Towards Concept-Based Language Models for Digital Phenotyping:

A Window into Users' Mental States

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Computational Science

presented by
Gualtiero Marenco Turi

under the supervision of
Prof. Mira Antonietta
co-supervised by
Ravenda Federico

January 2026

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Gualtiero Marenco Turi
Lugano, 20 January 2026

Abstract

Depression represents one of the most severe and widespread mental health challenges worldwide. According to estimates by the World Health Organization, approximately 332 million people are affected globally, and even in high income countries only about one third of individuals with depression receive adequate mental health treatment, partly due to limited access to mental health professionals and the difficulty of deploying scalable screening tools. In this context, the growing availability of user generated text on social media platforms offers a unique opportunity for early and large scale mental health screening.

Artificial intelligence is increasingly applied in psychological assessment, yet its limited interpretability remains a significant barrier to clinical integration. While many machine learning models have demonstrated strong predictive performance for mental health detection tasks, they often operate as black boxes, providing little insight into the reasoning behind their decision process. In psychological and clinical settings, this lack of transparency is particularly problematic, as explainability is essential for trust, ethical accountability, and informed decision making. As a result, despite the urgency for accessible and easy to deploy screening systems, AI based solutions are often met with skepticism and remain underutilized in practice.

Concept bottleneck models have introduced a promising paradigm by incorporating human understandable concepts as intermediate representations, but they often do so at the cost of predictive performance. Recent advancements in Concept Embedding Models aim to alleviate this tradeoff by relaxing the strict information bottleneck, enabling both competitive accuracy and more transparent model behavior.

In this work, we investigate the applicability of concept based learning approaches for mental health screening from social media data, with a primary focus on depression. Using user generated text from Reddit, we explore and compare several modeling strategies, including concept bottleneck models and concept embedding models, as well as different retrieval mechanisms for selecting relevant user posts. Depression related symptoms, derived from clinical manuals, are used to guide intermediate representations, allowing for an analysis of how clinically motivated concepts interact with noisy, real world language.

Our results highlight the practical limitations of strict concept bottlenecks in this setting and suggest that concept embedding models provide a more flexible and robust alternative for handling complex and weakly supervised textual signals, while still preserving interpretability hooks that are meaningful for mental health applications.

Acknowledgements

This thesis is the result of not only my efforts but also the support, guidance, and companionship of many remarkable individuals. I am grateful to each of you for being part of this journey.

I owe immense thanks to Professor Antonietta Mira for the opportunity and her advice, and to Federico Ravenda for his constant support and for being an exemplary T.A. over these two years, besides being the other biggest contributor to this thesis.

I'm also deeply grateful to all the wonderful and wonderfully skilled people I've met at USI, as well as to the vast majority of the faculty. My experience at USI has surpassed all my expectations by far and completely reframed my view of the academic world.

I can only offer a thank you and my friendship in return for putting up with me to my family and friends. I don't see many of you as often as I'd like, but I don't forget you.

A special thank goes to my girlfriend, for being there for every step of the long road done over the last five years.

Finally, I've enjoyed this process even more thanks to my beautiful cats, who sat on my lap during tireless nights of coding. You made it all a little warmer.

Contents

Chapter 1

Introduction

Mental health disorders, and depression in particular, represent one of the most pressing public health challenges of our time. According to the World Health Organization, 332 million people worldwide are affected by depression(Institute for Health Metrics and Evaluation [2024]), yet even in high income countries only about two third remain undiagnosed or untreated due to limited access to mental healthcare and the absence of scalable, low-cost screening tools (Evans-Lacko et al. [2018]). In recent years, the widespread adoption of social media platforms has generated unprecedented volumes of user-produced text reflecting personal experiences, emotions, and daily activities. Over the past decade, this data has attracted growing interest from the research community as a potential source of signals for mental health assessment, screening, and monitoring. Unlike traditional clinical settings, where information is elicited through structured interviews or validated questionnaires, social media content is produced spontaneously, without explicit diagnostic intent, and often outside of any clinical context. This fundamental difference both motivates and complicates the use of machine learning models for mental health inference from online text.

A substantial body of work has demonstrated that individuals affected by mental health conditions, particularly depression, exhibit statistically detectable differences in language use compared to control populations. These differences manifest in lexical choice, sentiment polarity, self-referential language, temporal focus, and discourse (Pennebaker et al. [2001]; Rude et al. [2004]; Alm et al. [2007]; Cohan et al. [2018]). Such observations have been documented across multiple platforms, including Twitter, Reddit, and online forums, and have motivated the formulation of supervised and weakly supervised classification tasks aimed at distinguishing affected from unaffected users (Losada and Crestani [2017]; Bao et al. [2023]). However, the presence of correlational linguistic patterns does not imply that mental health states are directly or consistently observable in text, and the extraction of clinically meaningful signals remains highly challenging.

One of the central difficulties lies in the inherently noisy nature of user generated content. Social media posts are shaped by informal language, slang, sarcasm, memes, and rapidly evolving cultural references. Users frequently discuss a wide range of topics unrelated to their mental health, even when participating in communities nominally associated with psychological support. As a result, relevant signals are often sparse and embedded within large volumes of irrelevant or ambiguous text. This sparsity is particularly pronounced at the user level, where a diagnosis label may be assigned to an entire posting history despite only a small subset of posts containing content indicative of psychological distress (Losada and Crestani [2017]).

Recent work (Torous et al. [2025], Stangl et al. [2023], Aragon-Guevara et al. [2023], Montag et al. [2024]) has emphasized that explainability is not merely a desirable add on, but a central requirement for any system operating in this domain. Clinicians and stakeholders require insight into why a model produces a given prediction, which aspects of the data contribute most strongly, and how these signals relate to established psychological constructs. Approaches that provide only aggregate risk scores without interpretable intermediate representations risk undermining trust and limiting real world applicability (Ravenda et al. [2025]). Consequently, the problem of mental health detection from social media cannot be reduced to a standard text classification task, but must be understood as a complex socio technical challenge involving noisy data, weak supervision, ethical constraints, and the need for human interpretable reasoning.

These concerns have motivated increasing interest in modeling approaches that incorporate interpretability as a structural property rather than an afterthought. Concept based learning (Kim et al. [2018]) represents one such paradigm, in which predictions are mediated through intermediate variables intended to correspond to human understandable concepts. In the context of mental health, these concepts may relate to symptoms, behavioral indicators, or clinically defined dimensions that are meaningful to practitioners. By explicitly modeling these intermediate representations, concept based approaches aim to provide explanations that are more directly aligned with psychological theory and clinical reasoning.

At the same time, imposing interpretability constraints introduces new challenges. Requiring a model to operate through a limited set of predefined concepts may restrict its expressive capacity, particularly when those concepts only partially capture the underlying phenomenon of interest. Prior work has shown that strict concept bottlenecks can lead to degraded performance, especially in complex and noisy domains (Seonghwan and Jueun Mun [2025], Espinosa Zarlenga et al. [2022]). This trade-off between interpretability and predictive accuracy lies at the heart of current debates around explainable artificial intelligence and is especially pronounced in mental health applications.

The present work is situated within this tension. Rather than treating explainability and performance as mutually exclusive objectives, the goal is to investigate modeling strategies that attempt to balance both. This involves examining not only whether concept based models can achieve competitive results on realistic datasets, but also how their internal representations behave, where they fail, and under which conditions interpretability constraints become detrimental. Importantly, this perspective shifts the emphasis from optimizing a single evaluation metric to understanding the trade offs induced by different modeling choices.

To test these hypothesis, we will use the eRisk benchmark as a case study for predicting the risk of mental health conditions, specifically depression, from Reddit posts. Its main challenges include noisy, indirect labels, a pronounced class imbalance, long sequences with sporadic relevant signals, and data that reflects realistic, weakly-supervised screening scenarios.

1.1 Research Objectives and Scope

The central objective of this thesis is to investigate whether concept based models can provide a viable balance between predictive performance and interpretability in mental health screening from social media data. Specifically, the work examines whether symptom level concepts derived from the Beck Depression Inventory–II (BDI-II) (Beck et al. [1996]) and annotated using large language models can support robust and interpretable user level prediction in the presence of substantial noise, weak supervision, and class imbalance. Rather than treating interpretability as a secondary constraint, the goal is to study how it can be integrated into the modeling process without rendering performance impractical for realistic screening scenarios. Beyond this primary question, the thesis also aims to analyze which machine learning techniques and design choices are most effective in this setting, and which prove insufficient or detrimental. This includes examining different strategies for concept aggregation, representation learning, loss design, and class imbalance handling, with particular attention to how these choices affect both predictive behavior and the interpretability of intermediate representations. The emphasis is therefore not on achieving state of the art benchmark scores, but on understanding the trade offs induced by concept based modeling under realistic and ethically constrained conditions.

Within this broader context, the definition and sourcing of concepts becomes a central methodological question. In clinical domains, access to expert annotations is limited, and the feasibility of concept based modeling depends critically on how such concepts are obtained and validated. The next section addresses this issue directly, discussing strategies for defining symptom level concepts under data scarcity and motivating the use of large language models as a practical tool in this process.

1.2 Overview of the Thesis

This thesis is organized to progressively move from conceptual and theoretical foundations to methodological design, empirical evaluation, and interpretation of results, with a consistent focus on balancing predictive performance and interpretability in mental health screening from social media data.

Chapter 2 introduces the theoretical background underlying the proposed approach. It reviews concept based learning and multiple instance learning, discussing their relevance for weakly supervised and noisy settings such as social media analysis. Particular attention is given to concept bottleneck models and concept embedding models, highlighting their strengths and limitations when applied to high dimensional text data. The chapter also introduces embedding based representations, including transformer based language models, and discusses design choices related to pooling, aggregation, and similarity based retrieval that are central to the proposed architecture.

Chapter 3 presents the modeling framework adopted in this work. The chapter describes the overall pipeline, including text embedding, concept retrieval, concept aggregation, and final user level prediction. Different aggregation strategies are discussed from a theoretical perspective, with particular emphasis on max based and sum based pooling formulations. While multiple variants were explored, the focus is placed on the max based approach, whose empirical advantages are analyzed later in the thesis. The chapter also clarifies the role of concepts as intermediate latent variables rather than standalone diagnostic predictors.

Chapter 4 addresses learning under class imbalance and weak supervision, which are intrinsic characteristics of mental health data. The chapter provides theoretical grounding for the techniques used to mitigate severe class imbalance, including loss reweighting, sampling strategies, and threshold optimization. Evaluation metrics are discussed in detail, with an emphasis on recall oriented assessment and the interpretation of trade offs between false positives and false

negatives in a clinical screening context.
Chapter 5 describes the experimental setup and implementation details. This includes dataset construction and splitting based on the eRisk benchmarks, reproducibility considerations, and descriptive statistics of the data. The chapter outlines how training, validation, and test sets are defined to ensure fair comparison with prior work, and motivates the choice of experimental protocols used throughout the evaluation. In Chapter 6 baseline model performance is analyzed both at the user level and at the concept level, with particular attention to recall, robustness under imbalance, and interpretability of intermediate representations. Here the results are discussed in relation to existing benchmarks, showing that the proposed approach achieves competitive performance while maintaining transparent and clinically meaningful explanations. Furthermore in this chapter are presented the experiments that have been tried to improve the baseline and their outcome, as well as their rationale and the plausible theoretical grounding regarding their outcome.
Finally, Chapter 7 concludes the thesis by summarizing the main findings, discussing their implications, and outlining directions for future work.

Chapter 2

Concept Based Learning: From Bottlenecks to Embeddings

2.1 Concept-based Explainable Artificial Intelligence

The drive for predictive accuracy in machine learning has historically led to the widespread adoption of complex, highly parameterized models. These models, particularly deep neural networks, often function as "black boxes," providing high-performance predictions but little to no insight into the underlying decision-making process. While this opacity may be tolerable in some applications, it becomes a critical liability in high-stakes domains such as healthcare, finance, and autonomous systems Rudin [2019]. As noted in the context of mental health screening from social media, the lack of transparency complicates model validation, erodes trust among clinicians and end-users, and obscures the potential reliance on spurious correlations inherent in noisy, weakly-supervised data Bao et al. [2023]; Lawrence et al. [2024]. This limitation has catalyzed the rapid development of the field of Explainable Artificial Intelligence (XAI).

Explainable Artificial Intelligence (XAI) encompasses a broad suite of methodologies and techniques designed to render the behavior and outputs of machine learning models comprehensible to human stakeholders. The core objectives of XAI extend beyond mere technical interpretability; they include fostering trust, ensuring accountability, facilitating regulatory compliance, enabling effective human-AI collaboration, and providing avenues for debugging and improving model performance Arrieta et al. [2020]; Doshi-Velez and Kim [2017]. In healthcare applications the imperative for explainability is especially pronounced. Clinicians are ethically and legally responsible for their decisions, and any algorithmic tool intended to support diagnosis or screening must provide reasoning that can be scrutinized, challenged, and integrated into clinical judgment Amparore et al. [2023]. A model that cannot articulate "why" it reached a conclusion is of limited utility, regardless of its statistical accuracy, as it cannot be properly validated against medical knowledge or patient context.

XAI approaches are commonly categorized along two primary axes: post-hoc versus ante-hoc (or intrinsic) explainability, and model-specific versus model-agnostic techniques Molnar [2022]. Post-hoc explainability involves applying an external procedure to a trained, and often opaque, model to generate explanations for its predictions. Prominent examples include Local

Interpretable Model-agnostic Explanations (LIME), which approximates the model locally with an interpretable surrogate Ribeiro et al. [2016], and SHapley Additive exPlanations (SHAP), which attributes prediction outcomes to input features based on cooperative game theory Lundberg and Lee [2017]. For deep learning models, saliency maps and gradient-based methods highlight which input features (e.g., pixels or words) were most influential for a given prediction Selvaraju et al. [2017]; Simonyan et al. [2013]. While powerful for analysis, post-hoc methods have significant limitations. They are not guaranteed to be faithful representations of the model's true inner workings; the explanation is a separate construct that may misrepresent the actual computation Rudin [2019]. Furthermore, they often explain predictions in terms of low-level features (e.g., the importance of the word "sad"), which may not map coherently to the high-level, abstract concepts used by human experts (e.g., the symptom "self-criticalness").
In contrast, ante-hoc or intrinsically interpretable models are designed from their inception to be transparent. Their architecture enforces a structure where the reasoning process is explicitly represented and can be directly inspected. Examples include linear models with meaningful coefficients, decision trees with clear rule paths, and rule-based systems Rudin [2019]. The trade-off, historically, has been that such models were often assumed to be less expressive and thus less accurate than their black-box counterparts on complex tasks. However, this trade-off is increasingly being challenged, with research focusing on designing models that do not sacrifice significant performance for transparency Chen et al. [2022].

Among the most promising paradigms within intrinsic interpretability is Concept-based Explainable Artificial Intelligence (C-XAI). C-XAI directly addresses the semantic gap between low-level feature explanations and human understanding by structuring the model's reasoning around intermediate, human-interpretable units called concepts. A concept, in this framework, is a semantically meaningful attribute that is relevant to the task, such as "wheezing" in a respiratory diagnosis model, "microcalcification" in mammogram analysis, or "social withdrawal" in mental health assessment Koh et al. [2020]; Kim et al. [2018]. Rather than mapping inputs directly to outputs, a C-XAI model first detects or infers the presence and intensity of these concepts from the raw data. The final prediction is then made based on these concept activations.

This two-stage reasoning, from data to concepts, and from concepts to prediction, fundamentally aligns with human cognitive processes. Experts in any field, including clinicians, rarely reason directly from raw sensory data; they identify meaningful patterns (concepts) and combine them using learned rules or knowledge Miller [2019]. By mirroring this process, C-XAI provides explanations that are native to the model's operation: the explanation is the set of concept activations and their inferred contribution to the outcome. This offers several critical advantages:

- High-Level Abstraction: Explanations are provided in terms of clinical or domain-relevant constructs, making them immediately actionable for experts. A psychiatrist can evaluate whether a model's prediction of depression was driven by concepts like "Worthlessness" and "Fatigue," which are directly related to diagnostic criteria.

- Inherent Faithfulness: Since the concepts are part of the model's forward pass, the explanation is guaranteed to reflect the actual computation, unlike post-hoc approximations which can be unfaithful Yeh et al. [2020].

- Support for Interaction and Debugging: If a model makes an error, a human can inspect which concepts were incorrectly detected or weighted. This allows for targeted correction, either by improving the concept detector, refining the concept set, or adjusting the downstream logic Yuksekgonul et al. [2022].

- Causal Intervention and "What-If" Analysis: Concept-based models naturally support counterfactual reasoning. One can perturb concept values (e.g., "what if the level of social withdrawal were lower?") and observe the effect on the prediction, enabling exploration of causal relationships and alternative scenarios Mahinpei et al. [2018].

C-XAI thus represents a paradigm shift from explaining to build a transparent model. It reframes interpretability not as an external add-on to be applied after training, but as a core architectural principle to be baked into the learning process. This is particularly vital for applications in mental health technology, where the need for ethical, trustworthy, and clinically-relevant tools is paramount Ravenda et al. [2025]; Torous et al. [2025]. By forcing the model to express its reasoning through a vocabulary of human-understandable concepts, C-XAI provides a robust framework for developing AI systems that are not only accurate but also accountable, auditable, and aligned with expert knowledge.

2.1.1 What is a Concept?

The term concept within C-XAI denotes a human-interpretable intermediate representation that captures a specific, meaningful attribute or idea relevant to the task at hand. Formally defining a concept can be hard (Genone and Lombrozo [2012]), Poeta et al. [2023] defines them as "Human-interpretable high-level features of the input data that are important for the model's decision-making process". A concept c is usually associated to a value that quantifies the presence, intensity, or probability of that attribute This value is typically binary (e.g., the concept "Crying" is either present or absent), but different formulation have been hypothesized.
Concepts act as a semantic bottleneck, translating complex, often uninterpretable data distributions into a vocabulary that humans can reason about. As outlined in the comprehensive survey by Poeta et al. ([Poeta et al., 2023, Section 2.1]), their defining characteristics are:

- Human-Intelligibility: The meaning of the concept should be readily understandable without requiring deep knowledge of the model's architecture. For example, "texture" in an image or "self-referential language" in a text are intelligible concepts, whereas "activation of neuron 143 in layer 7" is not.

- Semantic Coherence: The concept should correspond to a coherent, nameable attribute or idea. It should not be an arbitrary or entangled combination of unrelated features.

- Relevance: The concept should be pertinent to the downstream prediction task. In a medical diagnosis model, relevant concepts would be clinical symptoms or signs, not unrelated visual or linguistic artifacts.

Poeta et al. Poeta et al. [2023] propose a categorization of concepts into four primary typologies based on their source and representation:

1. Symbolic Concepts: These are human-defined symbols or high-level attributes of the task, such as diagnostic criteria or visual features like "beak" in a bird classification task. They are typically predefined by domain experts and require auxiliary data with concept annotations.

2. Unsupervised Concept Bases: These are clusters of similar samples learned by the network. While not designed to resemble human-defined concepts, they may capture abstractions more understandable than raw features, such as a network learning a clusters of animals of the same colour in an animal classification task.
3. Prototypes: These are representative (parts of) training samples that encode peculiar traits in the network's weights. For instance, a part prototype might be a specific "beak" shape. Prototypes are explicitly encoded and thus belong to explainable-by-design models.
4. Textual Concepts: These are short textual descriptions of main classes , often generated by external models like LLMs and used in the form of numerical embeddings.

Furthermore, the nature of the concept-based explanation itself can vary. Poeta et al. [Poeta et al., 2023, Section 2.2] identify three distinct categories:

- Class-Concept Relation: Explains the relationship (e.g., importance or logical rule) between a concept and an output class.
- Node-Concept Association: Explicitly associates a concept with an internal unit or filter of the neural network, enhancing transparency.
- Concept Visualization: Highlights the input features that best represent a specific concept, akin to saliency maps but at a conceptual level.

A single C-XAI method may provide one or more of these explanation types, tailoring its output to different interpretability needs. The choice of concept type involves a fundamental trade-off. Symbolic concepts, directly grounded in expert knowledge, offer high fidelity to human understanding but might be dependent on costly annotations and may be incomplete. Conversely, automatically learned concepts (encompassing unsupervised bases, prototypes, and textual concepts) offer greater flexibility and scalability, as they can be discovered from data with minimal or no supervision. However, this automation risks producing concepts that lack a clear semantic mapping to established knowledge, potentially compromising the explainability guarantee (Yeh et al. [2020]). The specific challenges and methodologies for learning concepts automatically will be explored in the dedicated following section.

2.1.2 Automatically Learned Concepts

The paradigm of using predefined, expert-defined concepts offers a direct and robust connection to established domain knowledge. However, obtaining comprehensive, high-quality concept annotations for large datasets is often prohibitively expensive, time-consuming, or simply infeasible, particularly in specialized fields like mental health (Rajpurkar et al. [2019]). This practical limitation has spurred significant interest in methods for automatically learning concepts directly from data, with the goal of preserving explainability while alleviating the annotation burden.

Automatically learned concepts are not provided a priori but are instead discovered or generated by the machine learning model itself during training. The core idea is to identify recurring, semantically meaningful patterns within the model's internal representations or the input data that can serve as explanatory units.

A prominent family of approaches focuses on post-hoc concept discovery in already trained models. Methods like Automatic Concept-based Explanations (ACE) (Ghorbani et al. [2019]) and its successors (e.g., invertible concept-based explanation (ICE) Zhang et al. [2021], Concept recursive activation factorization for explainability (CRAFT) Fei et al. [2023]) operate by segmenting input samples (e.g., image patches), projecting these segments into the model's latent space, and applying clustering algorithms to group similar activations. The resulting clusters are presented as "discovered concepts," with their importance to a prediction quantified via specialized metrics. The appeal of such methods lies in their applicability to any pre-trained model without requiring architectural changes or concept labels, making them highly versatile for analysis and debugging (Yuksekgonul et al. [2022]).

Another major direction is the integration of concept discovery into explainable-by-design architectures. Here, the model is explicitly constrained or regularized during training to develop an interpretable concept layer. For instance, Self-Explaining Neural Networks (SENN) (Alvarez-Melis and Jaakkola [2018]) and Bottleneck Concept Learners (BotCL) (Wang et al. [2023]) employ auto-encoder-inspired architectures with sparsity or contrastive losses to encourage the learning of a disentangled, concept-like latent representation. Similarly, prototype-based models like ProtoPNet (Chen et al. [2019]) explicitly store prototypical parts of training examples (like a characteristic beak shape) in their weights and make predictions based on similarity to these prototypes, offering an intuitive "this looks like that" explanation (Chen et al. [2019]).

The most recent advancement leverages the generative power of LLMs to produce textual concepts. Frameworks like Label-free Concept Bottleneck Models (Label-free CBM) (Oikarinen et al. [2023]) and LaBO (Yang et al. [2023]) use LLMs (e.g. GPT-3) to generate candidate descriptive phrases for classes. These textual descriptions are then aligned with visual features using vision-language models (e.g., CLIP) to create a concept bottleneck without manual annotation. This approach is highly scalable and can produce rich, language-grounded concepts.

Automatically learned concepts have demonstrated value in several scenarios:

- Exploratory Analysis: They can reveal what patterns a black-box model has inadvertently learned, potentially uncovering spurious correlations or novel, data-driven constructs not captured by existing theory.

- Domains with Ill-Defined Concepts: In areas where human-understandable concepts are fuzzy, subjective, or unknown, data-driven discovery can suggest meaningful intermediate abstractions.

- Scalability: They bypass the need for exhaustive concept labeling, making concept-based explainability applicable to vastly larger and more diverse datasets.

- Performance: Models employing learned concepts, particularly through embeddings or prototypes, often do not suffer the performance penalty associated with strict symbolic bottlenecks and can match or even exceed the accuracy of their black-box counterparts (Espinosa Zarlenga et al. [2022]; Chen et al. [2019]).

Despite their appeal and strong performance, the use of automatically learned concepts introduces a fundamental tension with the core objective of C-XAI: providing faithful and actionable explanations grounded in human knowledge. The primary critique is that while these concepts may be statistically salient to the model, their semantic meaning is not guaranteed to be intelligible or aligned with domain expertise (Yeh et al. [2020]). A cluster of image patches or a prototype may be described by the model as an important concept, but if a clinician cannot reliably map it to a known clinical construct, its explanatory power is severely diminished. The explanation risks becoming a translation from one black box (the original prediction) into another black box (the learned concept), failing to bridge the semantic gap.

In the context of high-stakes applications like mental health screening, this trade-off is decisive. The paramount goal is not just to expose the model's internal statistics, but to enable validation against clinical knowledge and ethical standards. A concept labeled "Cluster 42" or an uninterpretable embedding vector cannot be audited for clinical relevance or bias. Consequently, for this thesis, which prioritizes explanations that are directly meaningful to clinical reasoning and diagnostic criteria, approaches relying on automatically learned concepts are considered suboptimal. While their technical prowess is acknowledged, their potential to obscure rather than clarify the model's reasoning in terms of established symptoms leads to their exclusion from the primary modeling strategy. The focus remains on predefined, clinically-grounded symbolic concepts, accepting the associated annotation challenges in exchange for guaranteed conceptual fidelity and interpretability.

2.1.3 Concept and Symptom Definition under Data Scarcity

A central assumption underlying concept based learning approaches is the availability of a set of human understandable concepts that can meaningfully mediate between raw data and model predictions. In clinical and psychological domains, however, the definition and annotation of such concepts present substantial practical challenges. High quality clinical annotations require expert knowledge, are time consuming to produce, and are often constrained by privacy regulations and ethical considerations. As a result, large scale datasets with reliable symptom level annotations are rare, particularly when compared to the abundance of unlabeled or weakly labeled textual data available from online sources (Rajpurkar et al. [2019]).

This scarcity has motivated increasing interest in alternative strategies for obtaining clinically relevant concepts without relying exclusively on manual expert annotation. Recent work has shown that large language models(Brown et al. [2020]), trained on vast amounts of textual data, can serve as effective tools for extracting, generating, or annotating mental health related concepts at relatively low cost. Empirical studies comparing LLM generated annotations with clinician judgments have reported competitive, and in some cases superior, performance across a range of mental health assessment tasks (Gilardi et al. [2023]). These findings suggest that LLMs can capture clinically meaningful patterns in language, even in the absence of task specific fine tuning (Poeta et al. [2023]).

The use of LLMs in this context should be understood as a form of weak or auxiliary supervision rather than as a replacement for clinical expertise. By prompting an LLM to identify symptoms, behavioral indicators, or psychologically relevant dimensions in user generated text, it is possible to generate auxiliary symptom annotations for subjects whose diagnostic labels are already known, while remaining scalable to large datasets. This approach is particularly attractive in settings where direct access to clinicians is limited or infeasible, and where the goal is to approximate expert reasoning rather than to replicate it exactly.

At the same time, relying on LLMs as end to end predictors in mental health applications raises well known concerns. Large language models are computationally expensive, difficult to deploy in resource constrained environments, and prone to producing outputs that are hard to audit or control. (Bommasani et al. [2021]; Bender et al. [2021]) Moreover, their internal reasoning processes are not transparent, and their predictions may vary depending on prompt formulation or contextual factors. These limitations complicate their direct use in high stakes decision making scenarios, where reproducibility, interpretability, and accountability are essential.

In this work, LLMs are therefore not used to predict mental health outcomes or user level labels, but solely to provide auxiliary annotations at the concept or symptom level on top of an existing, independently labeled dataset. The approach adopted in this thesis reflects instead a deliberate separation between concept generation and model inference. LLMs are employed offline to assist in the definition or identification of a limited set of clinically motivated concepts, drawing on established psychometric knowledge. Concretely, the LLM is used only during dataset construction to assign symptom level concept labels to existing already labeled eRisk subjects, while all user level predictions are produced exclusively by the concept based model described in later chapters. Once these concepts are fixed, the smaller and more compact model, which will be the primary focus of this thesis, is trained to operate on them, producing predictions through an explicit and interpretable intermediate representation. During inference, the LLM is no longer required, and the resulting system remains lightweight, efficient, and transparent in its decision making.

This hybrid strategy offers several advantages compared to approach entirely based on LLMs. It leverages the representational power of LLMs where they are most effective, namely in synthesizing knowledge and extracting high level semantic patterns, while avoiding their drawbacks at deployment time. At the same time, it preserves the benefits of concept based modeling, including the ability to inspect intermediate predictions, analyze failure modes, and relate model behavior to clinically meaningful constructs. From a practical perspective, this design also facilitates reproducibility and reduces computational overhead, making it more suitable for real world screening or monitoring applications.

Within this framework, concepts should not be interpreted as ground truth clinical diagnoses, but as structured proxies that encode relevant dimensions of mental health as expressed in language. Their role is to constrain and guide the learning process, not to provide definitive assessments. This perspective aligns with the broader view of mental health screening systems as decision support tools rather than autonomous diagnostic agents, and it emphasizes the importance of transparency and caution when operating under data scarcity and uncertainty.

A closely related challenge concerns the selection of the concept set itself. Defining an appropriate inventory of concepts is not a trivial design choice, particularly in mental health applications, where symptoms may overlap, manifest heterogeneously across individuals, or appear only implicitly in language. An arbitrary or purely data driven concept definition risks producing representations that are difficult to interpret clinically or that capture spurious correlations specific to the dataset, rather than stable and meaningful psychological constructs.

To address this issue, the concept space adopted in this work is grounded in the symptom definitions provided by the Beck Depression Inventory II (Beck et al. [1996]). This choice serves both interpretability and modeling considerations. On the one hand, BDI-II symptoms constitute a widely used and well validated psychometric vocabulary, facilitating direct interpretation of intermediate model representations and enabling meaningful communication with domain experts. On the other hand, these symptoms encode dimensions that are intrinsically relevant to depressive symptomatology and are therefore expected to provide a strong inductive bias for the learning task. Although symptom expression in social media text is often indirect and noisy, anchoring the concept set to BDI-II items constrains the model toward clinically meaningful patterns, reducing reliance on unconstrained latent representations while preserving flexibility in how symptoms are inferred from language.

2.2 Concept Bottleneck Models

A prominent family within this paradigm is that of Concept Bottleneck Models (CBM) (Koh et al. [2020]). These models are distinguished by an architectural constraint: all predictions must pass through an intermediate layer composed of (usually) predefined, symbolic concepts. The model is trained in two stages: firstly it learns to predict the concept values from the input data and then it uses those concept values to infer the target label. This architecture enables clear and tractable interpretations, since each decision is explicitly based on concept-level reasoning. For example, in the context of depression detection, the model may rely on intermediate variables such as "Self-dislike" or "Loss of interests", which mirror clinical diagnostic criteria and can be verified or contested by domain experts.

Formally, let $x \in \mathbb{R}^d$ denote an input (e.g. an image or a text representation), $c \in \mathbb{R}^k$ a vector of k human-specified concepts, and y the target label (which may be continuous or discrete). A concept bottleneck model consists of two functions:

- $g : \mathbb{R}^d \to \mathbb{R}^k$ that maps the input x to a predicted concept vector $\hat{c} = g(x)$,

- $f : \mathbb{R}^k \to \mathscr{Y}$ that maps the concepts to the final prediction $\hat{y} = f(\hat{c})$.

The overall model is therefore $\hat{y} = f(g(x))$, with the crucial property that the information from x flows entirely through the bottleneck layer $\hat{c}$.

During training, the model is provided with a dataset $\{(x^{(i)}, c^{(i)}, y^{(i)})\}_{i=1}^n$ containing input, concept, and label annotations. The goal is to learn g and f such that $\hat{c}$ aligns well with the true concepts c and $\hat{y}$ accurately predicts y. This is typically achieved by minimizing a composite loss:

$$\mathscr{L} = \mathscr{L}_y(\hat{y}, y) + \lambda \sum_{j=1}^{k} \mathscr{L}_c(\hat{c}_j, c_j), \tag{2.1}$$

where $\mathscr{L}_y$ is a task-specific loss (cross-entropy followed by softmax being a common choice in classification tasks (Goodfellow et al. [2016]; Bishop [2006])), $\mathscr{L}_c$ is a concept prediction loss, and $\lambda \geq 0$ controls the trade-off between task and concept accuracy. Common training strategies include independent, sequential, and joint optimization of g and f, which influence both predictive performance and the model's responsiveness to later intervention. Both intervention and the types of possible training strategies will be discussed in later dedicated sections.

The conceptual pipeline of a CBM is illustrated in Figure 2.1.

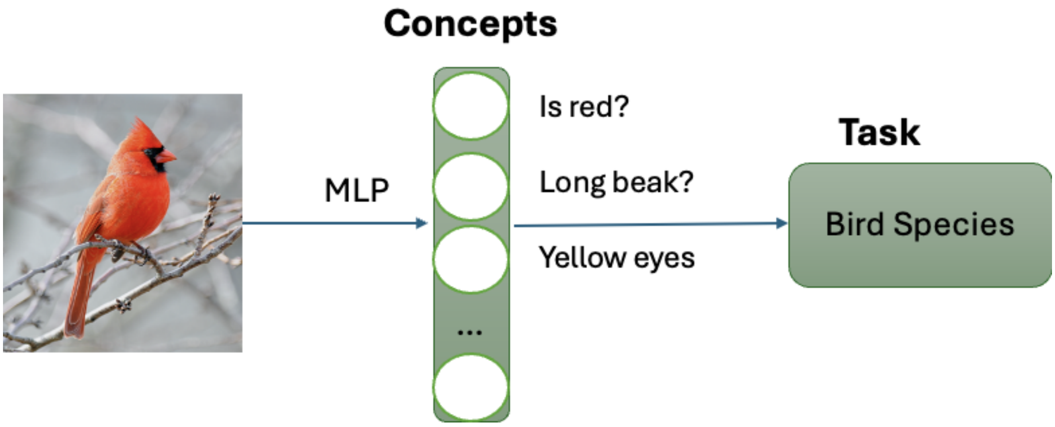

Figure 2.1. Concept Bottleneck Model for bird classification as described by Koh et al. [2020]

Despite their interpretability advantages, CBM exhibit several limitations that hinder their practical adoption. Most notably, the bottleneck constraint tends to degrade predictive performance, especially when the available concept set is incomplete, noisy, or insufficiently expressive (Koh et al. [2020]; Espinosa Zarlenga et al. [2022]). CBM assume that the entire predictive process can be mediated by symbolic concepts, which may not hold in domains like mental health where the mapping from data to concepts is subjective or complex. For instance, psychological constructs such as "anhedonia" or "negative self-perception" are inherently ambiguous and context-dependent; encoding them as fixed, discrete concepts may oversimplify their manifestation in natural language and ignore important nuances or interactions (Seonghwan and Jueun Mun [2025]). This reductionism can limit the model's ability to capture the full spectrum of symptom expression, particularly when signals are sparse, indirect, or embedded in noisy social media text.

Finally, the rigid bottleneck structure can inhibit the model's capacity to leverage information that is not encapsulated by the predefined concepts. In mental health, relevant signals may be subtle, multimodal, or expressed in ways that are not easily categorized into a fixed set of symptoms. By forcing all predictive information through a limited set of concepts, CBM may discard valuable evidence, leading to suboptimal performance compared to more flexible end-to-end architectures (Yuksekgonul et al. [2022]). This trade-off between interpretability and expressiveness is a central challenge in concept-based modeling and is especially salient in applications where both high accuracy and trustworthy explanations are required.

2.2.1 Different Kinds of Bottlenecks

The design and training of concept bottleneck models can vary significantly depending on how the mappings from input to concepts (g) and from concepts to target (f) are learned, as well as how the two stages interact during optimization. These variations lead to distinct modeling behaviors, particularly in terms of predictive accuracy, concept alignment, and suitability for later intervention. Below, we outline three primary formulations of concept bottlenecks commonly described in the literature: independent, sequential, and joint bottlenecks.

Independent Bottleneck

In the independent bottleneck approach, the two mappings g and f are learned separately and independently, each using the ground-truth annotations available in the training set. First, the concept predictor g is trained to predict the true concept values c from the input x by minimizing a concept-level loss:

$$\hat{g} = \arg\min_{g} \sum_{i=1}^{n} \sum_{j=1}^{k} \mathscr{L}_c(g_j(x^{(i)}), c_j^{(i)}). \tag{2.2}$$

Subsequently, the label predictor f is trained to predict the target y from the true concepts c, without any further reference to the input x:

$$\hat{f} = \arg\min_{f} \sum_{i=1}^{n} \mathscr{L}_y(f(c^{(i)}), y^{(i)}). \tag{2.3}$$

At test time, the model composes these functions: $\hat{y} = \hat{f}(\hat{g}(x))$. This decoupled training ensures that the concept predictor is optimized purely for concept accuracy, and the label predictor learns from clean, ground-truth concepts. However, a distributional mismatch arises at inference because $\hat{f}$ receives predicted concepts $\hat{c} = \hat{g}(x)$, which may differ statistically from the true c seen during its training. This mismatch can degrade task performance unless the concept predictor is exceptionally accurate.

Sequential Bottleneck

The sequential bottleneck approach also learns the two stages separately, but in a causally dependent manner. The concept predictor $\hat{g}$ is learned first, exactly as in the independent case. Then, instead of training $\hat{f}$ on ground-truth concepts, it is trained on the *predicted* concepts produced by $\hat{g}$:

$$\hat{f} = \arg\min_{f} \sum_{i=1}^{n} \mathscr{L}_y(f(\hat{g}(x^{(i)})), y^{(i)}). \tag{2.4}$$

This two-step procedure allows the label predictor to adapt to the specific errors and distribution of the concept predictor, potentially improving task performance under the actual inference conditions. However, because $\hat{f}$ never sees true concept values during training, the model may learn to rely on systematic biases in $\hat{g}$, making it less responsive to perfect concept corrections at test time.

Joint Bottleneck

In the joint bottleneck formulation, both g and f are trained simultaneously by optimizing a combined loss that balances concept prediction and target prediction:

$$\hat{f}, \hat{g} = \arg\min_{f,g} \sum_{i=1}^{n} \left[\mathscr{L}_y(f(g(x^{(i)})), y^{(i)}) + \lambda \sum_{j=1}^{k} \mathscr{L}_c(g_j(x^{(i)}), c_j^{(i)}) \right], \tag{2.5}$$

where $\lambda > 0$ controls the relative importance of concept alignment. This approach enables the concept representation to be refined in service of the final task, potentially improving task accuracy at the cost of deviating from the ground-truth concept definitions. As $\lambda \to 0$, the model reduces to a standard end-to-end model that ignores concept labels; as $\lambda \to \infty$, it approaches a purely concept-driven predictor similar to the independent bottleneck. The joint formulation offers a flexible trade-off between concept fidelity and task performance, but it also introduces the risk that the learned concepts become entangled or lose their intended interpretability if the task loss dominates.

Comparison and Trade-offs

Each bottleneck variant induces a different inductive bias and operational profile. Independent bottlenecks maximize concept interpretability and are most amenable to test-time concept intervention, because f is trained on true c and therefore expects clean concepts. However, they often suffer from compounded error due to the distribution shift between training and inference for f. Sequential bottlenecks mitigate this shift by training f on predicted concepts, usually improving task accuracy under normal operation but reducing intervention fidelity. Joint bottlenecks offer a tunable compromise, allowing the model to learn concepts that are both predictive and aligned, though the alignment may be partial and the concepts may drift from their original semantic definitions.

In the context of mental health screening from noisy text, the choice of bottleneck type interacts critically with the quality and completeness of the concept set, the noisiness of concept annotations, and the ultimate goal of supporting human-in-the-loop analysis. A model intended primarily for explanatory transparency might favor an independent or high-λ joint bottleneck, whereas a model optimized for standalone screening might benefit from a sequential or low-λ joint formulation.

2.3 Concept Embedding Models

While Concept Bottleneck Models enforce a strict architectural separation between concept prediction and target inference, a more flexible family of approaches, often termed Concept Embedding Models (CEM), proposes to represent each concept as a high-dimensional vector embedding rather than a binary representation. This extension, introduced by Espinosa Zarlenga et al. Espinosa Zarlenga et al. [2022], aims to overcome the accuracy-interpretability trade-off inherent in standard CBMs, particularly in settings where concept annotations are incomplete or noisy. By embedding concepts in a continuous semantic space, CEMs can capture richer concept semantics while preserving the ability to intervene and explain predictions in terms of human-understandable concepts.

In a Concept Embedding Model, each concept c_i is represented by two learned embedding vectors: $\hat{\mathbf{c}}_i^+ \in \mathbb{R}^m$ for the active state (concept true) and $\hat{\mathbf{c}}_i^- \in \mathbb{R}^m$ for the inactive state (concept false). These embeddings are generated from an intermediate latent representation $\mathbf{h} = \psi(x)$ via concept-specific linear transformations:

$$\hat{\mathbf{c}}_i^+ = \phi_i^+(\mathbf{h}) = a(W_i^+\mathbf{h} + \mathbf{b}_i^+), \quad \hat{\mathbf{c}}_i^- = \phi_i^-(\mathbf{h}) = a(W_i^-\mathbf{h} + \mathbf{b}_i^-), \tag{2.6}$$

where a is a non-linear activation function. A shared scoring function $s : \mathbb{R}^{2m} \to [0,1]$ then predicts the probability $\hat{p}_i = s([\hat{\mathbf{c}}_i^+, \hat{\mathbf{c}}_i^-]^\top)$ of concept c_i being active. The final concept embedding for c_i is computed as a convex combination of the two semantic embeddings, weighted by the predicted probability:

$$\hat{\mathbf{c}}_i = \hat{p}_i \hat{\mathbf{c}}_i^+ + (1 - \hat{p}_i)\hat{\mathbf{c}}_i^-. \tag{2.7}$$

All k concept embeddings are concatenated into a bottleneck vector $\hat{\mathbf{c}} = [\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_k] \in \mathbb{R}^{k \cdot m}$, which is then passed to a label predictor f to obtain the final output $\hat{y} = f(\hat{\mathbf{c}})$.
This architecture is trained end-to-end by minimizing a combined loss:

$$\mathscr{L} = \mathbb{E}_{(x,y,c)}\big[\mathscr{L}_{\text{task}}\big(y, f(\hat{\mathbf{c}})\big) + \alpha \mathscr{L}_{\text{CrossEntr}}\big(c, \hat{p}(x)\big)\big], \tag{2.8}$$

where $\alpha > 0$ balances task accuracy against concept prediction , in similar fashion to how discussed for joint bottlenecks in the previous section. The design encourages the model to learn disentangled and semantically meaningful embeddings: $\hat{\mathbf{c}}_i^+$ should capture features indicative of the concept's presence, while $\hat{\mathbf{c}}_i^-$ should capture its absence.
A key feature of CEMs is their support for test-time interventions. To correct a mispredicted concept c_i, one can simply set $\hat{p}_i := c_i$ (where $c_i \in \{0,1\}$ is the ground-truth) and recompute the mixture embedding accordingly. This switches the bottleneck representation from the predicted mixture to the embedding corresponding to the true concept state, directly affecting the downstream prediction. To improve intervention effectiveness, Espinosa Zarlenga et al. Espinosa Zarlenga et al. [2022] also propose RandInt, a regularization technique that randomly intervenes on concepts during training with probability p_{int}, exposing the model to corrected concept states and encouraging the embeddings to remain responsive to such changes.
The conceptual pipeline of a Concept Embedding Model is illustrated in Figure 2.2.

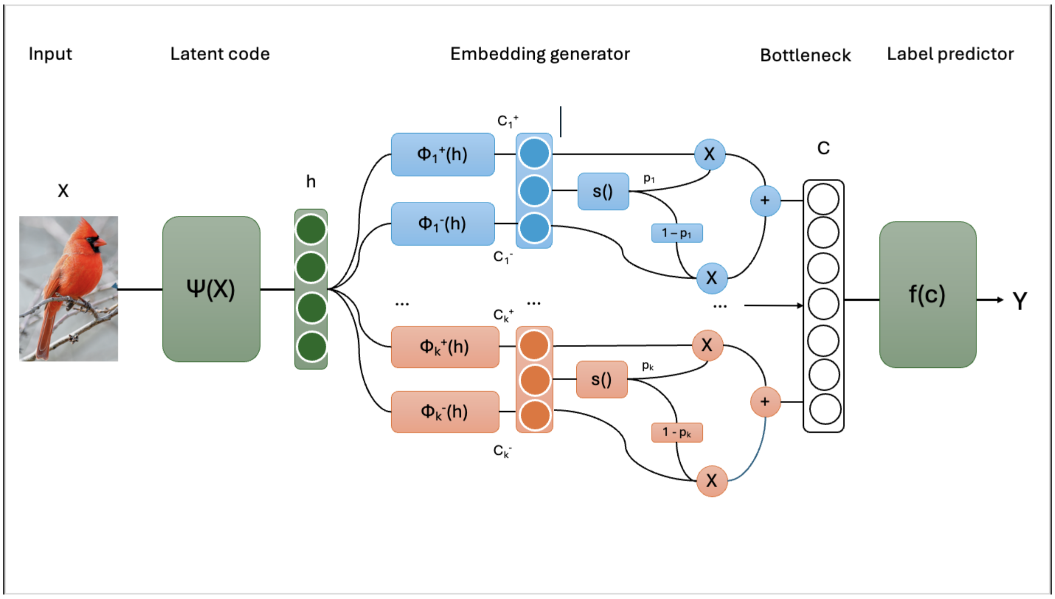

Figure 2.2. CEM pipeline as presented by Espinosa Zarlenga et al. [2022]

Compared to scalar-based CBMs, CEMs provide higher capacity in the bottleneck without sacrificing interpretability, as each dimension of the embedding remains supervised by concept labels. This allows the model to retain more input information, mitigating the information bottleneck effect observed in vanilla CBMs. Empirically, CEMs have been shown to achieve competitive or superior task accuracy compared to standard neural models, while maintaining high concept alignment and supporting more effective interventions Espinosa Zarlenga et al. [2022]. These properties make CEMs particularly suitable for applications where concept annotations are incomplete or noisy, such as mental health screening from social media text, where symptom concepts may be only partially observed or ambiguously expressed.

2.3.1 On Intervention

The ability to intervene on predicted concepts is a defining characteristic of concept-based models, distinguishing them from post-hoc interpretability methods. Intervention refers to the process of manually modifying the predicted concept values $\hat{c}$ at test time, typically by replacing them with ground-truth or alternative concept values, and observing how the final prediction $\hat{y}$ changes as a result. This mechanism enables human-in-the-loop correction, interactive debugging, and the generation of counterfactual explanations, making it particularly valuable in high-stakes domains such as healthcare.

From a causal perspective, interventions in concept bottleneck models can be understood as performing a *do-operator* on the internal concept variables Pearl [2000]. Rather than intervening on the real-world concepts (which would require altering the input itself), the intervention is applied to the model's internal representation, asking: “If the model believed concept c_i had a different value, how would its prediction change?” This allows users to probe the model's dependence on specific concepts and to test hypothetical scenarios without needing to collect new data. However, it is important to note that such interventions reflect the model's internal logic, not necessarily true causal relationships in the world. As noted by Koh et al. [2020], a concept bottleneck model may correctly update its prediction when a concept is corrected even if the concept is not causally related to the target, it needs only to be predictive.

Interventions can be performed in different ways depending on the model architecture. In scalar-based concept bottleneck models (Boolean or Fuzzy CBMs), intervening on a concept c_i means directly setting its predicted probability $\hat{p}_i$ to 0 or 1. In Concept Embedding Models, intervention involves swapping the mixed concept embedding $\hat{\mathbf{c}}_i$ with either $\hat{\mathbf{c}}_i^+$ or $\hat{\mathbf{c}}_i^-$, depending on the desired state. The effect of an intervention can be measured by the change in the final prediction, and when interventions are performed using ground-truth concept values, they often lead to improved task accuracy, simulating a collaboration between the model and a human expert.

A key factor influencing intervention effectiveness is the degree to which the model's label predictor f is trained on concept distributions that match the intervened distribution. In independent bottleneck models, f is trained on true concept values c, but at test time it receives predicted concepts $\hat{c}$, creating a distribution shift that can reduce intervention impact. In sequential and joint bottlenecks, f is trained on predicted concepts, aligning better with test-time inputs but potentially making the model less responsive to perfect corrections. This trade-off is illustrated in the empirical results of Koh et al. [2020], where independent bottlenecks performed best under full concept replacement, while joint models sometimes suffered when concept predictions were too heavily tuned for task performance at the expense of concept fidelity.

To enhance a model's responsiveness to interventions, training-time regularization techniques such as *RandInt* Espinosa Zarlenga et al. [2022] can be employed. During training, RandInt randomly replaces predicted concept probabilities $\hat{p}_i$ with ground-truth values c_i with a fixed probability p_{int}. This exposes the label predictor to both predicted and corrected concept distributions, encouraging it to rely meaningfully on concept values and improving its robustness to test-time corrections. However, the choice of p_{int} involves a balance: too low a rate may not sufficiently teach the model to respond to interventions, while too high a rate may degrade normal task performance by overly distorting the training signal. Empirical studies suggest that moderate values (e.g., $p_{\text{int}} = 0.25$) often yield a favorable compromise Espinosa Zarlenga et al. [2022].

In practice, interventions can be applied selectively, correcting only a subset of concepts, or exhaustively, replacing all concept predictions. The order in which concepts are intervened upon can also affect performance; a fixed ordering based on validation-set impact may be used to maximize accuracy gains with fewer queries. Moreover, interventions need not be limited to ground-truth corrections; they can also be used to explore "what-if" scenarios, such as toggling a concept to see whether the model's prediction flips, thereby generating local counterfactual explanations.

While interventions are a powerful tool for model interpretability and human–AI collaboration, their reliability depends critically on the alignment between predicted concepts and true concepts. If concept predictions are inaccurate or semantically misaligned, interventions may produce misleading changes in the output. Therefore, high concept accuracy remains a prerequisite for meaningful intervention. Additionally, in models with side channels (e.g., hybrid or embedding-based architectures), interventions on concepts may be partially offset by information flowing through unsupervised or direct pathways, reducing their effect. This highlights the importance of architectural choices when intervention fidelity is a primary objective.

2.3.2 Intervention-Aware Concept Embedding Models

The notion of intervention-aware training was introduced by Espinosa Zarlenga et al. building on Concept Embedding Models Espinosa Zarlenga et al. [2023]. The proposed Intervention-aware Concept Embedding Models (IntCEMs) extend the standard CEM framework with the explicit goal of improving a model's responsiveness to test-time concept interventions. Rather than treating interventions as an external correction mechanism applied after training, this paradigm incorporates simulated intervention trajectories directly into the learning process.

Formally, an IntCEM is defined as a triplet (g, f, ψ), where g denotes the concept encoder, f the label predictor, and ψ a learnable intervention policy. Given an input $\mathbf{x}$ and ground-truth concept labels $\mathbf{c}$, the model first produces concept embeddings $\hat{\mathbf{c}} = g(\mathbf{x})$ and concept probabilities $\hat{\mathbf{p}}$. During training, an intervention trajectory of length T is sampled. An initial intervention mask $\mu^{(0)} \sim p(\mu)$ is drawn, and for each step $t = 1, \ldots, T$, the policy ψ outputs a distribution over concepts from which the next intervention $\eta^{(t)}$ is sampled. The intervention mask is updated as $\mu^{(t)} = \mu^{(t-1)} + \eta^{(t)}$, and the concept bottleneck representation is updated through the CEM intervention mechanism

$$\hat{\mathbf{c}}_i := (\mu_i \tilde{c}_i + (1-\mu_i)\hat{p}_i)\hat{\mathbf{c}}_i^{+} + \big(1 - (\mu_i \tilde{c}_i + (1-\mu_i)\hat{p}_i)\big)\hat{\mathbf{c}}_i^{-}, \tag{2.9}$$

where $\tilde{c}_i$ denotes the ground-truth value for concept i when an intervention occurs, and $\hat{p}_i$ is the model's predicted probability otherwise.

Training optimizes a composite loss

$$\mathscr{L}(\mathbf{x}, \mathbf{c}, y, \mathscr{T}) := \lambda_{\text{roll}} \mathscr{L}_{\text{roll}}(\mathbf{x}, \mathbf{c}, y, \mathscr{T}) + \mathscr{L}_{\text{pred}}(\mathbf{x}, \mathbf{c}, y, \mu^{(0)}, \mu^{(T)}) + \lambda_{\text{concept}} \mathscr{L}_{\text{concept}}(\mathbf{c}, \hat{\mathbf{p}}), \tag{2.10}$$

where $\mathscr{T} = \{(\mu^{(t-1)}, \eta^{(t)})\}_{t=1}^{T}$ denotes the sampled intervention trajectory. The rollout loss $\mathscr{L}_{\text{roll}}$ trains the policy ψ to approximate an oracle intervention strategy, often defined as selecting the concept whose correction most increases the probability of the true class. The prediction loss penalizes misclassification of the target label y before and after the intervention sequence, assigning a higher weight to post-intervention predictions

$$\mathscr{L}_{\text{pred}} := \frac{\mathrm{CE}\big(y, f(\tilde{g}(\mathbf{x}, \mu^{(0)}, \mathbf{c}))\big) + \gamma^{T} \mathrm{CE}\big(y, f(\tilde{g}(\mathbf{x}, \mu^{(T)}, \mathbf{c}))\big)}{1 + \gamma^{T}}, \tag{2.11}$$

with $\gamma \geq 1$. Finally, $\mathscr{L}_{\text{concept}}$ is the standard cross-entropy loss on concept predictions. Differentiable sampling via Gumbel-Softmax (Jang et al. [2017]) enables gradients to propagate through intervention decisions, allowing joint optimization of the concept encoder, label predictor, and intervention policy.

While this framework is appealing in settings where concepts correspond to manipulable and causally meaningful variables, its applicability is less clear in clinical domains such as mental health. In this work, concepts are defined as symptom-level indicators grounded in established psychometric instruments and are intended to function as structured, interpretable proxies rather than as causal mechanisms. As discussed in the previous chapter, the objective is not to model how manipulating a symptom would causally alter an outcome, but to leverage clinically motivated concepts to constrain and interpret prediction.

From this perspective, concept interventions serve primarily as a diagnostic or stress-testing tool, allowing one to probe model sensitivity and, in some cases, marginally improve predictive performance by correcting unreliable concept predictions. They are not required to be optimal, nor to generalize across arbitrary intervention sequences. Very recent work by Espinosa et al. Espinosa Zarlenga et al. [2025] highlights that intervention-aware training can be fragile under distribution shift, showing that learned intervention policies may suffer from leakage poisoning when concept correlations differ between training and test data. This issue is particularly salient in weakly supervised settings with incomplete or noisy concept annotations, as is the case for social media-based mental health data.

In preliminary experiments, we explored adapting an intervention-aware training scheme to our Concept Embedding Model baseline. However, we observed no consistent performance gains over standard CEM training. Given the absence of a causal interpretation for concept interventions in this setting, and the additional sensitivity of learned intervention policies to distribution mismatch, we opted not to include an intervention-aware variant in the final experimental comparison.

Instead, this thesis treats interventions as an optional post-training mechanism for model analysis and performance probing, rather than as a core component of the learning objective. The development of intervention policies that remain robust under distribution shift and weak concept supervision remains an important open problem, but lies beyond the scope of the present work.

Chapter 3

Retrieval and Aggregation as a Theoretical Problem

3.1 User Level Prediction as a Multi Instance Learning Problem

Mental health screening from social media data is inherently a user level prediction task, while the available observations are typically post level textual fragments, as it is in the case of eRisk dataset(Losada and Crestani [2017]). Each individual is associated with a variable length sequence of posts, produced over extended periods of time and under heterogeneous contextual conditions. The learning problem therefore consists in mapping a collection of textual instances to a single binary label indicating the presence or absence of a mental health condition. This setting naturally corresponds to the multi instance learning (MIL) paradigm, in which labels are provided only at the bag level rather than at the instance level Dietterich et al. [1997b]; Amores [2013].

In classical MIL formulations, a bag is considered positive if at least one instance within it satisfies a target property, while negative bags contain no such instances. Although this assumption is a simplification, it closely reflects the structure of mental health signals in user generated text. Depressive symptoms are rarely expressed consistently across all posts produced by an individual. Instead, they tend to appear intermittently, often embedded in otherwise neutral or unrelated content, and may be expressed implicitly rather than explicitly. As a result, only a small subset of a user's posts may carry diagnostically relevant information, while the majority contribute noise or irrelevant background context.

A common baseline approach for aggregating post level representations into a user level embedding consists of naive pooling strategies, such as averaging or summing sentence embeddings across all posts. While computationally simple, these approaches implicitly assume that all instances contribute equally to the final prediction. In the context of mental health data, this assumption is problematic. Averaging tends to dilute sparse but highly informative signals, causing posts that strongly express depressive symptoms to be overshadowed by large volumes of neutral content. Sum based pooling partially mitigates this effect by accumulating evidence across posts, but it remains sensitive to the number of instances and favors users who express mild signals repeatedly over those who express strong signals sparsely.

More expressive aggregation mechanisms have been proposed to address these limitations, most notably attention based pooling Ilse et al. [2018]; Lin et al. [2017b]. Attention mechanisms assign instance specific weights, allowing the model to focus on a subset of posts deemed most relevant for the task. From a MIL perspective, attention can be interpreted as a soft instance selection mechanism that learns to approximate the latent relevance of each post. However, when attention weights are learned purely end to end from user level labels, they are prone to instability under class imbalance and weak supervision, and may converge to spurious heuristics that correlate with dataset specific artifacts rather than clinically meaningful signals.

In mental health screening scenarios, the difficulty of instance selection is exacerbated by the absence of reliable post level annotations. Posts are not labeled with symptoms, and even experts may disagree on whether a given fragment constitutes evidence of a particular condition. This makes it challenging to directly supervise attention mechanisms or instance classifiers. As a result, aggregation strategies that explicitly encode assumptions about signal sparsity and relevance are particularly appealing. Rather than treating all posts as equally informative, or relying entirely on learned attention, these approaches introduce inductive biases that favor the selection of posts exhibiting strong semantic alignment with clinically motivated criteria.

Viewed through this lens, user level prediction from social media is best understood not as a standard text classification problem, but as a structured MIL problem characterized by extreme instance imbalance, high noise, and sparse discriminative signals. This perspective motivates the need for principled retrieval and aggregation mechanisms that can isolate and amplify diagnostically relevant content before downstream modeling.

3.2 Semantic Retrieval with Concept Queries

Framing user level prediction as a multi instance learning problem raises a central question, namely how to identify which instances within a user's post history are potentially relevant for downstream modeling. In the absence of post level supervision, this problem cannot be solved reliably through direct instance classification. Instead, it requires an intermediate mechanism for estimating semantic relevance based on external structure or prior knowledge. In this work, this structure is provided by clinically motivated concepts, which are used as semantic queries to retrieve relevant textual instances.

Concepts in concept based models are typically introduced as intermediate variables mediating between inputs and predictions. However, they can also be interpreted as high level semantic descriptors that define regions of interest in the representation space. Under this interpretation, each concept corresponds not to a discrete label attached to individual posts, but to a semantic direction that captures the linguistic expression of a particular symptom or psychological dimension. Given a shared embedding space for posts and concepts, relevance can then be computed as similarity between their vector representations.

Embedding based retrieval offers a natural way to implement this idea. Sentence embedding models such as SBERT map variable length text to fixed dimensional vectors that preserve semantic similarity under cosine distance Reimers and Gurevych [2019]. By embedding both posts and concept descriptions into the same vector space, it becomes possible to compute a relevance score for each post with respect to each concept. This approach avoids brittle keyword matching and allows retrieval to account for paraphrasing, implicit expressions, and contextual variation, all of which are common in mental health related language.

Formally, let $\mathbf{p}_i \in \mathbb{R}^d$ denote the embedding of post i, and let $\mathbf{c}_j \in \mathbb{R}^d$ denote the embedding of concept j. A relevance score can be computed as the cosine similarity

$$s_{ij} = \frac{\mathbf{p}_i \cdot \mathbf{c}_j}{\|\mathbf{p}_i\| \, \|\mathbf{c}_j\|}. \tag{3.1}$$

These scores provide a continuous measure of semantic alignment between posts and concepts, which can be used to rank or weight instances within each user's post history. Importantly, this relevance signal is independent of the user level label and does not require concept annotations at the post level, making it well suited to weakly supervised settings.

A common strategy in embedding based retrieval systems consists of applying a threshold to similarity scores, retaining only posts whose relevance exceeds a predefined value. While intuitive, this approach is poorly suited to mental health data. First, similarity distributions vary substantially across users, concepts, and posting behavior, making it difficult to define a global threshold that generalizes across subjects. Second, thresholding introduces a hard decision boundary that discards potentially informative posts whose relevance lies just below the cutoff. Third, in highly imbalanced datasets, aggressive thresholding risks eliminating the very sparse signals that distinguish positive cases from controls.

An alternative approach consists of ranking posts according to their relevance scores and selecting a fixed number of top instances per user. This strategy ensures a consistent representation size and avoids reliance on absolute similarity values. More importantly, it shifts the focus from deciding whether a post is relevant in an absolute sense to deciding which posts are most relevant relative to the rest of a user's history. This relative notion of relevance is particularly appropriate in mental health screening, where symptom expression is subjective, contextual, and unevenly distributed across individuals.

By treating concepts as semantic queries and retrieval as a ranking problem in embedding space, this formulation establishes a principled interface between unstructured text and structured concept based models. However, it leaves open a critical modeling choice, namely how relevance scores across multiple concepts should be combined when ranking or weighting posts. Different aggregation strategies implicitly encode different assumptions about how mental health signals manifest in language.

3.2.1 BERT and Sentence-BERT for Semantic Retrieval

Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. [2018] represents a major milestone in natural language processing. By introducing deep bidirectional pretraining based on masked language modeling and next sentence prediction, BERT demonstrated that large transformer models pretrained on unlabeled text could achieve substantial performance gains across a wide range of downstream tasks. Formally, BERT takes a sequence of tokens $\mathbf{x} = [x_1, \ldots, x_n]$ as input and produces contextualized token representations $\mathbf{h}_1, \ldots, \mathbf{h}_n$. The model architecture consists of multiple transformer layers, each performing multi head self attention followed by position wise feedforward networks. The output for the special token [CLS] is often used as a sentence level representation, though it is not explicitly trained to capture semantic similarity. Since its introduction, BERT and its variants have become foundational components of modern NLP systems, and an extensive body of literature has explored its architectural properties, extensions, and applications. A comprehensive analysis of BERT and its successors is beyond the scope of this thesis, and the discussion here is limited to aspects directly relevant to semantic retrieval and representation learning.

In its original formulation, BERT is designed to produce contextualized token representations rather than fixed size sentence embeddings. While sentence level representations can be obtained by pooling token embeddings, for example by using the CLS token or mean pooling over tokens, these representations are not explicitly optimized to reflect semantic similarity between sentences. As a consequence, naïvely applying cosine similarity to pooled BERT outputs often yields suboptimal performance for retrieval and similarity based tasks. This limitation becomes particularly salient in scenarios that require comparing a large number of short texts, such as social media posts, against a set of semantic queries or concepts.

Sentence-BERT (SBERT) was introduced specifically to address this gap (Reimers and Gurevych [2019]). SBERT modifies the BERT architecture by placing it in a siamese or triplet network configuration and training it using objectives that directly optimize sentence level semantic similarity. In contrast to standard BERT, SBERT uses a siamese network structure where two identical BERT models share weights. Given a pair of sentences (s_a, s_b), each is passed through the shared BERT model to obtain pooled embeddings $\mathbf{u} = \text{pooling}(\text{BERT}(s_a))$ and $\mathbf{v} = \text{pooling}(\text{BERT}(s_b))$. The model is trained with objectives such as the softmax loss over the cosine similarity between $\mathbf{u}$ and $\mathbf{v}$, or with triplet loss that encourages semantically similar sentences to be closer than dissimilar ones. The resulting embeddings are optimized so that $\cos(\mathbf{u}, \mathbf{v})$ correlates with semantic relatedness, enabling efficient retrieval via nearest neighbor search in the embedding space. This design enables SBERT to produce fixed dimensional sentence embeddings that can be efficiently compared using cosine similarity. Importantly, SBERT decouples representation learning from pairwise comparison at inference time, allowing embeddings to be precomputed and reused, which makes large scale retrieval and nearest neighbor search computationally feasible.

For tasks involving semantic retrieval, SBERT offers two key advantages over standard BERT representations. First, the embedding space learned by SBERT is explicitly structured such that semantically related sentences are close under cosine similarity, while unrelated sentences are pushed apart. Second, SBERT enables efficient retrieval pipelines in which a large collection of candidate texts can be embedded once and subsequently queried without repeated forward passes through a cross encoder. These properties make SBERT particularly well suited for applications where relevance must be assessed across many short texts under computational constraints. SBERT implementation to compute cosine similarity can be seen in image 3.1

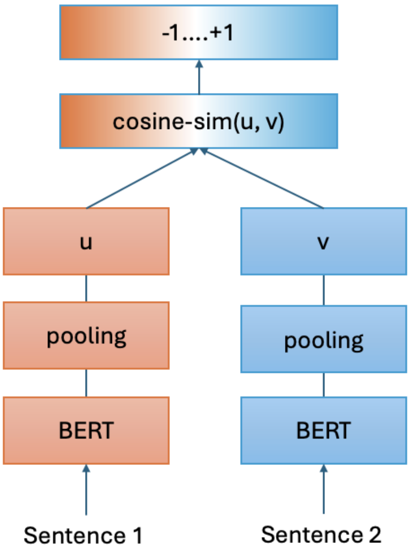

Figure 3.1. SBERT used to compute cosine similarity between sentences as presented in Reimers and Gurevych [2019].

In the context of this work, the task involves retrieving posts that are semantically related to a set of clinically motivated concepts and aggregating evidence across a user's posting history. This setting naturally aligns with an embedding based retrieval paradigm, in which posts and concepts are mapped into a shared semantic space and compared using a similarity measure. Using standard BERT representations would require either computationally expensive cross encoding or reliance on sentence embeddings that are not optimized for similarity comparison. SBERT, by contrast, provides a principled and efficient mechanism for representing both posts and concepts in a common embedding space, making it a natural choice for the retrieval and aggregation strategies explored in the remainder of this chapter. For these reasons, SBERT is adopted as the backbone for semantic representation in this thesis. This choice reflects a practical trade off between representational quality, computational efficiency, and methodological clarity, rather than a claim of universal superiority over alternative transformer based models. Other embedding methods and comparative approaches could be considered, but SBERT provides a well established and widely validated baseline for semantic retrieval in noisy, large scale text settings.

3.3 Aggregation Strategies: Sum vs Max Similarity

Once post level relevance scores have been computed with respect to a set of semantic concepts, a central modeling choice concerns how this information should be aggregated at the user level. In the context of user generated text, each individual is associated with a collection of posts that vary widely in length, topic, and psychological relevance. Aggregation therefore plays a critical role in determining which signals are amplified, which are suppressed, and how evidence accumulates across a user's posting history.

Two simple yet widely used aggregation strategies are mean and maximization of relevance scores. Despite their apparent simplicity, these operators encode fundamentally different assumptions about how information is distributed across instances and how it should contribute to the final representation. Understanding these assumptions is essential for interpreting their behavior in noisy and weakly supervised settings such as mental health screening from social media.

Summation based aggregation treats relevance as an additive quantity. Under this formulation, posts that exhibit moderate similarity to several concepts contribute cumulatively to the user representation. This implicitly favors users whose content exhibits diffuse or repeated weak signals across multiple posts or concepts. From a probabilistic perspective, summation can be interpreted as evidence accumulation under an assumption of conditional independence, where multiple weak indicators jointly increase confidence. Similar intuitions appear in traditional bag of words models, average pooling in neural architectures, and mean aggregation strategies commonly used in multi instance learning Dietterich et al. [1997b]; Amores [2013]. However, summation also introduces a sensitivity to volume. Users who post frequently, or whose posts contain broadly generic language, may accumulate higher aggregate scores even if no single post is strongly indicative of a clinically relevant concept. In domains characterized by high noise and topical diversity, this can lead to representations dominated by quantity rather than diagnostic salience. Prior work in multi instance learning has noted that such pooling strategies can obscure sparse but decisive instances when positive evidence is concentrated in only a small subset of examples Maron and Lozano-Pérez [1998].

Max based aggregation embodies a markedly different inductive bias. Instead of accumulating evidence across all posts, it selects the strongest signal observed for each user, effectively assuming that the presence of a single highly relevant instance may be sufficient to characterize the bag. This formulation aligns with the classical definition of multi instance learning, in which a bag is labeled positive if at least one instance is positive Dietterich et al. [1997b]. In this sense, maximization can be interpreted as modeling a disjunctive relationship between instances, prioritizing salience over frequency.

From a representational standpoint, max aggregation emphasizes specialization. It privileges posts that are highly aligned with at least one clinically motivated concept, even if the remainder of the user's content is unrelated or neutral. This behavior is particularly appealing in mental health settings, where symptom expression is often sparse, episodic, and context dependent. Individuals may produce long sequences of everyday content interspersed with occasional posts that directly or indirectly reflect psychological distress. In such cases, aggregation strategies that dilute strong signals with large amounts of irrelevant text risk suppressing precisely the evidence that is most diagnostically informative.

Related intuitions have emerged in attention based multi instance learning, where mechanisms are designed to assign disproportionate weight to a small number of salient instances Ilse et al. [2018]. While attention models generalize both summation and maximization through learned weighting, empirical studies have shown that, in highly noisy settings, attention often collapses toward a small number of dominant instances, effectively approximating a max like behavior. This observation suggests that explicit max aggregation may serve as a strong inductive prior when learning reliable attention weights is difficult due to limited supervision or severe class imbalance.

In this work, multiple aggregation formulations were explored, including mean/summation, maximization, and additional variants, as discussed later in the experimental section. Empirically, max based aggregation consistently yielded stronger performance than summation based alternatives. At this stage, the focus is not on reporting these results in detail, but on highlighting that this outcome reflects a meaningful alignment between the aggregation operator and the structural properties of the data. A more thorough quantitative and qualitative analysis of this behavior is provided in the later chapters, where the implications of max based retrieval for predictive performance and interpretability are examined in depth.

These considerations suggest that aggregation is a modeling decision that encodes assumptions about how evidence manifests in user generated text. In domains where relevant signals are sparse, concentrated, and heterogeneous, max based aggregation offers a principled mechanism for preserving salient information that might otherwise be obscured by more diffuse pooling strategies.

3.3.1 Other Aggregation Strategies

Beyond the sum (or mean) and max operators, numerous alternative aggregation strategies have been proposed in the literature, each encoding different assumptions about the relationship between instances and the bag label. While sum and max are the most commonly used due to their simplicity and interpretability, several other approaches have been explored, even if they may be suboptimal in certain settings.

One class of alternatives includes statistical pooling operators such as log-sum-exp (LSE) and generalized mean pooling. Log-sum-exp, defined as $\frac{1}{r}\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{r\cdot s_i}\right)$, provides a smooth interpolation between max and mean pooling as the parameter r varies (Pinheiro and Collobert [2015]). For large positive r, LSE approximates the maximum; for $r = 0$, it recovers the log of the mean. This allows the model to learn a pooling behavior that balances between emphasizing the strongest instance and incorporating contributions from all instances. Generalized mean pooling, defined as $\left(\frac{1}{n}\sum_{i=1}^{n} s_i^p\right)^{1/p}$, similarly interpolates between min, mean, and max as p varies (Boulch et al. [2017]). These operators offer flexibility but introduce additional hyperparameters and can be less interpretable than pure sum or max.

Another line of work employs attention-based aggregation, where instance-level scores are weighted by a learned attention mechanism before summation. Formally, the bag representation is computed as $\sum_{i=1}^{n} \alpha_i h_i$, where $\alpha_i = \mathrm{softmax}(w^\top \tanh(V h_i))$ or a similar parameterized function (Ilse et al. [2018]). This allows the model to dynamically emphasize instances deemed most relevant to the bag label. However, in noisy, weakly-supervised settings with severe class imbalance, attention mechanisms can struggle to learn meaningful weights and may collapse toward uniform or degenerate distributions, effectively reducing to sum or max pooling. Regularized or constrained attention variants have been proposed to mitigate this issue, but they increase model complexity.

Noisy-or aggregation is another probabilistic formulation commonly used in multiple-instance settings, particularly in medical imaging and document classification. It models the bag probability as $P(Y = 1|B) = 1-\prod_{i=1}^{n}(1-P(y_i = 1))$, where $P(y_i = 1)$ is the probability that instance i is positive (Maron and Lozano-Pérez [1998]). This assumes that instances contribute independently to the bag label and that the bag is positive if at least one instance is positive, a logical OR under uncertainty. While theoretically appealing, noisy-or can be sensitive to calibration of instance-level probabilities and may not capture interactions between instances.

Distribution-based aggregation methods represent the bag as a summary statistic of the instance-level feature distribution, such as the histogram of embeddings, moments (mean, variance), or parameters of a Gaussian mixture model (Gärtner et al. [2002]). These approaches can capture richer statistical properties beyond a single scalar aggregation but are computationally more expensive and may require larger amounts of data to estimate reliably.

Graph-based aggregation constructs a graph over instances within a bag (and potentially across bags) and uses graph neural networks to propagate information before pooling (Yan et al. [2018]). This can model relationships between instances, such as similarity or co-occurrence, but introduces significant architectural overhead and may not be justified when instances are assumed to be independent.

Finally, learning-to-aggregate methods treat the aggregation function itself as a learned module, often parameterized by a neural network. For example, the aggregation could be implemented as a recurrent neural network that processes instances sequentially or as a set transformer that models interactions among instances (Zaheer et al. [2017]; Lee et al. [2019]). While maximally flexible, such approaches require substantial training data and can overfit in low-resource scenarios.

In the context of this work, where instances are noisy social media posts, supervision is weak, and positive signals are sparse, simple aggregation strategies like max and sum offer transparency, stability, and a clear inductive bias. More complex alternatives, while theoretically possible, were not pursued because they introduce additional parameters, computational cost, and risk of overfitting without a clear empirical or theoretical advantage in this setting. The comparative analysis in later chapters therefore focuses on sum and max operators, which provide a robust baseline and facilitate interpretation of how evidence is accumulated across a user's posting history.

3.3.2 Attention Mechanisms in Multi-Instance Learning

Attention mechanisms in multi-instance learning (MIL) are computational strategies that assign importance weights to individual instances within a bag, allowing the model to focus on the most informative parts of the data Ilse et al. [2018]. Formally, given a bag of instances $B = \{x_1, x_2, \dots, x_n\}$ and their corresponding feature representations h_i, an attention mechanism computes a set of weights α_i such that $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$. The bag-level representation is then obtained as a weighted sum:

$$\mathbf{z} = \sum_{i=1}^{n} \alpha_i h_i. \tag{3.2}$$

The weights α_i are typically parameterized as a function of the instance features, often through a small neural network followed by a softmax normalization, ensuring that the aggregation remains permutation-invariant (Zaheer et al. [2017]).

Attention can be understood as a generalization of both sum and max pooling. When attention weights are uniform ($\alpha_i = \frac{1}{n}$), the mechanism reduces to mean pooling. When attention collapses to a one-hot vector (one weight equal to 1, others 0), it behaves like max pooling. In practice, learned attention allows the model to interpolate between these extremes and to assign different degrees of importance to different instances based on their relevance to the task.

In many MIL applications, especially those with sparse positive signals, attention mechanisms tend to learn highly skewed weight distributions, effectively approximating a max-like behavior (Ilse et al. [2018]). This occurs because when only a few instances in a bag are truly relevant to the label, the attention network learns to assign most of the weight to those instances, suppressing noise from irrelevant ones. This property makes attention particularly suited for domains like medical image analysis or mental health screening, where positive evidence is often concentrated in a small subset of the data.

The attention function is commonly implemented as:

$$\alpha_i = \text{softmax}\left(w^\top \tanh(V h_i)\right), \tag{3.3}$$

where V and w are learnable parameters. More advanced variants include gated attention, which adds a sigmoid gating term to improve gradient flow Ilse et al. [2018], and multi-head attention, which computes several sets of weights in parallel to capture different aspects of relevance (Vaswani et al. [2017]).

However, attention mechanisms are not without limitations. In settings with severe class imbalance or very weak supervision, attention weights can become unstable or collapse prematurely, leading to suboptimal aggregation. Moreover, attention introduces additional parameters and computational overhead, which may not be justified when simpler operators like max pooling already encode a suitable inductive bias. For these reasons, non-learnable attention, where weights are fixed a priori, such as in uniform or max aggregation, remains a viable and often more robust alternative in data-scarce, high-noise environments.

In this work, the primary aggregation strategy employs a non-learnable max operator, which can be viewed as a form of hard attention where all weight is assigned to the instance with the highest relevance score (Dietterich et al. [1997a]). This design choice is motivated by the sparsity and heterogeneity of mental health signals in social media text, where diagnostic evidence is likely to be concentrated in a small fraction of a user's posts. While learnable attention could theoretically adapt to more complex weighting schemes, the limited size and noisy nature of the dataset favor the simplicity and stability of a fixed max aggregation, which has been shown empirically to yield strong performance without the risk of overfitting or unstable training dynamics.

Chapter 4

Learning Under Class Imbalance and Weak Labels

4.1 Class Imbalance in Mental Health Data

A defining characteristic of mental health screening datasets derived from real world populations is the presence of severe class imbalance. In the context of depression detection from user generated text, the number of individuals exhibiting clinically relevant symptoms is typically much smaller than the number of control subjects. This imbalance is not an artifact of dataset construction or experimental design, but rather a direct consequence of underlying population prevalence and data collection mechanisms. Epidemiological studies consistently report that the prevalence of major depressive disorder in the general population is substantially lower than that of non affected individuals, even when considering underdiagnosis and reporting biases (Kessler et al. [2003]). As a result, any dataset intended to approximate realistic screening conditions is expected to reflect this asymmetry.

In social media based mental health datasets, class imbalance is further exacerbated by participation and disclosure effects. Users who do not experience significant psychological distress may nevertheless be highly active and verbose, while individuals experiencing depression may post less frequently or disengage entirely over time. Conversely, only a subset of affected users explicitly express symptoms in language that can be reliably detected, leading to weak or incomplete supervision even for positively labeled subjects. These factors jointly produce datasets in which minority class samples are not only fewer in number, but also more heterogeneous and noisier than their majority class counterparts. This asymmetry introduces substantial challenges for standard supervised learning algorithms, which implicitly assume balanced class distributions or symmetric error costs.

From a statistical learning perspective, class imbalance alters the effective training objective by biasing empirical risk minimization toward the majority class. When standard loss functions such as binary cross entropy are minimized on imbalanced data, the resulting decision boundary is often dominated by the majority class distribution, yielding models that achieve high overall accuracy while failing to detect minority class instances. This phenomenon is particularly problematic in mental health screening, where the primary objective is not to maximize aggregate correctness, but to identify individuals at risk with sufficient sensitivity to support early intervention or further assessment. In such settings, false negatives carry a substantially higher cost than false positives, yet this asymmetry is not reflected in accuracy based evaluation. The limitations of accuracy as a performance metric in imbalanced settings are well documented in the machine learning literature. Accuracy conflates performance across classes and can remain deceptively high even when a classifier completely ignores the minority class (He and Garcia [2009]). In extreme cases, a trivial classifier that predicts the majority class for all instances may outperform more nuanced models under accuracy, despite being useless for the task at hand. This issue is not merely theoretical, but has practical implications for mental health applications, where reporting high accuracy without appropriate context can obscure systematic failures to identify vulnerable individuals. As a consequence, reliance on accuracy as a primary evaluation criterion is inappropriate for imbalanced clinical screening tasks.

Beyond evaluation, class imbalance also affects representation learning and model calibration. Minority class samples may occupy sparse and fragmented regions of the feature space, making them difficult to model with standard parametric assumptions. In text based settings, this sparsity is compounded by linguistic variability, indirect symptom expression, and the presence of confounding factors such as irony, metaphor, or topic drift. Models trained under such conditions may learn spurious correlations that correlate with dataset specific artifacts rather than clinically meaningful patterns. This risk is particularly pronounced when powerful models are trained on limited minority class data, as they may overfit to superficial cues that do not generalize beyond the training distribution. This is the case for this work and is a limiting factor in the choice of the models, while there is abundance of posts, the users are rather few.

Importantly, the presence of class imbalance interacts with other sources of uncertainty that are intrinsic to mental health data, including weak labels, delayed manifestation of symptoms, and temporal inconsistency. Labels often reflect retrospective diagnoses or self reported conditions that do not align cleanly with the time span of available text. As a result, some users labeled as positive may exhibit little or no explicit symptom expression in their posts, while some users labeled as negative may nonetheless display subclinical or emerging signals. This label noise further complicates learning under imbalance, as minority class samples are not only scarce but also imperfect proxies for the underlying construct of interest. In the case of eRisk, positive labels are attributed to subjects exhibiting signs of depression, which avoids the risks of self-reporting, though it comes with its own challenges.

Class imbalance represents a fundamental challenge in machine learning, particularly in medical and psychological applications where positive cases (e.g., depression diagnoses) are naturally rare compared to negative cases. When the class distribution is skewed, standard learning algorithms tend to be biased toward the majority class, resulting in models that achieve high overall accuracy by simply predicting the majority class for all samples. This section provides the theoretical grounding for the three primary areas of improvement considered in this work to address class imbalance: loss function modifications, general techniques, and evaluation metrics.

4.2 Advanced Loss Functions for Class Imbalance

Beyond architectural modifications and sampling strategies, loss function engineering represents a fundamental approach to addressing class imbalance. This section provides theoretical analysis of two pivotal loss modifications: class-weighted loss functions and focal loss. These approaches operate at the optimization level, directly modifying how the learning algorithm penalizes errors on different classes.

4.2.1 Class-Weighted Loss Functions

Class-weighted loss functions represent the most straightforward theoretical approach to addressing class imbalance through explicit error cost differentiation. The fundamental insight, formalized by Elkan [2001b], is that misclassification costs are rarely symmetric in real-world applications, particularly in medical domains.

Theoretical Formulation

Given a standard binary cross-entropy loss:

$$\mathscr{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{4.1}$$

the class-weighted version introduces class-specific multipliers:

$$\mathscr{L}_{\text{wBCE}} = -\frac{1}{N}\sum_{i=1}^{N}[w_1 y_i \log(\hat{y}_i) + w_0(1-y_i)\log(1-\hat{y}_i)] \tag{4.2}$$

where w_0 and w_1 are weights for negative and positive classes respectively. The weights are typically set inversely proportional to class frequencies:

$$w_c = \frac{N}{C \cdot n_c} \tag{4.3}$$

where n_c is the number of samples in class c, N is total samples, and C is the number of classes (2 for binary classification).

Bayesian Interpretation

From a Bayesian perspective, class weighting can be interpreted as adjusting the prior class probabilities. Consider the posterior probability under the original model:

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1) + P(X|Y=0)P(Y=0)} \tag{4.4}$$

With class weights, we effectively modify the priors:

$$P_w(Y=1) = \frac{w_1\pi_1}{w_1\pi_1 + w_0\pi_0} \tag{4.5}$$

where $\pi_c = P(Y=c)$ are the empirical class frequencies. Setting $w_1/w_0 = \pi_0/\pi_1$ creates a balanced prior $P_w(Y=1) = P_w(Y=0) = 0.5$, theoretically eliminating the bias toward majority class predictions.

Gradient Analysis

The effect of class weighting on gradient updates reveals its theoretical mechanism. The gradient of weighted BCE with respect to parameters θ is:

$$\nabla_\theta \mathscr{L}_{\text{wBCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{w_1 y_i}{\hat{y}_i} - \frac{w_0(1-y_i)}{1-\hat{y}_i}\right]\nabla_\theta \hat{y}_i \tag{4.6}$$

For imbalanced data with $\pi_1 \ll \pi_0$, the term $\frac{w_1 y_i}{\hat{y}_i}$ for positive samples receives amplification through w_1, while $\frac{w_0(1-y_i)}{1-\hat{y}_i}$ for negative samples receives attenuation through w_0. This rebalancing ensures that gradient contributions from minority class samples maintain sufficient magnitude despite their scarcity.

Theoretical Limitations

Despite its intuitive appeal, class weighting has several theoretical limitations identified in recent literature(Cortes et al. [2025]):

- Scale sensitivity: The optimal weight ratio w_1/w_0 depends not only on class frequencies but also on the difficulty of classifying each class. Setting weights purely based on inverse frequency assumes equal classification difficulty, which rarely holds in practice.
- Gradient instability: Extreme weight ratios ($w_1 \gg w_0$) can destabilize optimization, causing gradient explosion for minority class samples or vanishing gradients for majority class samples.
- Overfitting risk: By amplifying the influence of minority class samples, class weighting can lead to overfitting on these samples, particularly when they are few in number.
- Hypothesis space distortion: Weighted loss functions change the optimization landscape, potentially causing convergence to different local minima than the original unweighted loss.

These limitations explain why more sophisticated approaches like LDAM and focal loss have gained prominence despite the simplicity of class weighting.

4.2.2 Focal Loss

Focal loss, introduced by Lin et al. [2017a] for object detection, represents a theoretically innovative approach that addresses class imbalance by dynamically modulating the loss based on prediction confidence. Unlike static class weighting, focal loss adapts to each sample's difficulty, providing a more nuanced handling of imbalance.

Theoretical Motivation

The fundamental insight behind focal loss is that imbalance manifests not only in sample counts but also in learning difficulty. In imbalanced datasets, the majority class often consists of "easy" examples that are quickly learned, while the minority class contains more "hard" examples. Standard cross-entropy loss treats easy and hard examples equally, allowing easy negatives to dominate the gradient.

Focal loss modifies cross-entropy by introducing a modulating factor $(1-p_t)^\gamma$:

$$\mathcal{L}_{\text{FL}} = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{4.7}$$

where:

- $p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases}$ is the model's estimated probability for the true class
- $\alpha_t \in [0,1]$ is a class balancing factor (similar to class weights)
- $\gamma \geq 0$ is the focusing parameter

Dynamic Sample Weighting Mechanism

The modulating factor $(1-p_t)^\gamma$ implements a theoretically elegant form of curriculum learning. For well-classified examples ($p_t \to 1$), $(1-p_t)^\gamma \to 0$, reducing their loss contribution. For misclassified or uncertain examples ($p_t \to 0$), $(1-p_t)^\gamma \to 1$, maintaining their full loss.

This creates a self-adjusting weighting scheme where:

$$\text{Effective weight} = \alpha_t \cdot (1-p_t)^\gamma \tag{4.8}$$

The gradient with respect to p_t reveals the adaptive nature:

$$\frac{\partial \mathcal{L}_{\text{FL}}}{\partial p_t} = -\alpha_t \left[\gamma(1-p_t)^{\gamma-1} \log(p_t) - \frac{(1-p_t)^\gamma}{p_t} \right] \tag{4.9}$$

For $\gamma > 0$, the gradient magnitude decreases as p_t increases, automatically down-weighting correctly classified examples during training.

Theoretical Analysis of the Focusing Parameter

The focusing parameter γ controls the rate at which easy examples are down-weighted. Several special cases illustrate its theoretical role:

1. $\gamma = 0$: Focal loss reduces to weighted cross-entropy: $\mathcal{L}_{\text{FL}} = -\alpha_t \log(p_t)$
2. $\gamma = 1$: The modulating factor becomes linear: $\mathcal{L}_{\text{FL}} = -\alpha_t(1-p_t)\log(p_t)$
3. $\gamma > 1$: Creates a super-linear down-weighting of easy examples

Lin et al. [2017a] found $\gamma = 2$ to work well in practice, corresponding to quadratic down-weighting. Theoretically, larger γ values increase the model's focus on hard examples but risk instability if too many easy examples are completely ignored.

Relationship to Hard Example Mining

Focal loss can be viewed as a continuous, differentiable alternative to hard example mining algorithms. Traditional hard mining selects a subset of difficult examples for training, which introduces discontinuities and requires heuristic thresholds. Focal loss implements soft mining by continuously varying example weights based on difficulty.

This differentiability property provides theoretical advantages:

- Smooth optimization landscape without discontinuities from example selection
- Gradient information from all examples, not just a hard subset
- Automatic adaptation to changing difficulty throughout training

Statistical Learning Theory Perspective

From statistical learning theory, focal loss implements a form of importance weighting where importance correlates with classification difficulty. Consider the population risk:

$$R(f) = \mathbb{E}_{(X,Y)}[\ell(f(X), Y)] \tag{4.10}$$

Focal loss modifies this to:

$$R_{\text{FL}}(f) = \mathbb{E}_{(X,Y)}[w(X,Y) \cdot \ell(f(X), Y)] \tag{4.11}$$

where $w(x, y) = \alpha_y(1 - p_t)^\gamma$ depends on both the label y and the model's current prediction $f(x)$.

This creates a moving target problem from a theoretical perspective: the weighting function $w(x, y)$ changes as the model learns, violating the standard i.i.d. assumption. However, Lin et al. [2017a] provide empirical evidence that this non-stationarity actually helps by creating a curriculum where the model increasingly focuses on remaining hard examples.

Comparison with Class-Weighted Loss

Focal loss generalizes class-weighted loss in several theoretically important ways:

- Two-dimensional adjustment: While class weighting adjusts only based on class membership, focal loss adjusts based on both class and difficulty.
- Dynamic adaptation: Class weights remain fixed throughout training, while focal loss weights evolve as predictions improve.
- Implicit hard example mining: Focal loss automatically identifies and emphasizes hard examples without explicit mining heuristics.
- Reduced hyperparameter sensitivity: The γ parameter in focal loss proves more robust across datasets than precise class weight ratios.

Theoretical analysis suggests focal loss should outperform simple class weighting when:

- Hard examples are not evenly distributed between classes
- Classification difficulty varies significantly within classes
- The model tends to quickly overfit on easy examples

However, focal loss introduces additional complexity and requires careful tuning of γ, particularly in very high imbalance scenarios where extreme down-weighting of easy majority examples might prematurely eliminate their gradient contributions.

4.2.3 Label-Distribution-Aware Margin (LDAM) Loss

The Label-Distribution-Aware Margin (LDAM) loss represents a theoretically rigorous approach to class imbalance that operates at the loss function level (Cao et al. [2019]). Traditional cross-entropy loss treats all misclassifications equally, regardless of class membership:

$$\mathscr{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{4.12}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability. This formulation implicitly assumes balanced class distributions, leading to suboptimal decision boundaries when $n_{\text{pos}} \ll n_{\text{neg}}$ (He and Garcia [2009]).

LDAM introduces class-dependent margins to rectify this imbalance. For a binary classification problem with classes $c \in \{0, 1\}$, the LDAM loss modifies the logits before applying the sigmoid function:

$$\mathscr{L}_{\text{LDAM}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{z_{y_i} - \Delta_{y_i}}}{e^{z_{y_i} - \Delta_{y_i}} + \sum_{j \neq y_i} e^{z_j}}\right) \tag{4.13}$$

where z_c represents the logit for class c, and Δ_c is the class-specific margin. The margin is computed as:

$$\Delta_c = C \cdot n_c^{-\frac{1}{4}} \tag{4.14}$$

Here, C is a hyperparameter controlling the maximum margin (set to 0.3 in our implementation), and n_c is the number of training samples in class c. The quarter-power scaling ($n_c^{-1/4}$) is derived from generalization theory for imbalanced datasets. Cao et al. (Cao et al. [2019]) proved that this scaling minimizes an upper bound on the generalization error for minority classes under certain assumptions about data separability.

To illustrate the impact of class-dependent margins on the decision boundary, consider the binary classification scenario depicted in Figure 4.1. Here, the decision boundary is adjusted to account for the imbalance between the two classes, with the margin for each class scaled according to its sample size. This adjustment ensures that the classifier allocates greater decision space to the minority class, thereby improving generalization performance on underrepresented data.

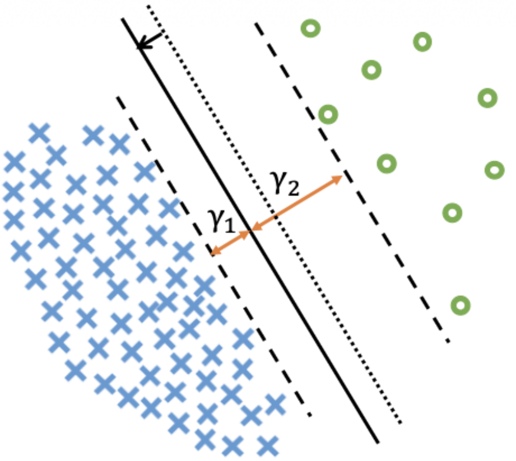

Figure 4.1. Visualization of the decision boundary in a binary classification task with class imbalance as presented by Cao et al. [2019]. The margin γ_i for each class i is defined as the minimum distance from the decision boundary to the nearest sample of that class. By optimizing the trade-off between γ_1 and γ_2, the classifier achieves a balance that favors improved generalization for the minority class, as theoretically motivated by the inverse quarter-root scaling of class-specific margins.

The theoretical justification stems from margin theory, which establishes that larger margins between classes lead to better generalization (Bartlett [1998]). By assigning larger margins to minority classes, LDAM effectively pushes the decision boundary away from minority class samples, forcing the classifier to learn more discriminative features for these underrepresented examples. This counteracts the natural tendency of gradient-based optimizers to focus on majority class samples, which contribute more to the overall loss due to their frequency.

Most implementations further include a scaling factor that amplifies the logits after margin subtraction:

$$\hat{z}_c = (z_c - \Delta_c) \cdot s \tag{4.15}$$

where s is the scaling factor. This scaling serves two purposes: First, it maintains numerical stability when margins are large. Second, it ensures that the gradient magnitude remains sufficient for effective optimization, preventing vanishing gradients when $z_c \approx \Delta_c$ (Wang et al. [2018]).

From an optimization perspective, LDAM can be viewed as a form of cost-sensitive learning where the cost matrix is not fixed but adaptively determined by the data distribution (Elkan [2001a]). Unlike simple class weighting approaches that multiply the loss by inverse class frequency, LDAM operates in the feature space by modifying the geometry of the decision boundary. This spatial approach is particularly effective for high-dimensional embeddings, where distance metrics and margin violations have clearer geometric interpretations.

4.3 Class Imbalance Handling Techniques

4.3.1 Weighted Random Sampler

The Weighted Random Sampler addresses class imbalance at the data level by altering the sample distribution during mini-batch construction (Chawla et al. [2002]). While this technique was ultimately disabled in the final, better performing models, its theoretical foundations warrant discussion as it represents a fundamentally different approach from LDAM.
Given a dataset with class counts n_0 (negative) and n_1 (positive), the sampler assigns to each sample a weight inversely proportional to its class frequency:

$$w_i = \frac{1}{n_{y_i}} \tag{4.16}$$

where n_{y_i} is the count of samples belonging to class y_i. During batch construction, samples are drawn with probability:

$$P(\text{select sample } i) = \frac{w_i}{\sum_{j=1}^{N} w_j} \tag{4.17}$$

This sampling strategy ensures that in expectation, each batch contains equal numbers of samples from each class, effectively creating artificially balanced mini-batches (Drummond and Holte [2003]).
The theoretical basis for this approach comes from importance sampling theory in stochastic optimization (Bottou et al. [2018]). Consider the empirical risk minimization objective:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i) \tag{4.18}$$

Under class imbalance, the gradient estimates are dominated by majority class samples:

$$\nabla_\theta \mathscr{L} \approx \frac{n_0}{N} \mathbb{E}_{(x,y)\sim p_0}[\nabla_\theta \ell] + \frac{n_1}{N} \mathbb{E}_{(x,y)\sim p_1}[\nabla_\theta \ell] \tag{4.19}$$

where p_c denotes the distribution of class c. Weighted sampling modifies this expectation by sampling from a new distribution q where $q(y = c) = 1/2$ for both classes. The gradient becomes:

$$\nabla_\theta \mathscr{L}_{\text{weighted}} \approx \frac{1}{2} \mathbb{E}_{(x,y)\sim p_0}[\nabla_\theta \ell] + \frac{1}{2} \mathbb{E}_{(x,y)\sim p_1}[\nabla_\theta \ell] \tag{4.20}$$

This equal weighting of classes in the gradient estimate reduces the variance of minority class gradient contributions, leading to more stable optimization for rare classes (Bottou et al. [2018]).
However, this approach has significant theoretical limitations. First, it violates the i.i.d. assumption fundamental to many convergence proofs for stochastic gradient descent (Bottou et al. [2018]). The altered sample distribution introduces bias into the gradient estimates unless proper importance weighting is applied in the loss function itself. Second, by oversampling minority class examples, each minority sample is seen multiple times during training, potentially leading to overfitting (Drummond and Holte [2003]). This is particularly problematic when

minority class samples are scarce, as is the case in our dataset with only 83 positive examples. The decision to disable Weighted Random Sampler in favor of using LDAM loss reflects an empirically driven yet theoretically grounded preference: model-level regularization (via margin adjustments) was preferred over data-level manipulation. This aligns with recent insights suggesting that loss function modifications generally outperform sampling strategies when combined with appropriate regularization techniques (Cao et al. [2019]; Kang et al. [2020]), especially in deep learning settings where model capacity is high.

4.3.2 Threshold Optimization

Decision threshold optimization represents a post-hoc approach to addressing class imbalance that operates on model predictions rather than the training process itself (Elkan [2001a]). Unlike the default threshold of 0.5 used in balanced scenarios, optimal thresholds for imbalanced problems typically differ significantly and must be determined systematically (Fawcett [2006]).
For a binary classifier producing probability estimates $\hat{p}_i \in [0,1]$, predictions are obtained by comparing to a threshold τ:

$$\hat{y}_i = \mathbb{I}(\hat{p}_i \geq \tau) \tag{4.21}$$

where $\mathbb{I}$ is the indicator function. The choice of τ directly controls the trade-off between false positives and false negatives, which is particularly critical in imbalanced settings.
The theoretical foundation for threshold optimization comes from decision theory and cost-sensitive classification (Elkan [2001a]). Consider a cost matrix C where C_{ij} represents the cost of predicting class i when the true class is j. The optimal threshold minimizes the expected cost:

$$\tau^* = \arg\min_{\tau} [\pi_0 \cdot \mathrm{FPR}(\tau) \cdot C_{10} + \pi_1 \cdot \mathrm{FNR}(\tau) \cdot C_{01}] \tag{4.22}$$

where π_c is the prior probability of class c, FPR is the false positive rate, and FNR is the false negative rate. In medical applications like depression detection, the cost of false negatives (missing a true depression case) typically exceeds the cost of false positives (unnecessary follow-up), justifying thresholds lower than 0.5 (Fawcett [2006]).
A common approach is to implement a threshold optimization procedure which searches over a grid of thresholds. For each candidate threshold τ, it computes a performance metric on the validation set and selects the threshold maximizing this metric:

$$\tau^* = \arg\max_{\tau \in \mathcal{T}} M(y_{\mathrm{val}}, \hat{y}_{\mathrm{val}}(\tau)) \tag{4.23}$$

where $\mathcal{T} = \{0.05, 0.06, \ldots, 0.95\}$ and M is a chosen metric. This provides flexibility in the optimization objective, supporting recall, F1 score, or Matthews Correlation Coefficient (MCC) (Matthews [1975]).
From an information-theoretic perspective, threshold optimization can be viewed as calibrating the classifier's confidence estimates to align with operational requirements (Guo et al. [2017]). Well-calibrated classifiers produce probability estimates that match true frequencies. In imbalanced settings, even well-calibrated classifiers may require threshold adjustment because the operating point differs from the natural probability cutoff.

The theoretical advantage of treating threshold optimization as an integral part of the model rather than post-processing is that it acknowledges the separation between model training and deployment objectives (Elkan [2001a]). This separation is particularly important when class distributions differ between training and deployment environments, a common scenario in real-world applications.

4.4 Evaluation Metrics for Imbalanced Classification

When evaluating classifiers on imbalanced datasets, traditional metrics like accuracy become misleading and potentially dangerous (He and Garcia [2009]). A classifier that simply predicts the majority class for all samples can achieve high accuracy while completely failing to identify minority class instances. This section provides a comprehensive theoretical analysis of the evaluation metrics implemented in the code, explaining their mathematical formulations, statistical properties, and appropriateness for imbalanced medical classification tasks.

4.4.1 Confusion Matrix and Its Derivative Metrics

The confusion matrix represents the fundamental building block for all classification metrics.Fawcett [2006]; Duda et al. [2001] For binary classification, it is a 2 × 2 contingency table:

$$\text{CM} = \begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix} \tag{4.24}$$

where:

- True Negatives (TN): Correctly predicted negative cases
- False Positives (FP): Negative cases incorrectly predicted as positive (Type I error)
- False Negatives (FN): Positive cases incorrectly predicted as negative (Type II error)
- True Positives (TP): Correctly predicted positive cases

In medical diagnostics, these cells have specific clinical interpretations: FNs represent missed diagnoses with potentially severe consequences, while FPs may lead to unnecessary interventions or anxiety. The relative importance of these errors depends on the clinical context and cost-benefit analysis of interventions.

From a statistical perspective, the confusion matrix represents the joint distribution of predicted and true labels. Its marginal distributions reflect both the classifier's behavior and the underlying data distribution. For imbalanced datasets, the matrix is inherently asymmetric, with much larger values in the majority class row/column.

4.4.2 Accuracy and Balanced Accuracy

Standard Accuracy

Standard accuracy measures the overall correctness of predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.25}$$

While intuitive and widely used, accuracy suffers from severe limitations in imbalanced settings. Consider a dataset with 95% negative samples and 5% positive samples. A trivial classifier that always predicts "negative" achieves 95% accuracy while completely failing its primary task of identifying positive cases. This paradox has been formally analyzed, showing that accuracy becomes increasingly misleading as class imbalance grows (Provost and Fawcett [1998]).
The theoretical problem stems from accuracy's sensitivity to class priors. Let $\pi_0 = P(Y = 0)$ and $\pi_1 = P(Y = 1)$ be the true class probabilities. The expected accuracy of a classifier can be decomposed as:

$$\mathbb{E}[\text{Accuracy}] = \pi_0 \cdot \text{Specificity} + \pi_1 \cdot \text{Sensitivity} \tag{4.26}$$

When $\pi_0 \gg \pi_1$, the first term dominates, making specificity overwhelmingly influential. This mathematical property explains why accuracy favors classifiers that perform well on the majority class while potentially neglecting the minority class.

Balanced Accuracy

Balanced accuracy addresses this limitation by averaging class-wise accuracies:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \tag{4.27}$$

This metric gives equal weight to both classes regardless of their frequencies, making it invariant to class imbalance. The factor $\frac{1}{2}$ ensures the metric ranges from 0 to 1, with 0.5 representing random guessing under balanced class priors.
Theoretical analysis reveals that balanced accuracy is equivalent to the arithmetic mean of sensitivity and specificity, or equivalently, the average of true positive rate (TPR) and true negative rate (TNR). This formulation connects balanced accuracy to the ROC curve, where the point closest to the top-left corner (0,1) maximizes both TPR and TNR simultaneously.
Balanced accuracy has been shown to possess several desirable statistical properties for imbalanced data:(Brodersen et al. [2010])

1. It is unbiased with respect to class prevalence
2. Its variance is more stable across different imbalance ratios
3. It provides better discrimination between classifiers on imbalanced data

4.4.3 Precision (Positive Predictive Value)

Precision, also known as positive predictive value (PPV), measures the reliability of positive predictions (Altman [1994]):

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4.28}$$

This metric answers the question: "When the classifier predicts positive, how often is it correct?" Precision is particularly important in applications where false positives have high costs, such as unnecessary medical treatments or false accusations.

From a Bayesian perspective, precision can be expressed using Bayes' theorem:

$$\text{Precision} = P(Y = 1|\hat{Y} = 1) = \frac{P(\hat{Y} = 1|Y = 1)P(Y = 1)}{P(\hat{Y} = 1|Y = 1)P(Y = 1) + P(\hat{Y} = 1|Y = 0)P(Y = 0)} \tag{4.29}$$

This formulation reveals that precision depends on three factors: the true positive rate (sensitivity), the false positive rate, and the class prior $P(Y = 1)$. When the positive class is rare (small $P(Y = 1)$), achieving high precision requires extremely low false positive rates or very high true positive rates.

The relationship between precision and prevalence creates what is known as the "precision paradox": as the positive class becomes rarer, maintaining high precision becomes increasingly difficult, even for perfect classifiers. Formally, for a classifier with fixed sensitivity Se and specificity Sp, precision decreases with decreasing prevalence:

$$\lim_{\pi_1 \to 0} \text{Precision} = \frac{Se \cdot \pi_1}{Se \cdot \pi_1 + (1 - Sp) \cdot (1 - \pi_1)} \to 0 \tag{4.30}$$

This mathematical property explains why precision is often low in highly imbalanced medical screening tasks, even when sensitivity is high.

In the context of depression detection, precision represents the proportion of individuals flagged for follow-up who actually require intervention. Low precision indicates many false alarms, which can strain clinical resources and cause unnecessary distress to patients. Despite this, catching positives is usually prioritized, therefore favoring metrics like Recall or F1.

4.4.4 Recall (Sensitivity) and Its Clinical Significance

Recall, also known as sensitivity or true positive rate (TPR), measures the classifier's ability to identify positive cases:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.31}$$

This metric answers the critical question: "Of all actual positive cases, what proportion did the classifier correctly identify?" In medical screening applications, recall represents the detection rate, the fraction of diseased individuals who receive appropriate follow-up.

From a statistical decision theory perspective, recall is intimately connected to Type II error. Given the null hypothesis H_0: "patient does not have depression" and alternative H_1: "patient has depression," recall corresponds to statistical power $1-\beta$, where $\beta = P(\text{accept } H_0 | H_1 \text{ is true})$ is the Type II error rate. High recall therefore indicates high statistical power to detect true cases.

The theoretical importance of recall in medical applications stems from the asymmetric costs of errors. Let C_{FN} denote the cost of a false negative and C_{FP} the cost of a false positive. In depression screening:

- C_{FN} includes untreated depression progression, increased suicide risk, reduced quality of life, and long-term healthcare costs

- C_{FP} includes unnecessary follow-up assessments, temporary patient anxiety, and resource utilization

Clinical guidelines typically prioritize minimizing false negatives due to the severe consequences of untreated depression. This preference manifests mathematically in the recall-precision tradeoff: improving recall generally decreases precision, as catching more true positives inevitably increases false positives.

The relationship between recall and threshold selection is monotonic but nonlinear. As the decision threshold τ decreases:

$$\frac{d\text{Recall}}{d\tau} \leq 0 \tag{4.32}$$

with equality only when the probability density functions of positive and negative classes are disjoint. The rate of change depends on the overlap between class-conditional score distributions.

Receiver Operating Characteristic (ROC) analysis provides further insight into recall's behavior. The ROC curve plots recall (TPR) against false positive rate (FPR = FP/(FP+TN)) across all possible thresholds. The area under this curve (AUC-ROC) represents the probability that a randomly chosen positive sample receives a higher score than a randomly chosen negative sample, independent of class priors.

However, ROC analysis has limitations in highly imbalanced settings Davis and Goadrich [2006]. When the negative class dominates, large changes in the absolute number of false positives produce small changes in FPR, potentially masking problematic classifier behavior. This mathematical property explains why precision-recall curves are often more informative for imbalanced medical classification.

Recall maximization must be balanced against resource constraints and false positive rates. Let N be the total population size, π_1 the disease prevalence, and B the screening budget. The maximum number of individuals who can receive follow-up assessment is B/c, where c is the cost per assessment. The achievable recall is then bounded by:

$$\text{Recall}_{\max} = \min\left(1, \frac{B}{c \cdot N \cdot \pi_1 \cdot \text{Precision}}\right) \tag{4.33}$$

This resource constraint creates a fundamental limit on recall even for perfect classifiers, highlighting the interplay between metric optimization and practical implementation.

4.4.5 F1 Score: Harmonic Mean of Precision and Recall

The F1 score represents the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4.34}$$

This formulation gives equal weight to precision and recall, positioning F1 as a balanced metric when both false positives and false negatives have similar costs. The harmonic mean, rather than arithmetic mean, possesses several theoretically desirable properties for combining rates. Mathematically, the harmonic mean is more sensitive to low values than the arithmetic mean. For two positive numbers a and b:

$$H(a,b) \leq A(a,b) = \frac{a+b}{2} \tag{4.35}$$

with equality only when $a = b$. This property penalizes classifiers that excel at one metric while performing poorly on the other, encouraging balanced performance.

The F1 score can be derived from a more general family of F_β measures:

$$F_\beta = (1+\beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{4.36}$$

where β controls the relative importance of recall versus precision. The standard F1 score corresponds to $\beta = 1$, indicating equal weighting.

From an information retrieval perspective, the F1 score represents a single-number summary of precision-recall trade-off. However, this compression loses information about the operating point and threshold selection. Two classifiers with identical F1 scores may have very different precision-recall profiles, making F1 insufficient for comprehensive evaluation.

The F1 score exhibits several mathematical properties relevant to imbalanced classification:

1. Scale invariance: F1 is invariant to class prevalence when classifier performance metrics remain constant.

2. Threshold dependence: F1 varies non-monotonically with decision threshold, often exhibiting a maximum at some optimal τ^*

3. Break-even point: When precision equals recall, F1 equals their common value.

A limitation of F1 emerges in extremely imbalanced scenarios. As the prevalence of a class approaches zero, maintaining both high precision and high recall becomes mathematically impossible for any non-perfect classifier. The F1 score naturally decreases in this regime, reflecting the fundamental difficulty of the classification task rather than classifier deficiency.

4.4.6 F1 Macro and Micro: Multi-Label and Multi-Class Evaluation

In multi-label or multi-class problems, the F1 score can be extended using different averaging strategies. F1 macro computes the F1 score independently for each label or class and then averages:

$$\text{F1}_{\text{macro}} = \frac{1}{L} \sum_{l=1}^{L} F_1^{(l)} \tag{4.37}$$

where L is the number of labels or classes and $F_1^{(l)}$ is the F1 score for label l.

Micro-averaging, by contrast, pools all true positives, false positives, and false negatives across labels or classes before computing the F1 score:

$$\text{F1}_{\text{micro}} = 2 \cdot \frac{\sum_l \text{TP}_l}{\sum_l (2\text{TP}_l + \text{FP}_l + \text{FN}_l)} \tag{4.38}$$

Key differences between these two approaches include:

- F1 macro treats all labels equally, so rare labels contribute as much as frequent labels.
- F1 micro treats all instances equally, so labels with more samples dominate the metric.
- F1 macro is more sensitive to performance on minority labels.
- F1 micro is more stable statistically due to larger sample sizes.

Choosing between macro and micro averaging depends on evaluation priorities. Macro averaging is appropriate when balanced performance across all classes is desired, while micro averaging better reflects overall classifier accuracy across all predictions.

4.4.7 ROC-AUC: Area Under Receiver Operating Characteristic Curve

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) provides a threshold-independent assessment of classifier discrimination ability. Formally, for a scoring classifier $f : \mathscr{X} \to \mathbb{R}$, ROC-AUC equals:

$$\mathrm{AUC} = P(f(X^+) > f(X^-)) \tag{4.39}$$

where $X^+ \sim p_{\mathrm{pos}}$ and $X^- \sim p_{\mathrm{neg}}$ are random positive and negative instances.
This probabilistic interpretation reveals that AUC measures rank correlation between scores and true labels. A perfect classifier (AUC = 1) gives all positive samples higher scores than all negative samples. A random classifier (AUC = 0.5) has no discriminative power.
ROC-AUC possesses several theoretically appealing properties:

1. Scale invariance: AUC is invariant to monotonic transformations of scores
2. Class prior invariance: AUC is independent of π_1, making it suitable for imbalanced data
3. Statistical consistency: The empirical AUC converges to the population AUC

However, ROC-AUC has known limitations in imbalanced settings Lobo et al. [2008]. The metric weights false positive rate (FPR) linearly, which may not correspond to clinical utility when negative cases vastly outnumber positives. Consider two classifiers:

- Classifier A: FPR = 0.01, TPR = 0.8
- Classifier B: FPR = 0.10, TPR = 0.9

Both might have similar AUC, but Classifier A's 1% FPR produces 10× fewer false alarms than Classifier B's 10% FPR when $\pi_0 = 0.95$. In population screening, this difference could be operationally critical despite similar AUC.
The mathematical relationship between AUC and balanced accuracy is informative. For a fixed threshold, balanced accuracy equals:

$$\text{Balanced Acc} = \frac{1}{2}(\mathrm{TPR} + \mathrm{TNR}) = \frac{1}{2}(\mathrm{TPR} + 1 - \mathrm{FPR}) \tag{4.40}$$

AUC integrates this quantity across all possible FPR values, providing a comprehensive picture of classifier performance across operating points.

ROC-AUC's independence from threshold selection makes it valuable for model comparison but less useful for operational deployment decisions. A classifier with high AUC may still require careful threshold tuning to achieve clinically acceptable performance at a specific operating point.

4.4.8 Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC), also known as the phi coefficient, represents a comprehensive metric for binary classification that accounts for all four cells of the confusion matrix. Formally defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{4.41}$$

MCC ranges from -1 to +1, where +1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between prediction and reality. This bounded range with clear interpretations at extremes provides intuitive assessment of classifier performance.
The theoretical foundation of MCC lies in its derivation as the Pearson correlation coefficient between observed and predicted binary labels. Consider binary variables Y (true labels) and $\hat{Y}$ (predicted labels), both taking values in $\{0,1\}$. Their Pearson correlation is:

$$\rho_{Y,\hat{Y}} = \frac{\mathbb{E}[(Y-\mu_Y)(\hat{Y}-\mu_{\hat{Y}})]}{\sigma_Y \sigma_{\hat{Y}}} \tag{4.42}$$

Expanding this expression for binary variables yields exactly the MCC formula. This connection to correlation theory imbues MCC with several desirable statistical properties, including invariance to label encoding (0/1 versus 1/-1) and symmetry with respect to class interchange.
Theoretical analysis(Chicco and Jurman [2020]) has established MCC as particularly informative for imbalanced binary classification. Their comprehensive comparison demonstrates that MCC provides a more reliable and informative evaluation than either accuracy or F1 score, especially when class distributions are skewed. The authors identify several key advantages:

1. Balanced consideration of all confusion matrix cells: Unlike F1 which ignores true negatives, MCC incorporates all four cells, making it sensitive to both Type I and Type II errors regardless of class balance.

2. Invariance to class swapping: MCC produces the same absolute value when positive and negative classes are swapped, unlike F1 which changes value.

3. Geometric interpretation: MCC can be interpreted as the cosine of the angle between the observed and predicted label vectors in high-dimensional space, providing geometric intuition about classifier alignment with truth.

4. Statistical robustness: MCC's sampling distribution is better behaved under imbalance, with variance less dependent on class ratios compared to other metrics.

Mathematically, MCC also exhibits superior behavior in extreme imbalance scenarios.
While both F1 and MCC correctly identify the classifier's failure, MCC maintains this identification across all imbalance ratios, whereas accuracy becomes increasingly misleading.
The relationship between MCC and Cohen's kappa κ is theoretically interesting. Both metrics

measure agreement corrected for chance, with MCC being essentially a binary special case of the multiclass κ statistic. However, MCC possesses computational advantages and clearer geometric interpretations.

MCC's denominator structure warrants particular analysis. The product under the square root represents the geometric mean of four products:

$$D = \sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})} \tag{4.43}$$

This denominator ensures MCC remains well-defined even when some marginal sums are zero, though extreme cases require careful handling. The geometric mean construction gives MCC desirable scale properties, making it equally sensitive to changes in any confusion matrix cell.

For medical applications like depression screening, MCC's balanced treatment of false positives and false negatives aligns with clinical decision-making where both error types carry costs. The metric's correlation interpretation also facilitates comparison across studies with different prevalence rates, as MCC is relatively invariant to class distribution changes.

However, MCC is not without limitations. Its interpretability requires understanding of correlation coefficients, which may be less intuitive to clinicians than sensitivity or precision. Additionally, MCC values tend to be more conservative than other metrics, with "good" classifiers typically achieving MCC values in the 0.3-0.7 range rather than the 0.8-0.9 common for accuracy or AUC.

Chapter 5

Methodology

5.1 Reproducibility and Experimental Setup

This section defines the experimental constraints and implementation level choices that are held fixed across all experiments presented in the remainder of this chapter. Its purpose is to establish reproducibility guarantees and computational assumptions, rather than to describe the modeling pipeline or justify architectural decisions, which are addressed in subsequent sections.

5.1.1 Code Availability and Version Control

All experiments were implemented using two computational notebooks, one dedicated to the retrieval component and one to the classification component. The results reported in this section refer to the baseline pipeline, while additional experimental variants and less successful attempts are also documented for completeness. The complete source code will be made publicly available upon publication of this thesis through a dedicated GitHub repository at:
`https://github.com/Gualtier0/Master-Thesis-CEM-Depression-etc-case-study`.
The repository includes all scripts required to reproduce data preprocessing, embedding computation, model training, validation, and evaluation. A requirements file specifying exact library dependencies and versions is provided to facilitate environment replication. No external proprietary software is required beyond standard deep learning and scientific computing libraries.

5.1.2 Determinism and Randomness Control

To ensure consistency across runs, all sources of algorithmic randomness are explicitly controlled. Fixed random seeds are applied globally to NumPy, PyTorch, and PyTorch Lightning components. (Paszke et al. [2019], Harris et al. [2020], Harris et al. [2020], Data splitting, batch shuffling, and sampling operations therefore follow deterministic behavior across executions.

Despite these measures, complete numerical determinism cannot be guaranteed due to low level nondeterminism inherent to modern GPU backends. However, observed variability across repeated executions with identical settings was negligible in practice and did not affect qualitative or comparative conclusions.

5.1.3 Pretrained Models and Frozen Representations

All pretrained language models used in this work are employed strictly as fixed feature extractors. In particular, Sentence BERT is used to compute sentence level embeddings and concept similarities, but no fine tuning is performed at any stage. The same pretrained checkpoint is reused consistently across all experiments.

This design choice serves two purposes. First, it isolates the contribution of the proposed aggregation and modeling strategies from representation learning effects. Second, it substantially improves reproducibility and portability, as results do not depend on stochastic fine tuning procedures or large scale retraining.

5.1.4 Hardware and Execution Environment

All model training and evaluation experiments are executed on a single consumer grade device, specifically an Apple MacBook equipped with an Apple M2 Pro GPU with 19 cores, using the Metal Performance Shaders backend. This configuration demonstrates that the proposed models can be trained and evaluated without access to high end server grade hardware.

Embedding computation, which represents the most computationally intensive preprocessing stage, is performed separately on a cloud based NVIDIA L4 GPU provided through the Lightning AI platform. The resulting embeddings are stored and reused across experiments, ensuring that all downstream models operate on identical inputs regardless of training hardware.

5.1.5 Deployment Considerations

The overall system is designed to remain lightweight and deployment friendly. By decoupling embedding extraction from model training and by relying on frozen pretrained encoders, the final predictive models operate on compact fixed size representations. This significantly reduces memory requirements and inference time, making the approach suitable for real world deployment scenarios where computational resources are limited.

Importantly, once embeddings have been computed, model inference does not require access to the original text data or large language models. This property facilitates both scalability and privacy preserving deployment, which are particularly relevant considerations in mental health applications.

5.1.6 Scope of Reproducibility Claims

The objective of this work is comparative and analytical rather than statistical, focusing on understanding the impact of modeling choices under controlled settings. Extensive hyperparameter sweeps, ensembling, or repeated random restarts are therefore intentionally omitted.

Within this scope, the experiments presented in the following sections are fully reproducible given access to the released code, specified dependencies, and described hardware environment.

5.2 Dataset Description and Experimental Splits

As already discussed, this work adopts the eRisk shared task datasets as a longitudinal case study for risk detection of depression from social media data. The eRisk benchmark provides a sequence of yearly releases that are temporally ordered and explicitly designed to evaluate early detection systems under realistic conditions (Losada and Crestani [2017]; Losada et al. [2018]).

5.2.1 Available Dataset Releases

Three dataset releases are considered in this thesis, which are informally referred to as eRisk 2017, eRisk 2018, and eRisk 2020. Each release consists of user level social media data accompanied by binary labels indicating the presence or absence of depression. These shared tasks were not always using the data as independent snapshots, but were inserted in a temporally consistent evaluation protocol, where models are trained on earlier data and tested on future unseen cohorts.

In the original eRisk 2017 task (Losada and Crestani [2017]), participants were provided with training data collected up to 2017 and evaluated on a held out test set released in 2018. Similarly, in eRisk 2018(Losada et al. [2018]), systems were trained on the 2018 dataset and evaluated on a subsequent test set released in 2020. This structure reflects a realistic deployment scenario in which predictive models are trained on historical data and applied to future populations.

The datasets supporting this work are from eRisk collections and are available for research purposes under signing user agreements.

5.2.2 Dataset Usage in This Work

In order to maintain direct comparability with the original benchmark and to align with its temporal evaluation philosophy, this thesis adopts a split strategy consistent with the eRisk 2017 task definition. Specifically, the eRisk 2017 dataset is used exclusively for model development, including both training and validation, while the eRisk 2018 dataset is reserved as a held out test set.

This choice is motivated by two considerations. First, using the 2018 data as test data enables a direct comparison between the proposed approach and the official eRisk 2017 leaderboard results, which were evaluated under the same temporal separation. This is of particular relevance to answer the research question and achieve a model that both provides explanation but also achieves performance comparable to state-of-the-art models for the task. Second, top performing systems in eRisk 2018 achieved performance levels comparable to those reported in eRisk 2017, suggesting that the task difficulty and label distribution remain broadly consistent across these releases. As a result, conclusions drawn from this evaluation setting remain meaningful and representative.

The eRisk 2020 dataset is not used for training or evaluation in the baseline experiment, however, as it has more data, requiring additional computational power, and even more severe unbalance ratio.

5.2.3 Training, Validation, and Test Splits

The eRisk 2017 dataset is further subdivided to support model selection and hyperparameter tuning. User level data is split into training and validation sets following an 80 percent to 20 percent ratio, stratified by class label. This split is performed at the user level rather than at the post level, ensuring that no information from a given individual appears in more than one subset.
The final experimental configuration therefore consists of three distinct partitions:

- Training set, comprising 80 percent of eRisk 2017 users
- Validation set, comprising the remaining 20 percent of eRisk 2017 users
- Test set, comprising all users from eRisk 2018

All reported hyperparameter choices, threshold selection procedures, and model selection decisions are based exclusively on training and validation data. The test set is accessed only once for final evaluation.

5.2.4 User Level Statistics and Class Imbalance

At the user level, both the training set corresponding to eRisk 2017 and the test set corresponding to eRisk 2018 exhibit a pronounced class imbalance, with depressed users forming a clear minority. The training set contains 486 subjects, of which 66 are labeled as depressed and 420 as control users. The test set contains 401 subjects, with 52 depressed users and 349 controls. Figure 5.1 illustrates the distribution of depressed and control users across the two datasets. The skewed nature of the label distribution highlights the limitations of accuracy based evaluation and motivates the use of alternative metrics that better reflect performance on the minority class, as discussed in Chapter 4.

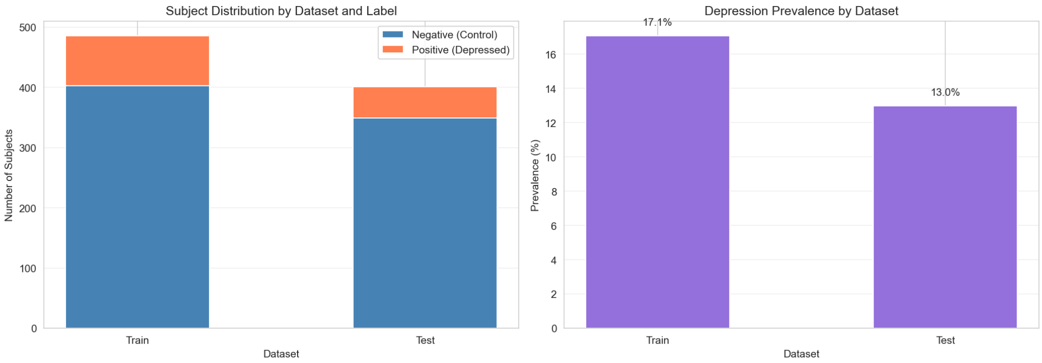

Figure 5.1. Distribution of Labels Across Datasets

5.2.5 Post Level Statistics and User Activity

Beyond class imbalance, the dataset displays substantial heterogeneity in user activity. The number of posts per user varies widely, ranging from a small number of posts to several hundred per subject. This variability is observed in both the training and test sets and affects both depressed and control users.

Figure 5.2 reports the distribution of posts per user on a logarithmic scale. The heavy tailed nature of this distribution reflects typical patterns of engagement in online forums, where a small number of users are highly active while most contribute sporadically. This variability motivates the treatment of each user as a bag of posts rather than as a fixed length sequence and supports the formulation of the task as a multiple instance learning problem.

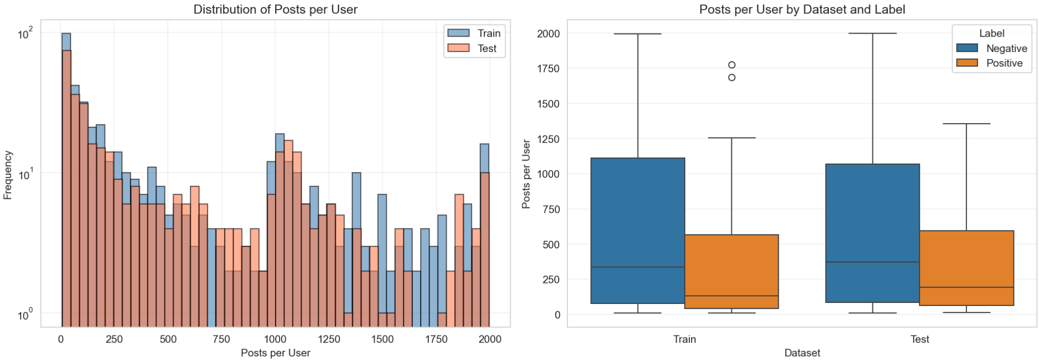

Figure 5.2. Distribution of posts per user

In addition to differences in posting frequency, individual posts vary substantially in length. While many posts consist of short messages, a non negligible fraction contains several hundred tokens. This variability is present across both datasets and labels. Figure 5.3 illustrates the distribution of post lengths measured in tokens.

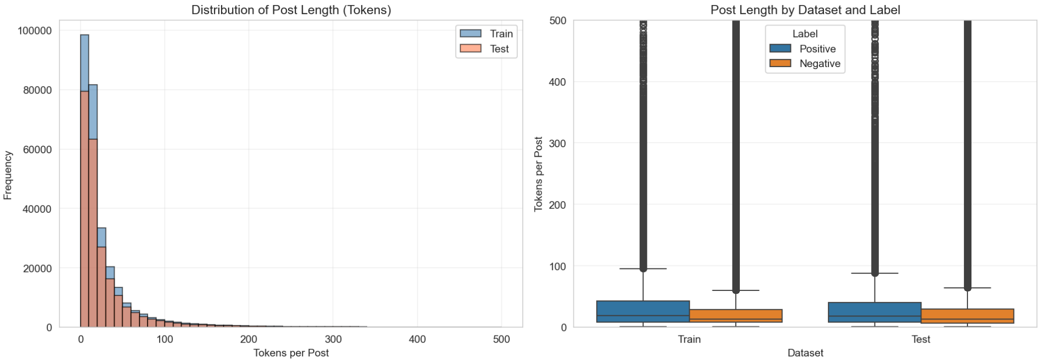

Figure 5.3. The distribution of post lengths by token count reveals that, while the majority of posts are relatively short, a notable proportion contains a significantly higher number of tokens.

The observed variability in post length further motivates the use of sentence or post level embeddings, which allow textual units to be represented independently and aggregated flexibly, rather than relying on monolithic document representations that would be sensitive to extreme length differences.

5.2.6 Concept Prevalence in the Training Data

For the training set only, binary concept annotations corresponding to the 21 symptoms of the Beck Depression Inventory–II are available. An analysis of concept prevalence reveals substantial imbalance across symptoms. Certain concepts, such as self-dislike, pessimism, past failure and loss of energy, occur relatively frequently, while others appear only sparsely.
Figure 5.4 reports the prevalence of each concept across the training users. This distribution highlights two important properties of the data. First, symptom occurrence is itself highly imbalanced, even within the positive class. Second, the presence of individual symptoms is not sufficient to reliably distinguish depressed from control users, as many symptoms may appear sporadically or implicitly in language.

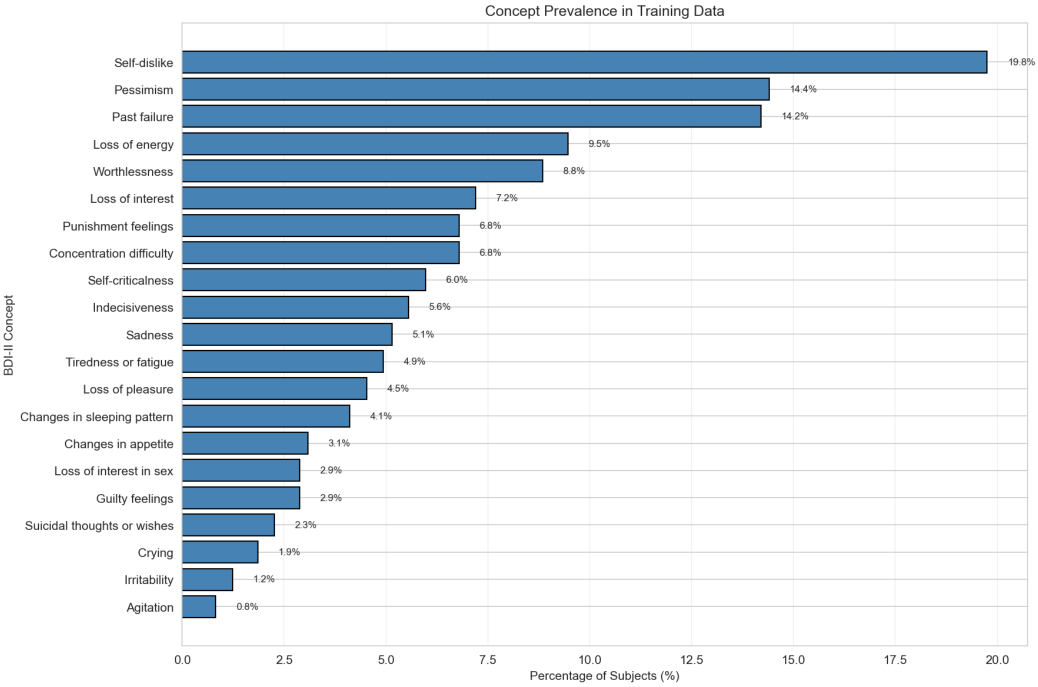

Figure 5.4. Distribution of Concepts by Percentage of Subjects Testing Positive

These observations motivate the use of models capable of operating under sparse and uneven concept supervision, rather than assuming dense or uniformly distributed concept activations.

5.2.7 Concept Co-occurrence and Correlation Structure

To assess whether certain symptoms tend to appear together, pairwise correlations between concept annotations were computed on the training data. Several symptom pairs exhibit moderate to strong positive correlations, reflecting clinically plausible co-occurrence patterns such as loss of energy and tiredness or fatigue.
Figure 5.5 shows the correlation matrix between concept annotations. While this analysis is primarily exploratory, it suggests that symptom independence should not be assumed a priori and that modeling approaches should allow correlated concept activations to coexist.

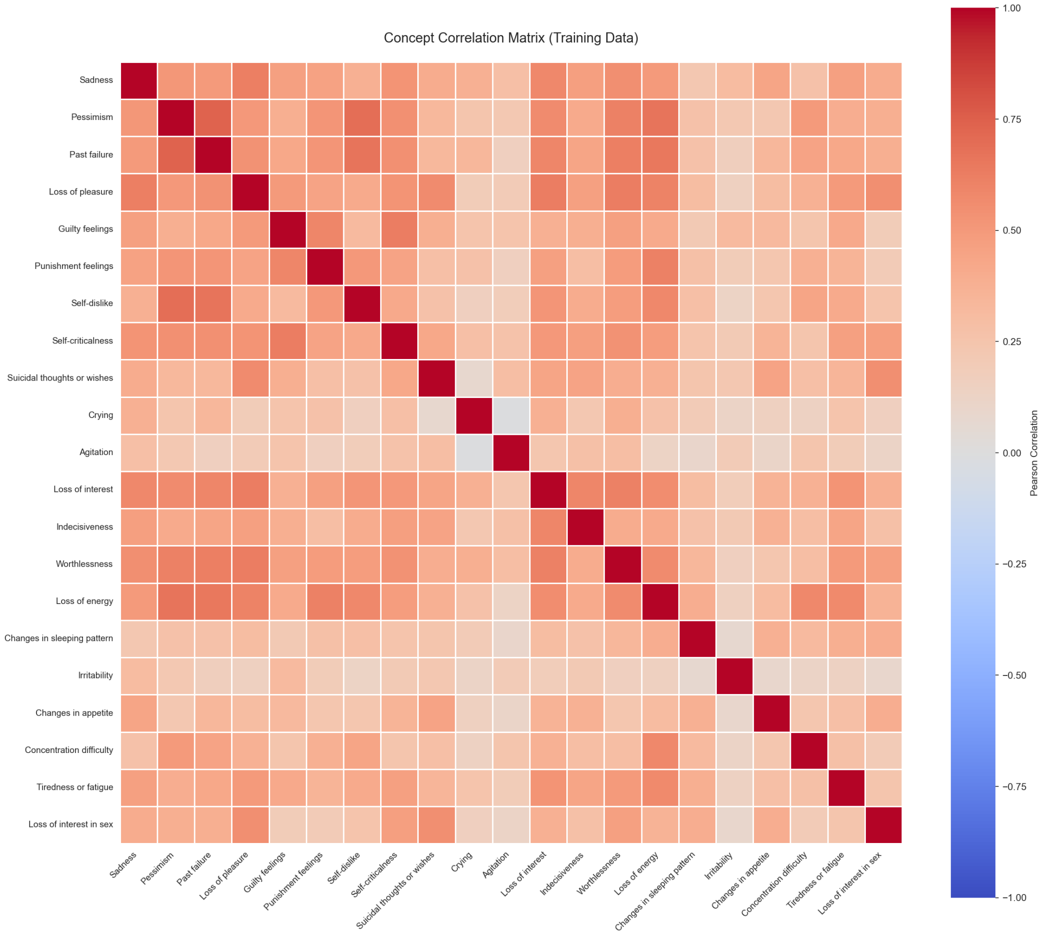

Figure 5.5. The correlation heatmap indicates predominantly positive correlations among the analyzed concepts.

5.2.8 Number of Concepts per User

Finally, the total number of active concepts per user was analyzed for the training set. On average, depressed users exhibit a larger number of active symptoms compared to control users, although substantial overlap exists between the two groups. Figure 5.6 illustrates the distribution of active concepts per user stratified by label. The analysis reveals a significant imbalance in the distribution of concept labels.

The overlap between classes reinforces the notion that concept presence alone is insufficient for perfect discrimination and that meaningful aggregation of textual evidence is required. This further motivates the retrieval and aggregation strategies proposed in this thesis, which aim to focus on the most informative textual signals rather than on raw symptom counts alone.

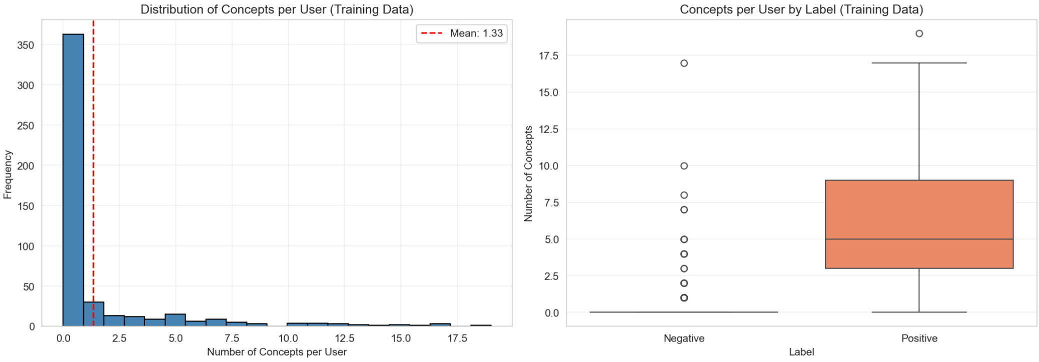

Figure 5.6. The distribution of concept labels, which is highly imbalanced.

5.3 Concept Generation

The clinical concepts used in this study were generated by Qwen2-72B (Yang et al. [2024]), a general-purpose large language model. The generation process is not free-form, but is explicitly structured and guided by the psychometric framework introduced by Ravenda et al.(Ravenda et al. [2025]). Instead of prompting the model for a direct diagnosis or open-ended symptom description, the generation is anchored to the items of standardized psychological questionnaires, our choice being the Beck Depression Inventory-II (BDI-II) (Beck et al. [1996]).

This method ensures that the generated concepts are interpretable and auditable, as each can be traced back to a formal clinical construct (the questionnaire item). The approach leverages the validated finding that LLMs guided by this adaptive, item-based retrieval mechanism produce more reliable and clinically-relevant outputs than direct prompting or non-retrieval methods, effectively translating noisy social media language into structured, assessment-ready clinical observations.

5.4 Retrieval and Pooling with Max-Based Attention

We implemented a retrieval and pooling pipeline designed to capture posts highly relevant to individual BDI-II concepts rather than generally relevant to multiple concepts. The key part is the use of maximum similarity scoring instead of sum-based scoring, which emphasizes posts with strong relevance to at least one specific clinical concept.

Retrieval Stage

Given a set of $K = 21$ BDI-II concepts $\{c_1, c_2, \ldots, c_K\}$ and a subject's posts $\{p_1, p_2, \ldots, p_N\}$, we computed the semantic similarity between each post and each concept using SBERT embeddings. For the standard approach, the relevance score for post p_i was computed as:

$$\text{score}_{\text{sum}}(p_i) = \sum_{k=1}^{K} \text{sim}(p_i, c_k)$$

where $\text{sim}(p_i, c_k)$ is the cosine similarity between the SBERT embeddings of post p_i and concept c_k.

In our max-based approach, we instead compute:

$$\text{score}_{\text{max}}(p_i) = \max_{k \in \{1,\ldots,K\}} \text{sim}(p_i, c_k)$$

This captures posts that are highly relevant to at least one specific clinical concept, even if they are not broadly relevant to multiple concepts.

For each subject, we retrieved the top $R = 50$ posts based on $\text{score}_{\text{max}}(p_i)$, using batched processing (batch size = 32 posts) to optimize memory usage on limited GPU resources.

Attention-Based Pooling

After retrieving the top R posts per subject, we pooled them into a single subject-level embedding using attention weights derived from the max scores.

First, we computed attention weights α_i for each retrieved post p_i:

$$\alpha_i = \frac{\exp\left(\frac{\text{score}_{\text{max}}(p_i)}{\tau}\right)}{\sum_{j=1}^{R} \exp\left(\frac{\text{score}_{\text{max}}(p_j)}{\tau}\right)}$$

where $\tau = 0.2$ is a temperature parameter that controls the sharpness of the attention distribution.

The final subject embedding e_{subj} was computed as a weighted sum:

$$e_{\text{subj}} = \sum_{i=1}^{R} \alpha_i \cdot \text{SBERT}(p_i)$$

where $\text{SBERT}(p_i)$ is the SBERT embedding of post p_i.

Implementation Details

The pipeline was implemented with the following specifications:

- SBERT model: `all-MiniLM-L6-v2` (384-dimensional embeddings)
- Retrieval: Top-50 posts per subject based on max similarity
- Batching: 32 posts per batch to prevent GPU memory exhaustion
- Memory management: Regular cache clearing every 10 subjects

This max-based approach produces embeddings that emphasize concept-specific signals. While previous approaches we've tested have relied on the psychological notion that depressed users display symptoms over extended periods, focusing on extreme or alarming social media posts proved more effective in the noisy social network environments.

5.5 Classifier/Training

5.5.1 Concept Embedding Model Architecture

We implemented a custom Concept Embedding Model based on the max-based attention pipeline. The full pipeline architecture can be observed from 5.7.
It is to be noted that it already exists a well documented library by Espinosa et al. (https://github.com/mateoespinosa/cem) which include several valid CEM implementations that proved helpful during initial testing. However, the unbalanced noisy dataset required several adjustments that would have been difficult to implement using a pre-made library.
The following architecture processes subject-level embeddings to predict both the primary depression classification task and the 21 BDI-II concept activations.

Model Components

The CEM consists of five key components:

1. Concept Extractor: A two-layer MLP that transforms the 384-dimensional subject embedding into 256-dimensional pre-concept features:

$$h_{\text{pre}} = \text{ReLU}(\text{Dropout}(W_2 \text{ReLU}(W_1 x + b_1) + b_2))$$

   where $W_1 \in \mathbb{R}^{256\times384}$, $W_2 \in \mathbb{R}^{256\times256}$.

2. Context Generators: 21 parallel networks (one per concept) that generate dual embeddings representing the true/false states of each concept:

$$e_k^{\text{dual}} = [e_k^{\text{true}}; e_k^{\text{false}]} \in \mathbb{R}^{256}$$

3. Probability Generator: A shared linear layer that predicts concept activation probabilities from dual embeddings:

$$p_k = \sigma(w_p^\top e_k^{\text{dual}} + b_p)$$

   where σ is the sigmoid function.

4. Convex Mixer: Mixes true/false embeddings based on predicted probabilities:

$$e_k = p_k \cdot e_k^{\text{true}} + (1 - p_k) \cdot e_k^{\text{false}}$$

5. Task Classifier: A two-layer MLP that predicts depression from concatenated concept embeddings:

$$\hat{y} = \sigma(w_2^\top \text{ReLU}(W_1[e_1; \ldots; e_{21}] + b_1) + b_2)$$

Intervention Mechanism

During training, we apply concept intervention with probability $\rho = 0.25$:

$$p_k^{\text{train}} = \begin{cases} c_k^{\text{true}} & \text{with probability } \rho \\ p_k & \text{with probability } 1 - \rho \end{cases}$$

where c_k^{true} is the ground truth concept label from the BDI-II questionnaires. This encourages the model to learn meaningful concept representations while still allowing gradient flow.

LDAM Loss for Class Imbalance

To address the class imbalance (83 positive vs 403 negative samples), we implemented Label-Distribution-Aware Margin (LDAM) loss. For a binary classification problem, LDAM applies class-dependent margins:

$$\mathscr{L}_{\text{LDAM}}(z, y) = \begin{cases} \text{BCE}(z - \Delta_1, y) & \text{if } y = 1 \text{ (positive class)} \\ \text{BCE}(z + \Delta_0, y) & \text{if } y = 0 \text{ (negative class)} \end{cases} \tag{5.1}$$

where the margins Δ_c are computed from class frequencies:

$$\Delta_c = \delta \cdot n_c^{-1/4}$$

with $\delta = 0.3$ (maximum margin) and n_c being the sample count for class c. The minority class (positive/depression) receives a larger margin, making its decision boundary more conservative.

Total Loss

The total training loss combines task prediction loss and concept prediction loss:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{task}}(\hat{y}, y) + \lambda \mathscr{L}_{\text{concepts}}(p, c)$$

where $\lambda = 1.0$ balances the two objectives, and $\mathscr{L}_{\text{concepts}}$ is binary cross-entropy loss for concept predictions.

5.5.2 Training Configuration

- Optimizer: Adam with learning rate $\eta = 0.01$ and weight decay $\lambda_{\text{wd}} = 4 \times 10^{-5}$
- Batch size: 32 for training, 64 for evaluation
- Epochs: 100 with early stopping based on validation loss
- Class balancing: WeightedRandomSampler to ensure balanced batches (disabled by default on baseline as LDAM loss handles imbalance)
- Hardware: Trained on Apple M1 GPU (MPS) with memory-optimized batching

5.5.3 Decision Threshold Selection

We optimized the decision threshold on the validation set using precision-recall analysis. For each candidate threshold $\tau \in [0.05, 0.95]$ (step 0.01), we computed:

$$\tau^* = \underset{\tau}{\operatorname{argmax}}\, \mathrm{Recall}(\tau) \tag{5.2}$$

The threshold maximizing recall is a sensible choice to prioritize sensitivity in detecting depression cases, which is clinically important for screening applications.

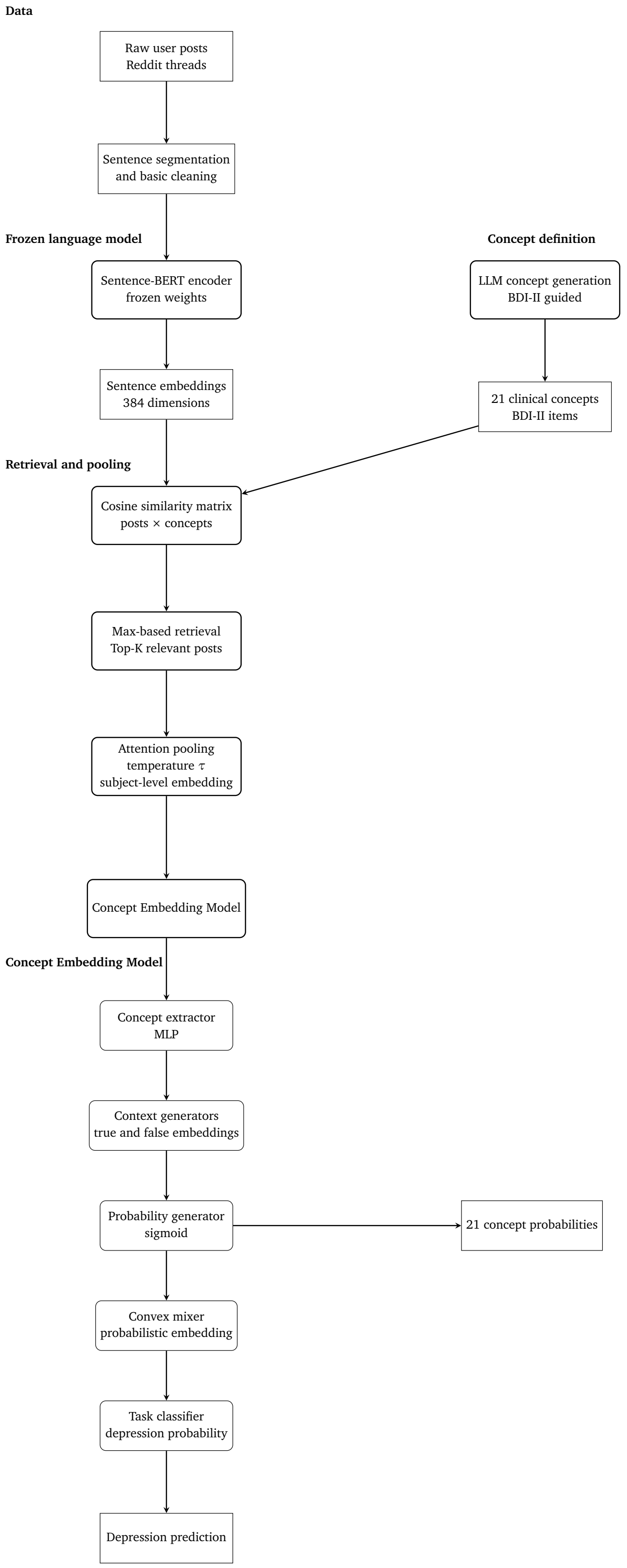

Figure 5.7. Overview of the full baseline pipeline, including concept generation, max based retrieval, attention pooling, and the Concept Embedding Model with concept level outputs and final depression prediction.

Chapter 6

Results and key findings

6.1 Evaluation Protocol and Results

Evaluating mental health screening systems from user generated text is inherently challenging. The task involves noisy language, sparse and indirect symptom expression, strong class imbalance, and labels that reflect clinical status rather than explicit linguistic markers. These difficulties have been explicitly acknowledged in the original eRisk shared tasks, where even top performing systems exhibit substantial error rates and strong tradeoffs between recall and precision (Losada and Crestani [2017]).

Within this context, the evaluation presented in this section should be interpreted as an assessment of whether the proposed concept based pipeline can operate effectively under realistic conditions, rather than as an attempt to maximize benchmark performance. In particular, all results reported here are obtained by optimizing the decision threshold for recall, reflecting a screening oriented use case where failing to identify a depressed user is considered more costly than producing false alarms.

6.1.1 Confusion Matrix Analysis

Standard binary classification metrics are reported, including accuracy, balanced accuracy, ROC-AUC, Matthews correlation coefficient, precision, recall, and F1 score. In addition, the full confusion matrix is provided to enable a transparent interpretation of error modes.

Table 6.1 reports the confusion matrix obtained on the test set.

	Predicted Negative	Predicted Positive
Actual Negative	268	81
Actual Positive	6	46

Table 6.1. Confusion matrix on the eRisk 2018 test set.

Out of 52 depressed users, 46 are correctly identified, corresponding to a recall of 88.5 percent. Only 6 depressed users are missed by the model. This level of recall is consistent with the objective of minimizing false negatives in a screening scenario.

At the same time, 81 control users are incorrectly flagged as depressed, corresponding to a false positive rate of 23.2 percent. While this represents a non negligible number of false alarms, it is important to note that such errors are expected in recall optimized systems operating under severe class imbalance. In a real world setting, these cases would typically be subject to further assessment rather than interpreted as definitive diagnoses.

6.1.2 Quantitative Performance Metrics

Table 6.2 summarizes the main evaluation metrics.

Metric	Value
Accuracy	0.7830
Balanced Accuracy	0.8263
ROC-AUC	0.8833
Matthews Correlation Coefficient	0.4712
Precision (Positive class)	0.3622
Recall (Positive class)	0.8846
F1 Score (Positive class)	0.5140
F1 Score (Macro)	0.6872

Table 6.2. Evaluation metrics on the eRisk 2018 test set.

The ROC-AUC of 0.883 indicates strong ranking performance independent of any fixed threshold, suggesting that the model is able to assign higher scores to depressed users even when precision is sacrificed for recall. Balanced accuracy further confirms that performance is not dominated by the majority class.

Precision on the positive class is comparatively low at 0.36, which is an expected consequence of recall oriented threshold optimization under heavy class imbalance. Importantly, this does not reflect a failure of the underlying model but rather a deliberate operating point choice aligned with the screening objective. The Matthews correlation coefficient of 0.47 indicates a meaningful positive association between predictions and true labels despite skewed class proportions.

6.1.3 Per Class Performance

A class wise breakdown of precision, recall, and F1 score is reported below for completeness.

Class	Precision	Recall	F1 Score
Negative	0.98	0.77	0.86
Positive	0.36	0.88	0.51

Table 6.3. Per class performance on the test set.

The model exhibits very high precision on the negative class, indicating that most users predicted as healthy are indeed controls. For the positive class, recall remains high while precision decreases, reflecting the intended asymmetry of the decision rule.

Taken together, these results indicate that the proposed concept based pipeline achieves competitive performance on a difficult benchmark while maintaining a clear and interpretable decision structure. The model performs well in identifying depressed users under realistic noise and imbalance conditions, consistent with the goals of this thesis and comparable to top performance levels reported in the eRisk shared tasks (Losada and Crestani [2017]).

6.1.4 Performance on Concept Predictions

Concept prediction represents a substantially more challenging problem than the downstream depression classification task. While the final label benefits from aggregation across many posts and multiple correlated signals, individual clinical concepts are sparse, highly imbalanced, and in some cases extremely rare, as detailed in the descriptive analysis section. All concept level evaluations reported here are computed on the validation set, as the hyperparameter optimization done does not affect them.

The model outputs continuous probabilities for each of the 21 BDI-II derived concepts, reflecting the degree to which each symptom is expressed in a subject's textual history. These outputs are not optimized as standalone classifiers, but rather as intermediate representations used by the final decision layer. Consequently, their evaluation should be interpreted primarily in terms of calibration, dynamic range, and discriminative signal, rather than raw thresholded accuracy.

Distribution of Concept Activation Probabilities

Across all concepts and validation samples, predicted probabilities exhibit a strongly right skewed distribution. The global mean probability is 0.0618, with a median of 0.0040, indicating that for most subjects and most concepts, the model assigns very low activation values. The maximum observed probability is 0.6550, and even for the most strongly expressed concepts, predicted values rarely exceed 0.6.

This behavior is consistent with the underlying data distribution. Most clinical symptoms are absent for most subjects, and even among depressed individuals, only a subset of concepts are typically expressed. The model therefore learns a conservative activation regime, reserving high probabilities for rare but strongly supported signals. This effect is particularly evident for concepts such as suicidal thoughts or wishes, agitation, and irritability, where both prevalence and textual evidence are limited.

At the per concept level, substantial heterogeneity is observed. Concepts such as past failure and self-dislike show higher means, larger variances, and broader probability ranges, suggesting that these symptoms are both more prevalent and more consistently expressed in language. In contrast, few other concepts exhibit near zero medians and very narrow interquartile ranges, reflecting extreme class imbalance and limited learnability.

Thresholded Concept Classification

For completeness, concept predictions were initially evaluated using a fixed decision threshold, treating each concept as an independent binary classification task. However, such threshold based evaluation was found to be fundamentally misleading in this setting. The underlying issue is not merely class imbalance, but the role these concepts are designed to play within the architecture. Concept outputs are learned as continuous latent variables whose primary function is to encode symptom related evidence in a graded and uncertainty aware manner, rather than to act as standalone diagnostic decisions.

When forced into a binary decision regime, standard classification metrics such as accuracy, precision, recall, and F1 score become dominated by prevalence effects. High accuracy values primarily reflect the overwhelming proportion of negative instances, while near zero precision and recall for most concepts obscure the presence of meaningful but low magnitude activations. Only a small subset of concepts that are both sufficiently prevalent and linguistically salient yield non zero F1 scores under a high threshold, while rarer symptoms are systematically suppressed. Lowering the threshold increases recall but does so by trading away precision in an uncontrolled manner, further highlighting the inadequacy of discrete decision boundaries for this task.

As a result, concept level performance is more appropriately assessed through distributional analysis of activation probabilities, dynamic range, and relative ordering across subjects, rather than through threshold dependent metrics. Concept prediction performance therefore reflects the intrinsic difficulty of the task. The model successfully learns structured, low noise intermediate representations that support strong downstream performance, while avoiding overconfident predictions on rare or weakly supported concepts. These findings align with the intended role of concepts in the proposed architecture, namely as interpretable latent variables rather than standalone diagnostic classifiers.

Interpretation and Probability Scaling

An important practical implication of these results concerns the interpretability of concept activations. While the raw probabilities are well suited for internal model reasoning, their compressed range may hinder human interpretability, particularly in clinical or explanatory contexts where values between 0 and 100 are more intuitive.

Rescaling or calibrating concept probabilities for reporting purposes may therefore be desirable. Such transformations would not alter the model's internal decision process, nor would they affect relative ordering between subjects, but could improve the readability of concept profiles. Importantly, any rescaling should be understood as a visualization and communication step, rather than a modification of the learned representations. The conservative nature of the raw outputs is itself a meaningful signal, reflecting uncertainty, sparsity, and the difficulty of inferring fine grained clinical symptoms from text alone.

6.2 Comparison with eRisk 2017 Results

This section compares the proposed approach with the official results reported in the eRisk 2017 shared task. The dataset used in this work is identical to that employed in the competition, ensuring a fair and direct comparison. While several eRisk submissions explicitly focused on early risk detection, evaluating performance after observing only a limited number of posts, it is noteworthy that the models achieving the best early detection scores are also those that perform best under standard binary classification metrics.

For this reason, and given the focus of this thesis on reliable screening rather than early intervention timing, the comparison is conducted using standard precision, recall, and F1 score for the positive class. ERDE metrics are therefore omitted from the table for clarity. Nonetheless, it

is worth noting that the submissions achieving the lowest ERDE scores also correspond to the top performing systems in terms of F1 and recall.
From a clinical screening perspective, recall is of particular importance. Missing at risk individuals comes with higher risk and recall values of at least 0.8 are commonly considered a desirable lower bound in screening scenarios. Precision, while still relevant, is typically secondary to recall in this context.

Table 6.4 reports the performance of all eRisk 2017 submissions alongside the baseline model proposed in this thesis.

Table 6.4. Comparison with eRisk 2017 submissions using standard classification metrics.

Model	**Precision**	**Recall**	**F1**
GPLA	0.22	0.75	0.35
GPLB	0.18	0.83	0.30
GPLC	0.42	0.50	0.46
GPLD	0.39	0.60	0.47
FHD0A	0.61	0.67	0.64
FHD0B	0.69	0.46	0.55
FHD0C	0.57	0.56	0.56
FHD0D	0.63	0.52	0.57
FHD0E	0.51	0.73	0.60
UArizonaA	0.31	0.58	0.40
UArizonaB	0.33	0.27	0.30
UArizonaC	0.21	0.92	0.34
UArizonaD	0.32	0.79	0.45
UArizonaE	0.34	0.63	0.45
LyRA	0.11	0.19	0.14
LyRB	0.11	0.29	0.16
LyRC	0.12	0.25	0.16
LyRD	0.13	0.17	0.15
LyRE	0.11	0.06	0.08
UNSLA	0.48	0.79	0.59
UQAMA	0.48	0.60	0.53
UQAMB	0.49	0.46	0.48
UQAMC	0.50	0.37	0.42
UQAMD	0.64	0.27	0.38
UQAME	0.45	0.35	0.39
CHEPEA	0.38	0.65	0.48
CHEPEB	0.37	0.63	0.47
CHEPEC	0.37	0.63	0.46
CHEPED	0.36	0.62	0.45
NLPISA	0.12	0.21	0.15
Proposed Baseline	**0.36**	**0.88**	**0.51**

Overall, the proposed baseline achieves recall well above the clinically desirable threshold, exceeding most eRisk 2017 submissions and approaching the highest recall reported in the competition. In particular, its recall is comparable to that of the top recall oriented systems, such as UArizonaC, which attains very high sensitivity at the expense of extremely low precision. In contrast, the proposed model maintains substantially higher precision and F1 score, indicating a more balanced trade off and a markedly lower rate of false positives.

When compared to the two ERDE winning systems of the eRisk 2017 challenge, which also represent the strongest overall performers according to standard classification metrics, the proposed baseline surpass them in terms of recall while operating with a significantly simpler and more lightweight architecture. Although the ERDE winners achieve higher F1 scores overall, they do so through substantially more complex modeling pipelines and without providing intermediate, human interpretable representations, and catching less positives.

These results suggest that the proposed approach occupies a favorable position in the performance spectrum, combining high recall, acceptable precision, and strong balanced accuracy, while simultaneously offering structured, concept level explanations. This balance is particularly relevant in mental health screening scenarios, where minimizing missed cases is critical, but excessive false alarms can lead to unnecessary clinical burden. The observed trade off is therefore consistent with the intended design goals of the concept based architecture and supports its suitability for real world screening applications.

6.3 Model Variants and other Experimental Findings

6.3.1 Concept Bottleneck Models did not work

Concept Bottleneck Models are based on the idea that predictions must pass through a layer of human defined concepts, which in this work correspond to clinically grounded depressive symptoms such as sadness, loss of pleasure, or changes in sleeping pattern. The architecture is divided into two mappings, from input texts to concepts, and from concepts to the final depression label. This design offers high transparency, since every prediction is explained through concept activations.

Multiple CBM variants were implemented in this project, including soft and hard bottlenecks, single stage and two stage training, and different regularization strategies. Despite extensive experimentation, CBMs systematically failed to reach acceptable task performance. Moreover, they did not learn stable or meaningful concept representations. Adjusting the decision threshold mainly produced trivial shifts from predicting almost everything negative to almost everything positive, rather than revealing a genuine discriminative structure.

The main reason lies in the interaction between modeling assumptions and the properties of the data. CBMs rely on the availability of reliable concept supervision. In the ideal scenario, each instance has ground truth annotations for all concepts, and the model can accurately learn both stages of the bottleneck. In this work, true manual concept labels are not available. Instead, concept supervision is derived automatically through similarity based scoring and distant supervision. These signals are useful as soft continuous indicators, but they are too noisy and weak to sustain hard concept supervision of the CBM type. As a result, the model is forced to learn a difficult target variable through an intermediate representation that is itself only approximately defined.

A second limiting factor is the nature of the eRisk data. User timelines are long, heterogeneous, and dominated by posts unrelated to mental health. Mentions of depressive symptoms are sparse, indirect, figurative, or expressed through slang and context dependent language. CBMs implicitly assume a clearer alignment between examples and concepts, whereas in this case concepts are rare and extremely imbalanced. The bottleneck therefore compresses the information into a small number of binary gates, which in practice either collapse to constant zero predictions for most concepts or become highly unstable under small threshold variations.

A third important issue is error amplification. In a CBM the entire information flow must pass through the concept layer. Once the representation is bottlenecked, the downstream classifier cannot correct for errors made in the concept predictor. In the two stage CBM used here, the first stage learns concepts and is then frozen, while the second stage learns the final label using only the predicted concepts. Any noise or systematic bias introduced in the first stage becomes structurally embedded in the second stage. Empirically, this resulted in low recall for the positive class or in degenerate solutions dominated by the majority class.

Different bottleneck configurations were tested to mitigate these problems, including joint optimization of concepts and label, discrete and continuous bottlenecks, LDAM loss for imbalance, temperature scaling, and alternative aggregation functions. None of these attempts led to consistent learning of useful concepts, nor to competitive downstream performance. The difficulty is therefore structural rather than just a question of hyperparameter tuning.

This negative result is nevertheless informative. It shows that CBMs, although attractive from the perspective of interpretability, are not well suited to settings where concept labels are weak, automatically generated, and embedded in highly noisy social media text. At the same time, this motivates the use of Concept Embedding Models. CEM relaxes the bottleneck by allowing concept information to be represented in a distributed manner rather than through discrete binary gates, and in the experiments of this thesis it succeeds in learning meaningful internal structure while preserving interpretability and achieving strong predictive performance.

6.3.2 Larger Models and Higher Dimensional Embeddings

A natural direction for improving performance is to increase model capacity. In neural models for text, this can be achieved in two complementary ways: by adopting richer sentence representations and by enlarging the downstream network operating on them. In principle, higher dimensional embeddings should allow the model to encode more nuanced semantic relations, while deeper or wider architectures should increase functional expressiveness.

However, in the context of concept based architectures this interacts directly with the information bottleneck. If the embedding dimension and the network size increase without a corresponding increase in the number of training samples, the model is prone to overfitting. Conversely, if the model capacity is deliberately constrained to enforce a bottleneck, richer embeddings cannot be fully exploited. This creates a structural trade off between expressiveness and regularization that is particularly acute in small, noisy datasets such as eRisk.

To investigate this systematically, additional data were incorporated. The 2020 corpus was merged with the 2017 data prior to constructing the train and validation split, in order to increase both the number of users and the lexical diversity of posts available during training. This mitigates overfitting risk and partially relaxes the data scarcity constraint.

On the representation side, a more powerful sentence embedding model was adopted, namely `all-mpnet-base-v2`, which provides dense contextual representations of substantially higher dimensionality than the SBERT variant used in the baseline model. To avoid creating an artificial information bottleneck, all layers in the concept extraction and task prediction modules were correspondingly widened, so that the network could process the additional representational capacity of the embeddings.

Multiple architectural configurations were explored, varying hidden layer widths, depth, and regularization strength. Across all tested configurations, the resulting models did not exhibit consistent performance improvements over the baseline presented in this thesis. In several cases, metrics remained statistically similar, while in others mild degradation due to overfitting was observed despite the enlarged training set. Best attempts managed to only slightly improve performance on the baseline method presented, with a ROC of 0.89 compared to baseline's 0.88.

This empirical result is informative. It suggests that limitations are not primarily due to insufficient model capacity, but rather to intrinsic properties of the data and the task. In particular, many users either display very weak linguistic evidence of depressive symptoms or express them in ways that remain semantically elusive to current NLP models. Under such conditions, simply scaling embedding dimensionality or network size does not unlock additional predictive signal. Instead, the key performance gains observed in this work arise from retrieval and aggregation strategies, especially MAX based selection of salient posts, rather than from brute force model enlargement.

6.3.3 Other unsuccessful variants

Beyond the CBM variants discussed above, several additional strategies were explored with the goal of exploiting stronger signals in the data or combining complementary sources of information. These approaches, while intuitively appealing, did not yield consistent improvements and in some cases degraded performance, further highlighting the difficulty of learning under noisy and weakly supervised conditions.

A first line of experimentation focused on amplifying extreme signals by manipulating the output distribution of the model through temperature scaling. The underlying intuition was that depressive language, when present, often appears in short bursts rather than as a uniform pattern throughout the user history. By sharpening the output probabilities, the model was encouraged to concentrate its mass on a smaller number of highly confident predictions, ideally corresponding to clearly pathological posts. In practice, the method failed to produce the desired effect. When the temperature was reduced, the model tended to collapse toward near binary predictions that were unstable and highly sensitive to noise. When the temperature was increased, predictions became uniformly flat and uninformative. In neither regime did the approach enable the model to identify discriminative extreme cases in a robust way. The empirical evidence therefore suggests that the problem is not a lack of confidence calibration, but the scarcity and ambiguity of truly unambiguous depressive signals in the data.

A second family of experiments attempted to combine information from different pooling and aggregation strategies. In particular, longer text embeddings were concatenated with both SUM based and MAX based similarity aggregations in order to exploit the potential complementarity between global similarity structure and salient outlying posts. The rationale was that SUM pooling might capture diffuse, low intensity signals spread across many posts, while MAX pooling might be better suited to detecting isolated acute expressions of distress. In practice, the combined representation did not improve prediction quality. The SUM component proved to be especially problematic. Since almost every post exhibits at least a small cosine similarity with most concepts, the summed similarity rapidly accumulates background noise and becomes dominated by dataset size rather than genuine conceptual evidence. The resulting representation is highly correlated with the number of posts and only weakly correlated with the presence of depressive symptomatology, which severely limits its discriminative value. Adding the MAX component on top of this noisy SUM representation did not compensate for the issue, and in some configurations it even hindered learning by forcing the classifier to disentangle heterogeneous and partially redundant features.

Overall, these negative results reinforce one of the central findings of the thesis. The challenge is not merely to extract stronger signals from the data by sharpening or combining representations. Rather, the fundamental difficulty lies in the intrinsic noisiness, sparsity, and imbalance of symptom level evidence in social media histories. Approaches that rely on diffuse accumulation of weak similarity signals tend to be dominated by noise, while those that attempt to aggressively amplify confidence often collapse or overfit.

Chapter 7

Conclusions and Future Work

This thesis investigated the applicability of concept-based learning approaches for mental health screening from social media text, with a specific focus on depression detection using the eRisk benchmark. The primary objective was not to maximize raw predictive performance, but to analyze whether clinically grounded, interpretable intermediate representations can be integrated into modern neural architectures without rendering performance impractical in a noisy, weakly supervised, and severely imbalanced setting.

7.1 Summary of Findings

The experimental results presented throughout this work lead to several key conclusions.

First, strict Concept Bottleneck Models were found to struggle in this application domain. While CBMs offer strong guarantees in terms of interpretability, their rigid information bottleneck proved ill-suited to the characteristics of social media data. Relevant mental health signals were sparse, indirect, and embedded within large volumes of unrelated content, and forcing all predictive information through a limited set of discrete symptom concepts led to a noticeable degradation in task performance. This confirms, in a real-world and clinically motivated setting, prior observations that strict bottlenecks can be overly restrictive when concept annotations are noisy, incomplete, or only weakly aligned with the underlying phenomenon.

Second, Concept Embedding Models emerged as a substantially more robust alternative. By relaxing the bottleneck constraint and allowing each concept to be represented as a continuous embedding rather than a scalar value, CEMs retained access to richer semantic information while preserving explicit concept-level supervision. When combined with appropriate mechanisms for handling class imbalance, including loss design and threshold optimization, CEM-based models achieved performance close to state-of-the-art baselines on the eRisk dataset. Importantly, this was accomplished without abandoning interpretability entirely. Intermediate predictions remained grounded in clinically meaningful symptom concepts derived from the Beck Depression Inventory–II, enabling inspection, analysis, and potential intervention at the concept level.

Third, the retrieval and aggregation strategy played a central role in overall system behavior. Treating user-level prediction as a multi-instance learning problem highlighted that not all posts contribute equally to mental health inference. Empirically, MAX-based retrieval and pooling strategies consistently outperformed sum-based or averaging approaches. This finding supports the hypothesis that, in highly noisy user histories, the presence of a small number of strongly indicative posts is more informative than weak, diffuse signals distributed across the entire posting history. From a practical perspective, this suggests that effective screening systems should prioritize the detection of salient, potentially alarming content rather than attempting to model subtle patterns across all available text.

Fourth, the baseline architecture adopted in this work, intentionally designed to be lightweight, fast, and easy to deploy, demonstrated favorable scaling behavior. Experiments with larger embeddings, more expressive encoders, and increased data availability indicated that performance improved consistently as model capacity and representational power increased. This suggests that the observed results are not a consequence of architectural limitations, but rather reflect the inherent difficulty of the task and the trade-offs imposed by interpretability constraints.

7.2 Limitations

Despite these encouraging results, several limitations must be acknowledged.

First, concept annotations were obtained using large language models rather than human experts. While this choice was motivated by scalability and supported by recent literature, it introduces an additional layer of noise and potential bias. The concepts should therefore be interpreted as structured proxies for clinical symptoms rather than ground-truth diagnostic indicators.

Second, this work deliberately set aside the temporal and early-detection aspects of the eRisk benchmark. While this allowed for a focused analysis of interpretability and representation learning, it limits direct comparability with systems optimized for early intervention scenarios.

Third, interventions at the concept level were analyzed conceptually but not exhaustively evaluated in practice. Although the adopted CEM architecture supports test-time interventions, the empirical study did not include systematic human-in-the-loop correction experiments.

Beyond technical challenges, the use of social media data for mental health assessment raises significant ethical concerns. Predictions derived from online text may influence clinical decision making, resource allocation, or individual self perception, yet they are based on data not originally produced for diagnostic purposes. Risks include false positives that may lead to unnecessary alarm or intervention, as well as false negatives that may provide unwarranted reassurance. Issues of consent, privacy, and data ownership are particularly salient, given that users may be unaware that their posts are being analyzed for mental health inference. These concerns underscore the importance of transparency, interpretability, and careful framing of model outputs as screening aids rather than diagnostic tools.

7.3 Future Work

Several promising directions for future research emerge from this work.

On the modeling side, more sophisticated retrieval strategies could be explored, including learned retrieval policies, attention-based instance selection, or hybrid MAX-attention mechanisms that combine interpretability with adaptive weighting. Similarly, alternative pooling strategies beyond simple operators may better capture intermediate regimes between sparse and diffuse evidence.

Extending the evaluation to additional benchmark datasets would help determine how far the results of this work generalize beyond eRisk and the specific task of depression detection. The eRisk collections offer several advantages, such as naturally noisy user generated text, strong class imbalance, and temporally structured data, all of which are highly representative of real screening scenarios and therefore valuable for stress testing interpretable models. At the same time, they also exhibit important limitations. The number of users is modest, which constrains statistical power. Moreover, results across this study suggest the existence of an empirical performance ceiling, likely reflecting the presence of users who either do not express symptoms linguistically or do so in ways that are extremely subtle. These characteristics make eRisk both an appropriate and a challenging testbed, but they also motivate future validation on larger and differently constructed datasets.

Applying the same framework to related conditions, such as anxiety or self-harm risk, would further test the robustness of clinically grounded concept spaces. Regarding this, the same methodology followed in this thesis can be replicated for different mental illnesses with minimal effort.

A particularly promising direction involves intervention-aware Concept Embedding Models. Recent extensions of CEMs explicitly incorporate simulated intervention trajectories during training, improving responsiveness to concept-level corrections.

Finally, future work could more directly involve domain experts in the evaluation loop. While this thesis emphasizes methodological analysis, the ultimate success of concept-based mental health screening systems depends on whether their explanations are perceived as meaningful, trustworthy, and actionable by clinicians and stakeholders.

7.4 Concluding Remarks

In conclusion, this thesis demonstrates that concept-based learning, when implemented with sufficient flexibility, can reconcile interpretability and performance in mental health screening from social media. While strict bottlenecks struggle under realistic noise and weak supervision, concept embedding approaches offer a practical and theoretically grounded alternative. By anchoring intermediate representations to clinically meaningful concepts while preserving expressive capacity, such models represent a promising step toward transparent, scalable, and ethically responsible mental health screening systems.

Bibliography

Cecilia Alm, David Roth, Timothy Sweeny, and David J Weiss. Emotional disclosure in essays and speech: Gender, depression, and linguistic differences. *Journal of Language and Social Psychology*, 26(3), 2007.

Douglas G. Altman. *Practical Statistics for Medical Research*. Chapman and Hall, 1994.

David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Jaime Amores. Multiple instance learning: A survey. *Pattern Recognition*, 46(5), 2013.

Elvio Amparore, Alan Perotti, and Paolo Bajardi. Who wants accurate models? arguing for a different metrics to take classification models to production. *arXiv preprint arXiv:2301.12563*, 2023.

Daniel Aragon-Guevara, Gregory Castle, Emma Sheridan, and Giacomo Vivanti. The reach and accuracy of information on autism on tiktok. *Journal of Autism and Developmental Disorders*, 2023.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 2020.

Eliseo Bao, Anxo Pérez, and Javier Parapar. Explainable depression symptom detection in social media. *IEEE Access*, 2023.

Peter L. Bartlett. Sample complexity of pattern classification under margin assumptions. *Journal of Machine Learning Research*, 1998.

Aaron T. Beck, Robert A. Steer, and Gregory K. Brown. *Manual for the Beck Depression Inventory–II*. Psychological Corporation, San Antonio, TX, 1996.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021. URL https://doi.org/10.1145/3442188.3445922.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Jeannette Boswell, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. URL `https://arxiv.org/abs/2108.07258`.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. *Optimization methods for large-scale machine learning*. SIAM Review, 2018.

Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. Generalized mean pooling for image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. *Proceedings of the 20th International Conference on Pattern Recognition*, 2010.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kaidi Cao, Colin Wei, Adrien Gaidon, Diego Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32, 2019.

Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16, 2022.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13, 2020. Key paper establishing MCC as the most informative metric for binary classification evaluation.

Arman Cohan, Nazli Goharian, Andrew Yates, Yi Hu, and Hal Daumé III. Detecting depression with social media: A multi-task multi-modal approach. *arXiv preprint arXiv:1807.03376*, 2018.

Corinna Cortes, Mehryar Mohri, and Yutao Zhong. Improved balanced classification with theoretically grounded loss functions. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. Curran Associates, Inc., 2025. URL `https://arxiv.org/abs/2512.23947`.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. doi: 10.48550/arXiv.1810.04805. URL `https://arxiv.org/abs/1810.04805`.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2), 1997a.

Thomas G. Dietterich, Richard H. Lathrop, and Tomas Lozano-Pérez. Solving the multiple instance problem: A review. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997b.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: why undersampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets*, 2003.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2001.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, 2001a.

Charles Elkan. The foundations of cost-sensitive learning. Technical report, University of California, San Diego, 2001b.

M. Espinosa Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, and M. Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35, 2022.

Mateo Espinosa Zarlenga, Katherine M. Collins, Krishnamurthy (Dj) Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jannik. Intervention-aware concept embedding models. *NeurIPS*, 2023.

Mateo Espinosa Zarlenga, Gabriele Dominici, Pietro Barbiero, Zohreh Shams, and Mateja Jamnik. Avoiding leakage poisoning: Concept interventions under distribution shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

S Evans-Lacko, S Aguilar-Gaxiola, A Al-Hamzawi, J Alonso, C Benjet, R Bruffaerts, W. T. Chiu, S Florescu, G de Girolamo, O Gureje, J. M. Haro, Y He, C Hu, E. G. Karam, N Kawakami, S Lee, C Lund, V Kovess-Masfety, D Levinson, F Navarro-Mateu, B. E. Pennell, N. A. Sampson, K. M. Scott, H Tachimori, M Ten Have, M. C. Viana, D. R. Williams, B. J. Wojtyniak, Z Zarkov, R. C. Kessler, S Chatterji, and G Thornicroft. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological Medicine*, 48(9), jul 2018. doi: 10.1017/S0033291717003336. Epub 2017-11-27.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 2006.

Zeyu Fei, Axel Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alex J. Smola. Multi-instance kernels. In *Proceedings of the 19th International Conference on Machine Learning*, 2002.

James Genone and Tania Lombrozo. Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, 25(5), 2012. doi: 10.1080/09515089.2011.627538.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:184487319.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), 2023. doi: 10.1073/pnas.2305012120.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Tyler R Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

Max Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*. PMLR, 2018.

Institute for Health Metrics and Evaluation. 2021 global burden of disease (gbd) online database. https://vizhub.healthdata.org/gbd-results/, 2024. Accessed 13 August 2025.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.01144.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Ronald C Kessler, Patricia Berglund, Olga Demler, et al. The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (ncs-r). *JAMA*, 289(23), 2003. doi: 10.1001/jama.289.23.3095.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*. PMLR, 2020.

Elizabeth Lawrence, Laura Robinson, and Clare Chambers. Opportunities and challenges in using social media for mental health research: A systematic review. *Journal of Medical Internet Research*, 26(1), 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017a.

Zhouhan Lin, Min Feng, Carlos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations*, 2017b.

Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2), 2008.

David E. Losada and Fabio Crestani. Overview of the clef 2017 early risk prediction on the internet (erisk) task. In *Proceedings of the CLEF 2017 Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings. CEUR-WS, 2017.

David E. Losada, Fabio Crestani, and Javier Parapar. Overview of the clef 2018 early risk prediction on the internet (erisk) task. In *Proceedings of the CLEF 2018 Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings. CEUR-WS, 2018.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Wei Pan. The promises and pitfalls of human evaluation for model interpretability. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, volume 10, 1998.

Brian W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, 1975.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 2019.

Christoph Molnar. *Interpretable machine learning*. Christoph Molnar, 2022.

Christian Montag, Florian J Stangl, René Riedl, and Raphael Kiemeswenger. Negative psychological and physiological effects of social networking site use: The example of facebook. *Frontiers in Psychology*, 14, 2024.

Tuomas Oikarinen, Subhrajit Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

James W Pennebaker, Martha E Francis, and Ryan J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahwah, NJ: Erlbaum*, 2001.

Pedro O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*, 2023.

Foster Provost and Tom Fawcett. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. Machine learning in medicine. *The New England Journal of Medicine*, 380(14), 2019. doi: 10.1056/NEJMra1814259.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. Are llms effective psychological assessors? leveraging adaptive rag for interpretable mental health screening through psychometric practice. *arXiv preprint arXiv:2501.00982*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. URL `http://arxiv.org/abs/1908.10084`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Stephanie S Rude, Emily-Maria Gortner, and James W Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(4), 2004.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Park Seonghwan and Namhoon Lee Jueun Mun, Donghyun Oh. An analysis of concept bottleneck models: Measuring, understanding, and mitigating the impact of noisy annotations. *arXiv preprint arXiv:2505.16705*, 2025. doi: 10.48550/arXiv.2505.16705. URL `https://arxiv.org/abs/2505.16705`.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Florian J Stangl, René Riedl, Raphael Kiemeswenger, and Christian Montag. Negative psychological and physiological effects of social networking site use: The example of facebook. *Frontiers in Psychology*, 14, 2023.

John Torous, Mary E Larsen, Munmun De Choudhury, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *Journal of Medical Internet Research*, 27(1), 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Bin Wang, Lijun Li, Yuta Nakashima, and Hajime Nagahara. Learning bottleneck concepts in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Hao Wang, Yitong Wang, Zheng Zhou, Xiangyu Ji, Dian Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yijun Yan, Li Chen, Jianping Fan, and Xiaohua Wu. Graph-based multiple-instance learning. *Pattern Recognition*, 74, 2018.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL `https://arxiv.org/abs/2407.10671`.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Chih-Kuan Yeh, Been Kim, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post hoc explanations may be ineffective for detecting unknown spurious correlations. *arXiv preprint arXiv:2210.12516*, 2022.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc Concept Bottleneck Models. *arXiv e-prints*, May 2022. doi: 10.48550/arXiv.2205.15480.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.