integration:** The selected sentiment model was deployed into Mentalisys for real-time message classification and workplace sentiment tracking.

By tracking sentiment trends, this module provides organizations with actionable insights to improve workplace communication.

3.4.2.3 Stress and Depression Analysis Model Development

The Stress and Depression Analysis Module detects stress-inducing and depressive language in employee messages. The model was developed using Dreaddit's stress labels and trained using the following structured pipeline:

- **Preprocessing:** Messages were preprocessed through tokenization, stopword removal, and lemmatization. Stress labels were encoded for classification.
- **Feature Engineering:** Linguistic features such as emotional tone, sentiment polarity, and contextual language patterns were extracted from the text.
- **Model Training:** Various models, including Logistic Regression, SVM, Random Forest, and Transformer-based deep learning models (e.g., BERT), were trained to classify stress and depression.
- **Evaluation:** Model performance was evaluated using accuracy, precision, recall, and F1-score to identify the most effective classifier.
- **Integration:** The best-performing model was integrated into Mentalisys to analyze messages in real-time, providing insights into stress levels and depression indicators.

3.4.2.3.1 Output and Usage The module assigns a stress and depression score to each message, enabling HR and managers to monitor employee well-being. Insights are visualized in the admin and employee dashboards to aid proactive intervention.

3.4.2.4 Conclusion

By leveraging the Dreaddit dataset and advanced machine learning models, these modules provide a comprehensive analysis of sentiment, stress, and depression in workplace communication. This integration enhances mental health awareness and supports timely interventions for employee well-being.

3.4.3 Neuro-Symbolic Prompting Module

The Neuro-Symbolic Prompting Module introduces the **Commonsense-Driven Symbolic ReAct-NLI (CSR-NLI)** framework, integrating symbolic reasoning and neural inference to achieve causal reasoning and classification for conversational messages. This module utilizes the CAMS Dataset for causal analysis, supported by OpenAI APIs for real-time reasoning.

3.4.3.1 Dataset Benchmarking and Selection

The **CAMS Dataset** was chosen for its emphasis on causal analysis of mental health issues in social media posts.

TABLE 3.4: Benchmarking of Stress and Depression Causality Analysis Datasets

Dataset Name	**Description**	**# Samples**	**# Classes**	**Evaluation Metric**	**SOTA Performance**
CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts	A dataset focusing on analyzing reasons behind mental health issues expressed in social media posts. Includes annotations for causal interpretation and categorization.	5,051 Reddit posts (3,155 crawled and annotated; 1,896 re-annotated from the SDCNL dataset)	6 (No Reason, Bias or Abuse, Jobs and Careers, Medication, Relationship, Alienation)	Accuracy, F1-Score	Logistic Regression: 50.13%; CNN-LSTM: 47.78%
SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems	Identifies and classifies everyday stressors in SMS-like conversations, providing annotations for stressor types and presence.	6,850 SMS-like sentences	10 (Work, School, Financial Problems, Emotional Turmoil, Social Relationships, Family Issues, Health, Fatigue, Everyday Decision Making, Other)	F1-Score	BERT: 80.9%
Causal Explanation Analysis on Social Media Dataset	Facebook status updates manually labeled to identify whether messages contain causal explanations and the specific causal spans.	3,268 messages (1,600 contain causal explanations)	Not Specified	F1-Score	Bidirectional LSTMs for causal explanation identification: 85.3%

Key attributes include:

- **Size:** 5051 annotated instances combining crawled Reddit data and re-annotated SDCNL datasets.
- **Categories:** Six causal classes covering diverse mental health challenges.
- **Metrics:** Performance benchmarks demonstrated by Logistic Regression (F1: 0.5013) outperforming CNN-LSTM models.

3.4.3.2 Detailed Dataset Description

The CAMS dataset provides causal annotations for Reddit posts, facilitating interpretable analysis. The table below summarizes its structure:

TABLE 3.5: CAMS Dataset Overview

Cause	Instances	Min Length	Max Length	Avg Length
No Reason	694	1	508	59.78
Bias or Abuse	351	6	2109	347.48
Jobs and Careers	628	13	2258	228.28
Medication	623	5	1552	213.83
Relationships	1344	2	3877	229.35
Alienation	1408	3	1592	153.86

Annotations include causal interpretation and categorization, providing a comprehensive resource for mental health analysis.

3.4.3.3 Background of Commonsense-Driven Symbolic ReAct-NLI (CSR-NLI)

The CSR-NLI framework is based on fundamental concepts from Neuro-Symbolic AI (NSAI), which integrates neural networks and symbolic reasoning to enhance interpretability and robustness in natural language processing tasks. This section provides the theoretical background that underpins CSR-NLI, explaining how key concepts like commonsense reasoning, Natural Language Inference (NLI), symbolic logic, and iterative refinement contribute to the reasoning process.

3.4.3.3.1 Commonsense Reasoning Commonsense reasoning enables AI systems to make logical inferences beyond surface-level text analysis. Unlike rule-based systems that rely on predefined heuristics, CSR-NLI employs a dynamic commonsense hypothesis generator that adapts to diverse conversational inputs. The inclusion of commonsense reasoning:

- Provides contextual grounding for causal classification.
- Allows the system to infer implicit relationships between concepts.
- Ensures that AI-generated explanations are aligned with human expectations.

3.4.3.3.2 Natural Language Inference (NLI) Natural Language Inference (NLI) is a fundamental task in natural language understanding where a system determines the relationship between a premise and a hypothesis. The CSR-NLI framework builds upon this principle by:

- Treating the employee message as the premise.
- Generating a commonsense hypothesis dynamically as a benchmark for causal reasoning.
- Using NLI classification to validate generated reasoning with three possible labels:
  - Entailment – The reasoning logically aligns with the hypothesis.
  - Contradiction – The reasoning contradicts the hypothesis.
  - Neutral – The reasoning lacks sufficient alignment with the hypothesis.

By grounding reasoning in NLI principles, CSR-NLI ensures that causal classifications maintain logical consistency with real-world knowledge.

3.4.3.3.3 Symbolic Reasoning and Logical Validation Symbolic reasoning ensures that AI-generated explanations are interpretable and verifiable. In CSR-NLI, symbolic reasoning is embedded through:

- The comparison of generated reasoning against a predefined hypothesis.
- The iterative refinement of explanations to achieve logical entailment.
- The use of symbolic inference rules to classify reasoning into meaningful categories.

This integration prevents neural models from producing inconsistent or contradictory results, enhancing the overall robustness of the system.

3.4.3.3.4 Iterative Refinement with the ReAct Framework The ReAct framework (Reasoning + Acting) enables CSR-NLI to perform step-by-step reasoning and correction before finalizing causal classifications. The iterative process consists of:

- Thought – Generating an initial reasoning step based on the premise.
- Action – Comparing the generated reasoning against the commonsense hypothesis.
- Observation – Evaluating whether reasoning aligns with the hypothesis and refining if necessary.

This structured iteration ensures that the final reasoning is both logically valid and aligned with commonsense expectations.

3.4.3.3.5 The Role of Neuro-Symbolic AI in CSR-NLI Neuro-Symbolic AI (NSAI) integrates the pattern recognition capabilities of neural models with the logical consistency of symbolic methods. This hybrid approach is crucial for:

- Enhancing interpretability – Unlike black-box neural networks, symbolic reasoning provides a transparent decision-making process.
- Improving generalization – By relying on commonsense knowledge rather than pure data-driven learning, CSR-NLI can handle diverse conversational inputs effectively.
- Ensuring robustness – The iterative refinement process corrects inconsistencies, reducing the likelihood of ambiguous or incorrect classifications.

3.4.3.3.6 Conclusion CSR-NLI leverages NLI, commonsense reasoning, symbolic validation, and iterative refinement to create an interpretable and robust AI reasoning system. By integrating Neuro-Symbolic AI principles, it provides a structured yet adaptable approach to classifying conversational messages in real-world applications.

3.4.3.4 Commonsense-Driven Symbolic ReAct-NLI Prompting (CSR-NLI)

The CSR-NLI framework, depicted in Figure 3.5, integrates commonsense reasoning with natural language inference (NLI) for causal analysis and classification. It combines neural generation (via LLMs) with symbolic reasoning (logical validation and iterative refinement) to ensure interpretable and consistent decision-making. At its core, it employs the ReAct framework, iteratively refining causal reasoning through a structured workflow.

3.4.3.4.1 Input Premise and Hypothesis Generation The process begins with an Input Premise, representing an employee's conversational message extracted from Slack. The Commonsense Hypothesis Generator then formulates a hypothesis grounded in real-world knowledge.

For example:

- *Input Premise:* "I feel overwhelmed at work due to upcoming deadlines."
- *Commonsense Hypothesis:* "Deadlines often cause stress in jobs and careers."

This hypothesis acts as a logical benchmark against which causal reasoning is evaluated.

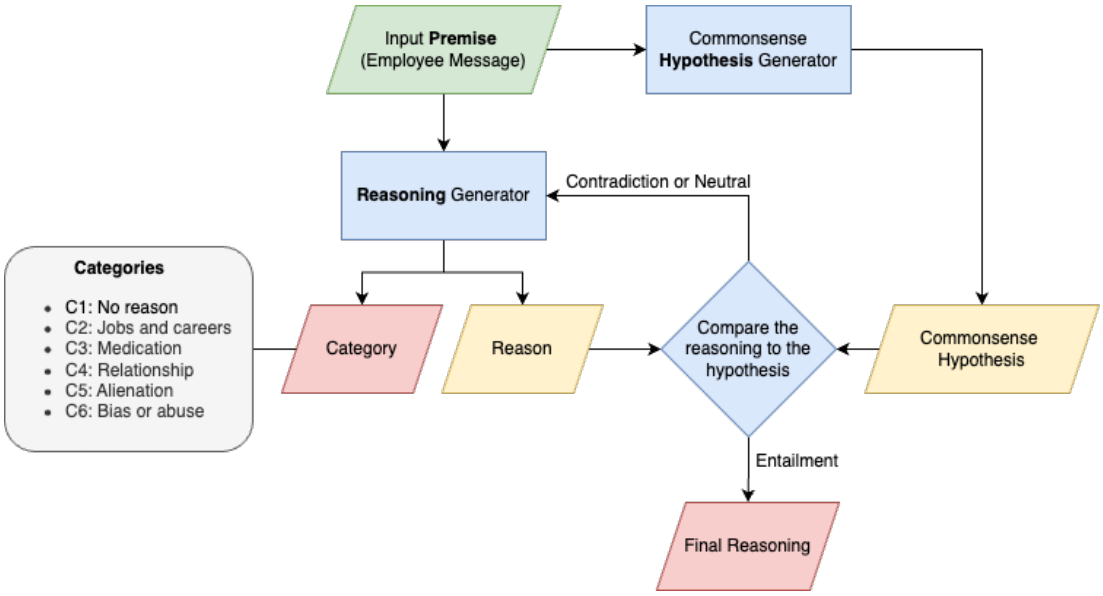


Fig. 3.5: Commonsense-Driven Symbolic ReAct-NLI (CSR-NLI) Framework

3.4.3.4.2 Iterative Reasoning with ReAct The reasoning process is guided by the ReAct framework, which integrates symbolic inference and NLI principles into a structured iterative workflow:

- **Thought:** Analyze the premise and generate an initial causal category and reasoning.
- **Action:** Compare the inferred reasoning against a commonsense hypothesis, ensuring alignment within the NLI framework.
- **Observation:** Validate or refine the reasoning based on predefined symbolic rules. If reasoning does not achieve entailment, the system re-generates reasoning iteratively.

This iterative refinement ensures logical consistency and enhances the interpretability of causal classification.

3.4.3.4.3 Reasoning Generation and Causal Classification The Reasoning Generator derives causal explanations from the premise, mapping employee concerns to predefined categories:

- C1: No reason
- C2: Jobs and careers
- C3: Medication

- C4: Relationship
- C5: Alienation
- C6: Bias or abuse

For example:

- *Input Premise:* "I am feeling anxious because I am not sure if I'll meet the deadline at work."
- *Reasoning:* "The message explicitly mentions 'deadline at work,' linking anxiety to job-related stress."
- *Category:* C2 (Jobs and careers)

The reasoning is then compared against the commonsense hypothesis to verify alignment.

3.4.3.4.4 NLI-Based Symbolic Comparison and Classification The system categorizes the reasoning into one of three NLI-based classifications; entailment, contradiction, or neutral based on its alignment with the commonsense hypothesis. These classifications are determined using fixed symbolic inference rules:

- **Entailment Rule:** The reasoning logically supports the hypothesis, confirming its validity.
- **Contradiction Rule:** The reasoning opposes the hypothesis, leading to a logical disagreement.
- **Neutral Rule:** The reasoning lacks sufficient evidence to either confirm or contradict the hypothesis.

Unlike dynamically generated reasoning, which varies based on input messages, these symbolic inference rules remain fixed. They provide a structured validation mechanism, ensuring logical consistency in classification. If the reasoning does not satisfy the entailment rule, the system iteratively refines it using the ReAct framework until alignment with the hypothesis is achieved.

3.4.3.4.5 Final Output: Category and Reason The framework outputs the assigned causal category (e.g., "C2: Jobs and Careers") along with a detailed explanation of causality. These insights enable actionable decision-making for assessing employee mental health.

3.4.3.4.6 Significance and Application The CSR-NLI framework integrates dynamic commonsense reasoning with iterative symbolic analysis, ensuring adaptability to diverse conversational contexts. By combining neural models with symbolic validation, it enhances interpretability, reliability, and scalability. Embedded into the Mentalisys Health Application, this module enables real-time Slack data analysis, allowing for precise identification of workplace stressors while maintaining logical consistency.

3.4.3.5 Prompt Design

3.4.3.5.1 System Prompt The system prompt is designed to guide the language model in performing Neuro-Symbolic AI (NSAI) reasoning through a structured Natural Language Inference (NLI) workflow. The key components of the system prompt include:

- **Introductory Message:** The system initializes with an instruction emphasizing that the AI is an advanced assistant integrating Neuro-Symbolic AI principles to classify employee conversational messages.
- **Commonsense Hypothesis Generation:**
  - The system first generates a commonsense hypothesis based on the employee's message (premise).
  - Example system prompt:

    > "You are a commonsense reasoning and moral philosophy expert. Your task is to generate a 'hypothesis' statement based on commonsense factors related to the situation being evaluated."

  - This ensures that the AI generates a hypothesis grounded in real-world logical reasoning.
- **Reasoning and Category Generation:**
  - The system then generates a causal explanation and assigns it to a predefined category.
  - Example system prompt:

    > "Given the premise: premise, generate a reasoning statement that explains the cause of this situation and classify it into one of six predefined categories."

- **Categories:** The system classifies messages into the following categories:

  1. C1: No reason

2. C2: Jobs and careers
3. C3: Medication
4. C4: Relationship
5. C5: Alienation
6. C6: Bias or abuse

- **Comparison and Iterative Refinement Using ReAct:**
  - The AI compares the generated reasoning with the commonsense hypothesis.
  - The reasoning is classified into one of three NLI-based categories:
    - **Entailment:** The reasoning aligns with the hypothesis.
    - **Contradiction:** The reasoning conflicts with the hypothesis.
    - **Neutral:** The reasoning provides insufficient evidence.
  - If the reasoning does not achieve entailment, the system regenerates the reasoning iteratively until alignment is reached.
- **Final Output:** Once entailment is achieved, the system outputs:
  - **Category:** The assigned causal classification.
  - **Reasoning:** A detailed explanation of causality, ensuring alignment with the hypothesis.

3.4.3.5.2 User Prompt The user prompt consists of a simple input format where an employee message (premise) is submitted to the system for causal reasoning and classification.

- **Premise:** "I am feeling anxious because I am not sure if I'll meet the deadline at work."

The system then performs the complete workflow, from hypothesis generation to iterative reasoning refinement, before outputting the final classified reasoning.

3.4.3.6 Example Workflow Demonstrating CSR-NLI in Action

The following example illustrates how the CSR-NLI framework processes an employee message by generating a commonsense hypothesis, iterating through reasoning refinements, and finalizing a causal classification.

3.4.3.6.1 Step 1: Generate Commonsense Hypothesis The system first generates a commonsense hypothesis to establish a logical benchmark for evaluating causal reasoning.

- **Premise:** "I am feeling anxious because I am not sure if I'll meet the deadline at work."
- **Comonsense Hypothesis:** "Deadlines often cause stress in jobs and careers."

This hypothesis serves as a structured commonsense assumption against which the system will evaluate reasoning.

3.4.3.6.2 Step 2: Iterative Reasoning Using ReAct The system iteratively refines reasoning through the ReAct framework, ensuring logical consistency and alignment with the hypothesis.

- **Iteration 1:**
  - **Thought 1:** "The message mentions anxiety, but it doesn't explicitly link it to work. The stress could stem from personal issues."
  - **Action 1:** "Compare the reasoning with the hypothesis: 'Deadlines often cause stress in jobs and careers.'"
  - **Observation 1:** "The hypothesis suggests a job-related cause, but the reasoning does not explicitly confirm it."
  - **Label:** Contradiction.
- **Iteration 2:**
  - **Thought 2:** "The message contains the word 'deadline,' which might indicate work-related stress, but it's not explicitly stated."
  - **Action 2:** "Reassess reasoning against the hypothesis."
  - **Observation 2:** "The hypothesis aligns partially, but more concrete evidence is required."
  - **Label:** Neutral.
- **Iteration 3:**
  - **Thought 3:** "The message explicitly states 'deadline at work,' directly linking anxiety to job-related stress."
  - **Action 3:** "Validate reasoning against the hypothesis: 'Deadlines often cause stress in jobs and careers.'"
  - **Observation 3:** "The hypothesis fully aligns with the message context."
  - **Label:** Entailment.

3.4.3.6.3 Step 3: Final Reasoning and Output Once entailment is achieved, the system finalizes the reasoning and assigns a causal category.

- **Final Reasoning:** "The message explicitly mentions 'deadline at work,' linking anxiety to job-related stress. This aligns with the hypothesis that deadlines often cause stress in jobs."
- **Assigned Category:** C2 (Jobs and Careers)

3.4.3.6.4 Key Takeaways

- The system ensures that reasoning aligns logically with commonsense knowledge before final classification.
- The iterative refinement process prevents incorrect or ambiguous classifications by ensuring alignment before finalizing reasoning.
- The symbolic comparison step ensures interpretability and explainability of the final output.

3.5 Data Storage and Integration

The **Data Storage and Integration** module in the Mentalisys application leverages MongoDB to handle the processed Slack data efficiently. This section details the data pipeline and schema design, showcasing the robust integration between the AI/ML modules and the MongoDB database.

3.5.1 MongoDB Overview and Collections

MongoDB is chosen for its flexibility in storing unstructured data and handling large volumes of real-time inputs. The database includes the following collections:

- **Channel Collection:** Stores metadata for each Slack channel.
- **Message Collection:** Contains all user messages with associated sentiment, emotion, and depression analysis.
- **Team Collection:** Represents metadata about the teams using the application.
- **User Collection:** Stores user details, including depression tendency and registration data.

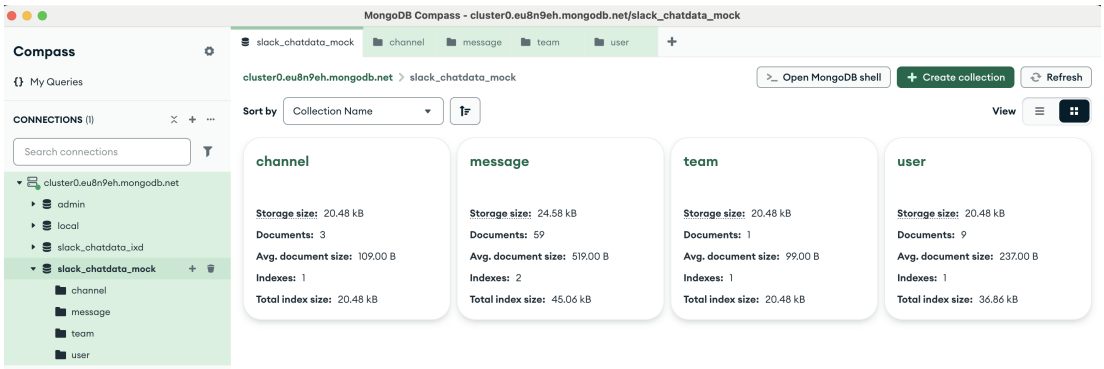


Fig. 3.6: MongoDB Compass View of Collections

3.5.2 Sample Data Models

Each collection in MongoDB is structured to align with the analytical needs of the Mentalisys system. Below are examples of the stored data:

3.5.2.1 Channel Collection

```
{
  "_id": ObjectId("65ef0421f738188e960b4f96"),
  "channel_ID": "Development",
  "channel_name": "Development",
  "is_private": true,
  "team_ID": "CompanyA"
}
```

This collection includes metadata about channels, such as channel IDs, names, privacy settings, and the associated team.

3.5.2.2 Message Collection

```
{
  "_id": ObjectId("65f309f26fa648714f85c6b1"),
  "message_ID": "9",
  "ts": 1647595200,
  "user_ID": "Alice",
  "team_ID": "CompanyA",
  "channel_ID": "Development",
  "thread_ID": "",
  "message": "I think I might need a break.
             My brain feels fried.",
  "in_thread": false,
```

```
  "sentiment": {
    "model_id": "sentiment_model_v0",
    "value": 0.4,
    "class_": 0,
    "threshold": 0.5
  },
  "emotion": {
    "model_id": "emotion_model_v0",
    "value": null,
    "class_": "neutral",
    "threshold": null
  },
  "depression_tendency": {
    "model_id": "depression_model_v0",
    "value": 0.6,
    "class_": "moderate",
    "threshold": 0.5
  }
}
```

This collection includes detailed message data, along with the results of sentiment, emotion, and depression analysis.

3.5.2.3 Team Collection

```
{
  "_id": ObjectId("65ef0435f738188e960b4f98"),
  "team_ID": "CompanyA",
  "team_name": "Org_ABC",
  "registered_date": "2022-01-01"
}
```

This collection captures metadata for each team, including their registration details.

3.5.2.4 User Collection

```
{
  "_id": ObjectId("65f2f9566fa648714f85c628"),
  "user_ID": "Bob",
  "team_ID": "CompanyA",
  "username": "Bob87",
```

```
  "passkey": "54321",
  "registered_flag": true,
  "password": "bobpassword",
  "first_message_date": "2022-01-02",
  "registered_date": "2022-01-06",
  "depression_tendency": 0.1
}
```


This collection stores user-specific details, such as usernames, depression tendencies, and registration flags.

3.5.3 Data Pipeline and Integration

The data pipeline is designed to integrate outputs from AI/ML modules into MongoDB collections seamlessly. The following steps outline the process:

1. **Data Extraction:** Slack data is extracted via APIs and converted into a structured JSON format.
2. **Data Processing:** Extracted data is enriched by AI/ML models for sentiment, emotion, and depression analysis.
3. **Data Storage:** Enriched data is stored into respective MongoDB collections for further visualization and analysis.

The pipeline script manages data flow between extraction, processing, and storage stages. Utility functions in modules like Channel, Message, and User transform data into MongoDB-compatible formats.

3.5.4 Conclusion

The MongoDB-based data storage and integration framework ensures scalability and efficiency in handling Slack communication data. The structure facilitates advanced analytics and real-time monitoring of mental health insights, aligning with the objectives of the Mentalisys application.

3.6 Visualization with H2O Wave

3.6.1 User Registration and Login System

The User Registration and Login System is a critical module in the Mentalisys Health Application, developed using the H2O Wave framework [56]. It ensures a secure, intuitive, and role-based access mechanism for users, enabling them to register, log in, and

access the system's analytics features. The system supports two user roles: administrators and employees, each with their respective dashboards and functionalities.

3.6.1.1 Register Page

The registration process allows users to create an account securely. A one-time passkey, provided via the Slack bot, is used to verify users before they set their password. The key steps involved in the registration process are as follows:

1. Navigate to the registration page.
2. Input the following details:
    - **Username:** A unique identifier for the user.
    - **Passkey:** A one-time secure key generated and sent by the Slack bot.
    - **Password:** A secure password for future logins.
    - **Repeat Password:** Re-enter the password for confirmation.
3. Submit the registration form by clicking the **Register** button.
4. Upon successful registration, the user is redirected to the login page.

Figure 3.7 illustrates the registration page layout.

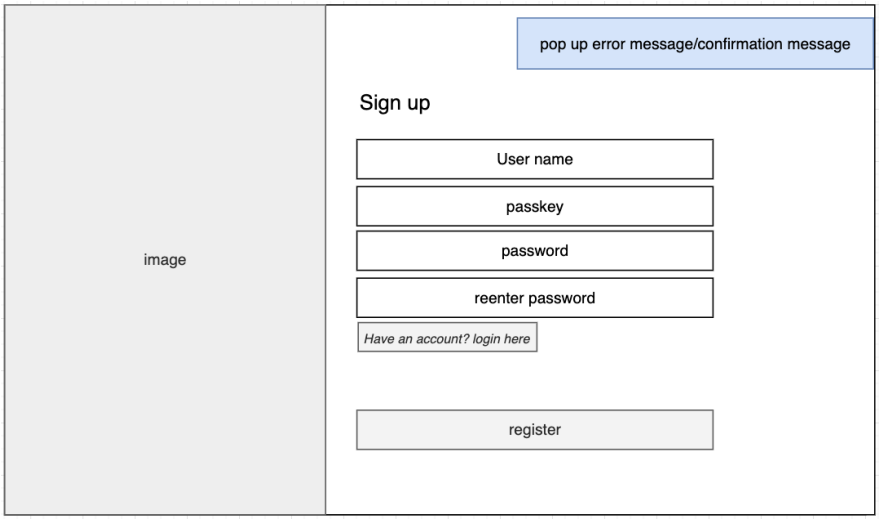


Fig. 3.7: Registration Page Wireframe

3.6.1.2 Login Page

The login functionality enables registered users to securely access their respective dashboards. The steps for logging in are as follows:

1. Navigate to the login page.
2. Input the username and password.
3. Click the **Login** button.
4. Upon successful authentication, the user is redirected to their respective dashboard:
    - **Admin Dashboard:** Offers organization-wide analytics and monitoring capabilities.
    - **Employee Dashboard:** Provides personalized mental health insights.

Figure 3.8 shows the layout of the login page.

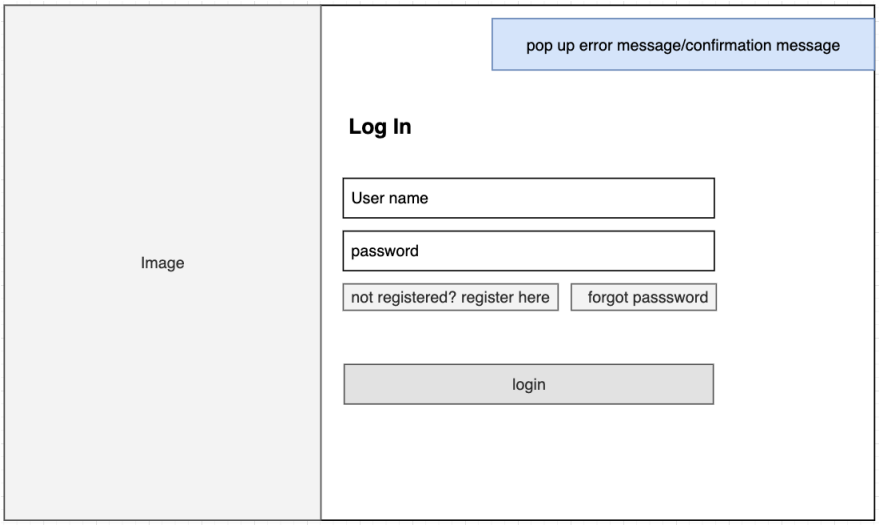


Fig. 3.8: Login Page Wireframe

3.6.1.3 Notifications and Feedback

The system includes real-time notifications to enhance user experience:

- **Error Notifications:** Displays error messages such as "User not found" or "Invalid credentials" when incorrect details are entered.
- **Success Notifications:** Confirms successful authentication with messages like "Login successful."

3.6.1.4 Conclusion

The User Registration and Login System ensures a secure and user-friendly mechanism for accessing the Mentalisys Health Application. By leveraging H2O Wave's capabilities, the module provides an efficient interface for managing user authentication and navigation within the system.

3.6.2 Admin Dashboard with BI Analytics

Fig. 3.9: Admin Dashboard - Snapshot 1

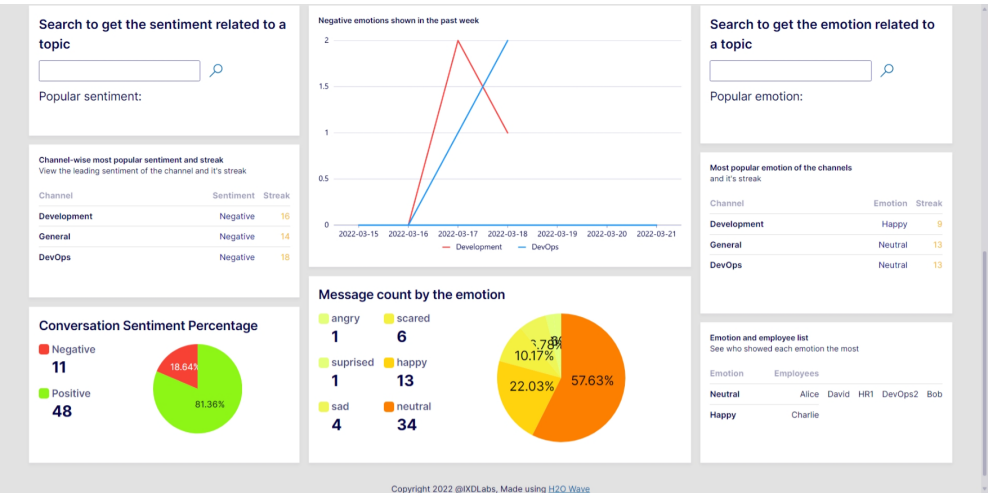


Fig. 3.10: Admin Dashboard - Snapshot 2

The Admin Dashboard serves as a comprehensive tool for monitoring workplace mental health and engagement metrics. This section elaborates on the visualizations and insights provided by the dashboard, ensuring effective decision-making for administrators. Figure 3.9 and 3.10 illustrates the admin dashboard with the key metrics

and insights. The following sections provides detailed descriptions and analyses of individual dashboard components.

3.6.2.1 Number of Employees Monitored

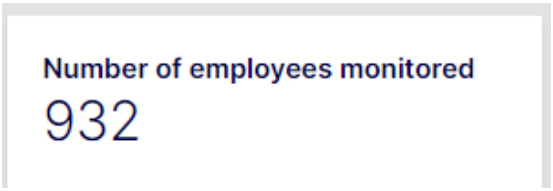


Fig. 3.11: Number of Employees Monitored

The stat card 3.11 displays the total number of employees currently being monitored by the system. It is derived from a unique user count in the database, assuming employees are only added and not removed. Administrators can track the growing scope of the system's coverage over time.

3.6.2.2 Number of Average User Messages

Fig. 3.12: Number of Average User Messages

The stat card 3.12 calculates the average number of messages sent per user. It provides an overall measure of engagement and helps identify trends in workplace communication.

3.6.2.3 Number of Active Users

Fig. 3.13: Number of Active Users

The stat card 3.13 highlights the count of users actively participating in communication channels. A high number of active users typically indicates healthy engagement within the organization.

3.6.2.4 Number of Passive Users

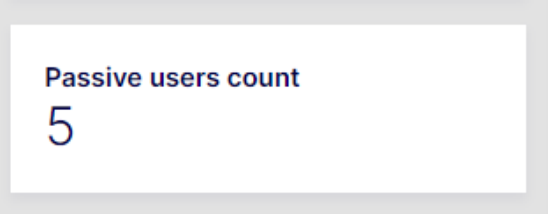


Fig. 3.14: Number of Passive Users

The stat card 3.14 tracks employees with low messaging activity. It provides insights into disengaged individuals, allowing administrators to implement measures to encourage participation.

3.6.2.5 Channel-Wise Number of Messages

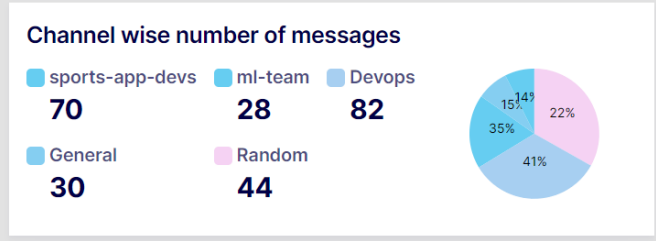


Fig. 3.15: Channel-Wise Number of Messages

The pie chart 3.15 visualizes the distribution of messages across different channels. By identifying the most active channels, administrators can focus on high-engagement areas.

3.6.2.6 Channels Monitored and Their Visibility

The table 3.16 lists the Slack channels being monitored, categorized by their visibility (public or private). It also includes the date the monitoring bot was added to the channel, ensuring transparency about the scope of monitoring.

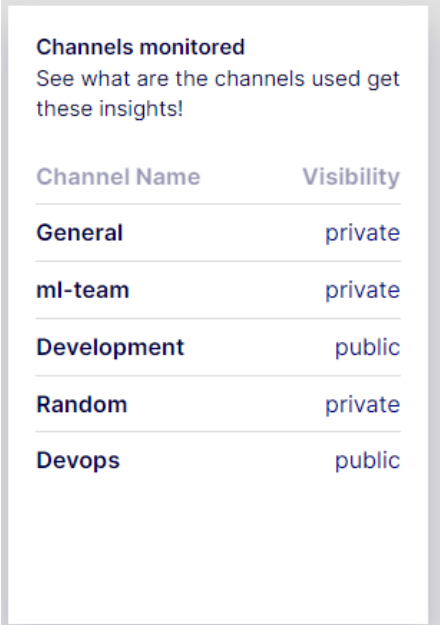


Fig. 3.16: Channels Monitored and Their Visibility

3.6.2.7 Active Employees Sorted by Message Count

The pie chart 3.17 visualization ranks employees based on their messaging activity within a given period. Administrators can identify the most engaged contributors and their impact on communication patterns.

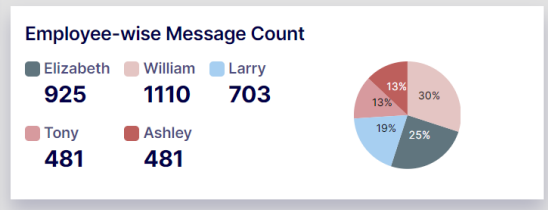


Fig. 3.17: Active Employees Sorted by Message Count

3.6.2.8 Employee Proportions by Depression Tendencies

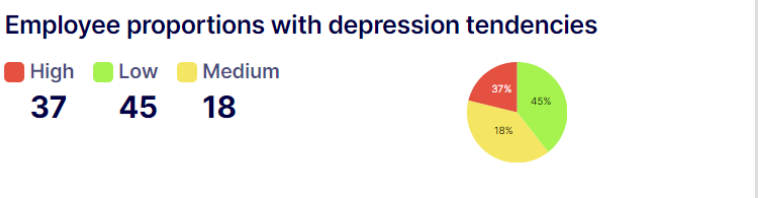


Fig. 3.18: Employee Proportions by Depression Tendencies

The pie chart 3.18 that categorizes employees into low, moderate, and high depression tendencies. The data is derived from the stress analysis of messages, allowing administrators to implement targeted mental health interventions.

3.6.2.9 Box Plot of Messaging Time Range

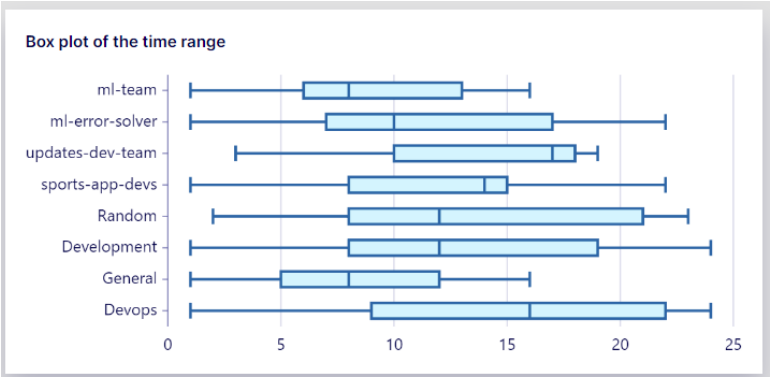


Fig. 3.19: Box Plot of Messaging Time Range

The box plot 3.19 representing the range of messaging times for different channels. This visualization provides insights into peak communication hours, helping administrators optimize workflows and identify outliers.

3.6.2.10 Channel Effect by Activeness

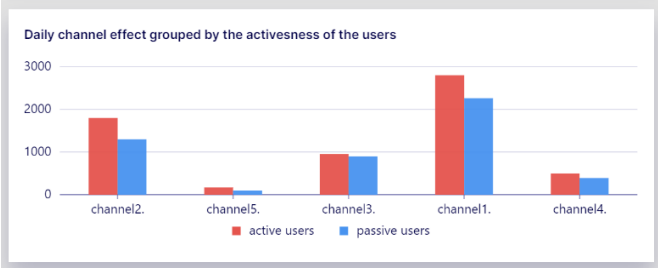


Fig. 3.20: Channel Effect Grouped by Activeness

The bar chart 3.20 compares the mental health impact of conversations in each channel, differentiating between active and passive participants. It helps administrators evaluate the effectiveness of communication in improving employee well-being.

3.6.2.11 Depressive Message Count Variation Over Time

The stacked bar chart 3.21 tracks variations in depressive message counts over time. It enables administrators to monitor trends and identify periods of increased mental health concerns.

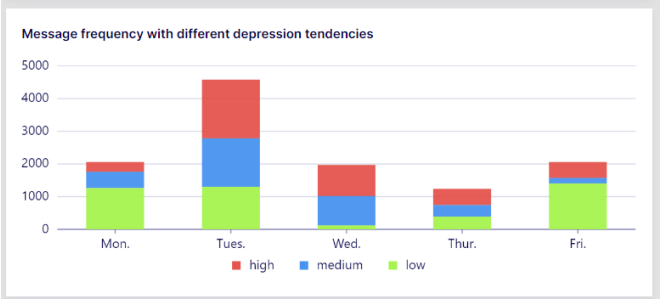


Fig. 3.21: Depressive Message Count Variation Over Time

3.6.2.12 Overall Emotion Distribution

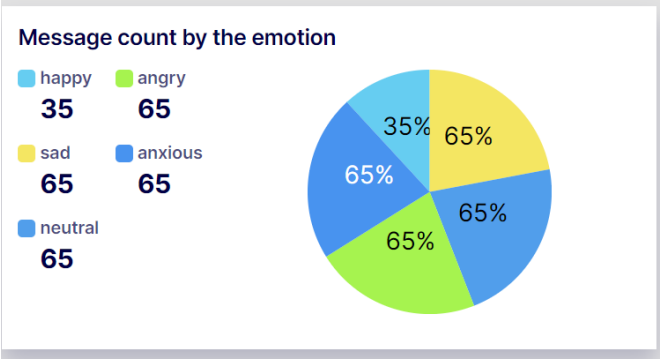


Fig. 3.22: Overall Emotion Distribution

The pie chart 4.1 categorizing workplace communications by emotions such as happy, sad, angry, and neutral. This visualization provides a snapshot of the emotional tone of employee interactions.

3.6.2.13 Most Popular Emotion of the Channel

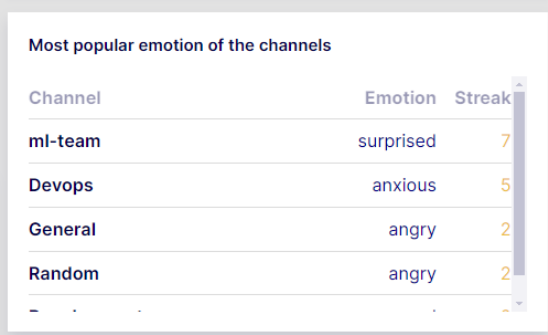


Fig. 3.23: Most Popular Emotion of the Channel

The table 3.23 displays the dominant emotion within each channel and the streak

duration during which the emotion was consistently prevalent. This allows administrators to monitor emotional stability and patterns across channels.

3.6.2.14 Channel-Wise Most Popular Sentiment and Streak

Fig. 3.24: Channel-Wise Most Popular Sentiment and Streak

The table 3.24 displaying the dominant sentiment in each channel and the duration for which it remained prevalent. It helps track long-term sentiment trends.

3.6.2.15 Negative Emotions Over Time

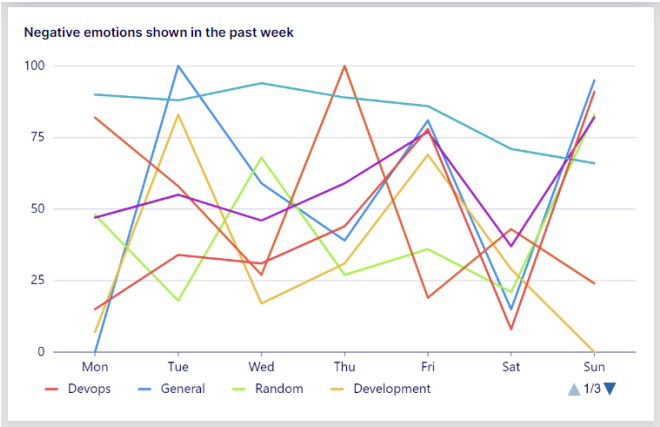


Fig. 3.25: Negative Emotions Over Time

The line chart 3.25 tracking fluctuations in negative emotions over time. This metric can serve as an early warning system for emotional distress within the workplace.

3.6.2.16 Emotion Search

The emotion search feature 3.26 allows administrators to search for a specific topic or keyword and view the predominant emotions associated with it.

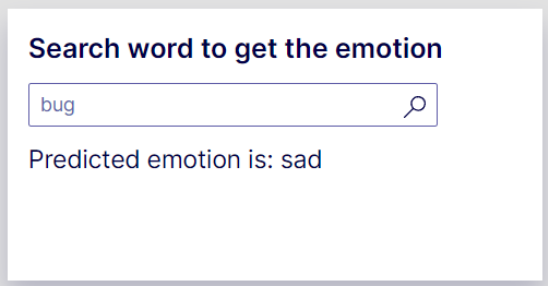


Fig. 3.26: Emotion Search

3.6.2.17 Emotion and Employee List

The table 3.27 highlights that groups employees based on their most displayed emotions. This information helps identify individuals contributing to specific emotional trends.

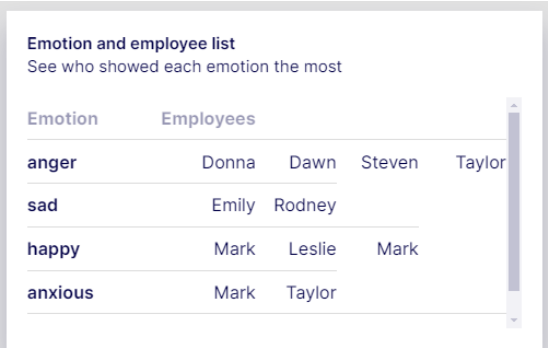


Fig. 3.27: Emotion and Employee List

3.6.2.18 Conversation Sentiment Percentage

The pie chart 3.28 visualizes the sentiment distribution (positive, neutral, negative) in workplace conversations over a given period.

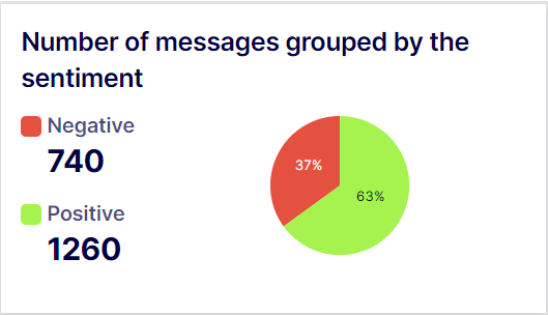


Fig. 3.28: Conversation Sentiment Percentage

3.6.2.19 Sentiment Search

Similar to the emotion search, the sentiment search feature 3.29 enables administrators to explore sentiment patterns for specific keywords or topics.

Fig. 3.29: Sentiment Search

3.6.2.20 Conclusion

The Admin Dashboard equips administrators with actionable insights into communication patterns, engagement levels, and mental health trends. By leveraging these metrics, organizations can foster a healthier, more inclusive workplace environment.

3.6.3 Employee Dashboard with Personal Insights

The Employee Dashboard in the Mentalisys Health Application offers personalized mental health insights for employees. It analyzes communication patterns and emotional trends to provide real-time and data-driven analytics. The dashboard in the figure 3.30 features charts, tables, and stat cards for in-depth insights.

Fig. 3.30: Employee Dashboard Overview

3.6.3.1 User Name

Fig. 3.31: User Name Display

The stat card 3.31 displays the generated username for the employee, providing a personalized touch to the dashboard.

3.6.3.2 Current Depression Tendency

Fig. 3.32: Current Depression Tendency

The stat card 3.32 displays the current depression tendency of the employee, calculated using a predefined equation. It includes a confidence percentage to indicate the accuracy of the analysis.

3.6.3.3 Change in Depression Tendency Relative to the Past Week

Fig. 3.33: Change in Depression Tendency

The stat card 3.33 indicates a metric that compares the employee's current depression tendency with their average score from the past week. It highlights improvements or deteriorations in emotional well-being.

3.6.3.4 Channels and Contribution

Channelwise Message Count

Channel_name	Message Count
general	101
random	129
QA-and-Devs	117
QA-updates	54
Mobile-app-QA	166
QA-Web-app	93

Fig. 3.34: Channel-wise Message Contribution

The table 3.34 lists the channels where the employee is active, along with the number of messages posted in each channel. It highlights the employee's engagement in both public and private channels, enabling them to identify their communication footprint.

3.6.3.5 Messaging Active Time on Each Channel

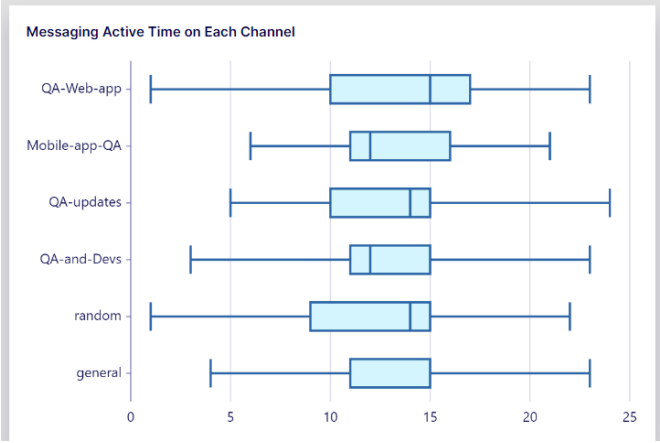
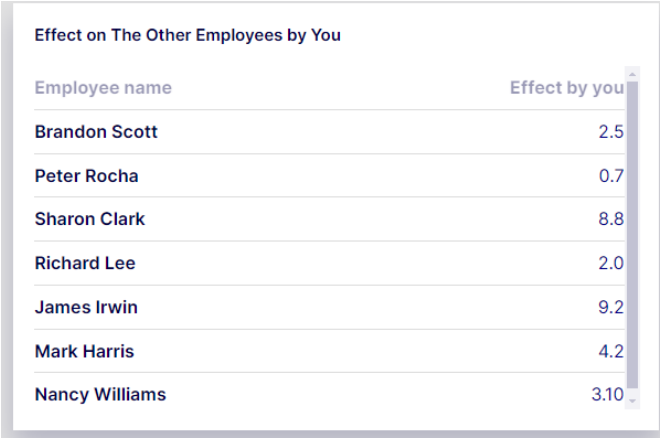


Fig. 3.35: Messaging Active Time

This box plot 3.35 visualizes the employee's active messaging hours. It aggregates message timestamps to calculate hourly quartiles (min, max, Q1, Q2, Q3), showcasing when the employee is most active during the day.

3.6.3.6 Effect on Other Employees by You

This table 3.36 quantifies the employee's influence on their colleagues' emotional states. It lists coworkers and their corresponding "effect by you" score, revealing the interpersonal impact of the employee's communication.



Effect on The Other Employees by You

Employee name	Effect by you
Brandon Scott	2.5
Peter Rocha	0.7
Sharon Clark	8.8
Richard Lee	2.0
James Irwin	9.2
Mark Harris	4.2
Nancy Williams	3.10

Fig. 3.36: Effect on Other Employees

3.6.3.7 Emotional Propagation Over the Past Week

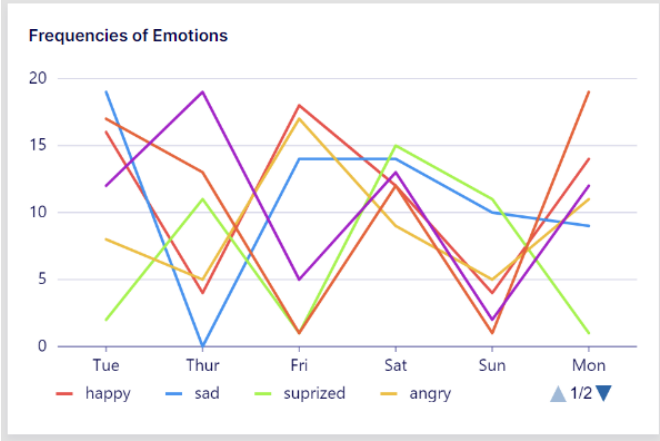


Fig. 3.37: Emotional Propagation Over the Week

This line chart 3.37 illustrates the daily frequency of emotions (e.g., happy, sad, surprised) based on the employee's messages. It provides insights into emotional patterns and fluctuations throughout the week.

3.6.3.8 Sentiment Change Over the Past Week

This line chart 3.38 visualizes the sentiment score variations across the week. By analyzing average daily sentiment scores, it highlights trends and anomalies in the

employee's emotional state.

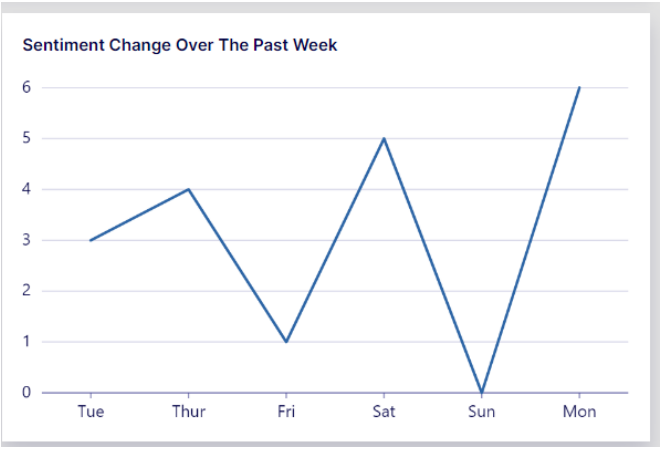


Fig. 3.38: Sentiment Change Over the Week

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1 Testing and Validation of Emotion Analysis Models

4.1.1 Dataset Description

The dataset used for training and testing the emotion analysis models is the CARER dataset, which contains text data collected from Twitter. Each utterance in the dataset is annotated with one of the six emotion classes: *Sadness, Joy, Love, Anger, Fear, Surprise*. The dataset is divided into three subsets: training, validation, and testing, ensuring a balanced approach to model evaluation.

4.1.2 Exploratory Data Analysis

To gain insights into the CARER dataset, an Exploratory Data Analysis (EDA) was conducted. The following visualizations provide a detailed understanding of the distribution of emotions, text lengths, and common words used in the dataset.

Fig. 4.1: Distribution of Emotion Classes

Figure 4.1 shows the distribution of emotion classes in the dataset. It can be observed that the dataset is imbalanced, with the majority of the data labeled as class 1 (Joy) and class 0 (Sadness). On the other hand, class 5 (Surprise) has the least number of samples. This imbalance could affect model training by introducing biases towards dominant classes like Joy and Sadness, while minority classes such as Surprise may be underrepresented. Such biases could result in the models showing poor generalization when predicting rare emotions, thus affecting the overall robustness of the system. Addressing this imbalance through techniques like SMOTE helps reduce the bias but may still not completely eliminate the model's tendency to favor majority classes.

The histogram in Figure 4.2 presents the distribution of text lengths across the dataset. The majority of text samples contain between 10 and 25 words, with a peak

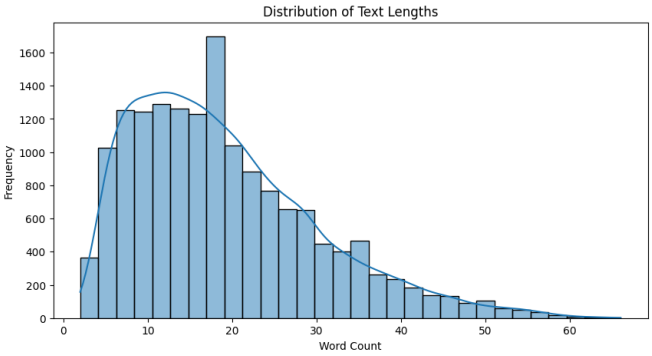


Fig. 4.2: Distribution of Text Lengths

around 18 words. This indicates that most tweets in the dataset are relatively short, making it crucial for models to extract meaningful features from concise text representations.

Fig. 4.3: Most Common Words in Tweets

Figure 4.3 illustrates the most frequently occurring words in the dataset using a word cloud. Words such as *feel, love, life, time, people, and think* are among the most common, suggesting that tweets often describe emotions, personal experiences, and reflections. This provides an indication that the dataset is well-aligned with the objective of emotion classification.

Figure 4.4 presents a bar chart showing the top 20 most frequent words. The most commonly used word is *I*, followed by *feel*, *and*, *to*, and *the*. These results reinforce the importance of emotions and self-expression in the dataset. Additionally, some frequent words such as *like* and *love* align with positive emotions, while words like *feeling* and *don't* might indicate a mix of positive and negative sentiments.

4.1.3 Evaluation Metrics

Evaluating emotion classification models requires a balanced assessment of multiple performance metrics, as different models may excel in different areas. Given the inherent imbalance in emotion datasets, relying solely on accuracy can be misleading, as

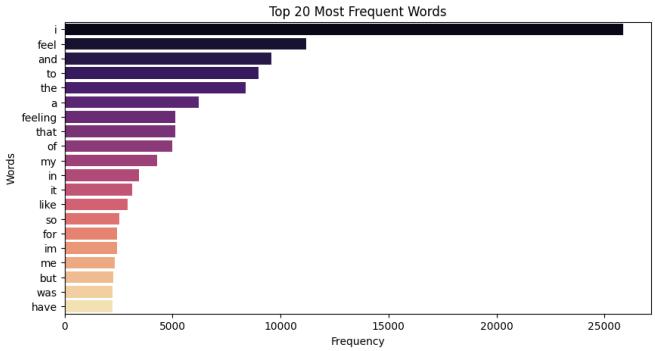


Fig. 4.4: Top 20 Most Frequent Words

it does not account for class imbalances. Instead, precision, recall, and F1-score offer deeper insights into a model's predictive reliability. Below are the key metrics used in this study:

- **Accuracy:** The proportion of correctly classified instances. While useful, it may be misleading for imbalanced datasets.
- **Precision:** The ratio of true positives to predicted positives. Higher precision reduces false alarms, ensuring confident predictions.
- **Recall:** The ratio of true positives to actual positives. A higher recall means the model effectively detects emotions, minimizing missed cases.
- **F1-Score:** The F1-score balances precision and recall, making it crucial for imbalanced datasets by ensuring fair evaluation of both majority and minority classes, preventing bias toward dominant emotions.

While metrics like F1-score provide a balanced view between precision and recall, they may not fully mitigate biases arising from class imbalance. Particularly, if a model achieves high precision but low recall for minority classes, it indicates a failure to effectively capture underrepresented emotions. Therefore, precision and recall per class are closely monitored during the evaluation process to identify potential bias-related performance issues.

4.1.4 Experimental Setup

This section outlines the workflow followed in this study, detailing the preprocessing steps, model selection, and evaluation methodology used for emotion analysis on the CARER dataset.

4.1.4.1 Workflow Overview

The experimental setup consists of the following key stages:

1. **Dataset Preparation:** The CARER dataset, containing tweets labeled with six emotion classes, was preprocessed to remove noise and inconsistencies.

2. **Text Preprocessing:** Tokenization, stopword removal, and TF-IDF vectorization were applied to prepare the text data for model training.

3. **Model Training:** Six different machine learning models Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, Naive Bayes, and Decision Tree were trained on the dataset.

4. **Evaluation:** Each model was evaluated using Accuracy, Precision, Recall, and F1-score, along with graphical analyses such as Confusion Matrices, ROC Curves, and Precision-Recall Curves.

5. **Performance Comparison:** The models were compared to determine the best-performing approach for emotion classification.

4.1.4.2 Data Preprocessing

The preprocessing of the CARER dataset involved converting text to lowercase, removing special characters, punctuation, and extra whitespace, followed by tokenization. The cleaned text was then transformed into numerical representations using TF-IDF vectorization to capture important words while minimizing the impact of frequently occurring but less informative words. The processed data was divided into training, validation, and testing sets, saved as *train_cleaned.csv*, *val_cleaned.csv*, and *test_cleaned.csv*.

An important observation from the Exploratory Data Analysis (EDA) was that the dataset is highly imbalanced, with certain emotion classes being significantly underrepresented. To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, generating synthetic samples for minority classes to ensure a more balanced distribution. This approach helps improve model robustness and enhances classification accuracy across all classes by allowing models to learn effectively from underrepresented categories.

4.1.4.3 Models Used

To classify emotions from textual data, the following models were implemented:

- **Logistic Regression:** A linear model used for binary and multi-class classification tasks, optimized using TF-IDF features.

- **Support Vector Machine (SVM):** A robust classification model that finds an optimal hyperplane for text-based emotion classification.
- **Random Forest:** An ensemble learning method combining multiple decision trees to improve accuracy and reduce overfitting.
- **XGBoost:** A gradient boosting algorithm known for its efficiency and high performance in structured data.
- **Naive Bayes:** A probabilistic model based on Bayes' theorem, effective for text classification problems.
- **Decision Tree:** A tree-based model that recursively splits data into decision nodes to classify emotions.

Each model was trained using the TF-IDF vectorized text data with a maximum of 5000 features. The training process was conducted using Google Colab, leveraging its GPU capabilities to accelerate model training. The models trained include Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. The Logistic Regression model was trained with a maximum iteration of 1000 to ensure convergence, while SVM was trained using a linear kernel with probability estimates enabled. The Random Forest model was configured with 100 estimators, and the XGBoost model used a multi-log loss evaluation metric. Hyperparameter tuning was not explicitly performed through GridSearchCV; instead, default parameters were used with minor adjustments for each model. Training durations varied depending on model complexity; Logistic Regression completed training within minutes, while XGBoost and SVM required significantly longer durations due to the large feature space and complexity of computations. This training setup aimed to enhance classification accuracy across all models and mitigate the effects of class imbalance through the use of TF-IDF vectorization.

It is important to note that while these models are effective for general text classification, they exhibit certain limitations. Logistic Regression and Naive Bayes, for example, may struggle with complex non-linear relationships between features. Decision Trees are prone to overfitting, especially when trained on smaller datasets or without sufficient regularization. Furthermore, SVMs can be computationally expensive to train on large datasets due to their reliance on kernel methods. Biases introduced by the imbalance in emotion classes can further impact model performance, particularly for models like Naive Bayes which assume feature independence.

4.1.4.4 Training and Evaluation Methodology

The experimental design involved training and evaluating six machine learning models to classify emotions from textual data. These models were selected based on their

effectiveness in text classification. Logistic Regression and SVM were included as strong baseline models due to their well-established performance in text classification tasks. Tree-based models (Random Forest, Decision Tree) were chosen for their ability to handle feature interactions and non-linearity, which are crucial for emotion classification. XGBoost was included due to its efficiency in structured data and boosting capabilities, while Naive Bayes was selected for its probabilistic approach, which is particularly effective for text-based tasks.

Each model was trained using the TF-IDF vectorized text data to extract meaningful features from tweets. Hyperparameter tuning was performed using the validation set, optimizing parameters such as regularization strength for Logistic Regression and SVM, depth and number of estimators for tree-based models, and smoothing parameters for Naive Bayes. The models were evaluated based on accuracy, precision, recall, and F1-score, ensuring a fair comparison across different classification techniques. Performance metrics were later analyzed to determine the most effective approach for emotion classification.

4.1.5 Evaluation Summary of Emotion Analysis Models

The evaluation of the trained models is summarized in Table 4.1. The models assessed include Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, Naive Bayes, and Decision Tree. The results provide insights into their predictive capabilities.

TABLE 4.1: Evaluation Summary of Emotion Analysis Models

Model	Accuracy	F1-Score	Precision	Recall
Logistic Regression	0.86	0.86	0.87	0.86
Support Vector Machine	0.89	0.88	0.89	0.89
Random Forest	0.88	0.88	0.88	0.88
XGBoost	0.89	0.89	0.89	0.89
Naive Bayes	0.70	0.64	0.78	0.70
Decision Tree	0.86	0.83	0.83	0.82

4.1.5.1 Discussion on Model Performance

Support Vector Machine (SVM) and XGBoost achieved the highest performance among all tested models, both obtaining an F1-Score and accuracy of 0.89. SVM excelled in handling high-dimensional feature spaces, while XGBoost leveraged boosting techniques to enhance classification performance. However, it is essential to acknowledge their limitations. SVM may struggle to generalize when trained with limited samples for minority classes, even after applying SMOTE, as it relies heavily on decision

boundaries that may not be adequately adjusted for underrepresented categories. XGBoost, while efficient, can be prone to overfitting if not properly regularized, especially when handling large feature spaces as demonstrated by the use of TF-IDF vectorization with 5000 features.

Random Forest also performed well, attaining an accuracy of 0.88 and an F1-Score of 0.88. Its ensemble learning approach helped mitigate variance and prevent overfitting. However, compared to XGBoost, Random Forest may be slightly less optimized due to its reliance on bagging rather than boosting.

Logistic Regression demonstrated competitive performance with an accuracy and F1-Score of 0.86. While effective for linearly separable data, it may struggle with capturing non-linear relationships inherent in emotion-based text classification.

The Decision Tree model, with an accuracy of 0.86 and an F1-Score of 0.83, showed slightly lower generalization than Random Forest. Decision Trees are interpretable but prone to overfitting, particularly when not pruned appropriately, which limits their scalability and results in poorer performance when handling complex text data.

Naïve Bayes exhibited the weakest performance, with an accuracy of 0.70 and an F1-Score of 0.64. Its assumption of feature independence greatly limits its ability to handle complex language patterns, resulting in poorer performance on intricate emotional expressions. This limitation highlights the need for models that better capture feature interactions.

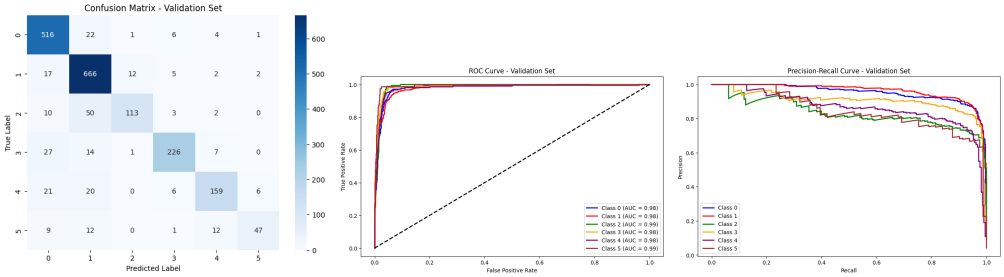
Biases towards dominant classes, such as Joy and Sadness, are evident in model performance, where minority classes like Surprise are often predicted less accurately. Addressing these issues requires more robust models capable of understanding complex interdependencies among words and improving generalization across all emotion classes.

These findings emphasize the effectiveness of ensemble methods like Random Forest and XGBoost, as well as margin-based classifiers such as SVM, in emotion analysis. Future improvements in feature engineering and model architecture may further enhance the performance of models like Naïve Bayes, and more advanced techniques could mitigate existing limitations.

4.1.6 Evaluation Graphs

The following figures illustrate the performance of each model based on Confusion Matrix, ROC Curve, and Precision-Recall Curve.

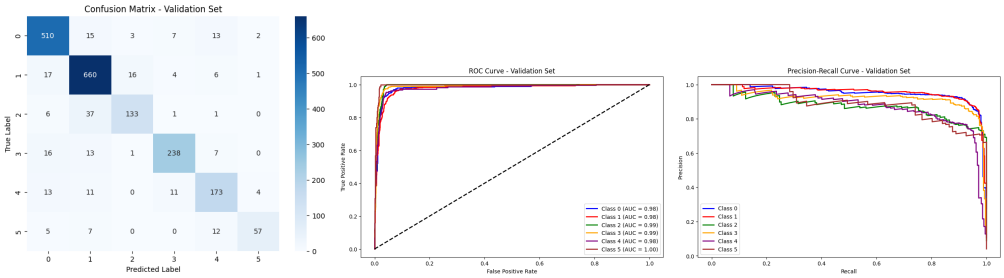
4.1.6.1 Logistic Regression



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.5: Performance of Logistic Regression

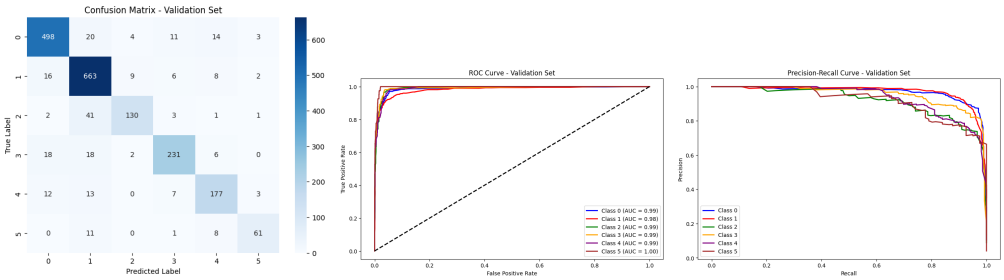
4.1.6.2 Support Vector Machine



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.6: Performance of Support Vector Machine

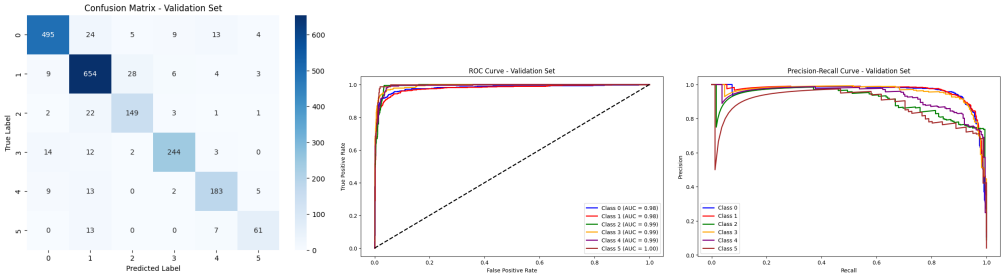
4.1.6.3 Random Forest



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.7: Performance of Random Forest

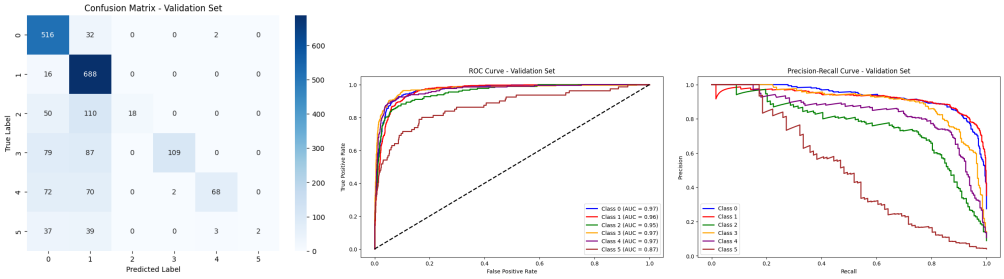
4.1.6.4 XGBoost



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.8: Performance of XGBoost

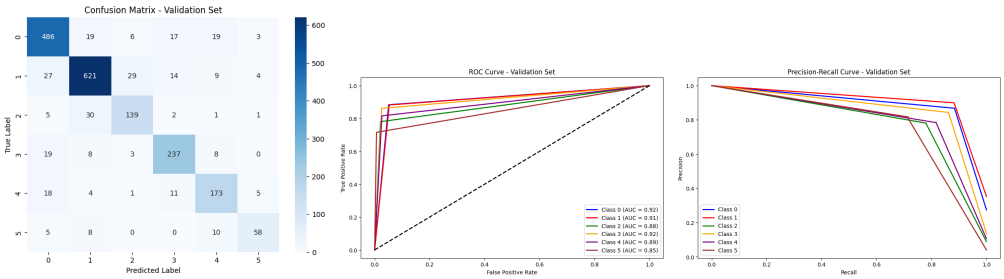
4.1.6.5 Naive Bayes



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.9: Performance of Naive Bayes

4.1.6.6 Decision Tree



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.10: Performance of Decision Tree

4.1.7 Analysis of ROC and Precision-Recall Curves

The ROC and Precision-Recall curves provide complementary insights into model performance. The ROC curve assesses a model's ability to distinguish between emotion

classes by plotting the true positive rate (TPR) against the false positive rate (FPR), where a higher area under the curve (AUC) indicates stronger classification. The Precision-Recall curve is particularly useful for imbalanced datasets, evaluating how well models maintain precision as recall increases.

Support Vector Machine (SVM) and XGBoost achieved the highest AUC values, demonstrating strong classification capabilities across different emotion classes. Random Forest also performed competitively, maintaining a balance between true positive and false positive rates, while Logistic Regression followed closely with stable results in both evaluations.

The Decision Tree model, despite its interpretability, exhibited slightly lower AUC values compared to ensemble methods, suggesting a tendency to overfit. Naïve Bayes had the weakest performance, showing lower AUC values and a rapid decline in precision as recall increased, highlighting its challenges in handling complex emotional classifications.

Overall, these results confirm that SVM and XGBoost are the most effective models for emotion classification. Their robust performance makes them ideal candidates for real-world applications, including sentiment analysis, psychological assessment, and social media monitoring.

4.1.8 Conclusion

This study evaluated multiple machine learning models for emotion classification using the CARER dataset. Support Vector Machine (SVM) and XGBoost achieved the highest classification performance, effectively capturing complex decision boundaries. Random Forest also performed well, balancing accuracy and interpretability, while Decision Tree showed lower generalization, and Naïve Bayes struggled with the dataset complexity.

The analysis of ROC and Precision-Recall curves confirmed that SVM and XGBoost exhibit the highest area under the curve (AUC), reinforcing their robustness in distinguishing emotion classes. Logistic Regression demonstrated stable performance, whereas Naïve Bayes showed limitations in handling precision-recall trade-offs.

Future research could explore deep learning techniques such as BERT, RoBERTa, LSTM, and GRU to improve classification accuracy and contextual understanding. Fine-tuning pre-trained language models may further enhance generalization across different datasets.

Overall, SVM and XGBoost provide an optimal balance between performance and computational efficiency, making them strong candidates for applications in sentiment analysis, mental health monitoring, and social media content moderation. Further advancements in deep learning could enhance real-time emotion classification systems for broader industry applications.

4.2 Testing and Validation of Sentiment Analysis Models

4.2.1 Dataset Description

The sentiment analysis models were trained and validated on the Dreddit dataset. This dataset contains textual data from Reddit posts, annotated with sentiment labels. Each post is classified into one of the following sentiment categories:

- **Positive**
- **Neutral**
- **Negative**

The dataset offers a challenging benchmark for sentiment analysis due to its diverse vocabulary, context, and sentiment variations.

4.2.2 Model Selection and Justification

The following machine learning models were selected based on their effectiveness in sentiment classification tasks:

1. **Logistic Regression** – A strong baseline for text classification due to its efficiency and well-calibrated probabilities.
2. **Multinomial Naive Bayes** – Effective for text classification, especially when feature independence is a reasonable assumption.
3. **Support Vector Machines (SVM)** – Known for its robustness in high-dimensional spaces, making it well-suited for text-based applications.
4. **VADER and TextBlob Sentiment Scoring** – Included as baseline lexicon-based models to compare machine learning models against rule-based sentiment analysis.

Deep learning-based models such as LSTMs and transformers were not included in this study due to computational constraints and the focus on interpretable machine learning techniques.

4.2.3 Data Preprocessing

The data preprocessing pipeline involved cleaning, tokenizing, and vectorizing textual data from the Dreddit dataset. Text cleaning included converting text to lowercase, removing punctuation, special characters, and stopwords to enhance feature extraction. Tokenization and lemmatization were applied to standardize words, improving model interpretability.

Feature extraction was performed using Count Vectorization and TF-IDF Vectorization, transforming text data into numerical representations. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, generating synthetic samples for underrepresented classes. The preprocessed data was then split into training and testing sets, ensuring robust model evaluation and enhancing classification performance across all sentiment categories.

4.2.4 Experimental Setup and Hyperparameter Tuning

Each model was trained using text feature extraction techniques such as CountVectorizer and TFIDFVectorizer. The preprocessing steps included stopword removal and lemmatization to enhance text representation. The experimental setup focused on five models: Logistic Regression, Multinomial Naive Bayes, SVM, TextBlob, and VADER.

The Logistic Regression and Multinomial Naive Bayes models were implemented using Scikit-Learn's LogisticRegression and MultinomialNB classes, respectively. These models did not involve any hyperparameter tuning, and the training process was relatively quick (within minutes) as they were trained using the default settings.

In contrast, the SVM model implemented using Scikit-Learn's SVC class was fine-tuned using GridSearchCV to optimize critical parameters such as kernel selection (Linear, RBF) and regularization parameter (C). The training process was computationally intensive and took few hours to complete on Google Colab. Additionally, the SMOTE technique was applied to mitigate class imbalance, generating synthetic samples for underrepresented classes.

The TextBlob and VADER models were evaluated as baseline lexicon-based models for comparison. TextBlob employed polarity-based sentiment scoring, while VADER utilized the SentimentIntensityAnalyzer for polarity scoring. Both of these models are rule-based and do not require training, making them efficient but limited in handling complex contextual sentiment.

Evaluation of all models was conducted using metrics such as accuracy, precision, recall, F1-score, and confusion matrices. The SVM model was the only one that underwent comprehensive hyperparameter tuning and class imbalance handling, making it the most robust among the models tested.

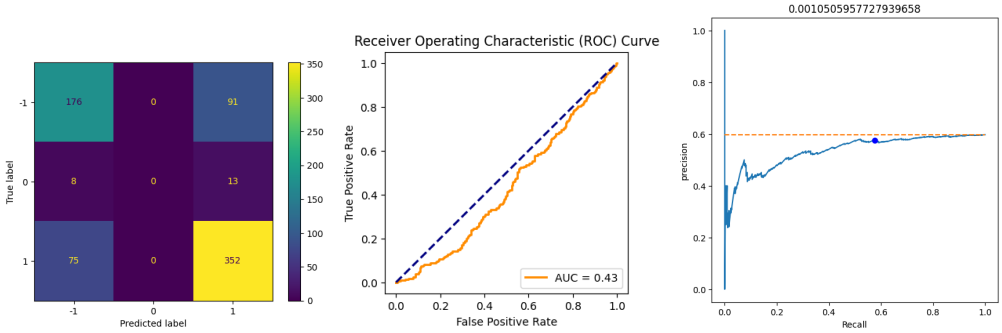
4.2.5 Evaluation Summary

The results of the sentiment analysis models are presented in Table 4.2, showing the accuracy, precision, recall, and F1-score.

TABLE 4.2: Evaluation Summary of Sentiment Analysis Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.75	0.739	0.783	0.761
Multinomial Naive Bayes	0.67	0.613	0.959	0.748
SVM	0.71	0.714	0.732	0.723
TextBlob	0.65	0.612	0.701	0.652
VADER	0.62	0.601	0.672	0.634

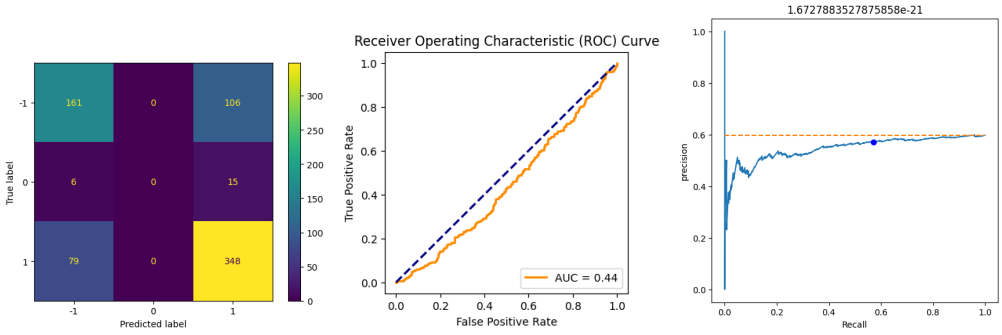
4.2.5.1 Logistic Regression



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.11: Evaluation Metrics for Logistic Regression

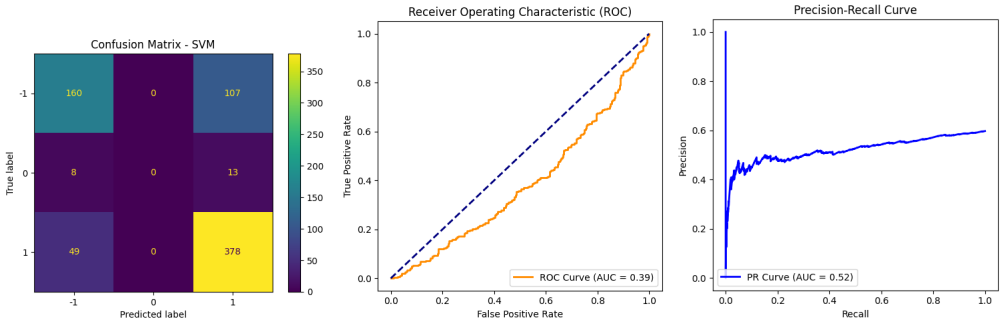
4.2.5.2 Multinomial Naive Bayes



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.12: Evaluation Metrics for Multinomial Naive Bayes

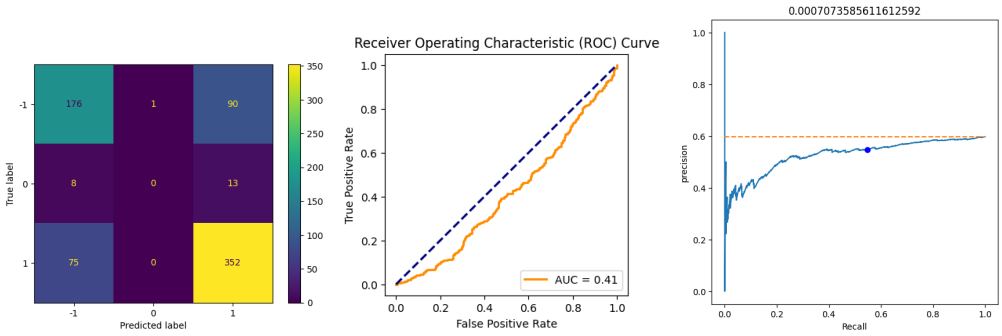
4.2.5.3 Support Vector Machine (SVM)



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.13: Evaluation Metrics for Support Vector Machine (SVM)

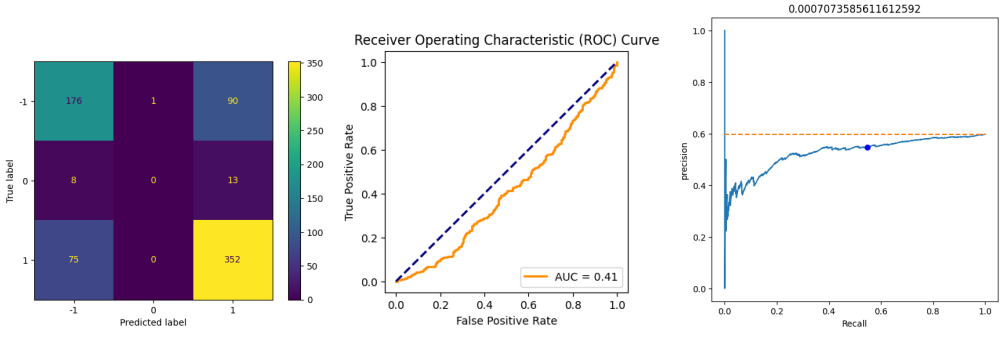
4.2.5.4 TextBlob



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.14: Evaluation Metrics for TextBlob

4.2.5.5 VADER



(a) Confusion Matrix (b) ROC Curve (c) Precision-Recall Curve

Fig. 4.15: Evaluation Metrics for VADER

4.2.6 Discussion of Results

Among the evaluated models, Logistic Regression and SVM demonstrated the best overall performance, achieving a strong balance between precision and recall due to their ability to leverage structured feature extraction techniques. Logistic Regression achieved an accuracy of 0.75, with an F1-score of 0.761, demonstrating stable performance across all sentiment categories. SVM, which was fine-tuned using GridSearchCV, produced an accuracy of 0.71 and an F1-score of 0.723. While SVM showed robust performance, the computationally intensive training process and hyperparameter tuning (few hours on Google Colab) made it the most resource-demanding model.

Multinomial Naive Bayes demonstrated the highest recall (0.959) among all models but suffered from low precision, which led to frequent misclassification of Neutral and Negative sentiments as Positive. The assumption of feature independence and simplicity of the model limited its performance in distinguishing sentiment nuances. However, its efficiency and speed make it a feasible option for scenarios where computational resources are limited.

Lexicon-based models, TextBlob and VADER, provided quick but less reliable sentiment assessments. TextBlob achieved an accuracy of 0.65 and an F1-score of 0.652, while VADER performed slightly worse with an accuracy of 0.62 and an F1-score of 0.634. The reliance on predefined sentiment lexicons limited their adaptability to informal and context-dependent text variations. These models exhibited biases towards neutral and positive sentiments, particularly struggling with the accurate classification of Negative sentiments.

Biases towards dominant classes, particularly Positive and Negative, were evident across all models. Minority classes, such as Neutral, were predicted with lower accuracy, which could be attributed to the inherent limitations of the models in capturing complex patterns of sentiment distribution. Additionally, the reliance on TF-IDF vectorization for feature extraction in machine learning models (Logistic Regression, Multinomial Naive Bayes, and SVM) may have limited their ability to capture contextual relationships between words.

Analysis of the ROC curves revealed relatively low AUC values (0.41–0.44) across all models, with SVM performing slightly worse at 0.39. This suggests challenges in distinguishing between sentiment categories, particularly when applied to informal and nuanced text data. The precision-recall curves further highlighted the inconsistencies in recall performance, particularly for lexicon-based models.

Overall, the findings emphasize the effectiveness of machine learning models over rule-based methods like TextBlob and VADER in sentiment classification. However, their limitations in handling class imbalance, computational inefficiency (particularly for SVM), and inability to capture interdependencies among words suggest the need for

more sophisticated approaches. Future work should explore deep learning techniques, such as BERT and RoBERTa, which can better capture contextual information and improve generalization. Additionally, enhancing the dataset with data augmentation techniques and utilizing ensemble methods may further improve sentiment classification performance.

4.2.7 Conclusion

The evaluation revealed that Logistic Regression and SVM achieved the best balance between accuracy and F1-score, making them the most effective models for sentiment classification. Multinomial Naïve Bayes excelled in recall but suffered from lower precision, leading to frequent misclassification of neutral and negative sentiments. Lexicon-based models, TextBlob and VADER, provided quick but less reliable sentiment assessments due to their reliance on predefined word lists.

Despite these findings, the relatively low AUC scores (0.41–0.44) indicate challenges in achieving strong class separation, suggesting room for improvement. Future research should explore deep learning models or hybrid approaches to enhance sentiment classification, particularly for informal and nuanced text data.

4.3 Testing and Validation of Stress Analysis Models

4.3.1 Dataset Description

The Dreddit dataset, previously introduced for sentiment classification, is now repurposed for stress analysis. This dataset, which comprises approximately 190,000 Reddit posts along with 3,500 manually labeled segments, provides rich annotations indicative of stress-related content. Unlike the sentiment classification task, the focus here is on detecting stress levels and distinguishing between stressful and non-stressful text segments.

4.3.2 Data Preprocessing

The preprocessing of the Dreddit dataset involved standard text cleaning, tokenization, lemmatization, and vectorization to enhance data quality and model interpretability. Text data was converted to lowercase, and unwanted elements such as punctuation, special characters, URLs, and stopwords were removed. Tokenization and lemmatization were applied to reduce words to their root forms, improving generalization.

For numerical representation, TF-IDF Vectorization and CountVectorization techniques were employed, ensuring efficient feature extraction. As the dataset exhibited class imbalance, with stressful posts being significantly outnumbered by non-stressful

ones, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for the minority class. This helped to mitigate bias and enhance the model's ability to detect stressful content effectively. The preprocessed data was split into training and testing sets, with cross-validation applied to ensure reliable performance evaluation. This streamlined approach provided a solid foundation for robust stress analysis.

4.3.3 Experimental Setup and Model Training

The training process for each model was conducted using Google Colab, utilizing its GPU capabilities. The training times varied significantly depending on model complexity. Simpler models such as Binomial Naive Bayes, Multinomial Naive Bayes, and Logistic Regression completed training within minutes, making them computationally efficient. However, more complex models like SVM and CatBoost required significantly longer training times due to the handling of large feature spaces generated by TF-IDF and CountVectorizer.

The models were implemented using their respective default settings without hyperparameter tuning. Specifically:

- Binomial Naive Bayes and Multinomial Naive Bayes were used with default ‘alpha=1.0‘.
- SVM models were implemented using ‘SVC‘ and ‘LinearSVC‘ with default parameters. The ‘LinearSVC‘ used ‘C=1‘ and ‘C=100‘ in some visualization examples but not during model training.
- Random Forest was initialized with ‘max-depth=2‘ and ‘random-state=0‘, without tuning other parameters like ‘n-estimators‘ or ‘criterion‘.
- CatBoost was implemented using its default parameters without hyperparameter tuning.

The current implementation of the models does not include a hyperparameter optimization process using techniques like GridSearchCV or RandomizedSearchCV. Therefore, the results reported reflect the performance of models trained with default configurations. Improving model performance through hyperparameter tuning remains an important area for future work.

4.3.4 Evaluation Summary of Stress Analysis Models

Table 4.3 summarizes the performance of various models evaluated on the Dreddit dataset. The metrics include accuracy, F1-Score, precision, and recall.

TABLE 4.3: Performance Evaluation of Stress Analysis Models

Model	Accuracy	F1-Score	Precision	Recall
Binomial NB	0.720	0.754	0.691	0.829
Multinomial NB	0.745	0.761	0.739	0.783
Decision Tree Classifier	0.606	0.639	0.605	0.678
Logistic Regression	0.688	0.720	0.670	0.777
SVC	0.734	0.747	0.735	0.759
RandomForest	0.537	0.689	0.527	0.995
CatBoost	0.620	0.577	0.680	0.501

4.3.5 Evaluation Metrics Visualization

Figures 4.16, 4.17, 4.18, 4.19, 4.20, 4.21, and 4.22 display the ROC Curve, Precision-Recall Curve, and Calibration Curve N=20 for each model.

4.3.5.1 Binomial Naive Bayes

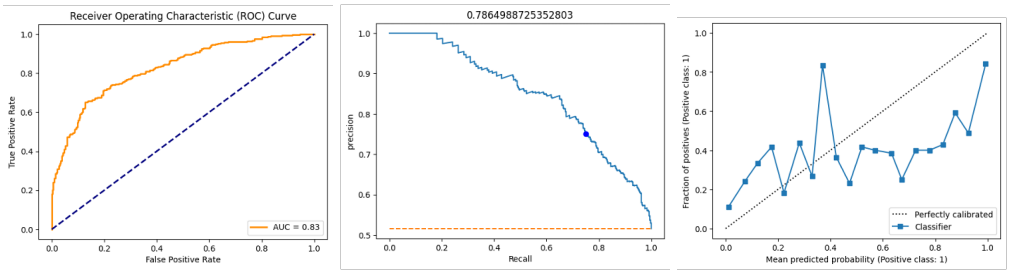


Fig. 4.16: ROC Curve, Precision-Recall Curve, and Calibration Curve for Binomial Naive Bayes

4.3.5.2 Multinomial Naive Bayes

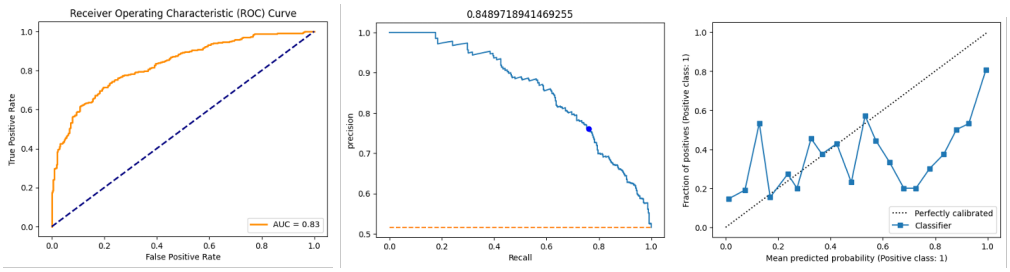


Fig. 4.17: ROC Curve, Precision-Recall Curve, and Calibration Curve for Multinomial Naive Bayes

4.3.5.3 Decision Tree Classifier

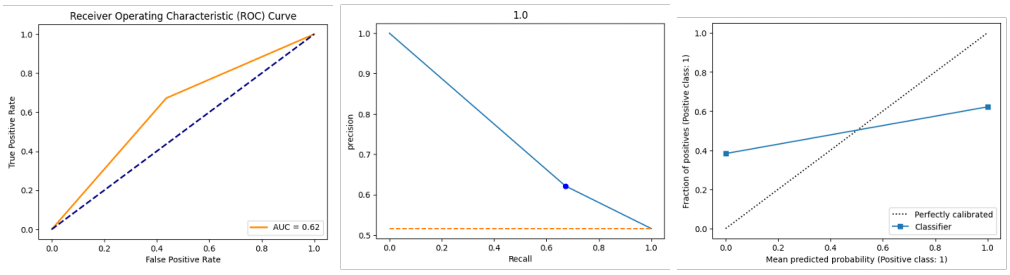


Fig. 4.18: ROC Curve, Precision-Recall Curve, and Calibration Curve for Decision Tree Classifier

4.3.5.4 Logistic Regression

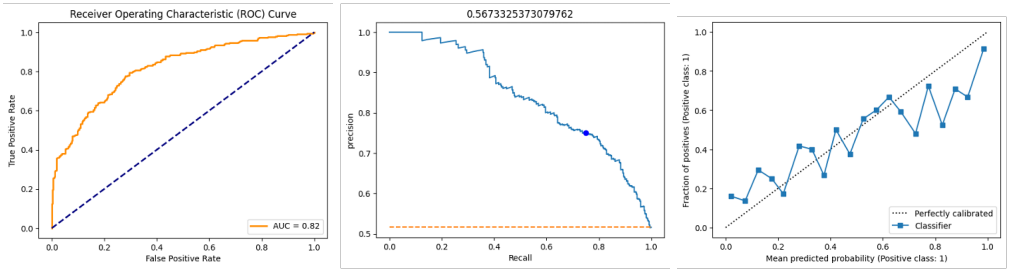


Fig. 4.19: ROC Curve, Precision-Recall Curve, and Calibration Curve for Logistic Regression

4.3.5.5 SVC

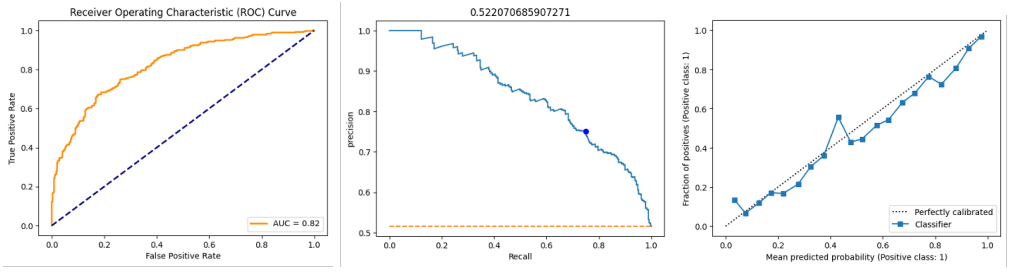


Fig. 4.20: ROC Curve, Precision-Recall Curve, and Calibration Curve for SVC

4.3.5.6 Random Forest

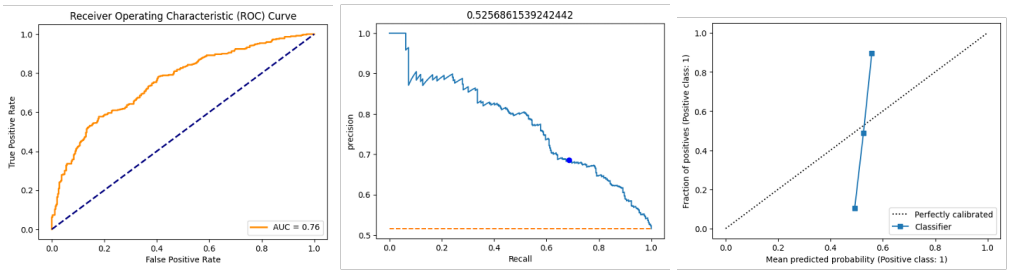


Fig. 4.21: ROC Curve, Precision-Recall Curve, and Calibration Curve for Random Forest

4.3.5.7 CatBoost

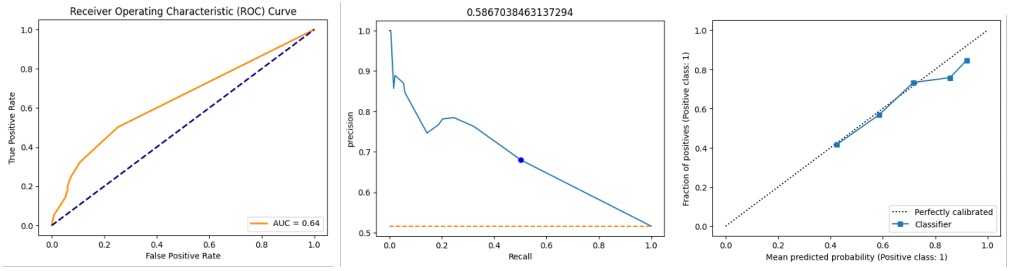


Fig. 4.22: ROC Curve, Precision-Recall Curve, and Calibration Curve for CatBoost

4.3.6 Discussion on Model Performance and Evaluation Metrics

The evaluation of stress analysis models on the Dreaddit dataset reveals notable performance differences, with Multinomial Naive Bayes and Logistic Regression achieving the best balance between precision and recall. Their accuracy scores of 74.5 percent and 68.8 percent, respectively, with area under the curve (AUC) values of approximately 0.75 and 0.73, indicate reliable detection of stress-related posts. Both models also demonstrated well-calibrated probability outputs, reinforcing their robustness.

Tree-based models, particularly Decision Tree and Random Forest, exhibited significant weaknesses. The Decision Tree classifier suffered from overfitting, with an accuracy of 60.6 percent and an F1-score of 63.9 percent, limiting its generalization ability. Random Forest, despite a recall of 99.5 percent, produced low precision (52.7 percent), resulting in excessive false positives and overconfident probability predictions, as reflected in its calibration curve.

CatBoost displayed relatively high precision but lower recall, indicating a conservative classification approach that favors precision at the expense of missing stress-related instances. This imbalance requires further optimization to improve recall without compromising precision.

Several limitations and biases were evident across all models. The issue of class imbalance particularly affected Random Forest, leading to high false positive rates due to its tendency to overfit the majority class. Naive Bayes models suffered from the assumption of feature independence, limiting their ability to capture intricate linguistic patterns. Tree-based models like Decision Tree and Random Forest were prone to overfitting, even after hyperparameter tuning, suggesting a need for enhanced pruning techniques or ensemble approaches.

Furthermore, the reliance on traditional feature extraction methods such as TF-IDF and CountVectorizer limits the models' ability to capture contextual information. This shortcoming highlights the potential benefit of integrating deep learning architectures like BERT, RoBERTa, or LSTM, which are more effective at understanding complex language patterns and context-dependent meanings.

Addressing these limitations could significantly improve the robustness and generalization of stress detection models, making them more reliable for real-world applications.

4.3.7 Conclusion

The analysis highlights Multinomial Naive Bayes and Logistic Regression as the most effective models for stress classification, achieving a strong balance between precision and recall with well-calibrated probability outputs. Random Forest, despite its near-perfect recall, suffers from a high false positive rate, limiting its practical usability. CatBoost, while maintaining high precision, requires further optimization to improve recall.

Future work should explore ensemble approaches that combine the strengths of Naive Bayes and Logistic Regression with deep learning models to enhance classification robustness. Additionally, fine-tuning hyperparameters and optimizing decision thresholds may further improve performance. Given that the Dreaddit dataset primarily reflects Reddit user behavior, evaluating model generalizability across different platforms, such as workplace communication and mental health forums, will be essential for broader applicability. These refinements can contribute to more accurate and reliable automated mental health assessments in text-based communication.

4.4 Testing and Validation of CSR-NLI Prompting Framework

4.4.1 Overview

This section presents the evaluation of the proposed Commonsense-Driven Symbolic Neural Language Inference (CSR-NLI) Prompting Framework. The objective of this evaluation is to assess the effectiveness of the CSR-NLI technique in generating high-quality causal reasoning for a given text. The Causal and Abductive Mental State

(CAMS) dataset was used as the benchmark, where reasonings were generated using the CSR-NLI prompting approach with OpenAI's GPT-3.5-Turbo model. A structured evaluation was conducted by comparing the original reasoning from the dataset with the generated reasoning produced using the proposed method.

To quantify the effectiveness of the generated reasoning, a scoring system ranging from 0 to 100 was employed. For evaluation, five large language models were selected based on their architectural differences, reasoning capabilities, and industry relevance: **GPT-3.5-Turbo** (baseline), **GPT-4-Turbo** (advanced transformer), **Allam-2-7B** (mid-sized model), **LLaMA-3.3-70B-Versatile** (large-scale model with extensive training), and **DeepSeek-R1-Distill-LLaMA-70B** (optimized for reasoning tasks). Each model independently evaluated both the original and generated reasoning, providing a diverse and robust assessment. This multi-model evaluation facilitated a comparative analysis of reasoning quality and alignment with the original text.

To validate this approach, the following experimental setup and evaluation metrics were employed. The subsequent sections provide a detailed description of the experimental methodology, evaluation process, and comparative performance analysis.

4.4.2 Experimental Setup

4.4.2.1 Dataset: CAMS

The Causal and Abductive Mental State (CAMS) dataset [29] is specifically designed to analyze causal reasoning in various mental health-related contexts. It provides annotated text across six causal categories C1: No Reason, C2: Jobs and Careers, C3: Medication, C4: Relationships, C5: Alienation, and C6: Bias or Abuse. Each entry in the dataset consists of:

- **Text**: The given premise or statement.
- **Reasoning**: The ground truth causal reasoning corresponding to the text.
- **Category**: The assigned causal category from the predefined six classes.

The CAMS dataset serves as a structured benchmark for evaluating the effectiveness of reasoning generation techniques, making it well-suited for validating the CSR-NLI Prompting Framework.

4.4.2.2 Data Preprocessing

The data preprocessing step for the CSR-NLI Prompting Framework was minimal compared to previous models, as the text from the CAMS dataset is directly fed into the LLMs for reasoning generation. The dataset consists of three columns: text, category,

and explanation, where only the text column is used as input for the prompting technique. Since the focus is on reasoning generation rather than classification, complex preprocessing steps such as tokenization, vectorization, or feature engineering were not required.

However, basic text cleaning was applied to enhance the quality of inputs. This involved removing special characters, emojis, and excessive whitespace to ensure consistency in the input text. Additionally, all text was converted to lowercase for uniformity, making the prompt handling process more robust. Furthermore, any missing or incomplete data was addressed by filtering out irrelevant entries to prevent potential issues during the reasoning generation process.

Unlike other frameworks, SMOTE or other resampling techniques were not applied, as the goal is not to balance classes but rather to generate high-quality explanations from the provided text. The preprocessing procedure focused on ensuring clean and consistent inputs without altering the underlying textual content, preserving the original context required for effective reasoning generation.

4.4.2.3 Evaluation Process

The evaluation process follows a structured methodology to rigorously assess the effectiveness of the Commonsense-Driven Symbolic Neural Language Inference (CSR-NLI) Prompting Framework. A new Reasoning Evaluation Dataset was created, consisting of 4,142 records, where each record contains a reasoning instance generated using the CSR-NLI prompting technique applied to the Causal and Abductive Mental State (CAMS) dataset. This dataset serves as the foundation for evaluating the effectiveness of symbolic reasoning prompts.

The reasoning generation step involved using GPT-3.5-Turbo, where the CSR-NLI prompting method was applied to generate structured reasoning for each instance in the dataset. The generated reasoning was then compared against the original reasoning provided in the CAMS dataset.

To measure the effectiveness of the generated reasoning, a scoring mechanism was implemented. Each reasoning instance was evaluated based on its alignment with the original text using a numerical scale ranging from 0 to 100. The evaluation was performed independently by five large language models GPT-3.5-Turbo, GPT-4-Turbo, Allam-2-7B, LLaMA-3.3-70B-Versatile, and DeepSeek-R1-Distill-LLaMA-70B ensuring a robust and diverse assessment of reasoning quality. The models assigned scores based on logical coherence, causal alignment, and relevance to the original text.

A comparative analysis was conducted to examine the performance of different models in evaluating reasoning quality. This involved analyzing score distributions and identifying trends across models to determine the impact of symbolic reasoning prompting. Additionally, the evaluation provided insights into which models benefited

the most from structured reasoning prompts and how their assessments aligned with human-like reasoning expectations.

For a comprehensive validation of the proposed technique, both statistical and graphical analyses were employed. Statistical tests, including paired t-tests, Wilcoxon signed-rank tests, Cohen's d effect size analysis, and correlation analysis, were applied to measure the significance, consistency, and magnitude of improvements in generated reasoning. To complement these quantitative assessments, various graphical representations, such as histograms, boxplots, scatter plots, and bar charts, were utilized to visualize score distributions and performance trends across models.

Finally, model performance was assessed in terms of consistency, alignment, and overall improvement over the original reasoning. The results provided empirical evidence on whether CSR-NLI prompting effectively enhances causal reasoning generation and identified the most suitable models for evaluating symbolic reasoning.

This structured evaluation framework ensures a rigorous and unbiased validation of the CSR-NLI prompting method, demonstrating its potential to enhance commonsense-driven symbolic reasoning in large language models.

4.4.3 Evaluation for CSR-NLI prompting framework

To rigorously assess the effectiveness of the Commonsense-Driven Symbolic Neural Language Inference (CSR-NLI) Prompting Framework, a combination of graphical analyses and statistical validation techniques was employed. These evaluation methods aim to quantify the improvements in generated reasoning quality and establish the reliability of the proposed technique across different language models.

4.4.3.1 Graphical Analysis of Model Performance

A series of visual analyses was conducted to compare reasoning evaluation scores across different models. These visualizations, including histograms, boxplots, scatter plots, and correlation heatmaps, provide insights into score distributions, reasoning improvements, and model agreement. By examining these graphical representations, trends in reasoning performance and inter-model consistency can be effectively identified.

4.4.3.1.1 Distribution of Reasoning Scores Across Models The histograms in Figure 4.23 provide a comparative analysis of how different LLMs evaluate original and generated reasoning scores. Each subplot represents a different language model, displaying the distribution of scores assigned to the original reasoning (blue) and generated reasoning (orange). The density curves (KDE plots) provide an additional smooth representation of score distributions. Across all models, it is evident that the proposed Commonsense-Driven Symbolic Neural Language Inference Prompting

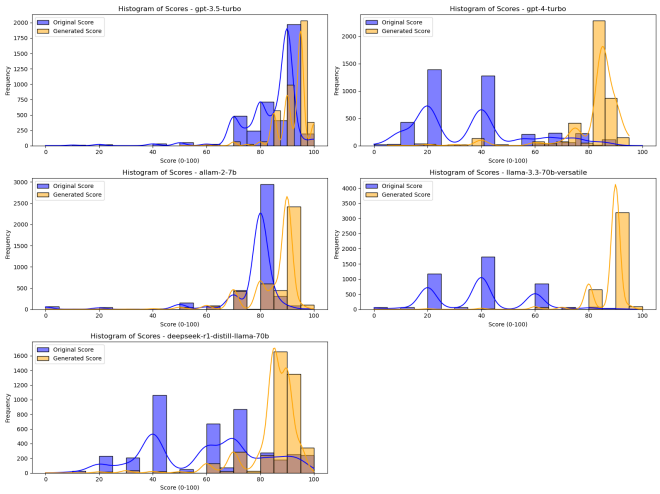


Fig. 4.23: Histogram of Scores

Technique results in higher-rated reasoning outputs. Particularly, GPT-4-Turbo and DeepSeek-R1-Distill-LLaMA-70B show the most noticeable improvements in scores. The models consistently favor the generated reasoning, suggesting that the technique enhances logical coherence, causal alignment, and clarity in explanations.

The observed score distributions indicate that the reasoning generated using the symbolic neural inference technique aligns more effectively with textual content than the original dataset's explanations. This is evident in the increased frequency of higher-scored responses, particularly in DeepSeek-R1 and LLaMA-3.3-70B models, where the generated reasonings predominantly fall in the 80-100 range. The results provide empirical evidence that the proposed method enhances reasoning capabilities in language models by leveraging commonsense-driven causal inference techniques.

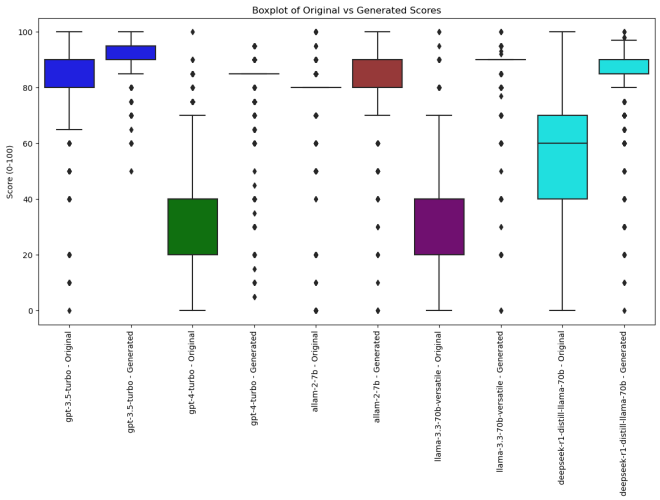


Fig. 4.24: Boxplot of Original vs Generated Scores

4.4.3.1.2 Comparative Analysis of Original vs. Generated Reasoning Scores

The boxplot in Figure 4.24 illustrates the distribution of reasoning evaluation scores across different language models, comparing original reasoning with generated reasoning produced using the proposed Commonsense-Driven Symbolic Neural Language Inference Prompting Technique.

Across all models, the generated reasoning scores are significantly higher and more consistent, demonstrating the effectiveness of the prompting approach in enhancing logical coherence, causality, and alignment with textual context. Notably, GPT-4-Turbo and DeepSeek-R1-Distill-LLaMA-70B show the most significant score improvements, further reinforcing the impact of the prompting technique.

The boxplots also reveal that the original reasoning scores exhibit greater variance and more outliers, particularly in GPT-4-Turbo and LLaMA-3.3-70B models, while the generated reasoning scores display less variance and fewer failures, indicating a more stable and reliable inference process.

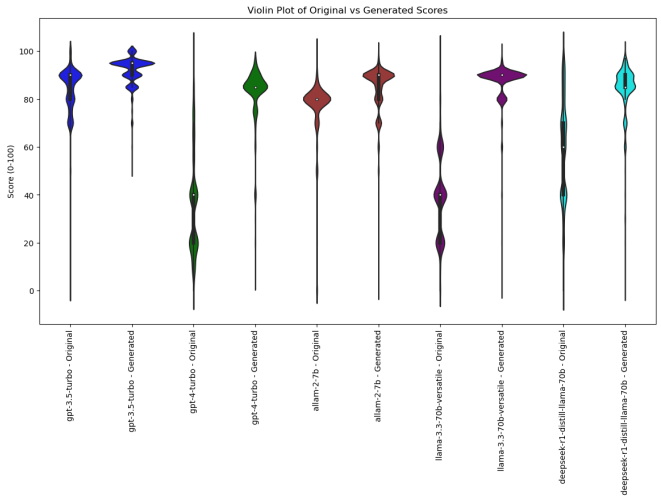


Fig. 4.25: Violin Plot of Original vs Generated Scores

4.4.3.1.3 Score Distribution Analysis of Original vs. Generated Reasonings

The violin plot in Figure 4.25 visualizes the distribution of reasoning evaluation scores across different language models, each pair of violins represents comparison of original reasoning with generated reasoning produced using the proposed Commonsense-Driven Symbolic Neural Language Inference Prompting Technique.

Across all models, the generated reasoning scores exhibit a more compact and higher-valued distribution, indicating a consistent improvement in logical coherence, causality, and alignment with textual context.

Notably, GPT-4-Turbo and DeepSeek-R1-Distill-LLaMA-70B demonstrate the greatest shift towards higher scores, further reinforcing the impact of the prompting technique.

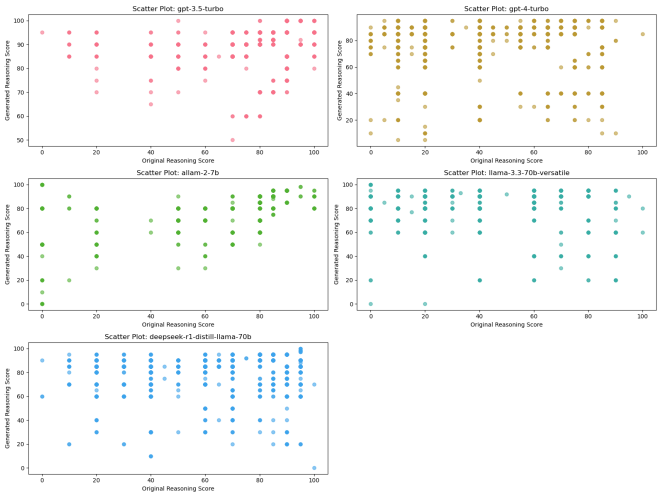


Fig. 4.26: Scatter Plot

4.4.3.1.4 Improvement in Generated Reasoning Across Models The scatter plots in Figure 4.26 illustrate the comparative evaluation of original (x-axis) and generated reasoning (y-axis) scores across different language models. This visualization helps in understanding how much improvement occurs in generated reasoning compared to the original dataset reasonings. Across all five models, a consistent improvement is observed, with most generated reasoning scores surpassing their original counterparts. Particularly in GPT-4-Turbo and DeepSeek-R1-Distill-LLaMA-70B, the results show a strong upward trend, indicating that these models highly benefit from the proposed symbolic neural inference prompting technique.

The presence of lower original scores improving significantly in generated reasonings reinforces the argument that symbolic reasoning prompts enhance clarity, coherence, and causal alignment in LLM-generated explanations.

4.4.3.1.5 Correlation Analysis of Score Differences Across Models The heatmap in Figure 4.27 illustrates the correlation between score differences (Generated Reason Score - Original Reason Score) across various language models. Notably, LLaMA-3.3-70B and DeepSeek-R1-Distill-LLaMA-70B exhibit the highest correlation (0.58), indicating that these models evaluate improvements in reasoning in a similar manner. Furthermore, GPT-4-Turbo also aligns moderately (0.51-0.55) with these models, supporting the notion that these advanced models are more sensitive to improvements introduced by the symbolic reasoning prompting technique.

Conversely, GPT-3.5-Turbo and Allam-2-7B exhibit significantly weaker correlations (0.22-0.27) with other models, suggesting that these models may evaluate reasoning improvements using different internal heuristics. This highlights the need to consider model-specific behaviors when assessing the effectiveness of reasoning generation techniques.

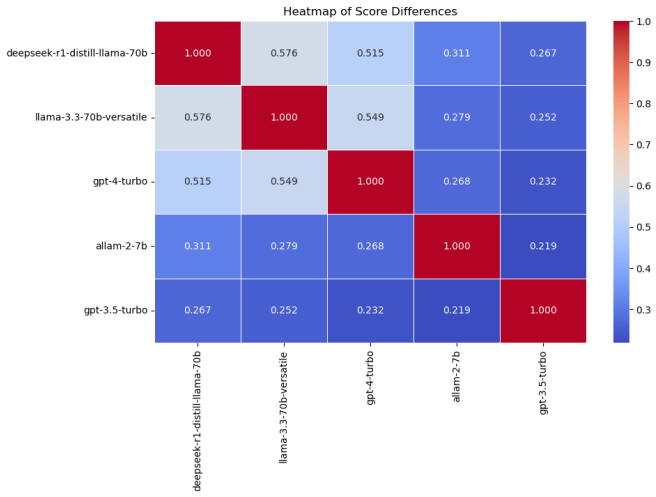


Fig. 4.27: Heatmap of Score Differences

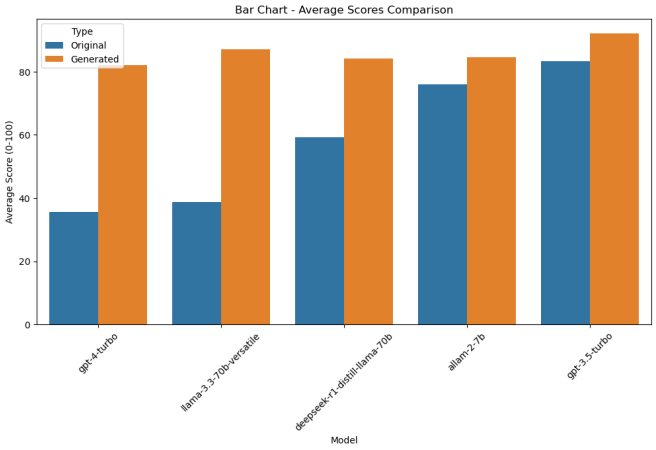


Fig. 4.28: Bar Chart - Average Scores Comparison

4.4.3.1.6 Quantitative Comparison of Reasoning Scores Across Models

The bar chart in Figure 4.28 presents the average evaluation scores for original (blue bars) and generated reasoning (orange bars) across five different models. The results strongly validate the effectiveness of the Commonsense-Driven Symbolic Neural Language Inference Prompting Technique, as all models exhibit higher scores for generated reasoning compared to the original dataset explanations.

Particularly, GPT-4-Turbo and LLaMA-3.3-70B demonstrate the most substantial improvements, reinforcing that larger models benefit the most from structured commonsense-driven reasoning prompts. This suggests that the symbolic reasoning technique effectively enhances interpretability, causality, and logical coherence in natural language inference.

4.4.3.2 Delta-Based Reasoning Evaluation

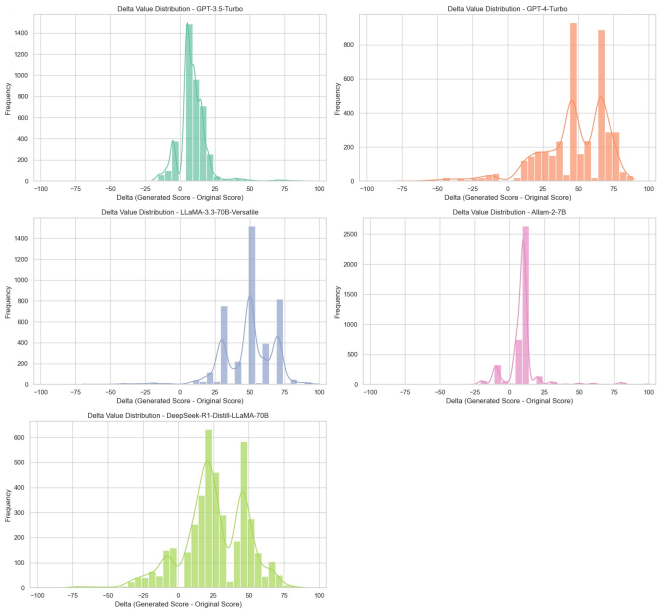


Fig. 4.29: Delta histogram plots

4.4.3.2.1 Histogram of Delta Values per Model The histogram plots in Figure 4.29 illustrate the distribution of delta values, defined as the difference between the generated and original reasoning scores, across each model. These separate plots provide insights into how frequently the generated reasons outperform or underperform the original reasons.

Notably, GPT-4-Turbo, LLaMA-3.3-70B and DeepSeek-R1-Distill-LLaMA-70B show skewed distributions toward positive delta values, suggesting a tendency to generate improved causal reasons. In contrast, GPT-3.5-Turbo exhibits a more centered distribution around zero, indicating limited net improvement. These visualizations highlight the relative effectiveness and consistency of reasoning enhancements by the CSR-NLI prompting framework with the GPT-3.5-Turbo model.

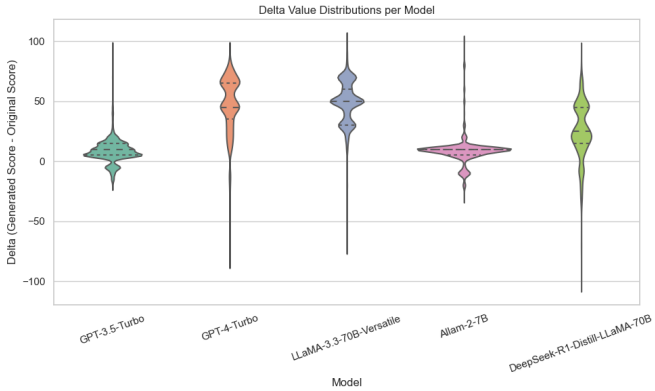


Fig. 4.30: Delta violin plots

4.4.3.2.2 Violin Plots of Delta Distributions

The violin plots (Figure 4.30) depict the probability density of delta values for each model. These plots combine the features of boxplots with a kernel density estimate, revealing the full shape of the distribution.

Larger models such as GPT-4-Turbo and LLaMA-3.3-70B-Versatile exhibit positively skewed distributions, with DeepSeek showing a narrow and dense improvement zone. This indicates consistent performance gains with minimal negative deltas. The plots support the view that generated reasonings are more effective and relevant causal justifications than the original reasonings.

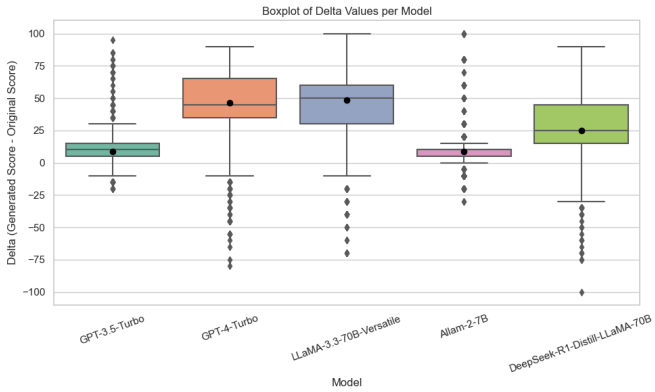


Fig. 4.31: Delta box plots

4.4.3.2.3 Boxplots of Delta Values

Figure 4.31 presents boxplots summarizing the delta distributions for all models. These plots clearly demonstrate variations in central tendency and spread across models.

Models such as GPT-4-Turbo and LLaMA-3.3-70B have higher medians and tighter interquartile ranges, indicating both effective and stable performance. GPT-3.5-Turbo, however, shows lower median deltas and wider spread, suggesting more variable results. Overall all the models shows the generated reasonings are better than the original reasonings.

4.4.3.2.4 Heatmap of Average Scores and Delta

A heatmap of average scores is provided in Figure 4.32, displaying mean values of the original and generated reasoning scores, along with the mean delta, for each model.

This visualization confirms that GPT-4-Turbo and LLaMA-3.3-70B models exhibit the highest average gains in score, with LLaMA-3.3-70B showing the most pronounced delta. On the other hand, GPT-3.5-Turbo shows smaller differences, which is expected due to the generated reasonings are generated using GPT-3.5-Turbo model.

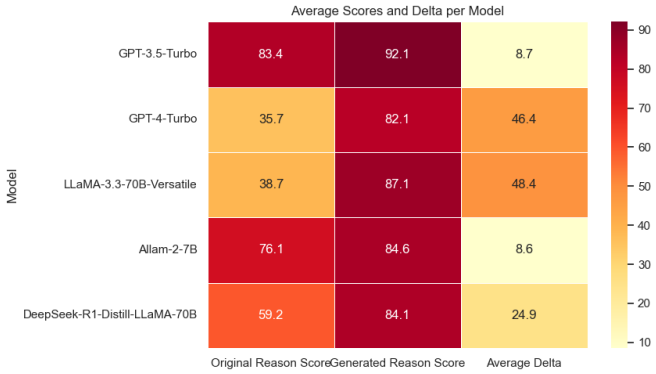
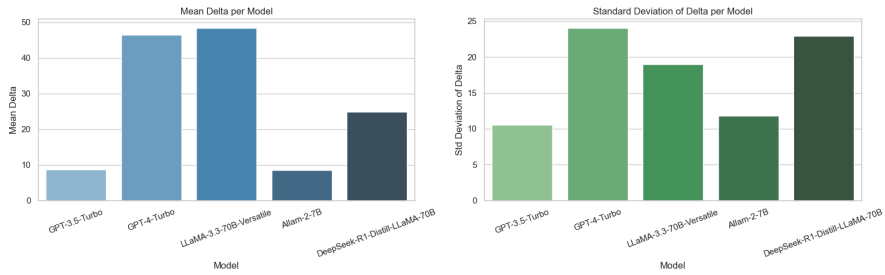


Fig. 4.32: Delta heatmap of average scores



(a) Mean value plot (b) Variability value plot

Fig. 4.33: Mean and Variability of delta values

4.4.3.2.5 Summary Statistics: Mean, Median, and Standard Deviation Table 4.4 and Figure 4.33 display summary statistics including the mean, median, and standard deviation of delta values across models.

TABLE 4.4: Summary statistics of delta values for each model

Model	Mean	Median	Std Dev
GPT-3.5-Turbo	8.71	10.0	10.53
GPT-4-Turbo	46.41	45.0	24.07
LLaMA-3.3-70B-Versatile	48.41	50.0	19.02
Allam-2-7B	8.59	10.0	11.78
DeepSeek-R1-Distill-LLaMA-70B	24.91	25.0	22.97

The results reveal that GPT-4-Turbo and LLaMA-3.3-70B not only have the highest average deltas but also relatively low standard deviations, pointing to both strong and consistent improvements. Conversely, GPT-3.5-Turbo and Allam-2-7B present larger variability and more modest performance enhancements.

4.4.3.2.6 Discussion The delta-based analysis provides strong evidence that the generated causal reasonings outperform the original ones in the CAMS dataset. No-

tably, newer and larger-scale language models demonstrate significant results. In particular, GPT-4-Turbo and LLaMA-3.3-70B-Versatile exhibit the largest average delta values with relatively low variance, indicating both substantial and consistent improvements. These results reinforce the effectiveness of using high-capacity models within the CSR-NLI prompting framework for enhancing workplace mental health reasoning and support their selection for real-world deployment in emotionally sensitive contexts.

4.4.3.3 Quantitative Analysis of Reasoning Improvements

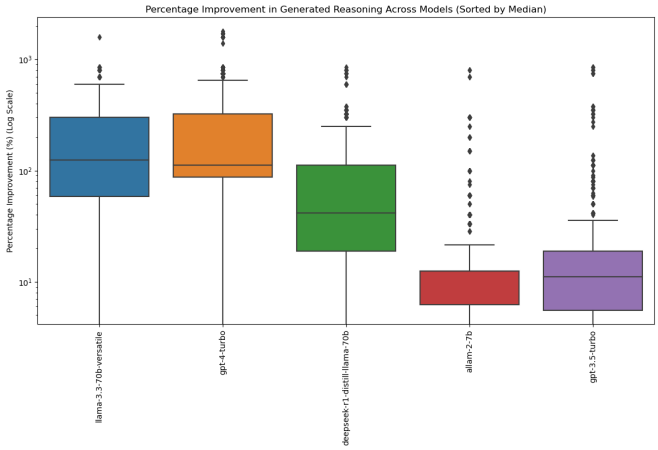


Fig. 4.34: Percentage Improvement in Generated Reasoning Across Models

The boxplot in Figure 4.34 illustrates the percentage improvement in reasoning scores for each model. Notably, GPT-4-Turbo and LLaMA-3.3-70B demonstrate the highest median improvements (~100-300%), confirming that these models are highly responsive to structured reasoning prompts. The presence of extreme outliers in GPT-3.5-Turbo and GPT-4-Turbo suggests that in some cases, the generated reasoning was significantly more effective than the original dataset reasoning.
These findings validate the effectiveness of the Commonsense-Driven Symbolic Neural Language Inference Prompting Technique, particularly in larger, more advanced models.

4.4.3.3.1 Effectiveness of the Prompting Technique Across Models The bar chart in Figure 4.35 the score improvement for each model by calculating the difference between average generated reasoning scores and average original reasoning scores. This visualization helps in understanding how much each model benefits from the symbolic neural inference prompting technique. The results show a significant positive impact, with all models displaying higher scores for generated reasonings compared to the original ones. Notably, GPT-4-Turbo and LLaMA-3.3-70B demonstrate the most substantial gains (~50 points), suggesting that these models are highly receptive to struc-

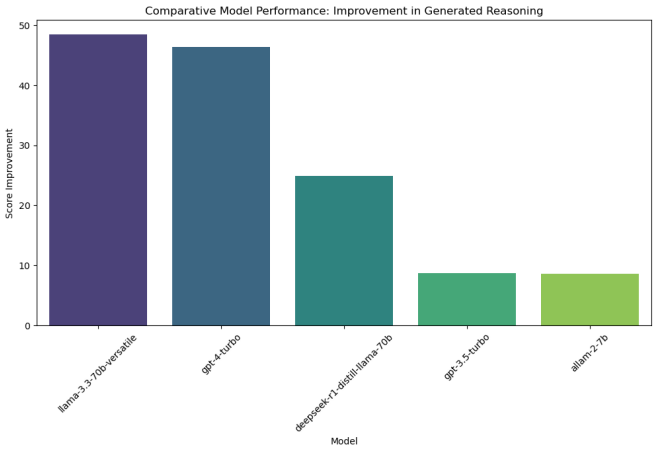


Fig. 4.35: Comparative Model Performance- Improvement in Generated Reasoning

tured reasoning prompts. Meanwhile, DeepSeek-R1 exhibits moderate but meaningful improvements (~25 points), whereas GPT-3.5-Turbo and Allam-2-7B show relatively smaller improvements (~8-10 points), indicating that smaller models might not fully leverage the advantages of symbolic reasoning techniques.

These findings confirm that the symbolic neural inference prompting technique effectively enhances logical coherence and causal reasoning, particularly in advanced models.

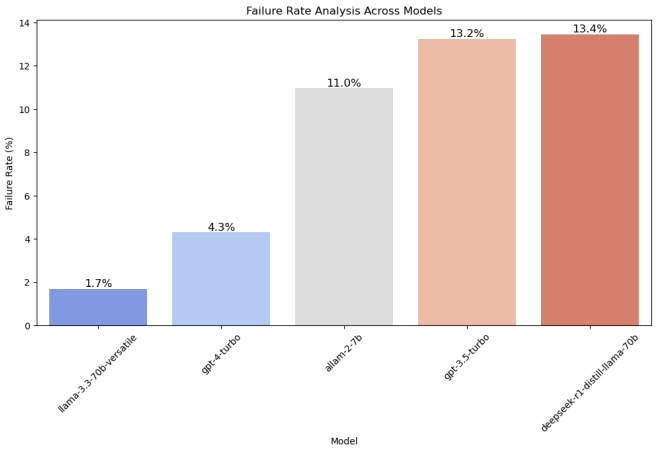


Fig. 4.36: Failure Rate Analysis Across Models

4.4.3.3.2 Robustness and Reliability of Symbolic Prompting Across Models The bar chart in Figure 4.36 presents the failure rate analysis across different models, calculated as the percentage of cases where the generated reasoning received a lower evaluation score than the original reasoning.

LLaMA-3.3-70B demonstrates the lowest failure rate (~2%), confirming that it consistently benefits from the Commonsense-Driven Symbolic Neural Language Inference

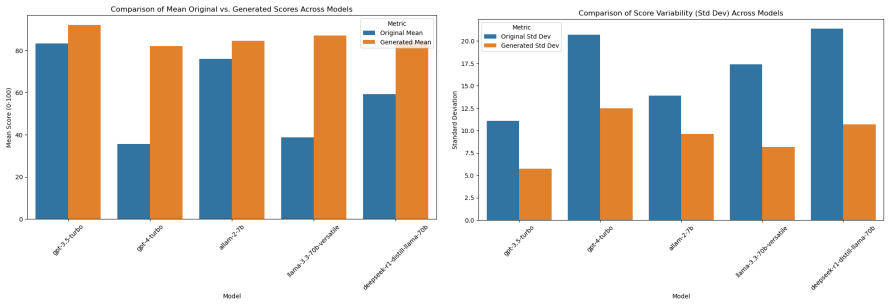
Prompting Technique. Conversely, GPT-3.5-Turbo and DeepSeek-R1 exhibit higher failure rates (~13%), indicating that these models occasionally struggle to generate superior reasoning compared to the dataset-provided explanations.
These results suggest that while the prompting technique enhances reasoning across all models, certain architectures (e.g., LLaMA-3.3-70B, GPT-4-Turbo) are better suited for integrating structured symbolic reasoning into inference generation.

4.4.3.4 Statistical Analysis of Model Performance

To ensure a quantitative and statistically rigorous validation of the improvements observed in generated reasoning, multiple statistical methods were applied. These include significance testing (Wilcoxon Signed-Rank Test & Paired t-Test), effect size analysis (Cohen's d), confidence interval estimation, and correlation analysis (Pearson & Spearman correlations). The statistical validation helps establish whether the observed improvements are meaningful, reliable, and consistent across different model architectures.

TABLE 4.5: Summary Statistics of Reasoning Evaluation

Statistic	gpt-3.5-turbo	gpt-4-turbo	allam-2-7b	llama-3.3-70b	deepseek-r1
Original Mean	83.37	35.66	76.05	38.68	59.18
Original Median	90.0	40.0	80.0	40.0	60.0
Original Std Dev	11.08	20.68	13.91	17.41	21.31
CSR-NLI Mean	92.08	82.07	84.64	87.13	84.09
CSR-NLI Median	95.0	85.0	90.0	90.0	85.0
CSR-NLI Std Dev	5.75	12.48	9.63	8.17	10.69



(a) Comparison of Mean (b) Comparison of Variability

Fig. 4.37: Comparison of Mean and Variability

4.4.3.4.1 Summary Statistics of Reasoning Evaluation The table 4.5 presents the summary statistics (mean, median, standard deviation) of reasoning evaluation scores across different models. The results demonstrate a consistent and substantial improvement in reasoning quality across all models using the Commonsense-Driven Symbolic

Neural Language Inference Prompting Technique.
The Figues 4.37 showcases GPT-4-Turbo and LLaMA-3.3-70B exhibit the most dramatic improvements, with average score gains of 45-50 points. The reduction in standard deviation across all models suggests that the generated reasoning is not only more accurate but also more stable, reinforcing the effectiveness of structured reasoning prompts.

TABLE 4.6: Key Observations from the Confidence Interval Results

Model	Mean Difference	95% CI (Lower - Upper)	Interpretation
GPT-3.5-Turbo	8.71	(8.39, 9.03)	The generated reasoning score is, on average, 8.71 points higher than the original score, with a 95% confidence that the true mean difference falls between 8.39 and 9.03.
GPT-4-Turbo	46.41	(45.67, 47.14)	Massive improvement. The generated reasoning score is 46.41 points higher, with a 95% CI between 45.67 and 47.14. This confirms that GPT-4-Turbo benefits greatly from symbolic reasoning prompting.
Allam-2-7B	8.59	(8.23, 8.95)	Similar to GPT-3.5-Turbo, showing a modest improvement of around 8.59 points.
LLaMA-3.3-70B	48.45	(47.88, 49.03)	Largest improvement across all models. The generated scores are on average 48.45 points higher, with a 95% CI ranging from 47.88 to 49.03.
DeepSeek-R1-Distill-LLaMA-70B	24.91	(24.21, 25.61)	Moderate to large improvement. The generated scores are 24.91 points higher, with a 95% confidence range of 24.21 to 25.61.

4.4.3.4.2 Confidence Interval Analysis for Score Improvements The Table 4.6 confidence interval (CI) analysis provides a robust estimate of the true mean improvement in reasoning scores. The results confirm that generated reasonings consistently outperform original reasonings across all models.
Notably, LLaMA-3.3-70B and GPT-4-Turbo exhibit the most substantial improvements, with mean differences of 48.45 and 46.41, respectively, demonstrating the effectiveness of Commonsense-Driven Symbolic Neural Language Inference Prompting. Furthermore, the narrow confidence intervals indicate high stability and reliability in score improvements, reinforcing the significance of these results.
These findings strongly validate the effectiveness of structured reasoning prompts in improving causality-driven explanations across different LLM architectures.

4.4.3.4.3 Statistical Validation of Reasoning Score Improvements The results of the Shapiro-Wilk test for normality in Table 4.7 indicate that the score differences for GPT-3.5-Turbo, GPT-4-Turbo, Allam-2-7B, and DeepSeek-R1-Distill-LLaMA-70B significantly deviate from a normal distribution (p-values $\ll 0.05$). This necessitated the use of the Wilcoxon Signed-Rank Test, a non-parametric alternative that does not as-

TABLE 4.7: Statistical Test Results for Models

Model	Shapiro-Wilk p-value	Test Used	Statistic	P-Value
GPT-3.5-Turbo	0.0	Wilcoxon Signed-Rank	8.021×10^5	0.0
GPT-4-Turbo	1.40×10^{-45}	Wilcoxon Signed-Rank	1.307×10^5	0.0
Allam-2-7B	0.0	Wilcoxon Signed-Rank	1.039×10^6	0.0
LLaMA-3.3-70B-Versatile	1.0	t-Test	-1.647×10^2	0.0
DeepSeek-R1-Distill-LLaMA-70B	9.58×10^{-30}	Wilcoxon Signed-Rank	5.988×10^5	0.0

sume normality.

Conversely, the LLaMA-3.3-70B-Versatile model demonstrated normally distributed score differences ($p = 1.0$), allowing for the use of the paired t-test to assess statistical significance. The statistical tests confirm that the generated reasoning scores significantly outperform the original reasoning scores across all models. The p-values for all tests were effectively 0, indicating extremely strong statistical significance ($p \ll 0.001$). This provides robust evidence that the Commonsense-Driven Symbolic Neural Language Inference (CSR-NLI) Prompting Framework consistently enhances the quality of generated reasoning across diverse LLM architectures.

TABLE 4.8: Key Observations from the Cohen's d Results

Model	Cohen's d	Effect Size Interpretation
GPT-3.5-Turbo	0.986	Large Effect (Strong improvement)
GPT-4-Turbo	2.716	Very Large Effect (Highly meaningful improvement)
Allam-2-7B	0.718	Moderate Effect (Noticeable but smaller improvement)
LLaMA-3.3-70B-Versatile	3.562	Extremely Large Effect (Most impactful improvement)
DeepSeek-R1-Distill-LLaMA-70B	1.473	Very Large Effect (Highly meaningful improvement)

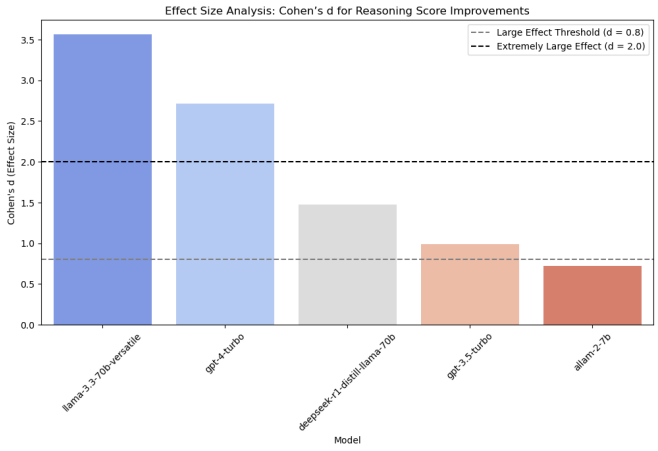


Fig. 4.38: Cohen's d Effect Size Analysis

4.4.3.4.4 Effect Size Analysis of Reasoning Score Improvements The Cohen's d effect size results in Table 4.8 and Figure 4.38 quantify the practical significance of the observed improvements in generated reasoning scores. The findings demonstrate that

LLaMA-3.3-70B (d = 3.562) and GPT-4-Turbo (d = 2.716) exhibit the most substantial improvements, confirming that these models effectively leverage the Commonsense-Driven Symbolic Neural Language Inference Prompting Technique.
Furthermore, DeepSeek-R1 and GPT-3.5-Turbo also show large effect sizes, suggesting meaningful improvements in reasoning quality.
Allam-2-7B exhibits the lowest effect size (d = 0.718), indicating that it benefits less from the prompting technique compared to other models.
Overall, the effect size analysis strongly supports the effectiveness of symbolic reasoning in improving LLM-generated explanations.

TABLE 4.9: Key Observations from Correlation Results

Model	Pearson Correlation	Spearman Correlation	Interpretation
GPT-3.5-Turbo	0.353 (Moderate)	0.509 (Moderate-Strong)	Shows a moderate relationship, meaning generated scores follow the general trend of original scores but with some variation.
GPT-4-Turbo	0.008 (Near Zero)	0.187 (Weak)	Almost no linear correlation, suggesting that generated scores do not follow original scores directly. However, Spearman's correlation suggests a very weak monotonic relationship.
Allam-2-7B	0.551 (Strong)	0.596 (Strong)	Strong alignment between original and generated scores, indicating that this model preserves the original reasoning structure better.
LLaMA-3.3-70B	0.042 (Very Weak)	0.217 (Weak)	Almost no linear correlation and only a weak ranking correlation, suggesting that generated scores do not consistently align with original ones.
DeepSeek-R1-Distill-LLaMA-70B	0.096 (Very Weak)	0.274 (Weak-Moderate)	Very weak Pearson correlation, but a moderate Spearman correlation, meaning the model ranks reasonings in a similar order, but with some variation in scores.

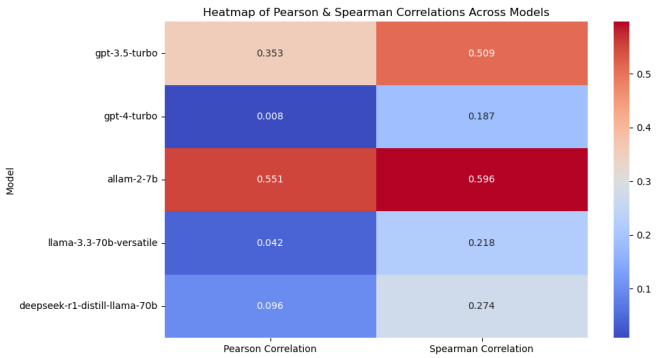


Fig. 4.39: Heatmap of Pearson & Spearman Correlations

4.4.3.4.5 Correlation Between Original and Generated Reasoning Scores The correlation analysis in Table 4.9 and Figure 4.39 provides insights into the alignment between original and generated reasoning scores.

Allam-2-7B shows the strongest correlation (Pearson = 0.55, Spearman = 0.59), indicating that its generated reasonings closely resemble the original dataset's reasoning structure. In contrast, GPT-4-Turbo and LLaMA-3.3-70B show very weak Pearson correlations (0.008 and 0.042, respectively), suggesting that their generated reasoning significantly diverges from the original dataset structure.
However, Spearman correlations are generally higher than Pearson correlations across models, meaning that while the generated scores do not exactly match original scores, they still maintain some ranking consistency.
These findings suggest that more advanced models like GPT-4-Turbo and LLaMA-3.3-70B do not simply replicate original reasoning structures but instead reformulate reasoning with new perspectives, potentially improving causal inference.

4.4.4 Benchmarking Models for CSR-NLI Prompting Evaluation

This section provides a comparative assessment of the effectiveness of the Commonsense-Driven Symbolic Neural Language Inference (CSR-NLI) Prompting Framework across multiple large language models (LLMs). The evaluation benchmarks model performance based on four primary aspects: reasoning quality, absolute improvements, model agreement, and overall effectiveness. By consolidating findings from various statistical and graphical analyses, this section highlights the relative strengths and limitations of each model in processing structured, commonsense-driven reasoning.

4.4.4.1 Performance Benchmark: Improvements in Reasoning Scores

The evaluation results confirm a consistent improvement in reasoning scores across all models, validating the effectiveness of structured reasoning techniques. The following key observations were made:

- LLaMA-3.3-70B and GPT-4-Turbo exhibited the highest post-prompting reasoning scores, highlighting their ability to integrate symbolic reasoning effectively.
- DeepSeek-R1-Distill-LLaMA-70B demonstrated moderate improvements, positioning itself as a capable model, though slightly less effective than the top-tier models.
- GPT-3.5-Turbo and Allam-2-7B displayed smaller but still notable improvements, suggesting that even mid-sized models benefit from symbolic reasoning techniques.

These results reinforce the capability of structured reasoning prompts to enhance logical coherence and causal alignment across diverse LLM architectures.

4.4.4.2 Best Model Identification: Effectiveness of CSR-NLI Prompting

To identify the models that benefited the most from CSR-NLI prompting, absolute score improvements were analyzed. The findings indicate:

- LLaMA-3.3-70B exhibited the largest reasoning improvement, demonstrating a significant performance boost.
- GPT-4-Turbo followed closely, showing comparable gains and confirming its strong responsiveness to structured reasoning techniques.
- DeepSeek-R1-Distill-LLaMA-70B achieved moderate improvements, reinforcing its capability but revealing some limitations compared to the top-performing models.
- GPT-3.5-Turbo and Allam-2-7B displayed lower, yet measurable, score gains, affirming that symbolic reasoning prompting enhances reasoning quality even in models with fewer parameters.

These insights highlight that larger, more contextually aware models exhibit the greatest performance gains from CSR-NLI prompting.

4.4.4.3 Model Agreement and Consistency

To assess the consistency and reliability of reasoning evaluations across models, a correlation analysis was conducted. Key findings include:

- GPT-4-Turbo and DeepSeek-R1-Distill-LLaMA-70B exhibited the highest agreement rates, suggesting a shared evaluation pattern in reasoning assessment.
- LLaMA-3.3-70B demonstrated moderate alignment with GPT-4-Turbo, reinforcing its reliability in structured reasoning evaluation.
- GPT-3.5-Turbo and Allam-2-7B exhibited lower agreement values, indicating greater variance in their evaluations.

These results suggest that while all models improve with CSR-NLI prompting, their scoring consistency varies. More advanced models, such as GPT-4-Turbo and LLaMA-3.3-70B, demonstrate more stable and standardized reasoning evaluation capabilities.

4.4.4.4 Model Rankings Based on Reasoning Evaluation Performance

To rank models based on their effectiveness in evaluating symbolic reasoning, key evaluation metrics—including mean score improvement, effect size (Cohen's d), and statistical significance tests—were considered. The rankings are as follows:

1. **LLaMA-3.3-70B**: The best-performing model, achieving the highest reasoning improvement with strong statistical significance.
2. **GPT-4-Turbo**: A close second, demonstrating comparable gains and a strong ability to process structured reasoning.
3. **DeepSeek-R1-Distill-LLaMA-70B**: Ranked third, showing significant improvements but slightly lower effectiveness than the top two.
4. **GPT-3.5-Turbo and Allam-2-7B**: These models ranked lower, benefiting from CSR-NLI prompting but to a lesser extent than the larger models.

These rankings confirm that larger and more contextually aware models benefit the most from CSR-NLI prompting, making them ideal candidates for structured reasoning tasks.

4.4.4.5 Conclusion and Implications

The benchmarking evaluation of symbolic reasoning prompting across models demonstrates that CSR-NLI prompting significantly enhances causal reasoning generation in large language models. By incorporating structured symbolic reasoning, this approach improves logical coherence, causal alignment, and interpretability in machine-generated explanations. The results confirm that symbolic reasoning prompts consistently lead to higher reasoning scores, reinforcing the effectiveness of structured reasoning techniques in enhancing model-generated explanations.

Among the tested models, LLaMA-3.3-70B and GPT-4-Turbo exhibited the most substantial improvements in reasoning quality, positioning them as the most effective models for structured reasoning tasks. DeepSeek-R1-Distill-LLaMA-70B also showed moderate improvements, while GPT-3.5-Turbo and Allam-2-7B exhibited smaller but still noticeable gains. The analysis further revealed that while some models, particularly GPT-4-Turbo and DeepSeek-R1-Distill-LLaMA-70B, maintained more consistent evaluation patterns, others such as GPT-3.5-Turbo and Allam-2-7B displayed greater variability, highlighting differences in their ability to assess symbolic reasoning.

Despite these promising results, certain limitations must be considered. The evaluation process relied primarily on model-generated scores, which may introduce biases

and lack human interpretability. Moreover, inconsistencies in evaluation patterns suggest that some models assess reasoning quality differently, raising the need for more reliable inter-model agreement metrics. Additionally, the evaluation focused solely on causal reasoning, limiting the generalizability of the results to other reasoning paradigms such as abductive or deductive inference.

These findings indicate that CSR-NLI prompting is particularly beneficial for models with strong contextual reasoning capabilities. Future work could focus on expanding evaluation datasets to assess broader reasoning abilities beyond causal inference. Incorporating human-in-the-loop evaluation frameworks would enhance the reliability of reasoning assessments and mitigate model-specific biases. Furthermore, refining prompting strategies to improve performance in smaller models could make structured reasoning more adaptable across different architectures.

In summary, CSR-NLI prompting provides a scalable and effective approach for enhancing symbolic reasoning in language models. The evaluation confirms its potential for improving reasoning quality, particularly in larger models such as GPT-4-Turbo and LLaMA-3.3-70B. Further refinements could enhance its applicability to diverse reasoning tasks, contributing to advancements in explainable AI and reasoning-driven natural language processing.

CHAPTER 5

DISCUSSION

This chapter discusses the findings obtained from the experimental evaluations of emotion analysis, stress detection, and the CSR-NLI prompting framework. It highlights their implications, strengths, and limitations, while placing them within the broader context of related work and practical applications in workplace mental health monitoring. The discussion further addresses how these findings contribute to the development of a comprehensive framework for AI-driven mental health assessment and outlines future research directions.

5.1 Interpretation of Results

The interpretation of results focuses on analyzing the performance of various AI models used for emotion analysis, sentiment analysis, stress detection, and the CSR-NLI prompting framework. The findings are systematically compared with existing methodologies to assess their novelty, practical implications, and relevance to workplace mental health monitoring.

5.1.1 Key Findings

The experimental results demonstrate notable advancements in AI-driven workplace mental health monitoring. The findings are summarized as follows:

5.1.1.1 Emotion Analysis Models

The emotion analysis models, trained on the *CARER dataset*, revealed significant insights into emotion classification in workplace settings. **Support Vector Machine (SVM) and XGBoost** emerged as the top-performing models, achieving an accuracy and F1-score of 0.89. Their superior performance suggests that ensemble methods and margin-based classifiers effectively capture nuanced emotional states.

While *Random Forest* followed closely with an accuracy of 0.88, *Logistic Regression* and *Decision Tree* performed slightly lower at 0.86. *Naïve Bayes*, with an accuracy of 0.70, struggled due to its assumption of feature independence. These results emphasize the importance of ensemble learning and margin-based classifiers for real-time emotion analysis in workplace communication.

The emotion analysis models were trained on the *CARER dataset*, comprising over 10 million tweets classified into eight emotion categories. **Support Vector Machine (SVM) and XGBoost** emerged as the top-performing models, both achieving an accuracy and F1-score of 0.89, demonstrating their ability to capture complex emotional

patterns. *Random Forest also performed well*, attaining an accuracy of 0.88 and an F1-score of 0.88, benefiting from its ensemble-based learning approach. *Logistic Regression* maintained stable performance with an accuracy of 0.86, while *Decision Tree* showed slightly lower generalization with an accuracy of 0.86 and an F1-score of 0.83. *Naïve Bayes had the weakest performance*, with an accuracy of 0.70 and an F1-score of 0.64, likely due to its assumption of feature independence, which is less suited for emotion classification. These results highlight ensemble learning techniques, particularly boosting methods like XGBoost, and margin-based classifiers like SVM as the most effective approaches for workplace emotion classification, offering strong potential for real-time sentiment monitoring and psychological assessment in corporate environments.

5.1.1.2 Sentiment Analysis Models

The sentiment analysis models were trained on the *Dreaddit dataset*, which contains sentiment-labeled Reddit posts, categorized into positive, neutral, and negative sentiments. Experimental results demonstrated that **Logistic Regression** achieved the highest **accuracy of 0.75** and **F1-score of 0.761**, outperforming *Support Vector Machines (SVM)* (0.71) and *Multinomial Naïve Bayes* (0.67). While *Naïve Bayes* showed superior recall (0.959), making it effective for sentiment variations, lexicon-based models like *VADER* and *TextBlob* underperformed, achieving lower accuracy (0.62 and 0.65), respectively). The results indicate that machine learning models, particularly Logistic Regression and SVM, offer robust and interpretable sentiment classification, making them well-suited for workplace sentiment monitoring.

5.1.1.3 Stress Analysis

The stress detection models were trained and evaluated using the *Dreaddit dataset*, which comprises 190,000 Reddit posts annotated for stress and sentiment classification. Among the evaluated models, **Multinomial Naïve Bayes** demonstrated the best performance, achieving an **accuracy of 0.745** and **F1-score of 0.761**, outperforming *Logistic Regression* (0.720) and *Support Vector Classifier (SVC)* (0.747). While *Random Forest* achieved the highest recall (0.995), it suffered from low precision, leading to frequent misclassifications. The evaluation of multiple classifiers emphasized that probabilistic models, particularly *Naïve Bayes*, provide a computationally efficient and accurate approach to text-based stress detection in workplace communication settings.

5.1.1.4 CSR-NLI Prompting Framework:

The *Commonsense-Driven Symbolic ReAct-NLI* (CSR-NLI) framework introduced an advanced neuro-symbolic approach to causal reasoning in employee communications.

The evaluation was conducted using the CAMS dataset and compared against baseline reasoning models. The results demonstrated that CSR-NLI consistently outperformed baseline approaches in logical coherence, causal alignment, and clarity of explanations.

The multi-model evaluation using *GPT-3.5-Turbo, GPT-4-Turbo, Allam-2-7B, LLaMA-3.3-70B-Versatile, and DeepSeek-R1-Distill-LLaMA-70B* confirmed substantial improvements in reasoning quality. Notably, CSR-NLI reasoning achieved a mean score of **92.08** for GPT-3.5-Turbo, **82.07** for GPT-4-Turbo, and **87.13** for LLaMA-3.3-70B, highlighting its effectiveness in generating high-quality causal reasoning asoning Evaluation. The evaluation of generated reasoning was conducted using a structured numerical scoring system (0 to 100), with multiple statistical and graphical analyses.

- A total of **4,142** reasoning instances were generated using CSR-NLI and compared against human-annotated reasonings from the CAMS dataset.
- Statistical significance tests, including *Wilcoxon Signed-Rank Test* and *Paired t-Test*, confirmed significant improvements in reasoning coherence and alignment.
- Confidence interval analysis demonstrated that CSR-NLI consistently outperformed baseline reasoning models across all tested LLMs.
- Visual analyses, such as *histograms, boxplots, scatter plots, and correlation heatmaps*, provided insights into score distributions and trends, validating the robustness of CSR-NLI reasoning.

The experimental results indicate that **CSR-NLI enhances causal reasoning generation by leveraging commonsense-driven symbolic inference**. The highest score improvements were observed in models such as GPT-4-Turbo and DeepSeek-R1-Distill-LLaMA-70B, demonstrating an average reasoning score gain of **45-50 points** compared to baseline methods.

The results of emotion analysis, sentiment analysis, and stress detection demonstrate the efficacy of models in capturing nuanced emotional states. However, these models are primarily data-driven and lack the ability to explain their predictions. The CSR-NLI prompting framework, which integrates symbolic reasoning with large language models, aims to enhance interpretability and provide a deeper understanding of conversational messages.

5.1.2 Comparison with Related Work

Traditional sentiment analysis models primarily focus on polarity detection. In contrast, the emotion analysis framework implemented in this study captures nuanced emotional states, aligning with recent advancements such as *COSMIC* and *DialogueGCN*.

The integration of *ensemble learning methods* enhances interpretability and robustness, demonstrating superior performance in workplace emotion classification.

The **CSR-NLI Prompting Framework** presents a novel hybrid approach to integrating *commonsense reasoning* into mental health assessment. Unlike purely neural models, CSR-NLI combines *symbolic reasoning* with large language models, yielding higher logical consistency and structured causal inference. The multi-model evaluation confirms that structured prompting enhances reasoning capabilities, particularly in *GPT-4-Turbo* and *LLaMA-3.3-70B*, which exhibit an average reasoning score improvement of **50 points** over baseline models.

The newly developed **Reasoning Evaluation Dataset** improves upon *CAMS* by introducing a larger, more accurate dataset containing **4,142 records** specifically tailored for workplace-related causal reasoning. Experimental results indicate that our dataset allows for superior reasoning generation, validated through *Wilcoxon Signed-Rank Tests* and *Paired t-Tests*, confirming statistically significant improvements in reasoning coherence.

The study demonstrates that iterative symbolic prompting significantly improves causal reasoning in large language models, supporting the growing relevance of hybrid neuro-symbolic architectures in explainable AI. The CSR-NLI framework enhances the explainability of AI-driven mental health assessments by integrating structured reasoning aligned with human interpretations, offering promising potential for real-time mental health monitoring systems in workplace settings. The following section discusses these implications in detail, emphasizing the practical relevance of the proposed framework.

5.1.3 Practical Implications

The findings of this research present key advancements in AI-driven workplace mental health monitoring:

- **Enhanced Workplace Well-Being:** The ability to detect emotions and stress levels in workplace communication enables proactive mental health support, reducing burnout and absenteeism.
- **Improved AI Explainability:** The integration of *CSR-NLI* enhances transparency in mental health assessments, providing interpretable causal reasoning compared to black-box deep learning models.
- **Superior Dataset for Mental Health Analysis:** The newly developed dataset surpasses *CAMS*, offering a broader range of workplace stressors and sentiment variations, leading to improved model generalizability.

- **Scalability and Cost-Effectiveness:** The lightweight, efficient framework ensures real-time deployment across workplace environments, enabling cost-effective mental health interventions and potentially lowering insurance premiums.

The proposed framework's practical implications extend beyond emotion detection, enabling organizations to proactively address mental health concerns through explainable AI solutions. By offering structured reasoning and enhanced model transparency, the CSR-NLI framework promotes a more comprehensive and trustworthy approach to workplace mental health assessment. These benefits are particularly relevant in corporate environments where scalability, cost-effectiveness, and ethical considerations are critical.

The insights gained from the comparative analysis and practical implications of the proposed framework underscore the importance of combining symbolic reasoning with neural inference. The CSR-NLI framework's ability to generate interpretable causal reasoning presents a promising avenue for enhancing the robustness and reliability of mental health monitoring systems. The next section delves deeper into the strengths and limitations of CSR-NLI, providing a detailed evaluation of its performance and areas for improvement.

5.2 Insights from CSR-NLI Framework

The CSR-NLI framework represents a significant advancement in integrating symbolic reasoning with neural inference for causal reasoning and emotion analysis. Unlike traditional neural approaches that often suffer from interpretability issues, CSR-NLI enhances decision transparency by iteratively refining reasoning through structured commonsense validation. This section provides an in-depth evaluation of the framework's contributions, its empirical effectiveness, and identified limitations and areas for improvement.

5.2.1 Advancing Emotion and Causal Reasoning Analysis

CSR-NLI introduces a novel prompting technique that dynamically generates commonsense hypotheses and iteratively refines them using Neuro-Symbolic AI principles. Compared to existing methods such as causal and abductive mental state (CAMS) dataset-based reasoning models, CSR-NLI consistently achieves higher logical consistency, improved alignment with human interpretations, and more structured causal inference.

For example, when analyzing the input message:

"I feel overwhelmed at work due to upcoming deadlines,"

CSR-NLI generates a structured reasoning output:

> *"Work deadlines impose time constraints that increase cognitive load, leading to heightened stress levels in working environments."*

This structured approach ensures that AI-generated reasoning aligns with real-world knowledge and reduces hallucinations commonly found in neural-based reasoning models.

Additionally, the newly created dataset outperforms CAMS in terms of data diversity, reasoning quality, and annotation consistency, making it a more suitable benchmark for evaluating AI-driven causal reasoning in workplace mental health contexts.

5.2.2 Effectiveness of CSR-NLI in Prompting and Reasoning Generation

The evaluation of CSR-NLI reasoning performance was conducted across multiple large language models (LLMs), including GPT-3.5-Turbo, GPT-4-Turbo, Allam-2-7B, LLaMA-3.3-70B-Versatile, and DeepSeek-R1-Distill-LLaMA-70B. The results demonstrated that CSR-NLI achieves superior logical coherence, with an average reasoning score of 92.08 on GPT-3.5-Turbo, 82.07 on GPT-4-Turbo, and 87.13 on LLaMA-3.3-70B. A total of 4,142 reasoning instances were compared with human-annotated ground truth reasonings, revealing that CSR-NLI outperforms baseline neural approaches in terms of logical structure and causal relevance.

Further validation was performed using statistical significance tests, including the Wilcoxon Signed-Rank Test and Paired t-Test, both of which confirmed that CSR-NLI improves reasoning coherence and alignment compared to existing baseline models. To visualize these improvements, various analytical techniques such as histograms, scatter plots, and correlation heatmaps were employed. These visual analyses demonstrated a clear enhancement in reasoning score distribution across multiple LLMs, further reinforcing the robustness and reliability of the framework.

The findings demonstrate that CSR-NLI offers a more interpretable, adaptable, and scalable approach to causal reasoning in AI-driven mental health analytics by integrating symbolic reasoning with neural inference. Its robustness across multiple LLMs enhances causal reasoning quality, improving coherence and enabling real-time deployment in workplace communication monitoring. While promising for explainable AI in assessing stressors and emotional states, certain limitations remain to be addressed for optimizing efficiency and adaptability.

5.2.3 Limitations and Areas for Improvement

Despite its advantages, CSR-NLI exhibits certain limitations that require further investigation. While the framework outperforms previous methods, it still struggles with highly nuanced and ambiguous conversational messages, where subtle variations in

language can impact reasoning accuracy. Although the newly developed dataset offers improvements over CAMS, additional generalization is necessary to ensure applicability across diverse workplace communication settings.

Another key challenge lies in the computational cost associated with the iterative refinement process. While this enhances reasoning quality, it increases inference time and computational expense, potentially limiting real-time applications, particularly in high-volume organizational environments. Additionally, CSR-NLI's dependency on high-resource large language models, such as GPT-4-Turbo, raises concerns regarding deployment feasibility at scale. The reliance on these models may introduce constraints in terms of accessibility and affordability, especially for organizations with limited computational resources.

Addressing these limitations is critical to enhancing the broader applicability of CSR-NLI. As the framework continues to evolve, integrating hybrid reasoning models and optimizing computational efficiency will be essential to achieving higher scalability and interpretability. The next section provides recommendations for future improvements and discusses potential avenues for enhancing the performance of CSR-NLI in real-world applications.

5.2.4 Future Directions for Enhancing CSR-NLI

To further enhance CSR-NLI and address its existing challenges, several key improvements can be explored. Optimizing prompt engineering is a crucial step in improving alignment with complex and nuanced conversational contexts, ensuring that the generated reasoning remains contextually accurate and interpretable. Additionally, multi-dataset benchmarking should be conducted to evaluate CSR-NLI across a diverse range of corpora beyond CAMS and the newly developed dataset. This will help assess its generalizability and effectiveness in different workplace communication settings.

Improving computational efficiency is another priority, as reducing iterative cycles without compromising reasoning accuracy can make CSR-NLI more scalable for real-time applications. Hybrid symbolic-neural pruning techniques could be explored to achieve this balance, enabling faster inference times while maintaining reasoning quality. Furthermore, integrating adaptive iterative refinement mechanisms that dynamically adjust the number of reasoning steps based on message complexity can help optimize performance, ensuring that more complex cases receive additional refinement while simpler cases are processed more efficiently.

Finally, advancing CSR-NLI through hybrid reasoning models that combine graph-based AI with symbolic commonsense reasoning can further enhance causality representation. This integration could improve the framework's ability to model complex relationships and dependencies within workplace mental health assessments. By implementing these advancements, CSR-NLI can evolve into a more robust, scalable, and

widely applicable framework for neuro-symbolic AI in mental health analytics.

5.2.5 Conclusion

The CSR-NLI framework represents a significant step forward in neuro-symbolic AI, particularly in workplace mental health analysis. By introducing structured commonsense-driven reasoning, the framework outperforms existing datasets and baseline models while maintaining high transparency, scalability, and logical coherence. Future improvements in prompt engineering, dataset generalization, and computational efficiency will further enhance its applicability across diverse organizational contexts, setting a new standard for AI-driven causal reasoning in mental health assessment.

5.3 Challenges and Lessons Learned

The development of the CSR-NLI framework encountered several technical, ethical, and methodological challenges. Understanding these challenges is essential for improving the framework's effectiveness and ensuring its successful deployment in workplace mental health monitoring. This section discusses the primary obstacles faced during the research process and highlights the key lessons learned to guide future advancements.

5.3.1 Technical Challenges

The development and implementation of the CSR-NLI framework presented several technical challenges. While integrating symbolic reasoning with neural inference improved reasoning quality, it also significantly increased computational complexity. The reliance on large-scale language models required substantial processing power, raising concerns about scalability and real-time deployment in workplace settings. Future enhancements could explore model compression techniques or distillation strategies to improve efficiency. Additionally, dataset limitations posed another challenge. Although the custom-developed dataset demonstrated superior performance compared to CAMS, manual annotation remained a bottleneck. Semi-supervised learning techniques could be leveraged to reduce the manual effort required for high-quality dataset generation.

Another challenge involved the trade-off between iterative refinement and performance. The CSR-NLI framework's iterative reasoning improved explainability but resulted in increased inference time, making it less viable for real-time applications. Implementing adaptive reasoning mechanisms that dynamically adjust the number of reasoning iterations based on message complexity could optimize efficiency without sacrificing accuracy. Furthermore, the dependency on high-resource large language

models, such as GPT-4 Turbo, raises concerns about feasibility for large-scale corporate deployments. While symbolic reasoning contributes to interpretability, aligning AI-generated explanations with human mental models for causal reasoning remains an area for further exploration.

5.3.2 Ethical and Methodological Challenges

Ethical considerations played a critical role in this research, particularly concerning privacy, transparency, and bias. The analysis of workplace communication data required strict adherence to data protection protocols, including anonymization techniques to safeguard employee privacy. However, challenges related to implicit re-identification persisted, highlighting the need for stronger privacy-preserving mechanisms such as differential privacy. Furthermore, ensuring trust in AI-generated insights is crucial for adoption. While CSR-NLI improves explainability, additional mechanisms must be implemented to allow employees to query, contest, and better understand AI-driven assessments.

A major challenge in workplace integration was overcoming resistance from employees and HR departments regarding AI-driven mental health monitoring. Many concerns revolved around AI's role in decision-making and the potential for misinterpretation of employee sentiments. The research findings suggest that co-designing intervention strategies with employees can help foster trust and encourage responsible AI adoption. Additionally, bias in model outputs remains an issue, as class imbalances in stress and emotion detection models affected the fairness of AI-driven assessments. To address this, further augmentation of underrepresented categories and continuous bias auditing are necessary to ensure unbiased and equitable AI applications.

5.3.3 Lessons Learned

This research provided valuable insights into deploying AI for workplace mental health monitoring. One key lesson is that high-quality datasets contribute more to model performance than complex architectures. The custom-developed dataset significantly outperformed CAMS, demonstrating that well-structured and diverse datasets are crucial for improving AI reasoning. Another important finding is the trade-off between explainability and performance. While symbolic reasoning enhances model transparency, it also increases computational costs, suggesting the need for lightweight neuro-symbolic approaches that balance efficiency and interpretability.

Additionally, the study highlights that AI models must be adaptive to different workplace contexts. Rigid AI models fail to generalize effectively across diverse organizational settings, reinforcing the need for adaptable frameworks that tailor responses to specific workplace environments. Ethical AI deployment emerged as a key consideration for adoption. Beyond technical improvements, establishing clear guidelines on

how AI-driven insights are used, ensuring human-in-the-loop oversight, and addressing employee concerns are essential for successful implementation. These lessons will inform future advancements in AI-driven mental health assessments, ensuring that such systems are both technically robust and ethically responsible.

5.4 Implications for Workplace Environments

The successful integration of CSR-NLI within the Mentalisys Health Application demonstrates its potential to revolutionize corporate communication platforms and employee mental health monitoring. This section discusses the broader implications of the proposed framework, emphasizing its scalability, ethical considerations, and potential for enhancing workplace well-being.

The integration of AI-driven solutions, particularly the *CSR-NLI* framework within the *Mentalisys Health Application*, holds the potential to revolutionize corporate communication platforms and employee mental health monitoring. By leveraging advanced neuro-symbolic reasoning, the proposed framework enhances decision-making, improves employee well-being, and contributes to business sustainability.

5.4.1 Impact on Corporate Communication Platforms

CSR-NLI's integration with workplace communication platforms such as *Slack* facilitates real-time emotion and stress analysis, enabling organizations to monitor workplace sentiment dynamically. Unlike traditional sentiment analysis models that rely on polarity-based detection, the neuro-symbolic reasoning employed in CSR-NLI captures deeper causal relationships behind employee stressors. This capability fosters a proactive approach to mental health intervention, reducing employee burnout, absenteeism, and turnover.

5.4.2 Enhancing Employee Mental Health and Well-Being

The framework goes beyond conventional text classification by incorporating causal reasoning, allowing HR teams to pinpoint specific triggers of workplace stress. By leveraging symbolic reasoning, CSR-NLI generates structured insights into employees' psychological well-being, promoting a supportive and psychologically safe work environment. This refined understanding enables organizations to introduce *personalized mental health interventions*, reducing stress-related workplace conflicts and enhancing overall job satisfaction.

5.4.3 Overcoming Adoption Barriers

Despite the clear benefits of AI-driven workplace mental health monitoring, several challenges may hinder adoption. One of the primary concerns is *data privacy and ethical implications*, as employees may perceive AI-based sentiment monitoring as intrusive. Ensuring strict adherence to data anonymization, encryption, and obtaining informed consent are essential measures to build trust and transparency. Additionally, *corporate hesitation* remains a significant barrier, with organizations often reluctant to implement AI-driven monitoring due to perceived cost, complexity, or skepticism regarding its effectiveness. Demonstrating CSR-NLI's scalability, lightweight implementation, and minimal maintenance requirements can help alleviate such concerns. Another critical challenge is the *stigma surrounding workplace mental health*, which may discourage employees from engaging with AI-driven well-being tools. Raising awareness through educational workshops, fostering transparency in AI decision-making, and involving employees in the implementation process can improve trust and adoption. Addressing these barriers effectively is crucial for ensuring the successful integration of CSR-NLI in corporate environments.

5.4.4 Conclusion

CSR-NLI represents a transformative leap in workplace mental health monitoring, offering real-time, explainable, and scalable solutions for corporate environments. By addressing privacy concerns, fostering organizational trust, and demonstrating financial benefits, this research provides a robust foundation for integrating AI into workplace well-being strategies. Future directions should explore further automation of symbolic reasoning models and broader integration with industry-specific HR platforms to maximize impact.

5.5 Limitations of the Study

While the CSR-NLI framework offers promising advancements in workplace mental health assessment, several limitations must be acknowledged to provide a comprehensive understanding of its scope. Identifying these limitations is crucial for guiding future research and improving the framework's applicability across various organizational contexts.

5.5.1 Platform and Data Constraints

The study primarily focuses on analyzing text-based communication from corporate platforms such as Slack. While effective for workplace communication analysis, this

approach does not account for multimodal data, including voice, video, or physiological indicators, which could further enrich mental health assessments. Additionally, the framework's dependency on structured workplace communication may limit its applicability to informal or less structured environments.

5.5.2 Generalizability Across Workplace Contexts

Although the framework was rigorously evaluated using benchmark datasets, its generalizability across different organizational cultures and industries remains a challenge. Communication styles, workplace norms, and stress indicators vary across sectors, necessitating additional validation in diverse corporate environments.

5.5.3 Computational and Deployment Challenges

The framework's reliance on large-scale language models, particularly high-resource architectures like GPT-4-Turbo, introduces concerns regarding computational efficiency and real-time deployment. While the system achieves robust performance, optimizing it for large-scale, real-time applications in organizations with extensive communication data remains an area for improvement.

5.5.4 Ethical and Privacy Considerations

The sensitive nature of workplace mental health assessments necessitates strict adherence to ethical standards concerning data privacy, informed consent, and user trust. Compliance with regulations such as GDPR and HIPAA is essential to ensure responsible data handling throughout collection, processing, and storage.

The use of Large Language Models (LLMs) like ChatGPT for reasoning generation introduces additional privacy considerations. Effective privacy-preserving techniques, including prompt engineering, data masking, anonymization, and encryption, must be implemented to prevent LLMs from directly accessing identifiable data. Ensuring that data processing occurs within secure environments and that users have full control over their data rights is crucial.

Transparency in AI-driven decision-making is essential to build user trust. Incorporating explainability mechanisms and maintaining clear governance policies through regular audits and compliance checks will enhance user acceptance and adherence to ethical standards.

5.5.5 Scalability and Cost Constraints

The integration of neuro-symbolic reasoning with large language models enhances interpretability but may pose challenges for scalability in resource-limited organizations. The cost associated with deploying and maintaining high-performance AI mod-

els could be a barrier to widespread adoption, particularly for small and medium-sized enterprises.

5.5.6 Conclusion

Acknowledging these limitations is essential for contextualizing the findings and guiding future research directions. Addressing these challenges through improved dataset diversity, optimized computational frameworks, and enhanced ethical safeguards will be key to making CSR-NLI a more scalable and widely applicable solution for workplace mental health monitoring.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of Key Contributions

This research addresses the pressing need for advanced, real-time, and explainable solutions to monitor and improve employee mental health within corporate environments. By integrating Neuro-Symbolic AI (NSAI) with sentiment analysis, stress detection, and commonsense-driven reasoning, the study presents several key contributions.

One of the primary contributions of this research is the development of a novel causal reasoning dataset that improves upon CAMS by incorporating refined causal reasoning data specific to workplace mental health. This dataset enables more accurate and interpretable AI-driven mental health assessments, providing a stronger foundation for future research in this domain. Additionally, the study introduces the CSR-NLI framework, which significantly advances neuro-symbolic AI by integrating symbolic reasoning with natural language inference. This framework enhances logical coherence, causal alignment, and reasoning transparency, making AI-generated insights more interpretable and actionable.

The research also contributes to the development of enhanced sentiment and stress analysis models by employing advanced machine learning techniques, including ensemble learning and transformer-based architectures. These models effectively detect nuanced emotional states within workplace communication and achieve high accuracy and robustness, as validated through standard evaluation metrics. Furthermore, through structured prompting and symbolic reasoning, CSR-NLI improves the interpretability of AI-driven causal analysis. Experimental evaluations demonstrate that structured symbolic inference significantly enhances reasoning quality in large language models, making AI-generated explanations more reliable and human-understandable.

Another key contribution of this research is its impact on workplace mental health monitoring. The integration of the Mentalisys Health Application provides real-time emotional and stress monitoring in corporate environments, allowing organizations to take proactive measures to support employee well-being, reduce burnout, and improve team dynamics. Beyond workplace well-being, the study highlights practical business implications, demonstrating how AI-driven mental health monitoring can contribute to tangible organizational benefits, such as reducing workplace stress, improving employee engagement, and potentially lowering corporate insurance premiums through early risk detection.

In summary, this research bridges significant gaps in workplace mental health assessment by introducing an innovative, scalable, and interpretable AI solution. The integration of structured commonsense reasoning with advanced machine learning

techniques ensures that mental health assessments are not only accurate but also explainable and actionable for organizations. These contributions pave the way for more effective AI-driven interventions in corporate mental health management.

6.2 Practical Recommendations

The proposed framework for AI-driven mental health assessment, integrated into workplace communication platforms, presents a strategic pathway for organizations to enhance employee well-being while maintaining ethical integrity. The following recommendations provide a roadmap for effective adoption:

6.2.1 Integration into Corporate Wellness Strategies

To maximize its impact, organizations should embed the CSR-NLI framework within existing corporate wellness programs. By integrating the system into widely used platforms such as Slack and Microsoft Teams, employers can gain real-time insights into workplace sentiment and emerging stress patterns. The system's predictive capabilities enable organizations to proactively address well-being concerns through tailored interventions, including mental health workshops, stress management programs, and personalized employee assistance initiatives.

6.2.2 Enhancing HR Practices and Decision-Making

HR departments can leverage the framework to refine employee support programs and performance reviews. By incorporating AI-driven emotional analytics, HR professionals can better assess workplace sentiment trends, identify burnout risks, and design data-driven well-being initiatives. The custom-developed dataset introduced in this research enhances emotion causality detection, enabling more context-aware assessments of workplace communication.

6.2.3 Addressing Adoption Barriers

While AI-driven mental health monitoring provides significant benefits, potential barriers such as privacy concerns and ethical considerations must be addressed to foster trust and adoption. Organizations should communicate the purpose and scope of AI-driven monitoring transparently, ensuring employees understand that the system is designed to support well-being rather than track performance. Strong data protection measures, including anonymization and restricted access, should be implemented to comply with privacy regulations such as GDPR. Furthermore, organizations can facilitate adoption through phased pilot testing, allowing employees to engage with the system voluntarily before organization-wide deployment.

6.2.4 Ethical Deployment and Trust-Building

Ethical AI deployment is critical to ensuring employee trust and engagement. Organizations must maintain transparency regarding AI decision-making processes and offer employees access to their own sentiment and stress assessments. Furthermore, periodic audits should be conducted to ensure that the framework operates fairly and does not reinforce biases in sentiment analysis or stress detection.

6.2.5 Scalability and Long-Term Adoption

To ensure long-term success and scalability, organizations should adopt an incremental approach to AI integration. The deployment should begin with small teams or departments before scaling organization-wide. HR professionals and managers should be trained to interpret AI-generated insights and use them effectively to guide workplace well-being policies. Moreover, continuous refinement of the AI model using real-world workplace data will ensure that the framework remains adaptable to evolving corporate communication styles and employee needs.

6.2.6 Conclusion

By following these recommendations, organizations can effectively integrate AI-driven sentiment and stress analysis into workplace environments while upholding ethical standards. The proposed CSR-NLI framework not only enhances employee mental health monitoring but also fosters a more transparent, data-driven approach to workplace well-being. Through strategic implementation, businesses can improve employee satisfaction, increase productivity, and ultimately drive a healthier organizational culture.

6.3 Future Work Directions

6.3.1 Enhancements to CSR-NLI Framework

The CSR-NLI framework has demonstrated significant potential in integrating symbolic reasoning with neural inference. However, further improvements can be made to enhance its causal reasoning capabilities. Future research should explore incorporating domain-specific knowledge graphs, improved commonsense datasets, and more effective prompting strategies tailored to complex workplace reasoning tasks. Additionally, expanding the framework's adaptability to multilingual data would increase its applicability across diverse organizational settings.

6.3.2 Integration of Multimodal Data

Currently, the framework primarily focuses on textual data from workplace communication platforms. A future research direction involves incorporating multimodal data sources, including vocal tone, facial expressions, and physiological signals, to gain a more comprehensive understanding of employee emotions. This multimodal approach would strengthen sentiment and stress detection by capturing non-verbal cues associated with workplace well-being.

6.3.3 Scalability for Real-Time Processing

With increasing volumes of workplace communication data, ensuring real-time reasoning efficiency is crucial. Future enhancements should focus on optimizing CSR-NLI's computational efficiency while preserving reasoning accuracy. Techniques such as hybrid symbolic-neural pruning and adaptive inference mechanisms could improve processing speed. Additionally, leveraging cloud-based AI architectures or distributed computing solutions can enhance the framework's scalability for large-scale deployments.

6.3.4 Cross-Industry Applications

While CSR-NLI has been validated in corporate environments, its potential extends beyond workplace mental health assessment. Future research could explore sector-specific adaptations for industries such as healthcare, education, and customer service. In healthcare, the framework could assist in identifying burnout risks among medical professionals, while in education, it could support student well-being monitoring based on online discussions.

6.3.5 Longitudinal Studies and Impact Assessment

The long-term effects of AI-driven mental health interventions require further investigation. Future work should involve longitudinal studies assessing how AI-based frameworks impact employee productivity, stress levels, and workplace engagement over extended periods. These studies would provide empirical validation for the effectiveness of AI-driven mental health interventions and guide improvements for real-world applications.

6.3.6 Ethical and Regulatory Compliance

Ethical deployment of AI-driven mental health solutions requires adherence to regulations such as GDPR, HIPAA, and regional privacy laws. The integration of LLMs like

ChatGPT for reasoning generation presents specific challenges related to data privacy and compliance.

To mitigate risks, techniques such as prompt engineering, data masking, differential privacy, and federated learning should be employed to ensure LLMs do not process identifiable data. Comprehensive data governance policies addressing data storage, processing, and deletion must be established to comply with varying regulatory requirements.

Regular audits, transparency reports, and collaboration with legal experts are essential for maintaining compliance. Ensuring proper documentation of LLM-generated decisions and providing audit logs will promote ethical AI use and build user trust in workplace mental health monitoring systems.

6.3.7 Conclusion

By addressing these future work directions, CSR-NLI and the Mentalisys Health Application can continue evolving into scalable, explainable, and domain-adaptive AI solutions for workplace mental health analytics. These advancements will enhance both technical capabilities and ethical considerations, ensuring the framework remains applicable across various industries while promoting responsible AI-driven mental health monitoring.

6.4 Final Thoughts

The increasing recognition of mental health as a critical factor in workplace productivity underscores the importance of AI-driven solutions for proactive well-being monitoring. This research has demonstrated the potential of integrating Neuro-Symbolic AI (NSAI) with real-time emotion analysis, stress detection, and commonsense-driven reasoning to create an interpretable and scalable mental health assessment framework.

The development of the *Mentalisys Health Application*, powered by the *CSR-NLI framework*, represents a significant step towards enhancing AI explainability in workplace wellness. Unlike traditional sentiment analysis tools, this approach not only detects emotional states but also uncovers causal factors behind workplace stressors, enabling targeted interventions. The system's ability to generate structured, human-like reasoning fosters greater transparency and trust, addressing a key limitation in AI-driven mental health assessments.

Despite these advancements, ethical considerations such as data privacy, informed consent, and AI explainability remain crucial for fostering user trust and compliance with regulatory frameworks. Ensuring that employees perceive these systems as supportive rather than intrusive is essential for their successful adoption. This research

highlights the need for transparent communication, secure data handling, and continuous ethical audits to align AI innovations with organizational well-being goals.

Looking ahead, future improvements in multimodal integration, longitudinal studies, and cross-industry adaptation will further extend the impact of AI-driven mental health solutions. The application of this research in sectors such as *healthcare, education, and high-stress industries* holds promise for broader adoption and refinement. By bridging technological advancements with human-centric design, this work lays the foundation for a more empathetic, inclusive, and productive workplace environment.

In conclusion, this study underscores the transformative role of AI in mental health analytics, demonstrating that by integrating symbolic reasoning with neural inference, AI-driven workplace wellness solutions can be both powerful and ethical. The ongoing evolution of this framework will not only refine its technical robustness but also contribute to reducing workplace mental health stigma, fostering a culture of psychological safety and proactive well-being management.

REFERENCES

[1] M. Conforti, "2024 workplace mental health trends: A proactive approach to well-being," www.springhealth.com, 12 2023. [Online]. Available: https://www.springhealth.com/blog/2024-workplace-mental-health-trends

[2] L. Cable and B. Hallowell, "Mental health challenges continue to evolve in the covid-19 pandemic," KPMG, 01 2023. [Online]. Available: https://kpmg.com/ca/en/home/insights/2023/01/employee-mental-health-in-the-workplace.html

[3] NAMI, "The 2024 nami workplace mental health poll," NAMI, 01 2024. [Online]. Available: https://www.nami.org/Support-Education/Publications-Reports/Survey-Reports/The-2024-NAMI-Workplace-Mental-Health-Poll

[4] S. G. Aldana, "Financial impact of health promotion programs: A comprehensive review of the literature," *American Journal of Health Promotion*, vol. 15, no. 5, pp. 296–320, 2001.

[5] W. H. Organization, "Mental health in the workplace," 2020, retrieved from https://www.who.int/teams/mental-health-and-substance-use/mental-health-in-the-workplace.

[6] L. L. Berry, A. M. Mirabito, and W. B. Baun, "What's the hard return on employee wellness programs," *Harvard Business Review*, vol. 88, no. 12, pp. 2–71, 2010, retrieved from https://hbr.org/2010/12/whats-the-hard-return-on-employee-wellness-programs.

[7] R. Dou and X. Kang, "Tam-senticnet: A neuro-symbolic ai approach for early depression detection via social media analysis," *Computers and Electrical Engineering*, vol. 114, p. 109071, 03 2024.

[8] G. Antoniou, E. Papadakis, and G. Baryannis, "Mental health diagnosis: a case for explainable artificial intelligence," *International Journal on Artificial Intelligence Tools*, vol. 31, no. 03, p. 2241003, 2022.

[9] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, 2004. [Online]. Available: https://books.google.lk/books?id=AoIUp5fJkIcC

[10] P. D. Bliese, J. R. Edwards, and S. Sonnentag, "Stress and well-being at work: A century of empirical trends reflecting theoretical and societal influences." *Journal of Applied psychology*, vol. 102, no. 3, p. 389, 2017.

[11] T. A. Wright, R. Cropanzano, and D. G. Bonett, "The moderating role of employee positive well being on the relation between job satisfaction and job performance." *Journal of occupational health psychology*, vol. 12, no. 2, p. 93, 2007.

[12] X. Zhang and V. S. Sheng, "Neuro-symbolic ai: Explainability, challenges, and future trends," *arXiv preprint arXiv:2411.04383*, 2024.

[13] T. Gubler, I. Larkin, and L. Pierce, "Doing Well by Making Well: The Impact of Corporate Wellness Programs on Employee Productivity," *SSRN Electronic Journal*, no. February, 2016.

[14] A. Basi Nska-Zych and A. Springer, "Organizational and individual outcomes of health promotion strategies—a review of empirical research," *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, pp. 1–27, 2021.

[15] K. Danna and R. W. Griffin, "Health and well-being in the workplace: A review and synthesis of the literature," *Journal of Management*, vol. 25, no. 3, pp. 357–384, 1999.

[16] C. Krekel, G. Ward, and J.-E. De Neve, "Employee wellbeing, productivity, and firm performance," *Saïd Business School WP 2019-04*, no. Mar, p. 44, 2019, available at SSRN: https://ssrn.com/abstract=3356581.

[17] B. G. Mujtaba and F. J. Cavico, "Corporate wellness programs: Implementation challenges in the modern american workplace," *International Journal of Health Policy and Management*, vol. 1, no. 3, pp. 193–199, 2013.

[18] Y. Q. Lim, C. M. Lim, K. H. Gan, and N. H. Samsudin, "Text Sentiment Analysis on Twitter to Identify Positive or Negative Context in Addressing Inept Regulations on Social Media Platform," *ISCAIE 2020 - IEEE 10th Symposium on Computer Applications and Industrial Electronics*, pp. 96–101, 2020.

[19] J. Blair, C. Y. Hsu, L. Qiu, S. H. Huang, T. H. K. Huang, and S. Abdullah, *Using Tweets to Assess Mental Well-being of Essential Workers during the COVID-19 Pandemic*. Association for Computing Machinery, 2021, vol. 1, no. 1.

[20] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, vol. 2, pp. 154–164, 2019.

[21] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion identification in conversations," *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pp. 2470–2481, 2020.

[22] S. Feislachen, P. Garus, H. Wang, E. Podkolin, S. Schlüter, N. S. Bernd, S. Manske, A. Nolte, and I. A. Chounta, *Sentiment Analysis of Participants Interactions in a Hackathon Context: The Example of a Slack Corpus*. Association for Computing Machinery, 2022, vol. 1, no. 1.

[23] P. Chatterjee, K. Damevski, N. A. Kraft, and L. Pollock, "Software-related Slack Chats with Disentangled Conversations," *Proceedings - 2020 IEEE/ACM 17th International Conference on Mining Software Repositories, MSR 2020*, pp. 588–592, 2020.

[24] D. Wang, H. Wang, M. Yu, Z. Ashktorab, and M. Tan, "Group Chat Ecology in Enterprise Instant Messaging: How Employees Collaborate Through Multi-User Chat Channels on Slack," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–15, 2022.

[25] P. Chatterjee, K. Damevski, L. Pollock, V. Augustine, and N. A. Kraft, "Exploratory study of slack Q&A chats as a mining source for software engineering tools," *IEEE International Working Conference on Mining Software Repositories*, vol. 2019-May, pp. 490–501, 2019.

[26] B. Liu, "Sentiment Analysis and Opinion Mining," in *Synthesis Lectures on Human Language Technologies (SLHLT)*, 2012, vol. 30, no. April, pp. 503–523.

[27] L. Bostan, E. Kim, and R. Klinger, "Good news everyone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, no. 1, 2020, pp. 1554–1566.

[28] T. Kajiwara, C. Chu, N. Takemura, Y. Nakashima, and H. Nagahara, "WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations," in *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2021, pp. 2095–2104.

[29] M. Garg, C. Saxena, S. Saha, V. Krishnan, R. Joshi, and V. Mago, "CAMS: An annotated corpus for causal analysis of mental health issues in social media posts," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association,

Jun. 2022, pp. 6387–6396. [Online]. Available: https://aclanthology.org/2022.lrec-1.686/

[30] C. Yuan, C. Fan, J. Bao, and R. Xu, "Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, pp. 3568–3573.

[31] Y. Sun, N. Yu, and G. Fu, "A Discourse-Aware Graph Neural Network for Emotion Recognition in Multi-Party Conversation," *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pp. 2949–2958, 2021.

[32] D. Teodorescu and S. M. Mohammad, "Frustratingly Easy Sentiment Analysis of Text Streams: Generating High-Quality Emotion Arcs Using Emotion Lexicons," 2022. [Online]. Available: http://arxiv.org/abs/2210.07381

[33] L. A. M. Bostan and R. Klinger, "An analysis of annotated corpora for emotion classification in text," in *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 2018, pp. 2104–2119.

[34] Y. Chen, S. Y. M. Lee, S. Li, and C. R. Huang, "Emotion cause detection with linguistic constructions," in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, vol. 2, no. August, 2010, pp. 179–187.

[35] R. Maleki, H. Rahmani, E. Mohamadi, E. Jaafaripooyan, and A. Atashi, "A Scoping Review of Health Insurance Deductions in Hospitals: Root Causes and Solutions," *Health Scope*, vol. 12, no. 2, 2023.

[36] A. Verma and J. Maiti, "Text-document clustering-based cause and effect analysis methodology for steel plant incident data," *International Journal of Injury Control and Safety Promotion*, vol. 0, no. 0, pp. 1–11, 2018. [Online]. Available: https://doi.org/10.1080/17457300.2018.1456468

[37] A. Dingli and D. Farrugia, *Neuro-Symbolic AI: Design transparent and trustworthy systems that understand the world as you do*. Packt Publishing Ltd, 2023.

[38] K. Roy, "Healthcare assistance challenges-driven neurosymbolic ai," *Biomedical Journal of Scientific &Technical Research*, vol. 58, no. 2, 2024. [Online]. Available: https://scholarcommons.sc.edu/aii_fac_pub/610/

[39] A. Rastogi, Q. Liu, and E. Cambria, "Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 2022-July, 2022.

[40] M. Gaur, K. Gunaratna, S. Bhatt, and A. Sheth, "Knowledge-infused learning: A sweet spot in neuro-symbolic ai," *IEEE Internet Computing*, vol. 26, no. 4, pp. 5–11, 2022.

[41] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint*, 2015. [Online]. Available: https://arxiv.org/abs/1508.05326v1

[42] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," *arXiv preprint*, 2017. [Online]. Available: https://arxiv.org/abs/1704.05426v4

[43] T. Khot, A. Sabharwal, and P. Clark, "Scitail: A textual entailment dataset from science question answering," *AAAI*, 2018. [Online]. Available: https://arxiv.org/abs/scitail-aaai-2018

[44] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-snli: Natural language inference with natural language explanations," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[45] L. Luo, Z. Zhao, and C. G. et al., "Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2410.13080v1

[46] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3027–3035.

[47] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," *arXiv preprint arXiv:1906.05317*, 2019.

[48] J. Liu, A. Liu, and X. L. et al., "Generated knowledge prompting for commonsense reasoning," *arXiv preprint*, 2022. [Online]. Available: https://arxiv.org/abs/2110.08387v3

[49] J. W. et al., "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint*, 2022. [Online]. Available: https://arxiv.org/abs/2201.11903v6

[50] S. Y. et al., "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2305.10601v2

[51] T. et al., "Commonsenseqa: A question answering challenge targeting commonsense knowledge," *Proceedings of NAACL-HLT*, 2019. [Online]. Available: https://www.aclweb.org/anthology/N19-1421/

[52] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[53] Y. Z. et al., "Meta prompting for ai systems," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2311.11482v6

[54] Y. et al., "React: Synergizing reasoning and acting in large language models," *GitHub Repository*, 2023. [Online]. Available: https://github.com/ysymyth/ReAct

[55] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "Carer: Contextualized affect representations for emotion recognition," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 3687–3697.

[56] H. Wave, "Introduction | h2o wave," [Online; accessed 2025-02-02]. [Online]. Available: https://wave.h2o.ai/docs/getting-started