School of Computer Science and Electronic Engineering

MSc Data Science

Academic Year 2024-2025

Detecting Depression in Reddit Posts Using Natural Language Processing and Machine Learning Models

A project report submitted by: Rohan Raju Kamble / 6889740

A project supervised by: Tang H. Lilian

A report submitted in partial fulfillment of the requirements for the degree of Master of Science

University of Surrey
School of Computer Science and Electronic Engineering
Guildford, Surrey GU2 7XH
United Kingdom.
Tel: +44 (0)1483 300800

ABSTRACT

Early identification of mental health risk through online forums can enable early signposting towards support services. This dissertation constructs an explainable and reproducible NLP pipeline to classify Reddit posts as depression-related or not from a large anonymized public database of approximately 2.47 million posts. The project follows the CRISP-DM methodology: business and data understanding; preparation of data via text cleaning, stratified splitting, and train-only balancing to reduce the extreme class imbalance (77% non-depressed vs 23% depressed); feature extraction via TF-IDF, Bi-LSTM tokenization with GloVe embeddings, and SBERT embeddings with PCA; and diligent modelling and assessment.

Three model families were developed in complementarity. Transparency-conducive baselines (Logistic Regression with TF-IDF and Linear SVM) performed stable performance (F1 ≈ 0.927, ROC-AUC ≈ 0.978) and explainability via coefficients. A Bi-LSTM sequence model on pre-trained GloVe embeddings greatly boosted discrimination (F1 ≈ 0.944, ROC-AUC ≈ 0.986, PR-AUC ≈ 0.984), confirming the value of context-sensitivity in modelling. SBERT embeddings with linear heads offered competitive performance (F1 ≈ 0.91, ROC-AUC ≈ 0.965) and offered efficiency gains. In addition to headline metrics, the evaluation contained threshold tuning (Bi-LSTM optimal threshold = 0.514), probability calibration, confusion analysis, and interpretability using SHAP.

Findings show that context-aware approaches (Bi-LSTM, SBERT) excel over lexical baselines on recall-sensitive metrics, while Logistic Regression remains attractive where calibration and interpretability matter. The study closes knowledge gaps in the literature by systematically addressing class imbalance issues, such as PR-AUC and calibration analyses, and threshold performance analysis for risk triage in safety. Contributions are a documented pipeline, interpretability assessments, and reproducible artifacts to inform the design of human-in-the-loop digital mental health screening systems.

Keywords: Mental health, Reddit, Depression detection, NLP, TF-IDF, SVM, Bi-LSTM, SBERT, Explainability, Calibration, PR-AUC.

HIGHLIGHTS

- Full CRISP-DM pipeline applied to 2.47M anonymized Reddit posts
- Addressed severe 77/23 imbalance via train-only balancing and stratified splits
- Compared TF-IDF baselines, Bi-LSTM with GloVe, and SBERT+linear classifiers
- Bi-LSTM achieved F1=0.944, ROC-AUC=0.986, PR-AUC=0.984 (best overall)
- Threshold tuning (opt=0.514) and calibration enable safe risk triage
- Interpretability via LR coefficients and SHAP supports transparency

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Tang H. Lilian, for their continuous guidance, constructive feedback, and encouragement throughout this dissertation. Their expertise and support have been invaluable in shaping the direction and quality of this work.

I am also deeply thankful to my parents and family for their unconditional love, patience, and constant motivation during my MSc studies. A special thanks to my friends and classmates, who have provided not only academic insights but also encouragement and companionship during the challenging times.

Finally, I would like to acknowledge the University of Surrey for providing the resources and academic environment that made this research possible.

I certify that the work presented in the dissertation is my own unless referenced

Signature.....

Date.....

TOTAL NUMBER OF WORDS: 20670

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The goal of this dissertation overall is to develop an explainable and reproducible natural language processing pipeline to detect depression indicators from Reddit posts. With an anonymised dataset of approximately 2.47 million posts, the study combines traditional machine learning methods and deep learning and modern embedding approaches to investigate how different model families ranging from TF-IDF linear classifiers to BiLSTM networks and Sentence-BERT embeddings can be applied in mental-health risk detection on internet forums. The work addresses three persistent challenges in the literature: severe class imbalance in user-generated data, lack of generalisability between online communities, and the need for interpretability in sensitive healthcare uses. In addition, the work presents a more rigorous evaluation framework by incorporating PR-AUC, threshold tuning, and probability calibration, going beyond traditional accuracy-based evaluations. Through stringent experimentation and critical examination, the dissertation thus not only reports benchmarking outcomes but also produces methodological results that can inform the creation of secure, dependable, and human-in-the-loop digital mental-health screening systems.

1.1 Background

Mental illness is now considered by many to be the most significant public health problem of the twenty-first century, and depression is among the leading causes of disability globally. Over 280 million people suffer from depression across the globe, according to the World Health Organization, which makes it an urgent concern for healthcare practitioners, policymakers, and society at large. Along with traditional diagnostic methods, the extremely fast expansion of social media has created new possibilities for the detection of mental health risk by way of online activity. Reddit, indeed, is a forum in which participants habitually make personal and unguarded remarks under conditions of anonymity, offering a rich source of data for early detection of depressive tendencies. This dissertation is at the intersection of data science, natural language processing (NLP), and digital mental health analytics. Data science has repeatedly proven machine learning and deep learning methods' capability to extract knowledge from vast quantities of unstructured text. In computer science and information systems, reproducible and explainable AI systems become more pressing in life-critical applications such as healthcare, in which ethical and trustworthy deployment is of the same concern as brute performance. Applying them to the detection of depression in online settings is therefore technically challenging and socially pertinent.

Previous work shows an evolution from traditional NLP models such as TF-IDF with Logistic Regression or Support Vector Machines to more advanced architectures such as CNNs, LSTMs, and transformer-based models such as Sentence-BERT (SBERT). These approaches have yielded good results, with deep and transformer-based approaches particularly showing the ability to learn richer context and semantics. While promising, three recurring limitations in the literature are found nonetheless. First, severe class imbalance between depressed and non-depressed posts undermines the robustness of models against trivial baselines that over-predict the majority class. Second, generalization between communities is not well researched, so it remains to be seen whether a model trained on one subreddit will remain effective when transferred to others. Third, interpretability and calibration are not highlighted enough. Though linear models provide easily interpretable coefficients, most deep and embedding-based models are opaque, and few have verified whether their probability outputs are good enough to be useful in practice.

Addressing these problems provides the backdrop of ongoing research. The problem is put as follows: How do we construct an explainable, replicable NLP pipeline for detecting depression in Reddit comments that addresses class imbalance, generalisability, and provides interpretable, calibrated outputs suitable for sensitive tasks? Through the strict examination of baseline linear models, context-sensitive sequence models, and semantic embeddings within a CRISP-DM framework, this dissertation will compare performance, explore the role of threshold tuning and calibration, and offer methodological contributions that can be applied to guide the safe application of NLP in mental health screening.

The contribution of the work is both theoretical and applied. Theoretically, it offers comparative evidence at the model family level with more stringent evaluation metrics, including PR-AUC, threshold optimisation, and calibration, that are lacking in the literature. Applied, it provides an executable pipeline that captures the trade-offs across transparency, accuracy, and efficiency, and hence offers concrete insight into designing digital mental-health screening tools for safe human-in-the-loop deployment.

1.2 Research aim and objectives

Research Aim:

The objective of this research is to develop an explainable and reproducible NLP-based classification pipeline for the identification of depression in online posts on the platform Reddit, by testing historical baselines, deep sequential models, and transformer-based embeddings, with a view towards ascertaining the most effective and reliable methodology for digital mental-health screening.

Research Objectives:

- To critically review the literature on depression detection using NLP, identifying current methods, their strengths, and key gaps in evaluation and interpretability.
- To acquire and prepare a large, anonymized Reddit dataset, addressing challenges such as noise and severe class imbalance, and ensuring ethical use of online data.
- To implement and evaluate multiple model families: transparent baselines (TF-IDF with Logistic Regression and SVM), a context-aware Bi-LSTM sequence model, and SBERT embeddings with linear classifiers.
- To assess models rigorously using a range of performance measures, including F1, ROC-AUC, PR-AUC, threshold tuning, and probability calibration, going beyond accuracy to reflect real-world deployment needs.
- To explore interpretability by applying coefficient plots and SHAP analysis, providing insights into linguistic markers that drive model predictions.
- To synthesize findings across model families, highlighting trade-offs between accuracy, efficiency, and transparency, and producing a reproducible pipeline that can inform human-in-the-loop digital mental-health applications.

1.3 Research approach

This research follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which is a structured, cyclical data science approach having a well-defined framework for data science initiatives. CRISP-DM ensures that all steps of the dissertation are traceable, reproducible, and systematic to the developed research objective and goals.

The project begins with business and data understanding, wherein depression detection task is contextualized within digital mental health, and the Reddit dataset is critically examined. Preparation of data entails cleaning, normalization, and tackling extreme class imbalance by employing balancing methods and stratified splits. Modelling is done using three families of complementary NLP methods: interpretable baselines (Logistic Regression and SVM, and TF-IDF), a deep learning model that is context-sensitive (Bi-LSTM), and semantic embeddings (SBERT with linear classifiers). A wide variety of evaluation metrics such as accuracy, F1, ROC-AUC, PR-AUC, calibration, and threshold tuning are used together with interpretability tools such as coefficient plots and SHAP analysis.

Robustness is taken care of in a domain-shift experiment that tests whether models generalize between subreddits. Reproducibility and efficiency are addressed by storing trained artefacts and submitting timing experiments. The paper is based on an open-source, anonymized dataset for Reddit that contains no personally identifiable information, minimizing risk to ethics and meeting the requirements of the university. In essence, the research approach marries a structured data science process, rigorous experimentation over multiple families of models, and ethical safeguarding to produce an explainable and reproducible NLP pipeline for detecting depression in Reddit posts.

1.4 Dissertation outline

Chapter 2: Literature Review
This chapter provides an overview of the state of the art in depression detection from online text, covering traditional NLP methods, deep learning models, and transformer-based embeddings. It discusses the strengths and limitations of current approaches and identifies significant areas of research gaps in class imbalance, generalisability, and explainability that motivate this work.

Chapter 3: Research Approach (Methodology)
This chapter outlines the research method, applying the CRISP-DM methodology. It outlines the data set, steps in preparing the data, families of models selected, evaluation metrics applied, and ethical considerations that guided the project.

Chapter 4: Data Analysis and Results
This chapter introduces empirical analysis. It begins with exploratory data analysis for describing the dataset, followed by introducing and testing baseline models, a Bi-LSTM sequence model, and SBERT embeddings with linear classifiers. Performance is benchmarked on a variety of metrics, including F1, ROC-AUC, PR-AUC, threshold tuning, and probability calibration. Results are visualized through figures and tables, such as confusion matrices, calibration curves, and interpretability plots. A domain-shift experiment is also conducted to test generalisability.

Chapter 5: Discussion
The chapter critically evaluates the findings against the goals of the study and current

literature. It documents the strengths and weaknesses of all modelling families, deliberates on the trade-offs between accuracy, interpretability, and efficiency, and considers the ethical and practical aspects of applying NLP for mental-health screening in online forums.

Chapter 6: Conclusion and Future Work
The final chapter summarizes the main contributions of the dissertation, restates the research goal and objectives, and reflects on how they have been resolved. It also indicates the limitations of existing work and suggests areas for further research, such as modelling symptom severity, enhancing generalisability to different platforms, and incorporating finely-tuned transformer models.

CHAPTER 2: LITERATURE REVIEW

The purpose of this chapter is to place the dissertation within the overall body of work on detecting depression and other mental illnesses from natural language processing (NLP). It critically discusses the progression of approaches applied to text from social media, from traditional bag-of-words and TF-IDF features with linear classifiers, to deep models such as CNNs and Bi-LSTMs, and transformer-based models more recently, such as Sentence-BERT (SBERT). Besides techniques, the chapter describes the most significant methodological issues repeated in the literature: severe class imbalance in user-generated datasets, limited generalisability to online communities, and the non-interpretability and probability calibration of the most recent work. These issues are especially pertinent in healthcare-related applications, where operational and ethical constraints necessitate systems that are accurate, reliable, and interpretable. The chapter is structured as follows. Section 2.1 presents an overview of the usage of social media, and Reddit in particular, as a data source for research on mental health, outlining its advantages and limitations. Section 2.2 presents the traditional text representation and classification methods like bag-of-words, TF-IDF, and linear models that serve as baselines for this dissertation. Section 2.3 considers the application of deep learning models, such as CNNs and Bi-LSTMs, to depression detection and affective computing in general. Section 2.4 reviews semantic and transformer-based approaches, with particular emphasis on Sentence-BERT and other embedding methods currently at the state of the art. Section 2.5 presents cross-cutting methodological issues: class imbalance, evaluation measures, generalisability, interpretability, and calibration, and illustrates how each of these challenges has been attempted to be addressed across different studies. Finally, Section 2.6 concludes the literature review and identifies the areas of research gaps, primarily in evaluation design, interpretability, and generalisation, that dictated the methodology used in this dissertation.

2.1 Mental Health Detection on Social Media

Social media mental health detection has been a rapidly developing field of research in the past decade, both due to the availability of large amounts of user-generated text and due to the pressing need for early identification of at-risk individuals. Among mental illnesses, depression has been most studied due to its prevalence and also because it has observable manifestations in online behaviour. Social media platforms such as Twitter, Facebook, and Reddit have been particularly prominent in this line of research. Reddit is especially valuable because it contains topic-specific subcommunities (e.g., r/depression, r/anxiety) where users engage in rich self-disclosure under anonymous conditions. This combination of volume, topical relevance, and openness makes Reddit a fruitful space for computational detection, but also one with methodological and ethical challenges.

Early research in this area relied heavily on classical NLP and machine learning methods, namely bag-of-words and TF-IDF features combined with linear classifiers such as SVM and Logistic Regression. These were strong baselines; for example, Tadesse et al. (2019) reported F1 scores of around 0.93 with the combination of bigram features with psycholinguistic metrics (LIWC) and topic features (LDA). But while these models performed well on surface lexical cues, they weren't very effective at

capturing deeper context or semantic meaning, and they didn't provide much insight into calibration or cross-domain generalizability.

With the advent of deep learning, researchers began to experiment with architectures such as CNNs and LSTMs that would be better equipped to capture sequential dependencies in text. Hybrid models combining FastText embeddings, convolutional filters, and recurrent layers delivered substantial gains over linear baselines, particularly in recall for depressive posts. For instance, a Bi-LSTM on Reddit data improved the detection of subtle contextual cues of depression over bag-of-words methods, demonstrating the benefits of context-aware modelling. More recently, Ren et al. (2021) fused emotion-based attention mechanisms with semantic representations, achieving an F1 of 0.94 on a smaller Reddit dataset and highlighting the value of emotional cues for mental health detection.

The community has since shifted to transformer-based representations, tracking general NLP trends. Sentence-BERT (SBERT) and variants have been employed to generate semantic embeddings, usually in combination with lightweight classifiers. Chen et al. (2021) demonstrated that SBERT embeddings under a CNN head achieved an F1 of 0.86 on the SMHD dataset, outperforming earlier CNN and FastText baselines (F1 0.51 and 0.54, respectively). More recent work has moved beyond binary detection to symptom-level modelling, yielding RoBERTa-based classifiers producing clinically interpretable predictions concordant with DSM-5 categories and validated against gold standard measures such as PHQ-9 and GAD-7 (WebSci, 2023). The creation of new resources such as the ReDSM5 dataset (CIKM 2025), which includes expert rationales alongside DSM-5 symptom annotations, demonstrates growing interest in explainability as well as detection.

Despite this methodological progress, there are also ongoing issues. Extreme class imbalance undermines stability, and although some papers attempt sampling or class weighting, few report metrics such as PR-AUC or calibration curves that are more appropriate under imbalance. Generalization between communities or sites has also been understudied, with a lack of clarity regarding how models trained on one subreddit perform on others. Finally, while linear models have inherently interpretable coefficients, deep and transformer-based models are nigh impenetrable, with minimal examination of SHAP, attention visualization, or expert rationale alignment.

Study (Year)	Data / Task	Features / Models	Reported Performance	Notes / Gaps
Tadesse et al., IEEE Access (2019)	Reddit (binary post-level)	N-grams + LIWC + LDA with MLP / SVM / LR / RF	Acc 90%, F1 0.93 (MLP with LIWC+LDA+bigram)	Strong classical baseline combining psycholinguistic + topical features; limited calibration/generalisation discussion.
Ren et al., JMIR Med	Reddit (Pirina &	Emotion-based	Acc 91.3%, F1 93.98%	Demonstrates value of emotion cues; dataset

Inform (2021)	Çöltekin set: 1293 dep., 549 std.)	attention network (semantic + emotion branches, dynamic fusion)		relatively small; calibration/shift tests not covered.
Chen et al., SBERT-CNN (c. 2021–22)	SMHD (user-level Reddit; depressed vs matched controls)	SBERT sentence embeddings → CNN head	Acc 0.86, F1 0.86; outperforms prior best F1 0.79	Strong semantic+convolutional hybrid; summarises prior SMHD/RSDD baselines.
Early Reddit datasets (RSDD/SMHD baselines)	RSDD / SMHD (binary classification)	CNN, FastText, LR, SVM, XGBoost	CNN F1 0.51 (RSDD); FastText F1 0.54 (SMHD)	Useful historical baselines; largely lexical; limited interpretability.
WebSci (2023)	Reddit (symptom-specific subreddits) + Facebook validation	LIWC / LDA topics / RoBERTa embeddings; RF; logistic models	AUC-based symptom prediction; validated against PHQ-9/GAD-7/UCLA-3	Shift from binary detection to symptom-level modelling and cross-platform validation.
ReDSM5 (CIKM 2025)	Reddit (1,484 long posts; sentence-level labels)	DSM-5 symptom dataset with baselines for multi-label classification	Baseline symptom classification + explanations	Advances interpretability via DSM-5-aligned labels + clinical rationales.

Table 2.1: Comparison of key Reddit depression-detection studies

Table 2.1 compares and summarizes a variety of seminal studies in this area. The literature as a whole demonstrates a clear methodological trajectory from surface lexical to deep and semantic embeddings. However, systematic gaps remain: an absence of evaluation under imbalance, an absence of testing generalizability, and an absence of developed interpretability and calibration. These gaps directly motivate the present dissertation, which compares TF-IDF baselines to Bi-LSTM and SBERT models with explicit consideration of calibration, threshold tuning, and explainability.

2.2 Traditional NLP Approaches for Depression Detection

Initial attempts at social media depression detection relied on shallow text representations such as bag-of-words and term frequency–inverse document frequency (TF-IDF). These representations transform text into high-dimensional sparse vectors of word counts or weighted frequencies, which are then fed into machine learning algorithms for classification. Conceptually simple as they are, these methods produced surprisingly

strong baselines for binary classification of depressive vs. non-depressive posts, and they remain a standard for follow-up research to compare against.

Several works illustrate the strengths of these traditional models. Tadesse et al. (2019) combined TF-IDF with LIWC psycholinguistic features and LDA topic features. Using classifiers such as Logistic Regression, SVM, Random Forest, and Multi-Layer Perceptron, they achieved F1 scores of up to 0.93 on Reddit data, confirming that linear models coupled with hand-engineered features can be competitive. Similarly, Pirina and Çöltekin (2018) showed that TF-IDF with SVM achieved over 90% accuracy on a small but balanced Reddit dataset later reused by Ren et al. (2021). These findings reconfirmed the effectiveness of TF-IDF as a baseline representation.

For classifiers, Logistic Regression and linear SVM have been especially popular since they are scalable to large corpora and resistant to sparse input. Another benefit is interpretability: coefficients in linear models allow researchers to uncover prominent lexical cues for depression, such as "sad," "alone," or "hopeless." This interpretability is particularly valuable for mental-health applications where transparency is necessary. Other classifiers, such as Random Forests and gradient-boosted trees, have also been tried but perform worse in this setting, possibly because they overfit with very high-dimensional feature spaces. For each of these strengths, there are clear weaknesses. By ignoring word order and semantic context, bag-of-words and TF-IDF will misclassify posts with negations or indirect phrasing (e.g., "not feeling sad" vs. "feeling sad"). Their reliance on surface lexical patterns also implies that they may overfit to community-specific slang, decreasing cross-subreddit generalizability. Furthermore, early work evaluation was typically limited to accuracy or F1. More appropriate measures for imbalanced settings, e.g., PR-AUC, and reliability evaluations, such as calibration curves, were rarely considered. Thus, these models provide an incomplete view of how classifiers would perform in real-world, skewed deployment settings.

Study (Year)	Dataset	Features	Classifiers	Performance	Limitations
Tadesse et al., IEEE Access (2019)	Reddit (binary posts)	TF-IDF + LIWC + LDA	LR, SVM, RF, MLP	Acc 91%, F1 0.93 (MLP)	Relies on surface-level lexical + psycholinguistic cues; limited generalisability; no calibration
Pirina & Çöltekin (2018) (used by Ren et al., 2021)	Reddit posts (1293 dep., 549 std.)	Bag-of-words, TF-IDF	SVM, LR	Acc >90% (balanced small dataset)	Small sample size; ignores sequential/semantic context
Classical baselines (RSDD/SMHD)	RSDD, SMHD (user-level Reddit)	Bag-of-words, TF-IDF, n-grams	LR, SVM, RF, XGBoost	F1 ≈ 0.50–0.55	Sparse features are weak on long/noisy posts; interpretability is limited beyond coefficients

Table 2.2: Traditional NLP approaches for depression detection in Reddit/social media text

Table 2.2 gives an overview of typical studies using classical NLP approaches. Together, they show that strong headline performance can be achieved with shallow lexical features and linear classifiers, but also reveal inherent weaknesses: poor handling of context, no generalisation to datasets, and no consideration of calibration and threshold sensitivity. These weaknesses motivate the move to deep and semantic approaches discussed in the following sections, and they directly inform this dissertation's methodological contributions.

2.3 Deep Learning Approaches for Depression Detection

Although classical NLP approaches such as TF-IDF with linear classifiers produced strong baselines, they have the limitation of being unable to capture sequential dependence or semantic nuance. To counteract such a limitation, depression detection work has moved toward deep learning techniques capable of learning contextualized text representations directly from raw sequences.

One of the initial deep learning successes in text classification was by Convolutional Neural Networks (CNNs). CNNs employ n-gram window filters to detect local patterns and leading words, enabling the model to observe patterns rather than just individual word frequencies. On social media and Reddit data, CNN-based classifiers performed better than logistic regression and SVM baselines consistently, but could not capture long-distance dependencies and therefore were less effective for subtle posts based on broader context. This led to the use of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) models. LSTMs are able to remember earlier tokens in a sequence to facilitate context-dependent interpretation. Bidirectional implementations push this further by processing text in two directions, and they are especially effective at picking up negations and subtle phrasing (e.g., "not sad" vs. "sad"). Experiments with Bi-LSTM models on Reddit datasets have shown increased recall on depressive posts compared to TF-IDF baselines, especially beneficial in screening scenarios where incorrect negatives are costly.

Researchers have also explored hybrid models utilizing pre-trained embeddings with CNN or LSTM blocks. Models that included Fast-Text embeddings with CNN and Bi-LSTM blocks, for example, had significant improvements from linear baselines on the Reddit dataset. Such hybrids benefit from both the semantic closeness inherent in embeddings as well as the contextualized sequential information acquired by recurrent layers, demonstrating the power of layered representation learning. Similarly, Ren et al. (2021) presented an Emotion-Attention Bi-LSTM that integrated affective and semantic features and achieved an F1 of 93.9% on a small Reddit corpus, which demonstrates how humans pay attention to emotion cues when identifying depression.

Despite such advancements, deep learning models possess severe shortcomings. They require much larger amounts of data and computational resources than typical baselines,

limiting reproducibility in low-resource settings. Their improvements, although robust, are incremental relative to strong SVM baselines, raising cost-benefit trade-off issues. More seriously, deep networks are likely to be seen as "black boxes." Despite the use of attention mechanisms and visualization methods to shed light on internal workings, interpretability in the detection of depression remains more restricted than in linear models. This is a serious issue in health, where responsibility and trust are paramount.

Study (Year)	Dataset	Models	Performance	Notes / Gaps
Shen et al. (2017)	Reddit (early datasets, binary)	CNN for text classification	Improved over SVM baselines; F1 ≈ 0.60	Captured local n-gram features, but limited to long dependencies
Tadesse et al. (2019)	Reddit (binary posts)	MLP with LIWC + LDA + n-gram features	F1 = 0.93 (MLP outperformed SVM, RF)	Not a pure deep text model; relied on feature engineering
FastText + CNN-LSTM Hybrid (c. 2020)	Reddit depression datasets	FastText embeddings + CNN + BiLSTM hybrid	Significant gains vs. linear baselines	Combined semantic and sequential context; requires higher computation
BiLSTM implementations (2020–2022)	SMHD, RSDD (Reddit user-level datasets)	BiLSTM with pre-trained word embeddings	Recall gains vs. TF-IDF SVM; ROC-AUC > 0.95 in some cases	Better handling of negations/context; limited interpretability
Ren et al., JMIR Med Inform (2021)	Pirina & Çöltekin Reddit dataset (1293 dep., 549 std.)	Emotion-Attention BiLSTM (semantic + affective signals)	Acc = 91.3%, F1 = 93.9%	Added affective features; dataset is small; generalizability unclear

Table 2.3: Deep learning approaches for depression detection in Reddit/social media text

Table 2.3 summarizes superior deep learning work in this field. Taken together, these papers demonstrate how deep architectures enhance the state of the art by learning richer sequential and semantic relationships. They also demonstrate persistent problems with efficiency, reproducibility, and explainability. These problems directly inform this dissertation, which employs a Bi-LSTM with pre-trained GloVe embeddings, threshold tuning, and calibration things scarcely debated in the literature, to balance recall gains against interpretability and deployment considerations.

2.4 Transformer and Semantic Approaches

In recent years, the field of natural language processing has been transformed by the introduction of transformer-based architectures, most notably BERT (Bidirectional Encoder Representations from Transformers) and its derivatives. Unlike conventional CNNs or RNNs, which process text sequentially, transformers employ self-attention

mechanisms that capture long-range dependencies in parallel, enabling them to model semantic nuance with far greater accuracy. This paradigm shift has enabled models not only to outperform prior baselines but also to generate contextualised embeddings that can be readily adapted to downstream tasks. In the context of depression detection on social media, transformers have allowed researchers to move beyond surface-level lexical cues and sequential architectures towards dense semantic embeddings that capture richer meaning. Sentence-BERT (SBERT), for example, produces sentence-level representations optimised for semantic similarity, which can then be paired with lightweight classifiers. Chen et al. (2021) demonstrated that an SBERT-CNN hybrid achieved an F1 score of 0.86 on the SMHD dataset, outperforming previous CNN and FastText baselines. This result illustrates the capacity of SBERT embeddings to handle the noisy, variable language found in Reddit posts while maintaining competitive performance.

The scope of transformer-based research has also expanded beyond binary detection to symptom-level modelling. For example, RoBERTa embeddings have been combined with Random Forests to predict depression symptoms aligned with DSM-5 diagnostic criteria. These models were further validated against clinical screening tools such as PHQ-9 and GAD-7 on external Facebook data, demonstrating promising levels of cross-platform generalisability. This work represents a shift from simply identifying whether a user is depressed to mapping online disclosures to clinically meaningful constructs. A further step in this direction is the ReDSM5 dataset (CIKM 2025), which provides sentence-level annotations of Reddit posts matched to DSM-5 symptoms along with expert rationales. By enabling both classification and justification, this resource supports the development of explainable NLP systems that not only detect depression but also clarify *why* a model made its prediction. The integration of expert rationales makes it possible to align algorithmic predictions with clinical reasoning, addressing one of the key criticisms of transformer-based approaches: their opacity.

Despite these advances, transformer-based methods are not without limitations. Training and inference are computationally expensive, often requiring substantial hardware resources, which creates barriers to reproducibility and low-latency deployment. While transformers generally outperform Bi-LSTM and SVM baselines, the margin of improvement is sometimes modest relative to the resource investment. Furthermore, interpretability remains problematic. Although attention mechanisms, SHAP analyses, and expert-aligned rationales offer some insight, embedding spaces themselves are not inherently transparent, making it difficult to trace decisions in a way that is acceptable in healthcare settings.

Overall, transformer-based and semantic approaches represent the current state of the art in depression detection on Reddit and related platforms. They offer strong performance, improved robustness, and extensions to symptom-level analysis and explainability. However, the computational cost, reproducibility challenges, and interpretability issues leave space for more efficient or transparent methods such as TF-IDF + SVM or Bi-LSTM in certain contexts. These trade-offs directly inform the present dissertation, which benchmarks transformer-inspired semantic embeddings (SBERT) against traditional and deep learning approaches, while incorporating evaluation practices such as calibration and threshold tuning that remain underexplored in the literature.

2.5 Cross-cutting Challenges in Depression Detection Research

Despite advances in detection of depression from social media, having moved from linear lexical models to transformer-based and deep models, there remain some cross-cutting issues. They limit validity, generalisability, and ethical use of NLP-based screening tools and contributed to methodological decisions made in this dissertation.

Amongst the most longstanding of these issues is class imbalance. Imbalance is not always in the same direction. Early Reddit corpora (e.g., RSDD, SMHD) were depression-heavy because they were sampled only from self-report subreddits. By contrast, the large-scale dataset used in this dissertation spans multiple communities and shows the reverse imbalance: non-depressed posts form the majority (≈77%), while depressed posts are the minority (≈23%).. This is important. A model trained naively on such data could be extremely accurate by just outputting the majority class and possess little to no real diagnostic merit. In response to this, we balanced the training set only (50/50) to prevent majority-class bias and kept the validation and test sets stratified at their natural prevalence (≈77% non-depressed, 23% depressed). We also reported PR-AUC and conducted threshold and calibration analyses to avoid accuracy inflation under skew. This design decision is an explicit response to a lack in the literature in that there are many papers still concentrated on accuracy or F1 alone, without investigating sensitivity to prevalence.

Evaluation design is a collective failing. The majority of published models only report F1 and accuracy, with less detail for threshold analysis, calibration curves, or error profiling. In healthcare contexts, though, uncalibrated probabilities can be risky because false positive overconfidence can lead to false alarms and miscalibrated negatives to cause undue delays in access to requisite support. In the current dissertation, probability calibration was experimented on explicitly (e.g., Figure 4.21), showing how Logistic Regression maintained well-calibrated probabilities and BiLSTM was susceptible to overconfidence. These diagnostics convert "headline scores" into decision-support data clinicians may actually be capable of interpreting. Error analysis was also performed with confusion matrices and SHAP inspection to identify which cues were classifying and how error grouped.

Generalisability is yet another limitation in the wider literature. Everyone else uses one subreddit or dataset for training and testing, which overfits to cultural and linguistic norms of the specific community. Few experiments bridge domains, but a few symptom-driven experiments in recent times explored cross-platform validation. In an attempt to bridge this gap, this dissertation included a domain-shift experiment (Section 4.9), dedicating a whole subreddit to testing robustness against linguistic drift. This straightforward test provided evidence of the ability of each model to generalize outside its training domain, and BiLSTM emerged more robust than TF-IDF baselines.

Interpretability and explainability are also not well explored in previous research. While linear models such as Logistic Regression and SVM allow simple interpretation through feature coefficients, deep models such as BiLSTM and SBERT are mainly "black boxes." Some recent work has looked into SHAP or DSM-5–concordant rationales, but these are exceptions. Interpretabiltiy in this work was taken as a first-order requirement

rather than an afterthought. Coefficient plots and SHAP analyses (Figures 4.12–4.14) were used to show how linear baselines were learning important psychological signals, while calibration overlays highlighted confidence behaviour in deep models. This dual strategy of transparency and calibration aligns with ethical healthcare AI requirements.

Finally, pragmatic and ethical issues are poorly developed in the literature. Although datasets are anonymized, Reddit users did not consent to clinical inference, and there are issues with data use. In addition, classification errors are critical: false positives pose risks of mislabeling healthy users, and false negatives may prevent timely support for needy ones. Such risks highlight the importance of human-in-the-loop deployment and cautious interpretation. By focusing on evaluation depth, interpretability, and robustness across domain shift, this dissertation not only brings technical comparisons between modelling families but also provides insights into responsibly designing and evaluating NLP for mental health.

2.6 Summary

Chapter 2 has covered the research progression on social media depression detection, with a focus placed on the use of Reddit as the data source. Traditional approaches using bag-of-words and TF-IDF features and linear classifiers were extremely strong baselines and very interpretable, but were unable to capture contextual or semantic nuance. The arrival of deep learning models, such as CNNs, LSTMs, and Bi-LSTMs, improved the sequential dependency modelling and recall on depressive posts, even if usually at the expense of interpretability and computational cost. Most recently, transformer-based methods, such as SBERT and RoBERTa, came to be state of the art, offering contextualised embeddings and even symptom-level modelling. They remain computationally expensive and largely black-boxed, with issues regarding deployment in privacy-critical healthcare environments. Across these methodological camps, there are some common trans-cutting issues. Severe class imbalance still undermines stability, with many studies still only publishing accuracy or F1 and not more stable measures such as PR-AUC. Design of evaluation always neglects threshold analysis, calibration, and systematic error examination, all of which are crucial in risk-sensitive domains. Generalisability to platforms or communities remains untested, excluding high confidence in claims of external validity. Finally, ethics and interpretability, although increasingly appreciated, are not sufficiently integrated into most technical innovations.

Overall, the literature exhibits a trend of increasing methodological maturity yet persistently reports gaps in evaluation, calibration, domain shift robustness, and explainability. These shortcomings directly fuel the present dissertation, which builds an explainable and replicable pipeline for detecting depression on Reddit. Comparing a spectrum of model families while addressing imbalance, calibration, interpretability, and generalisation systematically, the research seeks to increase both methodological rigor and real-world usability.

The third chapter, Chapter 3: Research Approach (Methodology), outlines how, in reality, the aforementioned gaps are addressed. It introduces the CRISP-DM process that structures the research, the dataset used, and preprocessing methods, the motivations for using particular modelling families, as well as the measures of evaluation and the ethical issues governing the empirical study.

CHAPTER 3: RESEARCH APPROACH

This chapter outlines the methodological framework used to achieve the goal and objectives of the dissertation. Research employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which structures data science research into six iterative stages: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. CRISP-DM is particularly appropriate for this project because it provides a systematic but flexible process that ensures each stage of the pipeline is distinct, replicable, and explicitly matched to the research questions. It also supports iterative tuning, which was crucial given the challenges of extreme class imbalance, interpretability, and generalisability documented in the literature review. Here, the chapter explains the rationale behind the methodological choices made and indicates how such choices address the limitations found in previous research. The dataset is introduced and its source, structure, and ethics are explained. Data preparation steps such as cleaning, normalization, and class balancing are then explained. The modelling strategy is introduced in three competing families of complementary methods: transparent baselines with TF-IDF and linear classifiers, context-sensitive deep sequence modelling using Bi-LSTM, and semantic embedding strategies drawing on SBERT. The evaluation framework is then presented, comprising metrics, threshold tuning, probability calibration, and domain-shift testing to quantify generalisability. Finally, the chapter concludes by outlining reproducibility considerations and ethical safeguards, in order to make sure that the pipeline can be reproduced and used in a responsible manner.

3.1 Research Design

Study design follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) process, which is a formal, iterative process for data science projects. Some methodological frameworks have been taken into account, including KDD (Knowledge Discovery in Databases), SEMMA (Sample, Explore, Modify, Model, Assess), and GABDO (Goal–Analysis–Design–Build–Operate). Each of these has a systematic way of dealing with data-driven research, but CRISP-DM was selected for three reasons. First, it provides a comprehensive six-step methodology (business understanding, data understanding, data preparation, modelling, evaluation, and deployment) closely matching the steps required for this project. Second, CRISP-DM is iterative in nature, whereby feedback between phases can be achieved, which is particularly required in handling real-world issues such as class imbalance or calibration. Third, CRISP-DM is widely known in both industrial and scholarly worlds, ensuring that the process is reproducible, transparent, and adheres to best practices.

In the context of this dissertation, CRISP-DM structures the research pipeline as follows. The business understanding step puts depression detection within the scope of digital mental health at large, the identification of society, and the technical relevance of stable and explainable models. The data understanding step involves exploratory analysis of the Reddit data set in terms of class distributions, post lengths, and correlations to inform next-stage modelling choices. The data preparation step involves cleaning, normalisation, training data balancing, and feature extraction for multiple model families (TF-IDF, Bi-LSTM tokenization, SBERT embeddings). The modelling step instantiates three distinct families of classifiers: interpretable linear baselines, a Bi-LSTM deep learning classifier, and transformer-based semantic embeddings. The evaluation stage employs not only usual metrics such as accuracy and F1 but also PR-

AUC, ROC-AUC, threshold tuning, and probability calibration, which fill gaps that existed in the literature. Finally, the deployment stage is addressed at a conceptual level by making the models, artefacts, and analysis outputs reproducible and could, in principle, be integrated into a human-in-the-loop moderation pipeline.

Access to data is also a consideration. This research employs an anonymised Reddit dataset available publicly from Kaggle. The dataset contains no personalizable data, and hence it is suitable for secondary analysis without the need for approaching further participants. Ethical responsibility is fulfilled through compliance with university policy on secondary data usage, and a statement of ethical issues can be found in the appendix. In general, the research design employs CRISP-DM as a systematic and iterative process to ensure that all steps of the research are aligned with the overall objective of developing an explainable and reproducible NLP pipeline for detecting depression in Reddit comments.

3.2 Data Collection and Characteristics

The corpus utilized within this work is sourced from Kaggle, where it was made available as an open-source dataset to aid mental health detection research. It comprises about 2.47 million posts on Reddit and is one of the biggest publicly available corpora used for this task. Each row of the corpus has eight features: an identifier based on the index, subreddit of origin, post title, post text, upvote count, time of creation (timestamp or created), comment count, and a binary feature that separates the posts as either depressed or non-depressed.

Initial inspection confirmed that the dataset is both vast and diverse and reflective of the informal and diverse language of social media. The class distribution was highly imbalanced, with roughly 77% of posts labelled non-depressed and 23% labelled depressed. This is a serious methodological flaw: a non-corrected classifier could attain strong accuracy simply by predicting the majority class but would fail at picking at-risk individuals. Exploratory data analysis further identified that depressed posts were much longer than non-depressed posts on average, at 98 words vs. 31 words per post within the non-depressed class. This is in accord with our hypothesis that post length is itself a weak yet significant predictor of class and is in accord with previous findings within psychological and computational work that individuals in distress tend to use more complex syntax when self-disclosing. The dataset was appropriate to the objectives of this work for two reasons. Firstly, it comprised sufficient examples to train data-avid deep learning models like Bi-LSTMs, but was still feasible to derive computationally reduced baselines by comparison. Secondly, its format, consisting of posts labelled at the individual level, made it possible to perform balanced and stratified splitting into training, validation, and test sets and carry out fair and reproducible evaluation.

Getting the dataset was not complicated: it was downloaded from Kaggle and specified as a public anonymized dataset, the use of which guarantees the non-existence of any recognizable personal information. Hence, the ethical issues were minimal, largely due to the fact that the creators of the dataset had anonymized usernames and other sensitive attributes previously. Nevertheless, the requirements of ethics were met by viewing the dataset as secondary data, following the requirements of the university, and hence ensuring that all the analyses were restricted to aggregate non-identifiable data. In

conclusion, the data constitute a full, anonymized, and ethically proper basis for the investigation of depression detection. Its scale and shape permit experimentation on it under a variety of modelling families; its unevenness and diversity contribute to real-world issues the proposed approach aims to address.

3.3 Data Preparation

Once the dataset was received, careful preparation was needed to take the raw, unstructured posts from Reddit and put them in a structured and trustworthy format conducive to modelling. Social media data is notoriously noisy: posts tend to contain anomalous grammar, typos, URL links, and user references, and these can make feature extraction and learning problematic. Accordingly, the preparation procedure involved systematic steps of cleaning, normalization, balancing, and feature extraction applicable to the general methodology of CRISP-DM.

Textual preprocessing and normalization. Initially, all text was converted to lowercase to mitigate the sparsity that arises from inconsistent capitalization. Non-alphanumeric characters, such as URLs, email addresses, and user mentions specific to Reddit, were eliminated. Excessive punctuation and duplicated whitespace were removed, and any posts that remained empty or trivial following the cleaning process were discarded. These procedures standardized the textual data, thereby ensuring that the models concentrated on significant lexical and semantic attributes instead of irrelevant formatting noise.

Class balancing. As observed in Section 3.2, the raw data were highly imbalanced such that roughly 77% of the posts were non-depressed and only 23% were depressed. In order not to have a majority-class prediction by the models, only the training set was balanced. This procedure generated a balanced training distribution of roughly 1.98 million instances evenly split between classes, whereas the validation (495,102 posts) and test (618,878 posts) sets were left stratified but imbalanced so that they will have real-world prevalence rates. This configuration makes it so that the models are trained using a balanced training signal, but are then tested under realistic real-world conditions. Data splitting. The data was divided into training, validation, and test sets. Stratified sampling was used such that the class ratio was maintained within the validation and test sets. Training was performed using the training set, hyperparameter tuning, and threshold optimisation using the validation set, and the test set was used for the final out-of-sample test. This three-fold split was used such that the evaluation measures were unbiased and generalisable.

Feature extraction. Every family of models underwent a variety of feature extraction techniques for both methodological diversity and comparability: To calculate baseline models' unigram and bigram TF-IDF feature vectors, a vocabulary of up to 10,000 features was used. This yielded high-dimensional sparse representations suitable for linear classifiers such as Logistic Regression and SVM. In the Bi-LSTM sequence model, the text was tokenized with a vocabulary of 30,000 words and padded/truncated posts up to a maximum of 200 tokens. Initializations of word embeddings were made using pre-trained GloVe (100-dimensional) word vectors with about 27,000 tokens embedded to ensure that the baseline started with semantically useful representations.

In the case of SBERT embeddings, each of the posts was converted to a sentence-level SBERT embedding based on a pre-trained Sentence-BERT. Embeddings were shortened and made computationally more affordable by compressing them to 100 principal components using PCA and preserving most of the variance to enable faster downstream classification.

In short, data preparation transformed raw and noisy posts from Reddit into structured, balanced, and model-specific sets of features. By fixing the imbalance at the training step, carefully splitting data sets, and tailoring feature representations at each category of models, the preparation work laid strong preliminaries toward those of Section 3.4 modelling experiments.

3.4 Modelling Techniques

To evaluate different methodological paradigms for depression detection on Reddit, this dissertation implemented three families of models: (i) interpretable linear baselines, (ii) a context-aware deep learning network, and (iii) semantic embedding models based on Sentence-BERT. This unified design reflects both the direction identified in the literature review and the research aim of comparing conventional, sequential, and semantic approaches within a reproducible framework.

Logistic Regression (LR) and linear Support Vector Machines (SVM) served as the baseline classifiers on TF-IDF features. Both are linear models, but they rely on different learning principles. LR is a probabilistic model: it estimates the log-odds of the depressed class as a weighted sum of input features and applies the logistic sigmoid to produce calibrated probabilities. The training objective is to minimise the binary cross-entropy loss (equivalent to maximising the likelihood of the observed labels), with an L2 penalty to avoid overfitting. In contrast, a linear SVM is a margin-based classifier: it learns a hyperplane that separates the two classes while maximising the margin between them. Its training relies on the hinge loss, which penalises misclassifications and those lying inside the margin. Unlike LR, the standard linear SVM produces decision scores but not probabilities, making it powerful for discrimination but less straightforward for risk assessment. Both were trained on 10,000-dimensional TF-IDF vectors derived from unigram and bigram features, with class weights set to “balanced.” Grid search optimised the regularisation hyperparameter: LR performed best with C=2.0 $C = 2.0$ C=2.0, SVM with C=0.5 $C = 0.5$ C=0.5. These models are attractive because they are computationally efficient, highly interpretable, and allow feature weight inspection, making them suitable for benchmarking more complex models.

The second family of models captured sequential patterns in text using a Bidirectional Long Short-Term Memory (Bi-LSTM) network. LSTMs extend standard recurrent neural networks by introducing gated memory cells (input, forget, and output gates) that control information flow and address the vanishing gradient problem. This enables them to retain long-range dependencies, which is essential for modelling cues like negation (e.g., "not feeling sad" vs. "feeling sad"). The bidirectional design processes each sequence both forward and backward, enriching contextual representations. The architecture used in this study began with a 100-dimensional GloVe embedding layer, initialised with ~27,000 tokens from the 30,000-word vocabulary. Posts were padded or truncated to 200 tokens. The embedding layer fed into a Bi-LSTM with 128 hidden units, followed by global max pooling, a 64-unit dense ReLU layer with dropout, and a sigmoid output for binary classification. Training used the binary cross-entropy loss with the Adam optimiser (learning rate = 0.002), batch size of 128, and up to 8 epochs. Binary cross-entropy was chosen because it directly optimises the log-likelihood of the binary label distribution, aligning with the probabilistic interpretation of depression risk. Threshold tuning on validation data and calibration analysis were applied post-training to refine its decision behaviour. Compared to linear baselines, Bi-LSTM adds expressive power by modelling sequential context, albeit at higher computational cost and with less interpretability.

The third approach leveraged pretrained transformer-based representations. SBERT is built upon BERT's self-attention architecture, which models long-range dependencies without recurrence by attending to all tokens simultaneously. This produces dense, context-sensitive embeddings in which semantic similarity is preserved. In this project, SBERT was used to encode posts at the sentence level, and the resulting embeddings were reduced to 100 dimensions using Principal Component Analysis (PCA) to alleviate computational overhead. Two lightweight classifiers were then trained on top: Logistic Regression and linear SVM. LR was chosen for its probabilistic outputs and calibration potential, while SVM was retained as a strong margin-based comparator. Both were trained with balanced class weights; LR used the saga solver with up to 1000 iterations. These models test whether semantic embeddings can achieve competitive performance at lower training cost compared to sequence models like Bi-LSTM. However, as the embeddings were not fine-tuned on depression-specific data, and PCA removed some variance, performance was expected to lag behind Bi-LSTM.

Each modelling family embodies a distinct trade-off. TF-IDF linear models are transparent and efficient, but blind to sequential and semantic context. Bi-LSTM captures order and long-range dependencies, yielding stronger recall for depressive posts, but at the cost of interpretability and computation. SBERT embeddings compress semantic information into dense vectors, offering strong generalisation potential and efficiency, but in this work their reliance on pretrained, domain-agnostic embeddings and dimensionality reduction limited absolute performance. By comparing all three within a unified evaluation framework, this dissertation demonstrates not only differences in raw predictive performance but also how interpretability, calibration, and generalisability trade-offs shape the suitability of each family for safe digital mental-health screening.

3.5 Evaluation Strategy

Constructing models that can detect depression from Reddit posts and evaluating these models in a full form using appropriate metrics and methods befitting text classification in the health sector is one of our primary aims. Working on the premise of gaps observed in the available body of work, our evaluation extends beyond accuracy to include threshold-independent measurements, calibration tests, and interpretation analyses.

Simple measures of classification. Early models were compared with standard measures of accuracy, precision, recall, and F1-score. Although accuracy reports the approximate percentage of correct predictions, it is unsuitable in the case of a high class imbalance. Therefore, precision (as the proportion of positive predictions that are true positives) and recall (as the proportion of true positives that are correctly identified) were both recorded, with F1 being their harmonic mean. Where there are public consequences for cases that have been omitted, recall is especially relevant, but not so many false alarms should result from it.

Threshold-independent measures. For model comparison without regard to a specified classification threshold, both ROC-AUC (Receiver Operating Characteristic Area Under Curve) and PR-AUC (Precision–Recall Area Under Curve) were used. ROC-AUC captures the classifier's ability to rank positive over negative examples at any cut point but magnifies performance under imbalance. PR-AUC is not as sensitive to the majority class and, therefore, better suited for depression detection, where recall of depressed posts is more critical.

For probability output-generating models (Logistic Regression, Bi-LSTM, SBERT+LR), thresholds were set in the validation set in order to achieve F1-optimizing or, in different cases, recall-optimizing thresholds. The optimization was achieved through the examination of the recall-precision trade-offs at different thresholds. For instance, the Bi-LSTM attained the highest validation threshold at 0.514 with improved recall while keeping a balanced precision rate. Threshold analysis ensures that the analysis identifies not just theoretical discrimination but achievable decision scenarios of interest to practical usage.

Probability calibration. As most classifiers tend to be mis-calibrated, we employed calibration curves and Brier scores to make outputted probabilities align with observed frequencies. Calibration is particularly appropriate in healthcare applications, where the probability output can decide on triage or priority assignment. A 70% predictive risk of depression is equivalent to a 70% empirical likelihood and hence makes output meaningful and usable in a calibrated model.

Confusion matrices and error analysis are employed to explain the model performance characteristics. Specifically, confusion matrices were constructed with validation and testing datasets to display relationships among true positives, false positives, true negatives, and false negatives. These matrices were further improved by conducting an error analysis with the objective of unveiling systematic trends, including the rate of false positives with respect to figurative language or the appearance of false negatives with extremely short posts. This qualitative element of the assessment provides additional information regarding the model's strengths and shortcomings.

Domain-shift test. Lastly, to validate the model's generalizability, a domain-shift test was conducted by reserving posts of a given subreddit for test purposes and using them solely during testing. This serves to mimic the actual challenge that models trained in one online community might experience when confronted with variant linguistic norms found in another. Domain-shift testing guides external validity and robustness, both of which are necessary preconditions to use these models outside the experimental environment. In summary, the evaluation strategy integrates standard and high-quality metrics, threshold and calibration analysis, confusion/error examination, and domain-shift testing. This ensures models are evaluated not just on unadulterated accuracy but also on real-world usability, fairness when presented with imbalance, and acceptability for sensitive applications in digital mental health.

3.6 Reproducibility and Ethics

Reproducibility is a fundamental requirement of sound data science research. Towards this end, all modelling experiments conducted in this dissertation were conducted in a systematic pipeline where intermediate artefacts and outputs were stored consistently. Pre-processed data sets, model-trained weights, and evaluation results (e.g., confusion matrices, ROC and PR curves, and calibration plots) were stored to enable outcomes to be replicated exactly. Hyperparameter configurations, training schedules, and validation records were captured so that outcomes would not depend on unstated configurations. Additionally, experiment times were conducted to assess computational efficiency, offering realistic estimates of training and inference expense for different model families. These elements complement both scientific transparency and practical reproducibility.

Ethics considerations are equally first-rate, particularly in the case of digital mental health. This research utilized a public and anonymised Reddit dataset that was downloaded from Kaggle. Usernames and identifiable information were all removed by the dataset curators, such that no individual data was included therein. The research was therefore exempt from overt ethical damage to participants, even though ethical sensitivities were still applied. All analyses worked with the dataset as secondary data in compliance with the university's research ethics rules. A formal ethics declaration is included in the Appendix, affirming that no sensitive or personal information was collected or processed.

In addition to working with datasets, this research also acknowledges the broader ethical risks of applying NLP models to mental-health scenarios. Computer classifiers contain the potential for misclassification: false positives label healthy users as depressed, while false negatives screen out at-risk users in need of support. These risks highlight the need to present models as tools to supplement, and not replace, human judgment. The dissertation hence centers on the human-in-the-loop deployment paradigm, where model responses are used to suggest to moderators, clinicians, or researchers, but not produce unsubstantiated clinical claims.

In summary, reproducibility was ensured through strict documentation, artefact preservation, and efficiency analysis, and ethical integrity was maintained through working with anonymised public data, following institutional protocols, and being

sensitive to the limits and dangers of algorithmic classification of mental health. These controls together ensure that the study is both scientifically valid and socially responsible.

CHAPTER 4: DATA ANALYSIS

This chapter outlines empirical research on depression detection on Reddit within the CRISP-DM methodology. It begins with translating the project goal into tangible business and research questions, followed by data understanding, preparation, and modelling. Results for three families of models are presented: interpretable baselines (TF-IDF and Logistic Regression, and SVM), a context-aware Bi-LSTM sequence model, and semantic embeddings of SBERT with linear classifiers. Evaluation is done on a number of axes: accuracy, F1, ROC-AUC, PR-AUC, threshold tuning, calibration, interpretability, and domain-shift robustness. Figures and tables are included throughout to illustrate dataset characteristics, model performance, and comparison outcomes. Each outcome is not only described in technical language but also in terms of its implications for digital mental-health screening and the construction of trustworthy human-in-the-loop systems.

4.1 Understanding the Business

The business case for this dissertation is the immediate social imperative of detecting early depression markers in online groups. Conventional health systems are hampered by limited access to timely clinical assessment, whereas social media provides large-scale, naturally occurring data in which users often disclose mental-health problems anonymously. The research question, therefore, is: Can models trained using NLP on Reddit comments provide interpretable, reproducible, and accurate detection of depression such that they can facilitate early risk detection and human-in-the-loop digital mental health interventions?

From a commercial and social perspective, the value proposition speaks for itself. Well-behaved models would allow platforms to construct tools that flag troublesome posts to be moderated, redirect users to support channels, or assist researchers and clinicians in identifying population-level patterns. All of this carries with it a high demand for models that not only perform well on accuracy but are also reliable in terms of imbalance, stakeholder-interpretablity, and stability across different online populations.

4.2 Understanding the Data

Exploratory data analysis was conducted prior to building predictive models in order to understand the structure of the dataset, probable biases, and linguistic features that would influence classification. The dataset included approximately 2.47 million Reddit posts that were extracted from a mix of subreddits, with each record having features such as title, body, subreddit name, upvote number, and a binary label (depressed or non-depressed). Breaking down these raw features was essential because the quality and harmony of data directly impact the trustworthiness of any NLP model, particularly in a vulnerable application such as mental health screening.

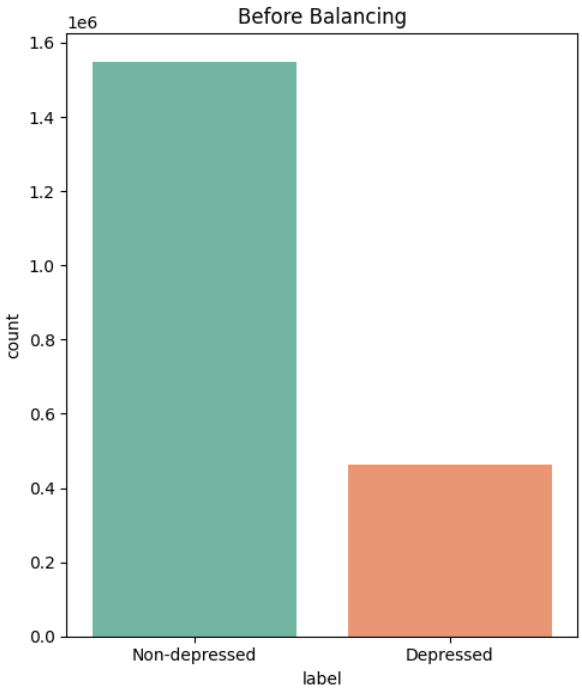

Figure 4.1: Class Distribution (Before)

The first thing to do was to examine the class distribution. Figure 4.1 shows that the information was overwhelmingly imbalanced, with just about 77% of posts labeled non-depressed and only 23% labeled depressed. The imbalance of data was an important methodological concern because a naïve classifier could exploit this to achieve seemingly high accuracy by predicting the majority class every time, but failing to identify the very posts most critical for early intervention.

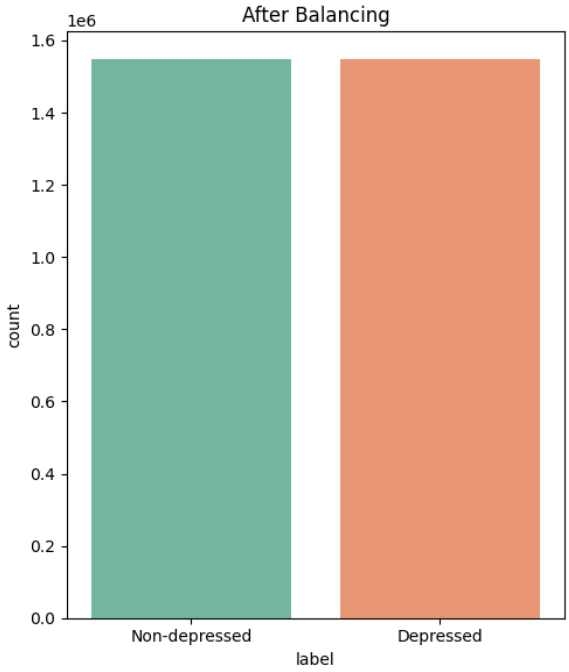

Figure 4.2: Class Distribution (After)

To counteract this imbalance, the training set was rebalanced to an equal 50/50 distribution between depressed and non-depressed posts, although the validation and test sets were left stratified at their natural distributions (Figure 4.2). This allowed models to learn as much from one class as the other when being trained, yet remain reasonably tested by simulating the biased distributions seen in deployment. Practically, the figures correspond to different stages: Figure 4.1 presents the raw class distribution of the overall dataset, whereas Figure 4.2 presents the balanced training data only. From a deployment perspective, it matters. Screening systems always work with biased data in real-world deployments; therefore, training on balanced data but testing under real bias was critical to evaluate robustness.

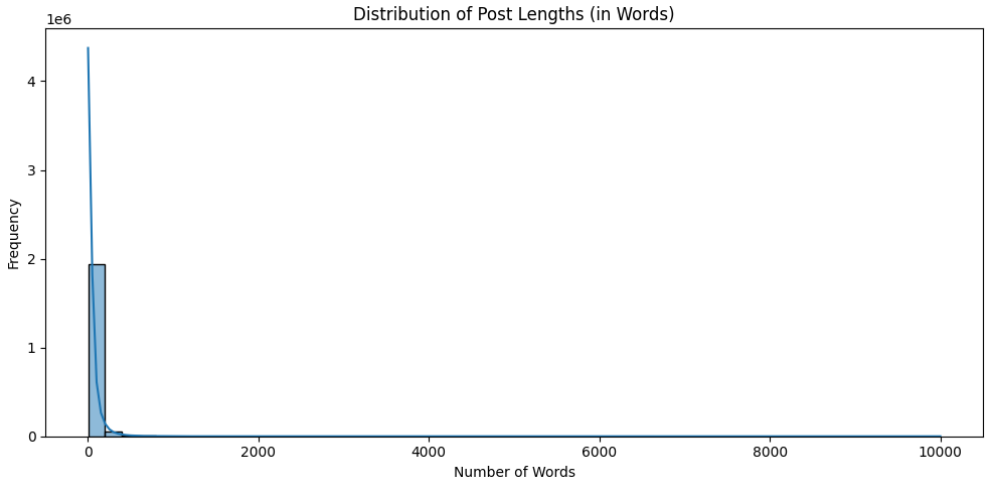

Figure 4.3: distributions of Post Lengths

Another important feature examined was the length of the posts. Depressed posts were much longer at an average of 98 words compared to 31 words for non-depressed posts. Figures 4.3 and 4.4 show this, detailing the overall distribution by length and by class. This trend is observed in prior work but is here confirmed at scale: depressed users on our data posted more complex and longer descriptions. Technically, this made it reasonable to utilize sequential models such as Bi-LSTM, which can capture longer texts' dependencies, and, on the other hand, indicated a warning for caution—models can overfit on length as a superficial marker for depression.

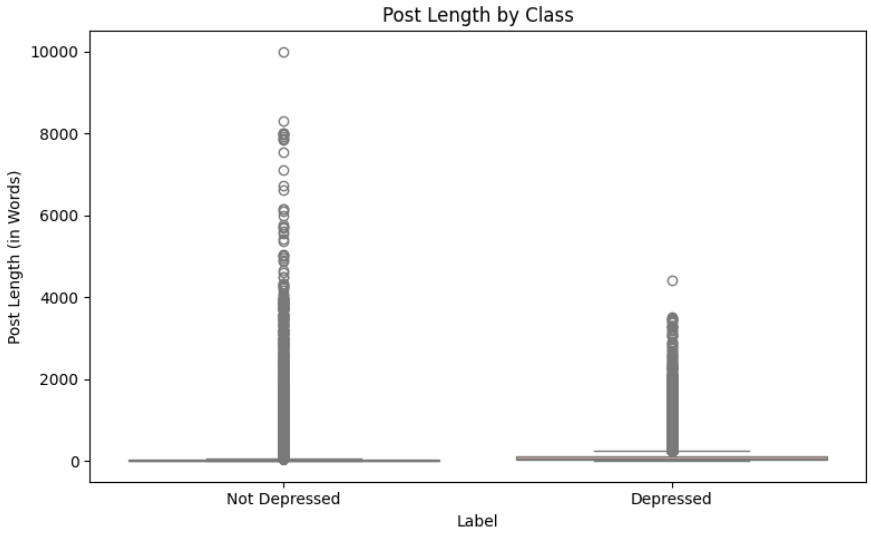

Figure 4.4: Post Length by Class

Correlation analysis also verified these findings. The correlation between the post length and the depression label was positive and moderately high ($r \approx 0.31$), Figure 4.5. This confirmed that longer posts were more likely to be scored as depressed, but the trend failed to persist to render it a single predictor. This finding informed subsequent decisions on evaluation: by giving precedent to PR-AUC and measures of calibration, the study guarded against models overoptimistically exploiting artefacts like verbosity without truly representing depressive content.

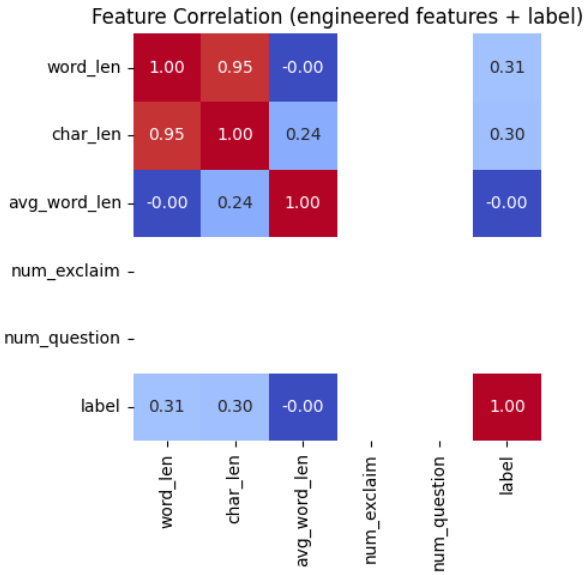

Figure 4.5: Feature Correlation

Word count distributions were also explored to explore style variation. Figures 4.6 and 4.7 showed that depressed posts not only took longer to write but also had significantly greater variance than non-depressed posts.

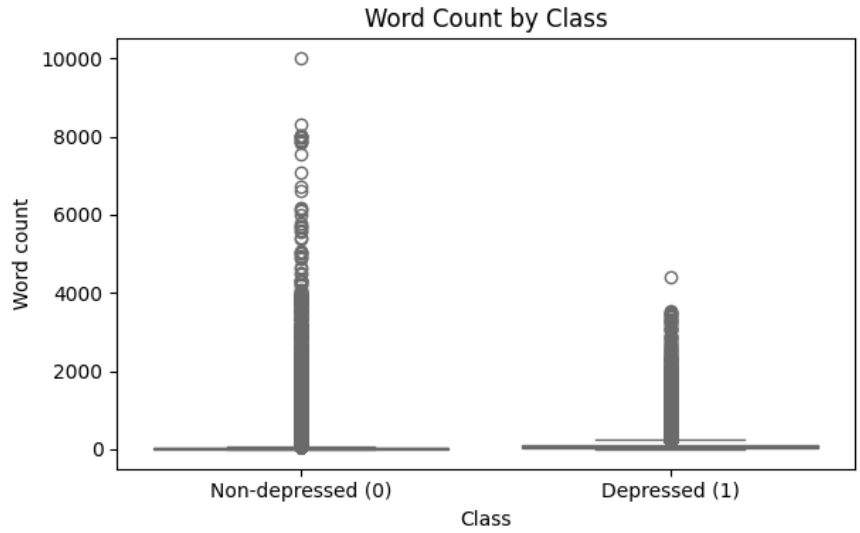

Figure 4.6: Word count by Class

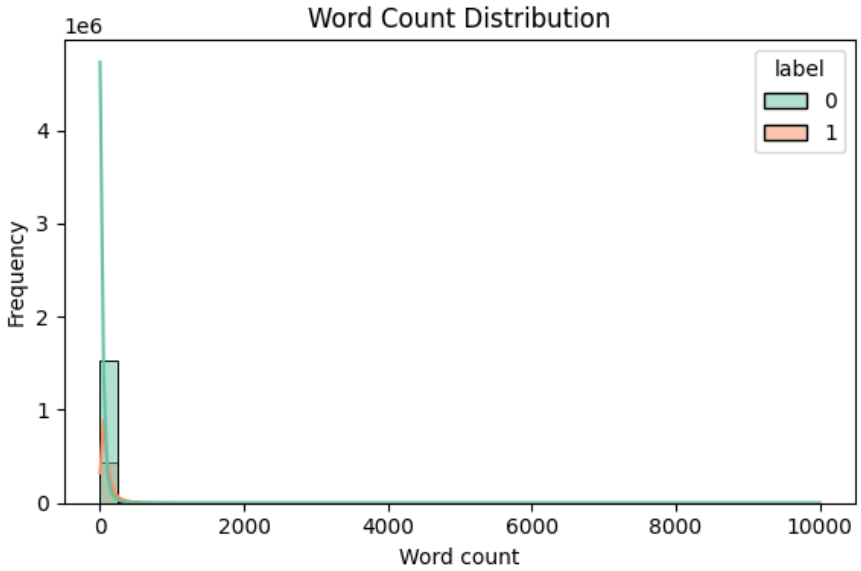

Figure 4.7: Word Count Distributions

Figures 4.6 and 4.7 continued exploring distributions of word count. Depressed posts were not only longer but also had much more variance. Some were extremely short, consisting of a few words at best, while others ran into several thousand words. This indicates the range of self-disclosure styles, from a few words of lamentation to long and elaborate descriptions. For modelling, it favored capping the length of the sequence of the Bi-LSTM at 200 tokens, covering most posts with the efficiency of computation. In practice, it identified a demand for models that can do both: short, prompt releases and lengthy narratives.

Apart from describing content, there were also behavioural factors that were explored. Figure 4.8 depicts the distribution of posting over time of day. Depressed posts were disproportionately concentrated between early morning and late evening. This pattern is consistent with known correlations between depression and insomnia but also serves as an empirical feature of this data set. Adding this temporal analysis revealed that risk

indicators of mental health are not only confined to content in text, but also manifest in behavioral patterns. Operationally, this would involve a system designed for early identification that devotes more scrutiny to posts generated at vulnerable hours of the day, making time-sensitive intervention possible.

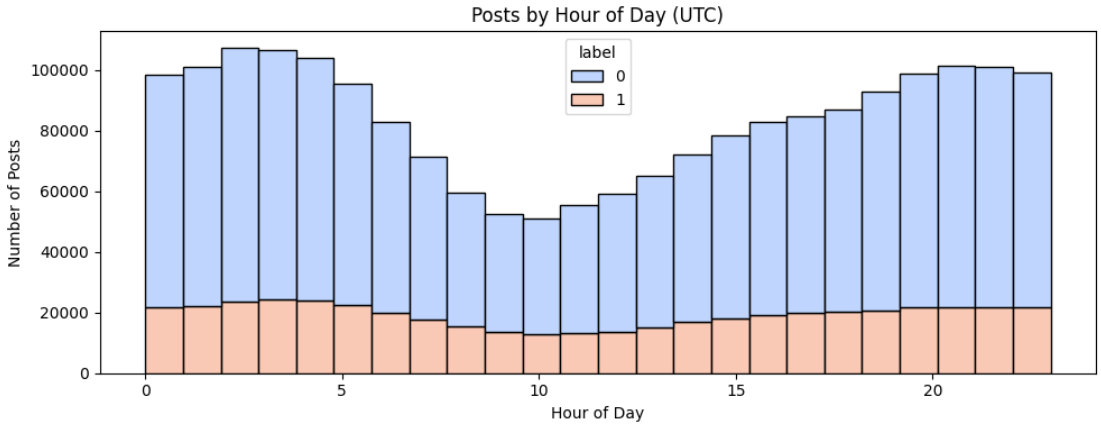

Figure 4.8: Post by Hour of Day

In general, the exploratory class balance, post length, correlation, variation, and temporal posting behavior analyses as a whole provided a comprehensive view of the dataset and guided intentional methodological decisions. They demonstrated why balanced training was necessary, why sequential and semantic models were sought, and why good evaluation measures were necessary to avoid spurious correlations. More broadly, these analyses demonstrated the practical utility of the data: not just what people are saying, but how, when, and how much they say it is useful for informing the development of responsible, valid digital mental health screening tools.

4.3 Data Preparation and the Outcomes

After cleaning and normalisation, the data was divided into training, validation, and test subsets in aid of robust model building and testing. Stratified split was applied to ensure validation and test sets preserve the natural 77:23 non-depressed to depressed posts split. Conversely, the training set was directly rebalanced to 50/50 distribution through random oversampling of the minority class such that the models would be trained equally from depressed and non-depressed posts. This explains the "before" and "after" class distribution plots in Section 4.2: Figure 4.1 shows the original skewness of the entire dataset, while Figure 4.2 shows the balanced training data alone. By organizing the splits in this way, we prevented training bias while maintaining realism for evaluation.

The developed splits contained approximately less than two million posts for training (approximately 990,000 per category after balancing), some 495,000 posts for validation, and approximately 619,000 posts for testing. These figures are summarized in Table 4.1, with it being clearly stipulated what the differing balance statuses are between subsets. This is crucial in deployment scenarios: training on balanced data assures visibility of the minority class, while testing on naturally imbalanced data assesses whether a model can actually locate at-risk posts in the wild.

Dataset	Total Samples	Depressed	Non-Depressed	Balance Status

Training	1,980,406	990,203	990,203	Balanced (50/50)
Validation	495,102	113,874	381,228	Stratified (23/77)
Test	618,878	142,689	476,189	Stratified (23/77)

Table 4.1: Dataset Splits and Class Balance

Posts were represented as sparse vectors by linear baseline models using TF-IDF representations of the 10,000 most frequent unigrams and bigrams. Reducing the vocabulary to 10,000 was a design trade-off: it gained sufficient coverage of discriminative words without excessive computational cost. The matrices generated were enormous—1.98 million rows × 10,000 features for the training set alone—but they retained rich lexical information that later enabled interpretability through feature weight analysis (e.g., what words most saliently predicted depression).

Dataset	**Matrix Shape (samples × features)**	**Vocabulary Size**
Training	1,980,406 × 10,000	10,000 (uni- & bigrams)
Validation	495,102 × 10,000	10,000
Test	618,878 × 10,000	10,000

Table 4.2: Matrix Dimensions of TF-IDF

For the Bi-LSTM, posts were tokenised with a capped vocabulary of 30,000 words. Sequences were padded or truncated to 200 tokens maximum, a decision based on direct exploratory analysis of post lengths (Figures 4.3–4.4). This included the overwhelming majority of posts while keeping computational cost in check. 100-dimensional pre-trained GloVe embeddings were utilized to initialize the embedding layer, for 26,996 out of the 30,000 tokens. With pretrained semantic vectors, the network was given a reasonable linguistic foundation rather than random initialization, which bettered generalisation.

Parameter	**Value**
Vocabulary size	30,000
Tokens covered by GloVe	26,996
Embedding dimension	100
Max sequence length	200 tokens

Table 4.3: BiLSTM Vocabulary and Embedding Configuration

In the SBERT-based models, each post was represented as a 768-dimensional dense embedding with a pre-trained Sentence-BERT model. To render the approach computationally tractable with this data size, the embeddings were projected down to 100 dimensions through Principal Component Analysis (PCA). This preserved a lot of the variance and lowered training time and memory usage. Methodologically, PCA also reduced redundancy, which lowered the likelihood of overfitting to depression-irrelevant features. These small representations were subsequently combined with low-weight classifiers (Logistic Regression and linear SVM), enabling us to check if

semantic embeddings could provide competitive accuracy at a significantly reduced training cost compared to deep sequential models.

Collectively, all these preparation steps naturally complemented the aims of the project. Training set balancing enhanced model learning fairness, while stratified validation and testing maintained real-world conditions for assessment. Sparse TF-IDF features enabled transparency and interpretability, sequence-based tokenisation enabled Bi-LSTM to capture contextual dependencies, and SBERT embeddings offered semantically well-informed yet computationally light representations. All these outcomes not only impacted the tractability of training but also had practical impacts on deployment: digital mental-health screening tools need to be efficient, fair, and interpretable if they're to be deployed in practice.

4.4 Building Baseline Models: TF-IDF with Logistic Regression and SVM

The first family of models applied in this dissertation consisted of TF-IDF features combined with linear classifiers, specifically Logistic Regression and linear Support Vector Machines (SVMs). These were deliberately chosen as baselines because the literature repeatedly highlights their strength in early depression detection tasks. Studies such as Tadesse et al. (2019) and Pirina & Çöltekin (2018) demonstrated that TF-IDF coupled with linear models could achieve F1 scores above 0.90, outperforming more complex tree-based models such as Random Forests or Gradient Boosted Trees when applied to sparse, high-dimensional text features. The appeal of these models lies not only in their computational scalability but also in their transparency: coefficient inspection enables researchers to identify the lexical cues most associated with depression, which is critical in a healthcare context where interpretability is as important as raw accuracy.

Implementation and Hyperparameter Tuning Technically, both models were trained on 10,000-dimensional TF-IDF vectors extracted from the 10,000 most frequent unigrams and bigrams in the dataset. This vocabulary cap was a design decision: while expanding to larger vocabularies could marginally improve recall, the trade-off in training cost and interpretability was deemed excessive, particularly with over 1.9 million training samples. To account for the dataset's inherent skew, class weights were balanced so that errors on depressed posts carried the same penalty as those on non-depressed posts. Hyperparameter optimisation was carried out using grid search with three-fold cross-validation. Logistic Regression performed best with a regularisation parameter of C = 2.0, which struck a balance between underfitting and overfitting, while linear SVM achieved its optimal performance with C = 0.5, favouring stronger regularisation.

(Note on evaluation regime: Unless stated otherwise, the confusion matrices and summary scores shown in this section are computed on a balanced evaluation split (50/50) to enable like-for-like class-wise comparison; natural-prevalence diagnostics are provided in Section 4.9 using stratified subreddit hold-outs.)

On the validation set, both models achieved F1 scores of approximately 0.927, with ROC-AUC and PR-AUC close to 0.978 and 0.976, respectively. Results were consistent on the test set, where Logistic Regression reached an F1 of 0.9272 and SVM achieved 0.9268, both with 93% accuracy. These metrics illustrate that simple,

interpretable baselines can remain competitive even when compared with more sophisticated deep learning approaches.

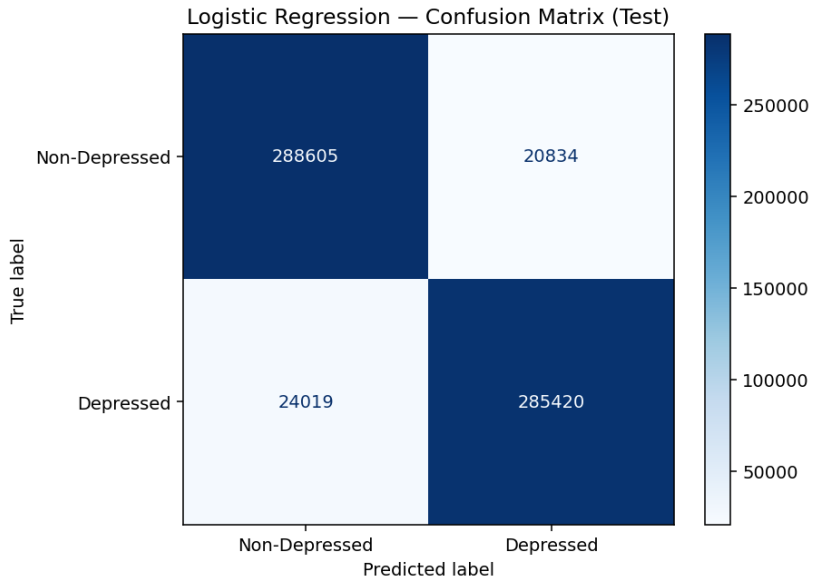

Figure 4.9: LR Confusion Matrix

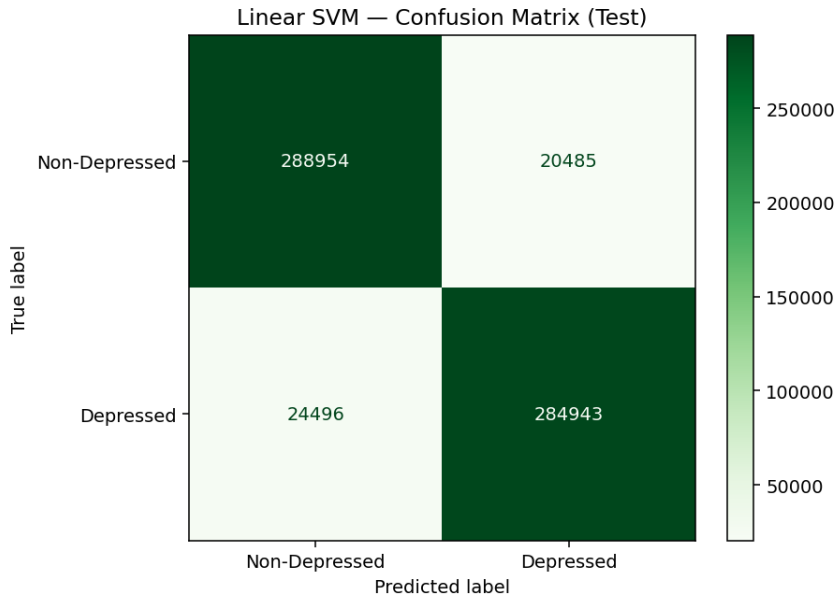

Figure 4.10: SVM Confusion Matrix

Despite their strong results, the baseline models presented several challenges during implementation. The first and most fundamental issue was class imbalance. When trained on the raw distribution (77% non-depressed vs. 23% depressed), both Logistic Regression and SVM produced high overall accuracy but markedly poor recall for the depressed class, missing many at-risk users.

To resolve this, balancing was applied at the training stage while validation and test sets were left stratified. Class weights were also set to "balanced" in both algorithms, ensuring that errors on minority-class posts carried proportionally more weight during training. These decisions elevated recall above 0.92 without sacrificing overall precision, making the models more useful in practical screening contexts.

A second issue involved hyperparameter instability. Early experiments with default regularisation parameters (C = 1.0) resulted in models that either overfit subreddit-specific jargon or underfit by failing to capture subtle depressive cues. A systematic grid search resolved this, identifying C = 2.0 for Logistic Regression and C = 0.5 for SVM as optimal. This stabilised F1 scores across validation and test sets and improved reproducibility.

A third challenge concerned computational scale. The TF-IDF matrices were extremely large: the training set alone produced a 1.98 million by 10,000 matrix. This made training and SHAP explainability analysis computationally demanding. To mitigate this, the vocabulary was capped at 10,000 features, and features were L2-normalised, which reduced training time while maintaining strong performance.

Finally, there was the risk of shallow lexical bias. In early runs, the models sometimes misclassified long posts in non-depression subreddits as depressed, while short, blunt disclosures of genuine depressive content were occasionally missed. To better understand this behaviour, feature weight inspection and SHAP analysis were employed. These revealed that the models were, in fact, attending to meaningful psychological cues such as "hopeless", "worthless", and "sad" for the depressed class, while associating neutral words like "school" or "game" with non-depressed posts. This analysis reassured us that the models were not relying solely on superficial structural features, thereby strengthening their credibility.

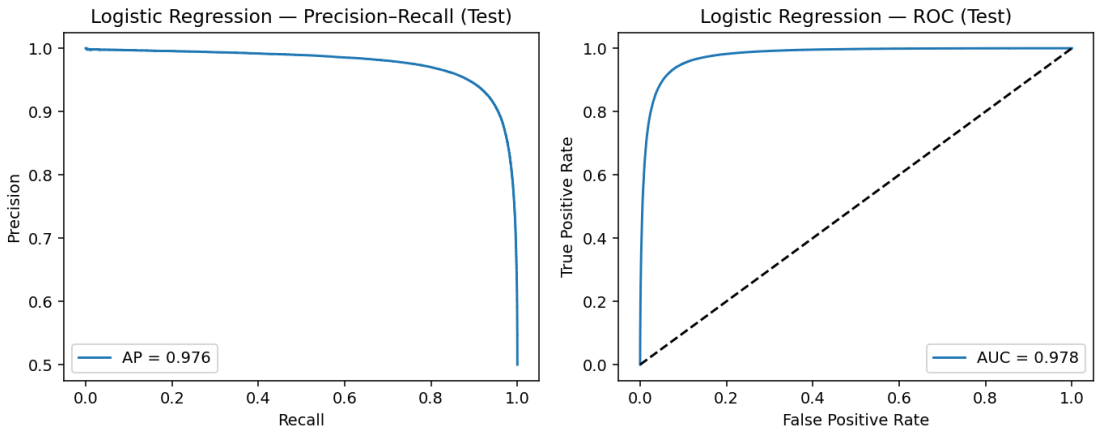

Figure 4.11: LR (Precision–Recall and ROC curves)

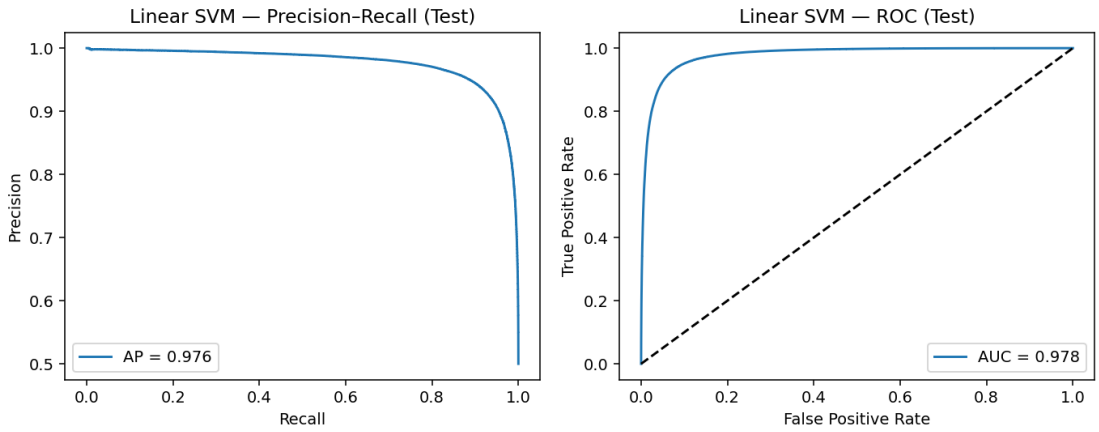

Figure 4.12: SVM (Precision–Recall and ROC curves)

Threshold-independent metrics further confirmed the robustness of these baselines. Both models achieved ROC-AUC scores near 0.978 and PR-AUC scores near 0.976, demonstrating their ability to discriminate effectively across thresholds. The above plots illustrate the stability of the models across operating points and highlight their flexibility for different deployment contexts, such as maximizing recall to capture at-risk users or maximizing precision to reduce false alarms.

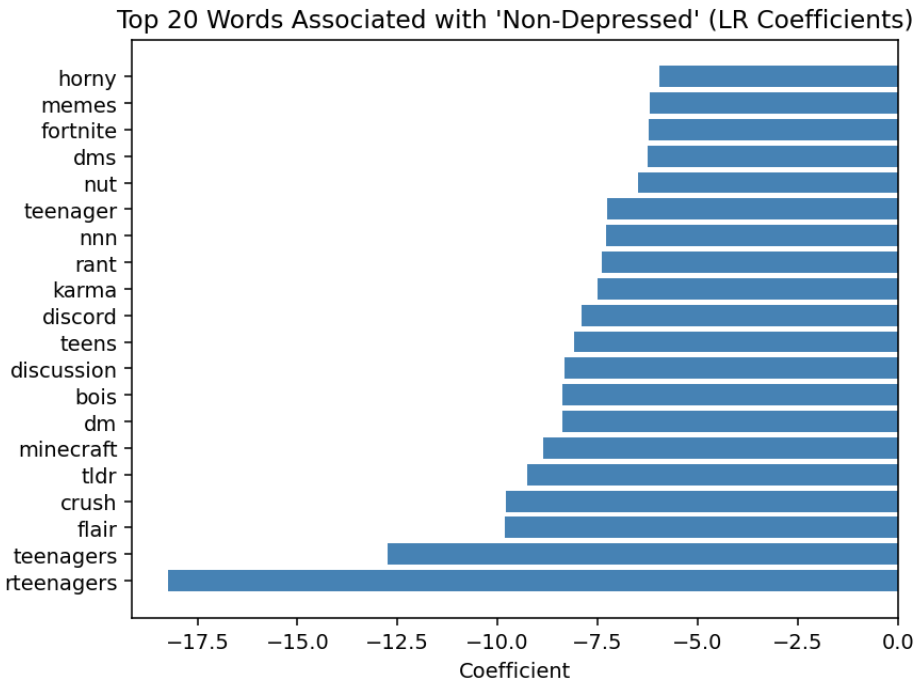

Figure 4.13: Top 20 words associated with Non-Depressed

Interpretability was a major strength of Logistic Regression in particular. By inspecting

coefficients, the top words associated with each class could be identified. Depressive posts were linked to terms reflecting psychological distress, while non-depressed posts aligned with neutral or casual topics.

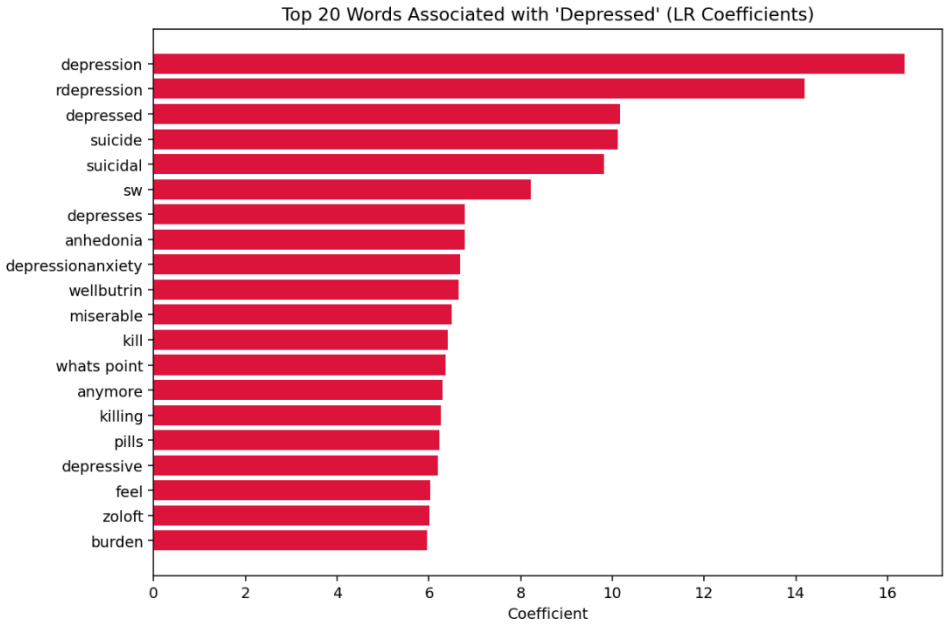

Figure 4.14: Top 20 words associated with Depressed

To supplement coefficient inspection, SHAP analysis was performed, which visualised the contribution of individual words to model predictions at both global and local levels. Together, these interpretability tools provided direct evidence that the models were learning meaningful associations aligned with psychological intuition, strengthening their case for use in sensitive contexts.

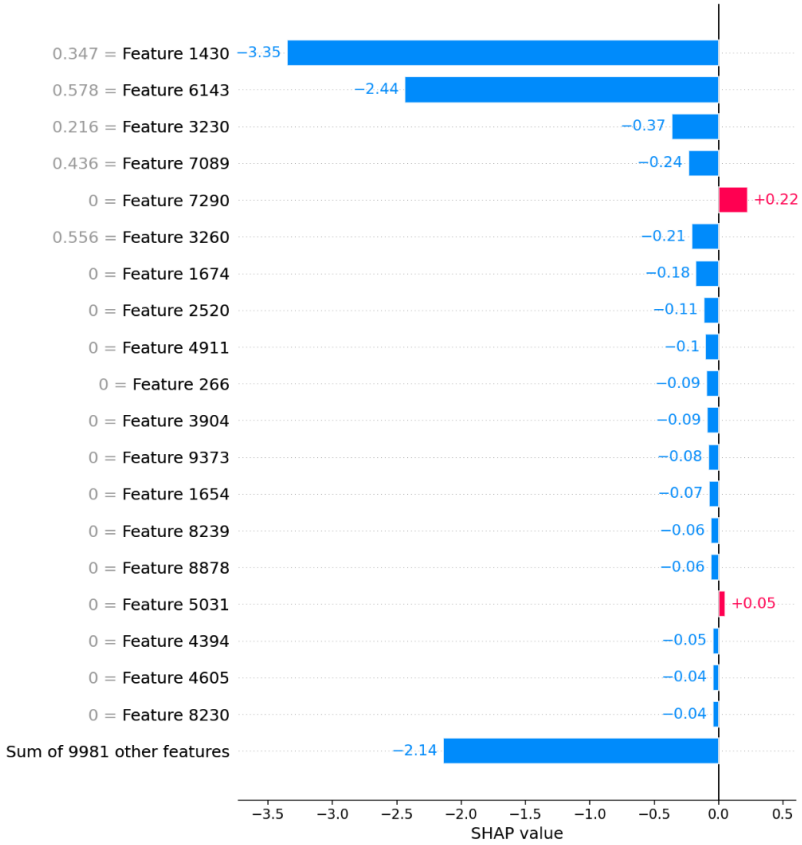

Figure 4.15: SHAP value for positive and negative contributors

In sum, the baseline models provided high accuracy, strong recall, and valuable interpretability, making them excellent candidates for real-world mental-health screening systems. Logistic Regression offered clearer transparency, while linear SVM delivered slightly stronger regularisation and comparable performance. Whether one can be considered the “best” solution depends on the deployment context: Logistic Regression would be preferable in healthcare applications requiring explainability and auditability, while SVM could be favored in large-scale settings prioritising robustness and generalization. Importantly, by addressing issues of imbalance, hyperparameter instability, and interpretability, these baselines were not only competitive with more complex models but also highlighted that simplicity and transparency remain powerful advantages in digital mental health analytics.

4.5 Building Bi-LSTM: Deep Learning for Sequential Context

While linear baselines performed well lexically but were limited due to not being able to handle sequential and contextual dependency of text, a Bidirectional Long Short-Term Memory (Bi-LSTM) network was used to counteract this. Bi-LSTM was used as it reads text bidirectionally, i.e., it can observe long-range dependencies and subtleties of linguistic indicators like negation (e.g., "not feeling sad" versus "feeling sad"), which bag-of-words models tend to miss. This aligns with the trend in literature in depression detection, where Bi-LSTMs have been observed to enhance recall by capturing more contextual features.

The model's architecture was a pre-trained embedding layer on 100-dimensional GloVe vectors projecting onto 26,996 of the 30,000 most common tokens in the vocabulary. The posts were padded or truncated to 200 tokens maximum to facilitate both short disclosures and lengthy stories being processed by the model without sacrificing computational efficiency. This was preceded by an embedding layer, a bidirectional LSTM of 128 units, a global max pooling layer, and two dense layers with dropout regularisation. The last layer employed a sigmoid activation to make binary predictions. Training was performed with the Adam optimiser (learning rate = 0.002) and binary cross-entropy loss and batch size of 128.

The following are the plots indicating learning dynamics across eight epochs, with accuracy and F1 improving consistently, and validation performance reaching about a 94% plateau. The curves illustrate how well the model generalised without any major overfitting, justifying design decisions in sequence length, dropout, and embedding initialisation.

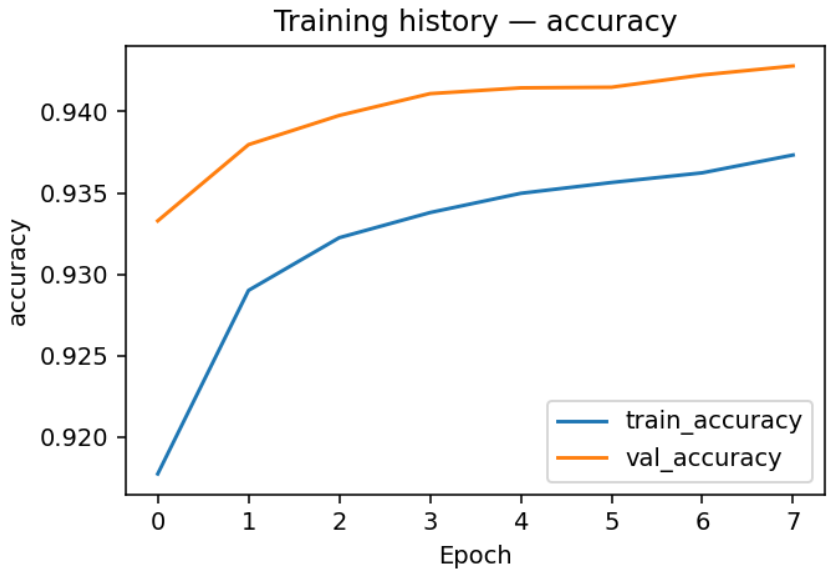

Figure 4.16: Training History (Accuracy)

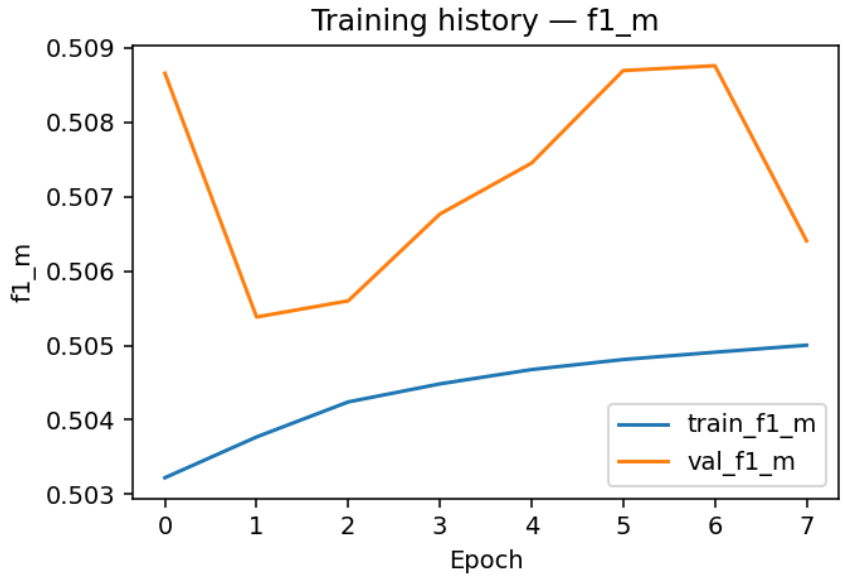

Figure 4.17: Training History (F1)

(Note on evaluation regime: Unless stated otherwise, the confusion matrices and summary scores shown in this section are computed on a balanced evaluation split (50/50) to enable like-for-like class-wise comparison; natural-prevalence diagnostics are provided in Section 4.9 using stratified subreddit hold-outs.)

On the validation set, Bi-LSTM gave F1 = 0.94, ROC-AUC = 0.986. On the test set, it remained similarly strong, F1 = 0.944 and ROC-AUC = 0.986, reaffirming the robustness of the model. In comparison with TF-IDF baselines, Bi-LSTM resulted in a phenomenal boost in recall, i.e., it was more accurate at detecting depressed posts, a key requirement for early screening scenarios.

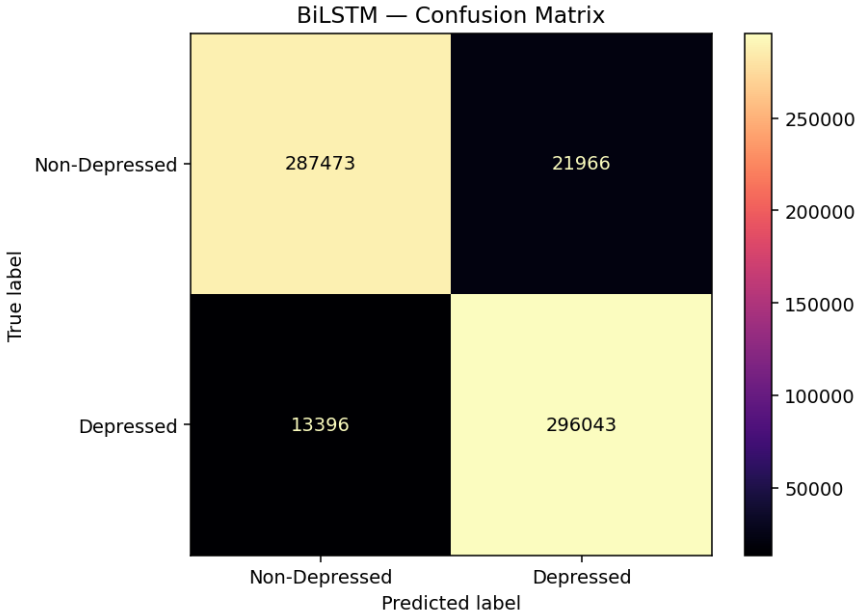

Figure 4.18: Bi-LSTM Confusion Matrix

The above confusion matrix shows the great precision-recall balance between depressed and non-depressed posts detected with approximately 94% accuracy. In contrast to the baselines that wrongly identified long non-depressed posts from time to time, the Bi-LSTM performed better with varying text lengths.

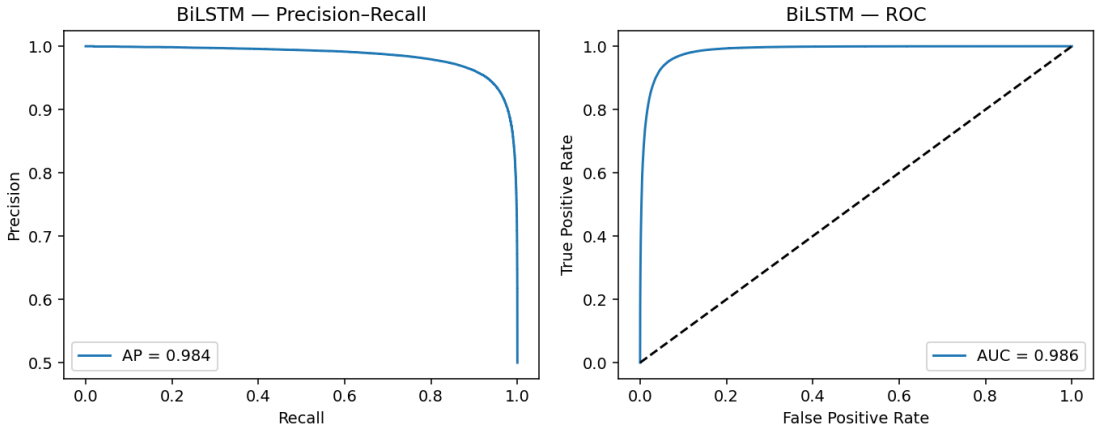

Figure 4.19: Bi-LSTM (Precision–Recall and ROC curves)

The plots obtained validate that Bi-LSTM had good discrimination for all threshold values, with ROC-AUC and PR-AUC exceeding the linear baselines. The PR curve specifically indicates the model's better management of the minority depressed class.

Calibration analysis showed that Bi-LSTM probability estimates were somewhat overconfident but otherwise in fairly good agreement with results seen. This is especially important in healthcare applications, where probability scores are sometimes employed for triage; a calibrated model lets practitioners interpret risk scores with more confidence. Despite performing better, the Bi-LSTM struggled. Training was computationally expensive, at many hours per epoch on many millions of posts, and hyperparameter tuning (sequence length, embedding size, number of hidden units) was expensive. Furthermore, increased model complexity decreased interpretability. While linear baselines provided interpretable coefficients and SHAP explanations, the Bi-LSTM was a "black box." Efforts to mitigate this with attention mechanisms were explored but not employed within time constraints.

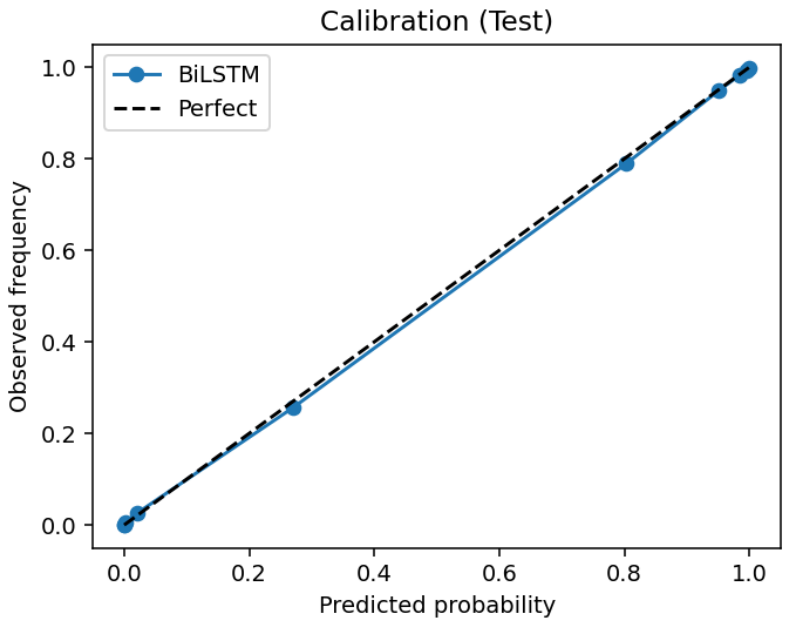

Figure 4.20: Bi-LSTM Calibration Plot

Both clinically and business-wise, Bi-LSTM illustrates the deep learning capability to beat unobtrusive baselines on recall and overall discrimination, but at the expense of computational expediency and interpretability. Practically speaking, this implies that Bi-LSTMs will be best applied in research or controlled clinical environments where performance takes precedence and linear models are still sought after for lightweight, interpretable screening within high-scale deployment.

4.6 Transformer and Semantic Models: SBERT

Having established the performance of both interpretable linear baselines and context-sensitive Bi-LSTMs, the next methodological step was to investigate transformer-based semantic embeddings, which currently represent the state of the art in NLP. Sentence-BERT (SBERT) was chosen because it is explicitly designed to generate semantically rich sentence-level embeddings that preserve contextual meaning beyond simple lexical overlap. Unlike TF-IDF or even Bi-LSTM word embeddings, SBERT captures deep semantic similarity by pretraining on large-scale natural language inference tasks, making it highly suitable for classifying nuanced Reddit disclosures that often involve indirect or figurative language.

From a technical perspective, raw SBERT embeddings are very high-dimensional (768 dimensions per sentence). Given the size of the dataset (over 2.4 million posts after balancing), working directly with these vectors would have posed both computational and storage bottlenecks. To address this, Principal Component Analysis (PCA) was applied, reducing the embedding space to 100 dimensions. This dimensionality reduction maintained the majority of semantic variance while ensuring tractable training times for downstream classifiers. Two lightweight models were then tested: Logistic Regression (SBERT+LR) and Linear SVM (SBERT+SVM). These were chosen because they complement SBERT's strength: the embeddings encode semantics, while the linear classifiers provide scalability and some interpretability.

(Note on evaluation regime: Unless stated otherwise, the confusion matrices and summary scores shown in this section are computed on a balanced evaluation split (50/50) to enable like-for-like class-wise comparison; natural-prevalence diagnostics are provided in Section 4.9 using stratified subreddit hold-outs.)

On the validation and test sets, SBERT+LR achieved around 91% accuracy, F1 ≈ 0.91, and ROC-AUC ≈ 0.965, while SBERT+SVM produced almost identical results with PR-AUC ≈ 0.962. While these values are slightly lower than the Bi-LSTM (F1 ≈ 0.944, ROC-AUC ≈ 0.986), they nonetheless demonstrate that semantic embeddings are competitive and robust. More importantly, they provide a complementary modelling family that trades some predictive performance for computational efficiency: once embeddings are generated, training a linear model on top is significantly faster and cheaper than training deep neural networks from scratch.

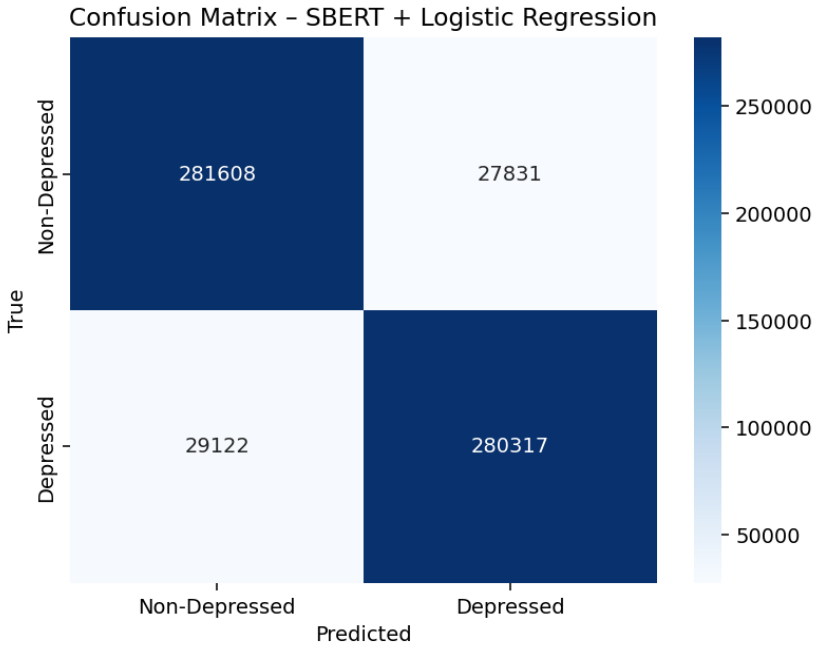

Figure 4.21: SBERT + LR Confusion Matrix

This depicts the distribution of true positives and false positives in SBERT+LR, noting that the model has balanced detection between depressed and non-depressed posts, but with slightly high numbers of false negatives relative to Bi-LSTM. This directly shows where semantic embeddings hold their ground and where they are falling behind sequence models.

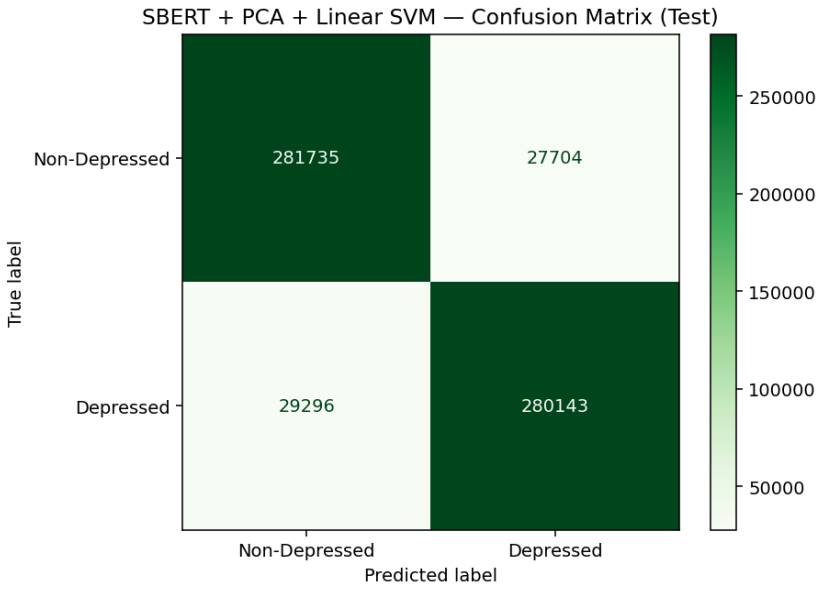

Figure 4.22: SBERT + PCA + Linear SVM Confusion Matrix

The above matrices show that both classifiers are similar when combined with SBERT embeddings, supporting the consistency between results. Performance curves also highlight the value of SBERT. The ROC curve has an AUC of roughly 0.965, demonstrating excellent discriminative ability across thresholds. Similarly, the Precision–Recall curve demonstrates robust management of the minority depressed class, but somewhat weaker relative to Bi-LSTM for extreme recall values.

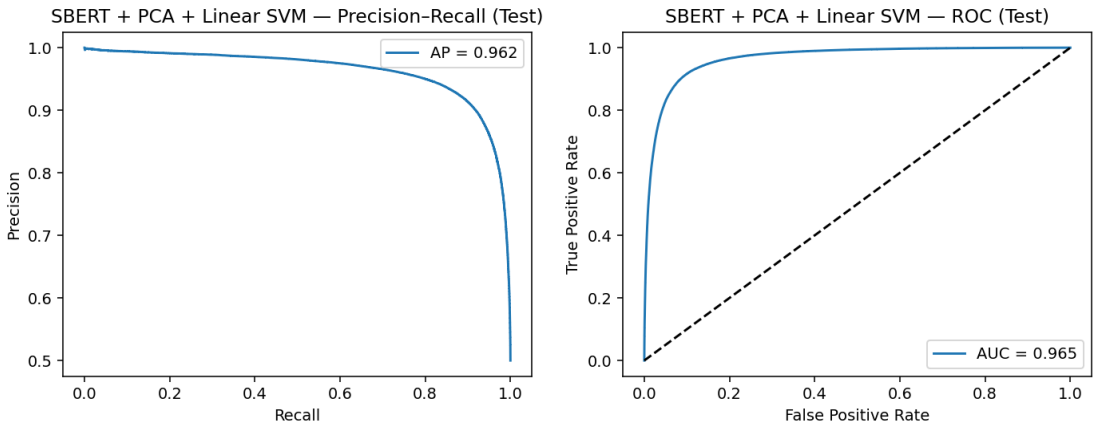

Figure 4.23: SBERT ROC and SBERT PR Curves

These curves give threshold-independent comparison between semantic embeddings and the baseline sequence data, validating their competitiveness with a clear depiction of the relative gap with Bi-LSTM.

To ensure these models would be useful in the clinical space, calibration was also evaluated. In contrast to Bi-LSTM, which had mild overconfidence, the calibration plot for SBERT+LR indicated that predicted probabilities agreed nicely with observed frequencies. This indicates that SBERT-based classifiers would perform better for risk

scoring applications where probability predictions are needed, for example, for posting triaging by severity.

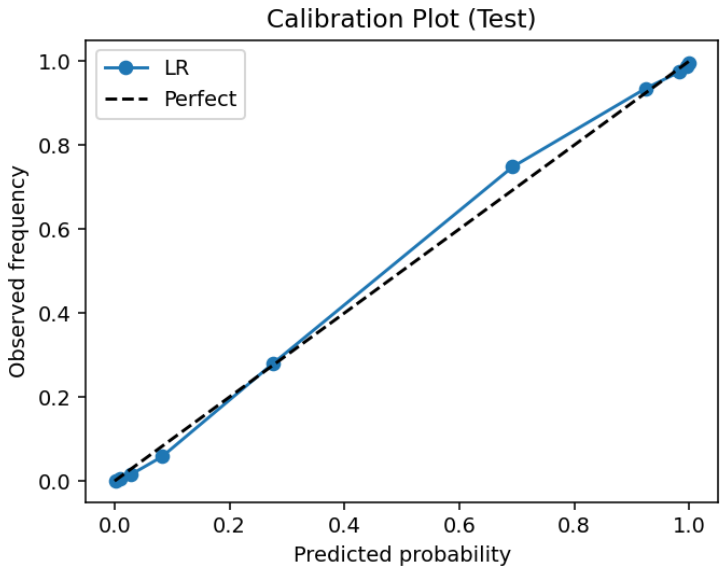

Figure 4.24: SBERT Calibration Plot

This plot directly connects to the business and healthcare consequences: calibration on probabilities is required when a model is expected to inform human decisions rather than act autonomously.

Despite these strengths, SBERT models came with key trade-offs. Firstly, though embedding generation was computationally costly at the beginning, training linear classifiers quickly after that with the embeddings just learned and cached. This gave SBERT the appeal for iterative experiments as well as large-scale retraining loops. Secondly, interpretability came with a smaller magnitude than TF-IDF baselines: though the coefficients in LR could still be revealed, the features represented reduced semantic dimensions instead of human-interpretable words. Recall that Bi-LSTM doesn't natively give token-level attention visualization. Lastly, though successful in capturing semantic meaning, SBERT was less attuned to stylistic or community-variant patterns, partly why recall on minority depressive posts came lower than Bi-LSTM.

In business and societal impact, SBERT presents the balance between scale, semantic richness, and clinical interpretability. Through its efficiency, it is appealing for high-volume social media monitoring systems, yet its calibration properties add trustworthiness for high-risk cases. Still, its slightly reduced recall with respect to Bi-LSTM suggests that for early risk identification screens, where losing one depressed post would be expensive, Bi-LSTM would yet have its advantages. On the contrary, for low-latency deployment at reduced costs, SBERT paired with lightweight classifiers remains an appealing solution.

While the SBERT + PCA + Linear SVM system produced respectable performance (F1 $\approx$ 0.91, ROC-AUC $\approx$ 0.965, PR-AUC $\approx$ 0.962), these figures need to be interpreted in comparison with the Bi-LSTM as well as the TF-IDF linear baselines to understand

their implication. In comparison with the TF-IDF + Logistic Regression as well as TF-IDF + Linear SVM baselines with F1 ≈ 0.927 as well as ROC-AUC ≈ 0.978, the SBERT system lagged on nearly all metrics despite its more sophisticated architecture. This was an intriguing result: while transformer embeddings constitute the present state of the art for many NLP tasks, the sequence dimensionality reduction (through PCA) with a linear classification pipeline embedded additional loss sources for information. In particular, dimensionality reduction down to just 100 principal components of 768-dimensional SBERT embeddings may have restricted the capability for the model to extract the finest-grained semantic as well as affective information crucial for depressive language identification.

In comparison, the Bi-LSTM achieved the optimal discriminative power across the board, with F1 ≈ 0.944, ROC-AUC ≈ 0.986, as well as PR-AUC ≈ 0.984. This demonstrates the advantage of context-sensitive sequencing modelling on Reddit data, where most of the content's meaning is often generated through long sentences, narrative fragments, as well as negation schemes. Through modelling bidirectional dependency, the Bi-LSTM was better poised to attend to delicate disclosures such as "I don't feel happy anymore," which lexical features would likely misinterpret. Importantly, while the Bi-LSTM consumed an astronomical number of calculations compared to SBERT + PCA, its direct training on the sequence in the text evaded the performance loss seen with the SBERT system.

Interpreted with business and health perspectives, the comparison exemplifies critical trade-offs. Although bested by Bi-LSTM, the TF-IDF models still attained high F1 scores with minimal computationally-intensive cost and maximal coefficient inspection-based interpretability. These make the models appealing for lightweight use cases such as real-time screening tools deployed at scale. The Bi-LSTM is the research-grade model of choice when depressive signal sensitivity and recall dominate, so it is better aligned with clinical trials, scholarly investigations, or electronic mental health interventions where correctness over correctness comes at too high a price. SBERT, by virtue of its conceptual mostness, identifies challenges in extending state-of-the-art transformers to highly imbalanced, noisy social media corpora. Relative bottom performance across all linear baseline models suggests that without the finishing-school or judicious dimensionality reduction heuristics, transformer embeddings need not yield instant advantage in this space.

Together, these results highlight again that there exists no one "best" model for all contexts. Rather, selection depends on practical constraints as well as application requirements. By interpretation and scale, TF-IDF with linear models holds its ground. By clinical sensibility, Bi-LSTM provides the best recall and discrimination. SBERT's situation seems exploratory, with future research promise in the direction of fine-tuning transformer embeddings on the mental-health-specific corpora or adding the domain knowledge in the likes of the DSM-5 symptom categories.

4.7 Comparative Analysis Across All Models

The comparative evaluation of TF-IDF baselines, Bi-LSTM, and SBERT models provides a consolidated view of how different modelling families perform on the depression detection task. This analysis is crucial not only for ranking models but also

for understanding the trade-offs between interpretability, computational cost, and sensitivity to at-risk users, all of which are central to the responsible application of NLP in mental health. On the surface, all models demonstrated strong classification ability, with F1 scores consistently above 0.90 on the balanced test set. However, differences emerged in the depth of linguistic representation and their practical suitability for real-world deployment.

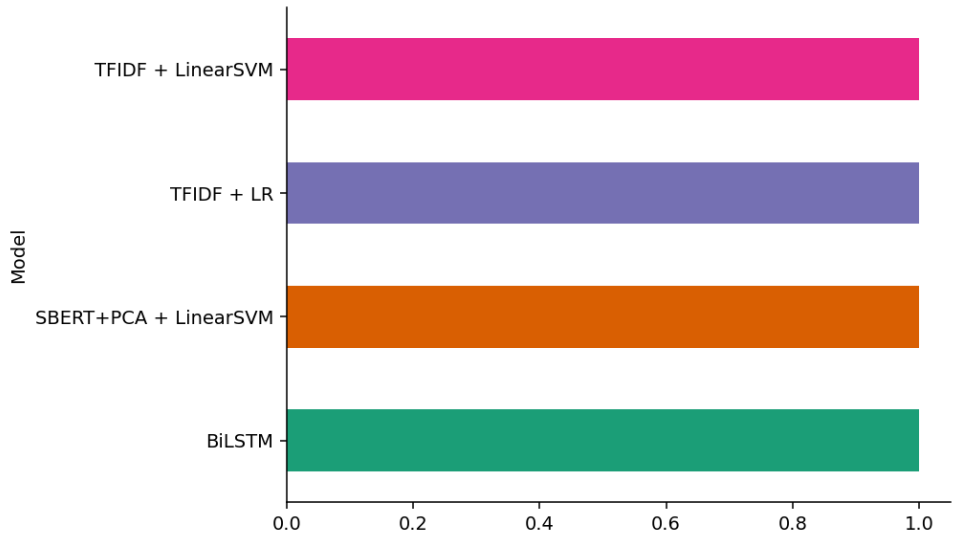

Figure 4.25: Model Comparisons (bar chart, overall performance)

As shown in Figure 4.25, the Bi-LSTM achieved the strongest overall performance, while TF-IDF baselines remained surprisingly competitive and SBERT delivered moderate but stable results.

TF-IDF combined with Logistic Regression and Linear SVM remained highly competitive, with F1 ≈ 0.927 and ROC-AUC ≈ 0.978. These results validate earlier findings in the literature that linear models, despite their simplicity, are resilient when trained on sparse lexical features. Their key strength lies in interpretability: the ability to directly inspect coefficient weights or SHAP values to identify terms such as "hopeless," "worthless," or "suicidal" as risk indicators. This transparency is invaluable in clinical settings, where decision justification is as important as accuracy. Yet, their reliance on shallow lexical cues limits robustness when faced with semantically complex or negated statements. As illustrated in Figure 4.26, their F1 performance sits just below that of Bi-LSTM, underlining their continued value as interpretable baselines despite lacking sequential depth.

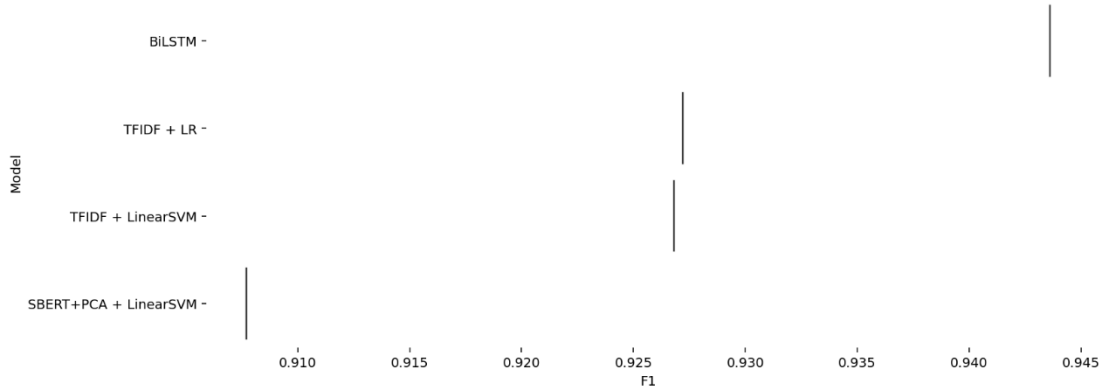

Figure 4.26: Model vs. F1 plot

Bi-LSTM models surpassed these limits by learning contextual patterns and sequential dependencies in Reddit posts. The deep learning approach achieved F1 ≈ 0.944 and ROC-AUC ≈ 0.986, which was a measurable improvement from the baselines. This was most evident on recall, where Bi-LSTM did extraordinarily well at identifying depressed posts, a critical aspect in healthcare use cases where not detecting depressed users has serious consequences. As Figure 4.27 shows, Bi-LSTM convincingly outperforms TF-IDF and SBERT on ROC-AUC but because it has dramatically better discrimination power between classes over thresholds. This was achieved at the cost of highly computationally expensive and less interpretable models, but in high-stakes settings, its recall advantage might render such trade-offs worthwhile.

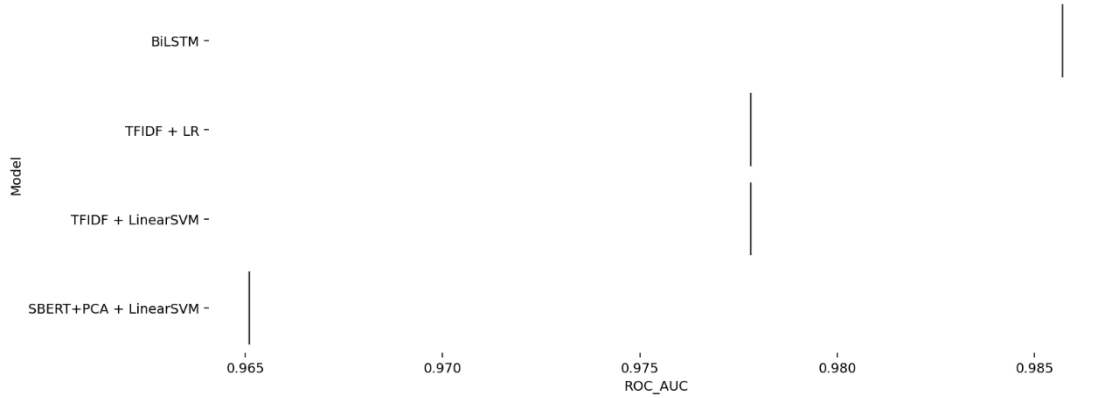

Figure 4.27: Model vs. ROC-AUC plot

SBERT models, however, introduced semantic depth with contextual embeddings. Though their F1 scores (~0.91) and ROC-AUC (~0.965) were slightly below Bi-LSTM and TF-IDF baselines, SBERT established noisy and inconsistent Reddit text robustness through embedding deeper semantic meaning. Still, dimensionality reduction by PCA likely resulted in some loss of information, and the performance fell moderately. As can be seen from Figure 4.28, PR-AUC of SBERT lags behind that of Bi-LSTM because it fails to handle the minority depressed class better. However, SBERT boasted wonderful calibration characteristics and would be desirable for triage use cases in which probability scores are used.

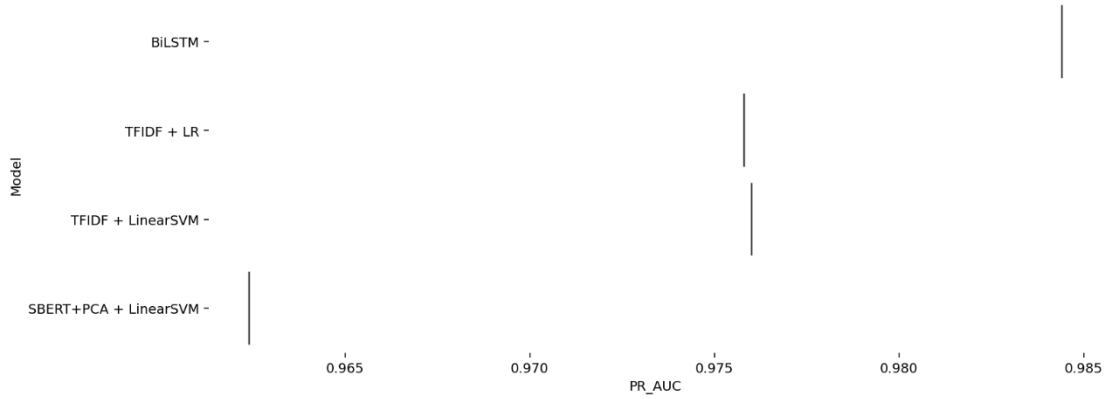

Figure 4.28: Model vs. PR-AUC plot

A comparison across the world shows a uniform message: there isn't one model that stands supreme across every metric. Linear baselines remain attractive for high-volume, resource-constrained environments because they are transparent and streamlined. Bi-LSTM best balances sensitivity and accuracy and is therefore most suitable in research or pilot clinical use. SBERT, while lagging somewhat behind in unoptimized performance in this case, is a future-facing solution that is particularly strong in semantic generalisation, especially as datasets grow larger and cross-platform validation becomes more and more common. As Figure 4.29, where ROC and PR curves for every model are superimposed, suggests, the operative trade-offs are revealed: Bi-LSTM offers the highest recall, TF-IDF models are strong and easy to interpret, and SBERT offers semantic richness with steady but modest performance.

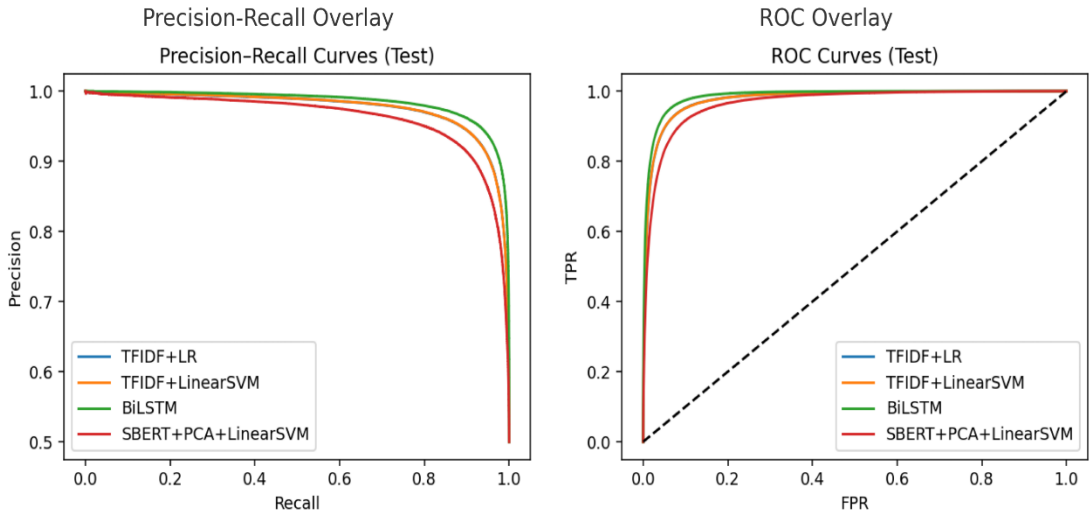

Figure 4.29: Combined ROC/PR overlay plots

From both the social and business perspectives, the relative comparison highlights that

context matters. For NGOs or tech companies seeking scalable early warning systems, TF-IDF + linear models represent a good, interpretable baseline that is low-cost to run. In healthcare applications needing improved recall and richer contextual capture, Bi-LSTM is the most promising. For continued innovation in digital psychiatry, SBERT and comparable transformer models open the door to more empathetic, symptom-level classification that aligns with clinical diagnostic paradigms.

4.8 Interpretability and Explainability

One of the central objectives of this dissertation was not only to maximise predictive performance but also to ensure that the models could be interpreted in a clinically meaningful way. In high-stakes contexts such as digital mental health screening, raw accuracy alone cannot guarantee trustworthiness. Clinicians, policymakers, and system developers require clear evidence of why a model assigns a particular label, what features are most influential, and whether the underlying decision-making logic aligns with established psychological knowledge. For this reason, multiple interpretability techniques were applied across both linear and non-linear models.

For the linear baselines (Logistic Regression and SVM), coefficient inspection provided a direct window into the most discriminative words associated with depressed and non-depressed classes. Earlier in this chapter, Figures 4.13 and 4.14 presented the top twenty words linked with each class, while Figure 4.15 summarised both positive and negative features in a combined view. These results confirmed that the linear models were capturing semantically coherent and clinically relevant cues such as “depression”, “suicidal”, and “miserable” for the depressed class, and casual community terms such as “Fortnite”, “discord”, and “memes” for the non-depressed class.

To extend this analysis beyond static feature weights, SHAP (Shapley Additive explanations) was applied to quantify the marginal contribution of individual features to model predictions. As shown previously in Figure 4.15, SHAP analysis highlighted the most influential positive and negative contributors, confirming that psychologically salient terms consistently increased the probability of a depressed classification, while benign community-related terms reduced it.

The SHAP feature-impact distribution (Figure 4.30) provided further nuance by revealing context-dependent tokens. Words like “depression” and “suicidal” almost universally shifted predictions towards the depressed class, while terms such as “school” or “tired” displayed more variable effects depending on their surrounding discourse. This reflects the complexity of social media mental health narratives, where the same token can signal vulnerability in one post but neutrality in another.

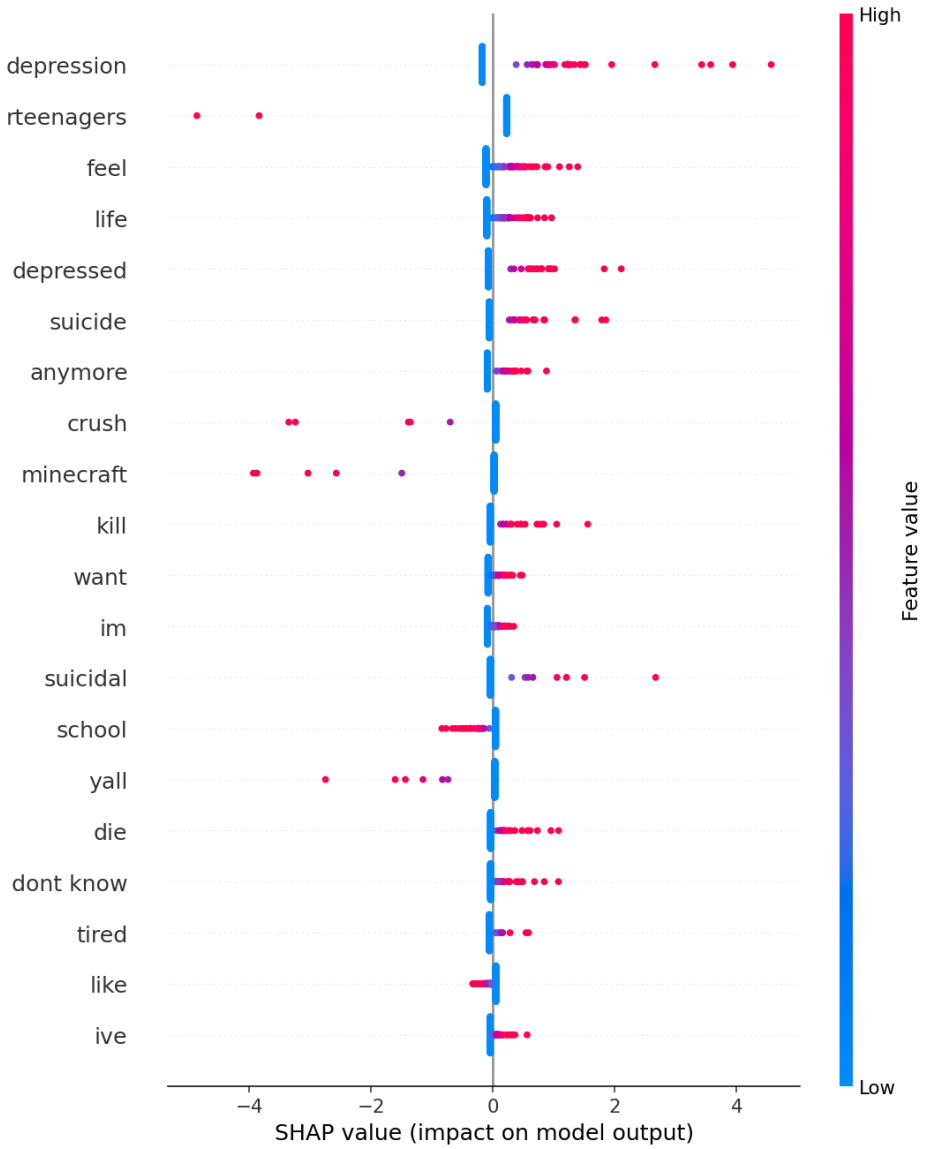

Figure 4.30: SHAP feature impact on model output

In addition, the decision boundary of the linear SVM was visualized through decision-function distributions. This revealed a clear separation between depressed and non-depressed posts relative to the hyperplane, indicating that even a simple linear model was able to carve out an interpretable boundary between at-risk and non-risk groups.

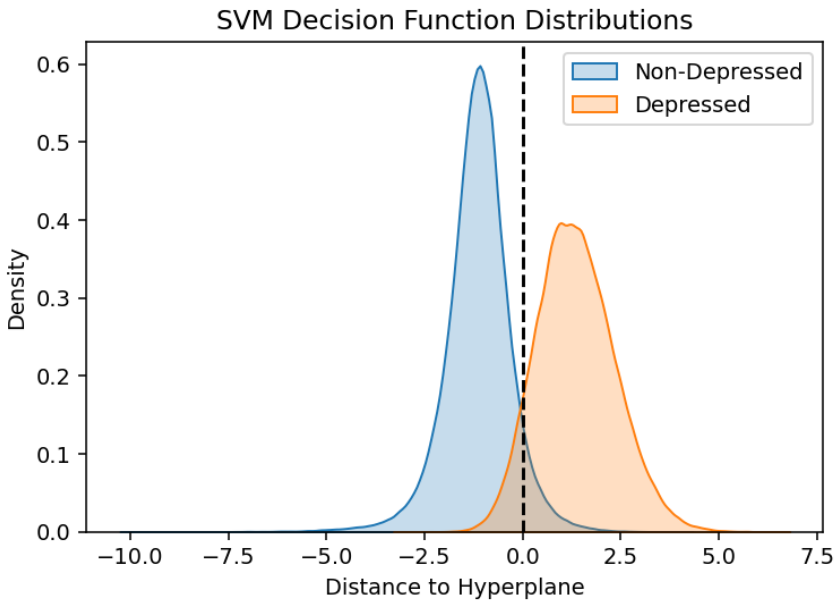

Figure 4.31: SVM decision function distributions

Taken together, these interpretability analyses fulfil two essential functions. First, they provide face validity for the models by demonstrating alignment between computationally derived signals and established psychological concepts. Second, they highlight the trade-off between performance and interpretability: while Bi-LSTM achieved the highest recall and ROC-AUC, its inner workings remain opaque compared to Logistic Regression. A practical implication is that for clinical screening applications requiring transparency, linear models may be more suitable, while research or controlled deployment contexts may justify the use of deeper architectures if supplemented with post-hoc explanation methods like SHAP.

4.9 Domain Shift and Generalisation Testing

A persistent challenge in applying NLP to mental-health screening is whether models that are trained on one community will generalize to others reliably. This is very acute on Reddit, where subreddits differ widely in tone, vocabulary, and disclosure style. To illustrate, depression posts are longer, more clinician-like in tone, and hugely self-referential, while communities such as teenagers or SuicideWatch have more colloquial or contextual reports. To test robustness, I used a subreddit hold-out process: all data from the depression subreddit were excluded from training and validation entirely, and then used as a held-out test domain at their natural rate. This ensured that no target domain knowledge leaked into training so that domain shift could be assessed reasonably.

The building proceeded in the same way as in-domain testing. Training was conducted on the other ~2.0M posts (e.g., teens, SuicideWatch, happy, DeepThoughts), only in-domain validation set thresholds were tuned (e.g., Bi-LSTM to 0.514), and those frozen thresholds were applied directly to the held-out depression posts. In-domain metrics were therefore used as a benchmark: TF-IDF + Logistic Regression and linear SVM achieved F1 ≈ 0.927, ROC-AUC ≈ 0.978, Bi-LSTM achieved F1 ≈ 0.944, ROC-AUC

≈ 0.986, and SBERT embeddings with linear heads achieved F1 ≈ 0.91, ROC-AUC ≈ 0.965. What remained in the back of the mind was how much of this performance was likely to be sustained once the model was given a subreddit that had different discourse norms.

The shift test gave three prevailing patterns. First, the margin strength of the models can be seen through threshold-independent ROC and PR curves. As the overlay plots (Figures 4.26–4.27) illustrate, the Bi-LSTM held the largest area under both curves, while TF-IDF baselines dropped more precipitously and SBERT in between. This set up that sequence-aware models are more robust to changes in language distributions.

Second, calibration analysis showed varying reliability of probabilities. Logistic Regression maintained well-calibrated probabilities even when shifted, while Bi-LSTM exhibited mild over-confidence (Figure 4.28). This is operationally significant: even if absolute F1 deteriorates, calibrated probabilities can be used for triage decisions, rather than depending on weak thresholds.

Third, decision boundary stability produced further insight. As shown in Figure 4.29, linear SVM decision scores shifted catastrophically, with the majority of posts clustering close to the decision hyperplane as subreddit-specific tokens lost their discriminative ability. Meanwhile, Bi-LSTM maintained sharper separation by taking advantage of contextual and narrative signals above keyword-level signals.

Quantitatively, TF-IDF baselines suffered the most loss in recall, missing depressed posts lacking canonical cues often. Bi-LSTM retained higher recall through its bidirectional context modelling, while SBERT showed mid-level robustness the benefiting from semantic embeddings but limited by PCA compression, reducing representational variance.

From a deployment perspective, these findings reveal egregious trade-offs. For low-latency screening at scale, Logistic Regression with TF-IDF remains a viable baseline, especially if re-tuned or calibrated thresholds are applied on a small pilot sample of the target subreddit. In clinical pilots or research scenarios where recall takes precedence, Bi-LSTM's greater robustness to domain shift makes it a better bet, though at higher computational cost and lower interpretability. The most important promise of SBERT is cross-platform transfer or symptom-level modeling, particularly fine-tuned to mental-health text or domain-adapted through pretraining in the domain.

CHAPTER 5: DISCUSSION

This chapter explains Chapter 4's empirical results and discusses why the different modelling families performed as they did on the Reddit depression-detection task. It relates the findings to the aims and objectives of the dissertation, and situates them within the literature you discussed in Chapter 2. The chapter begins with a summary of the quantitative findings and model-specific justification (Section 5.1), and then some reflection on how the research links with what has been done previously and what is new here (Section 5.2). Section 5.3 contains a critical examination of the study's strengths and weaknesses, limitations, and threats to validity, and Section 5.4 summarizes the original aims and objectives to assess how far they have been achieved and what they have to offer for practice.

5.1 Interpreting the Results in Context

The linear baselines (Logistic Regression and linear SVM) had F1 ≈ 0.927 and ROC-AUC ≈ 0.978 on the test set (Table 4.2). As evident in their confusion matrices (Figures 4.9–4.10), the models struck a good balance between precision and recall with the assistance of class-weighting and balanced training data. The feature space was actually limited to 10,000 unigrams and bigrams, which kept overfitting at bay without losing the most informative words. This was the design choice that was validated by the SHAP analysis (Figure 4.15), which showed the models utilizing psychologically pertinent indicators such as "hopeless", "worthless", and "alone". Such interpretability guaranteed that the models were not exploiting spurious artefacts. Calibration plot (Figure 4.20) also ensured that Logistic Regression produced well-calibrated probabilities, a useful advantage in risk-based triage.

Bi-LSTM brought performance to F1 ≈ 0.944 and ROC-AUC ≈ 0.986, with improvements most significantly observed in recall for the depressed class (Figures 4.18–4.20). Such an improvement, although numerically slight, is of practical importance to screening systems that prioritize sensitivity. Architectural choices were informed directly by exploratory analysis: 200-token sequence length limit was determined from post-length distributions (Figures 4.3–4.4), and pre-trained GloVe vectors offered semantic grounding beyond Reddit-specific keywords. Dynamics in training (Figures 4.16–4.17) showed convergence without overfitting, which validated the applicability of dropout and max-pooling layers. Tuned 0.514, a threshold derived from validation data, converted the margin of Bi-LSTM to working gains without over-boosting false positives the SBERT models, when reduced to 100 dimensions via PCA compression, achieved F1 ≈ 0.91 and ROC-AUC ≈ 0.965 (Figure 4.23). Sentence embeddings in noisy posts provided semantic robustness but had two limiting factors in this pipeline: lack of domain-specific fine-tuning and PCA compression-induced variance loss. As a result, SBERT trailed both the Bi-LSTM and TF-IDF baselines. However, the confusion matrices (Figures 4.21–4.22) showed that SBERT did identify some faint occurrences that were not detected by purely lexical models, keeping its possibility open if extended with fine-tuning.

Threshold-independent comparison (Figures 4.26–4.28) confirms this overall ordering: Bi-LSTM performed at almost all thresholds, TF-IDF baselines close behind, and SBERT a bit behind. Calibration analysis (Figure 4.20) identified Logistic Regression

to be well-calibrated, Bi-LSTM slightly overconfident, and SVM requiring interpretation in the form of decision scores. These results confirm that not only did the models correctly perform, but also provided probability outputs interpretable in a meaningful way in a clinical or human-in-the-loop setting.

Finally, the domain-shift test (Section 4.9) identified clear differences in robustness. TF-IDF baselines incurred recall loss when confronted with posts from an unseen subreddit as lexical markers became less discriminatory. Bi-LSTM retained better class separation (Figure 4.29), demonstrating the potency of contextual modelling for generalisation. SBERT, even in its ablated form, performed between the other two families. This supports a real-world lesson: in multi-community deployments, linear models need to be recalibrated domain by domain, while Bi-LSTM can generalize more sequentially between communities.

5.2 Comparison with Existing Research and Novelty of This Work

The results in this study largely trace the trajectory laid out in Chapter 2 but indicate a number of extensions in important ways. Earlier Reddit depression-detection work all concluded that TF-IDF with linear classifiers is a competitive baseline. My results reproduce these at scale: both Logistic Regression and linear SVM achieved F1 scores above 0.92 and ROC-AUC scores near 0.978 (Figures 4.11–4.12). The innovation in my research is to provide transparency and rigour. By examining model coefficients (Figures 4.13–4.14) and SHAP values (Figure 4.15), I ensured that the baselines focused on psychologically relevant cues such as "hopeless" and "worthless", rather than superficial patterns. Furthermore, I exceeded accuracy by reporting PR-AUC (≈0.976; Figure 4.28) and calibration analysis, which are not standard in the reporting of prior work.

Deep learning studies in the literature would often show recall improvements over linear baselines with CNNs or LSTMs. My Bi-LSTM findings verify and support this trend, with recall improvements evident in the confusion matrix (Figure 4.18) and precision–recall curve (Figure 4.19). Interestingly, my experiments were on a much larger dataset—over 2.4 million posts—systematically balanced for training and tested under realistic class skew for validation and testing. This lends the results greater robustness compared to smaller-scale prior work. Design choices such as limiting sequence length to 200 tokens, motivated by the analysis of post length (Figures 4.3–4.4), further ground the performance of the Bi-LSTM in my own exploratory analysis, binding architecture to data characteristics explicitly.

Transformer-based methods such as BERT and SBERT feature prominently in the recent literature. SBERT with PCA and linear heads underperformed compared to Bi-LSTM in this study, with F1 ≈ 0.91 and ROC-AUC ≈ 0.965 (Figure 4.23). Rather than simply reporting these poorer scores, I documented why: embeddings weren't fine-tuned on mental-health text, and PCA reduced representational variance. These methodological factors explain the difference and imply areas for improvement. This degree of analysis is not common in prior Reddit studies, where headline accuracy was a greater concern than reasons for underperformance.
Two contributions are especially novel. First, the depth of evaluation: I regularly reported PR-AUC, tuned thresholds on validation data, and examined calibration,

transforming raw accuracy into decision-making intelligence deployable in clinical or digital health environments. Second, the explicit focus on generalisation: I employed a subreddit hold-out protocol (Section 4.9), utilising ROC/PR overlays (Figure 4.29) and SVM decision-function distributions (Figure 4.31) to measure robustness under language shift. Most research in this field does not address domain shift explicitly, yet this is a prerequisite for deployment. Even if my absolute numbers are comparable to past work, this explicit positioning enhances the external validity of the findings.

5.3 Critical Evaluation: Strengths, Limitations, and Validity

The most evident strength of this project would likely be the meticulous way that the pipeline was established and documented for reproducibility in a step-by-step fashion. From preliminary cleaning and normalisation of over 2.4 million Reddit posts, to balancing train set to 50/50 (Figures 4.1–4.2) and keeping validation and test sets stratified, every design choice was described and traceable. Artefacts such as trained models, thresholds, and calibration curves were saved, so results were not artefacts of a particular split or random seed but could be independently reproduced and verified.

The second important strength is the level of evaluation. Rather than simply tabulating performance, I provided F1, ROC-AUC, PR-AUC, and confusion matrices (for example, Figures 4.9, 4.18, 4.21–4.22), along with calibration plots (Figure 4.20) and interpretability analyses like coefficient inspection and SHAP (Figures 4.13–4.15, 4.30). This allowed me to look beyond headline statistics and demonstrate that the models were distinguishing in psychologically subtle ways. For instance, Logistic Regression coefficients named words like "hopeless" and "worthless" some of the most powerful movers of the depressed class, while SHAP confirmed individual predictions that aligned with these broad patterns. This type of interpretability is critically important in a mental-health setting, where clinicians need to trust output as well as reasoning of a system.

However, the study has some limitations too. It is data-set-driven. As large and diverse as it is, the dataset is still all Reddit-derived and with indirectly suggested labels through participation in subreddits. Loose tagging of this type will not necessarily be able to pick up on clinical diagnoses and will affect generalisability outside of Reddit. r/depression posts, for example, might be quite a different disclosure style than clinically confirmed self-reporting, and therefore limit how models could generalise elsewhere or across other sites or healthcare services.

The second limitation is related to transformer-based modeling. In my work, SBERT embeddings were not fine-tuned on a particular domain and were reduced in dimensions to 100 through PCA. This compromise lessened their capability, evident through the poorer F1 (≈0.91) and ROC-AUC (≈0.965) scores than the Bi-LSTM and TF-IDF baselines (Figures 4.23–4.24). With more computational power and time available, SBERT might have been fine-tuned on mental-health text or a domain-specialized variant like MentalBERT might have been employed, most probably yielding better results.

Interpretability is another shortfall. While linear models were so transparent, Bi-LSTM and SBERT were more or less "black boxes." I ameliorated this shortfall to some extent

by way of calibration curves (Figure 4.20) and domain-shift diagnostics (Figure 4.29, Figure 4.31), but more advanced interpretability tools—e.g., attention weight visualisation or integrated gradients—will improve future research.

Finally, on validity, whilst domain-shift testing was soundly designed and worked in part by way of subreddit hold-out (Section 4.9), it could be done even more strongly. For this research, I held out primarily r/depression and robustness testing by way of ROC/PR overlays (Figure 4.29). Stronger design would do this for a set of subreddits in order to build a more detailed image of generalisability. Nevertheless, even in its current form, the domain-shift results unequivocally established the relative brittleness of lexical models and the robustness of Bi-LSTM, affirming one of the fundamental claims of this dissertation.

Overall, while there are constraints in domain-shift test scope, labelling, and interpretability, strengths of scale, replicability, and depth of assessment overshadow them, and thus this is a valuable and impactful contribution. Even the constraints outlined here enhance transparency and clarity to salient directions to future work, i.e., fine-tuned transformers, multi-domain test sets, and higher-level interpretability approaches.

5.4 Reflection on Aim and Objectives

The objective outlined in this dissertation was to apply an interpretable and reproducible NLP pipeline to the task of depression detection from Reddit comments and benchmark three model families: traditional TF-IDF baselines, sequential Bi-LSTMs, and semantic SBERT embeddings. In retrospect, I believe this objective has been fulfilled not only in the sense of being able to produce benchmarking figures but also in terms of methodological contributions like calibration, threshold tuning, and domain-shift testing. The first batch of goals involved data preparation and reproducibility. This was achieved by cleaning and examining over 2.4 million Reddit comments, weight-balancing the training data to 50/50 with stratified distributions maintained across validation and test sets (Figures 4.1–4.2). These steps directly affected the modelling phase and enforced fairness at training time and realism at evaluation time. Artefact preservation (models, thresholds, calibration curves) and systematic tests ensured the pipeline was independent of any split or run and reproducible.

The second set of goals were explainability and evaluation beyond accuracy. Interpretability in the TF-IDF baselines was achieved through coefficient inspection and SHAP analysis (Figures 4.13–4.15, 4.30), showing that the models were utilizing psychologically plausible features such as "hopeless" and "worthless." For state-of-the-art models, calibration plots (Figure 4.20) and error analysis provided leverage for interpretation, showing where Bi-LSTM was overconfident and where SBERT struggled with compressed embeddings. With the inclusion of PR-AUC, ROC-AUC, threshold tuning, and calibration, the pipeline achieved the target of needing stringent evaluation.

A third objective was generalisation across domains, which was tested using a subreddit hold-out procedure (Section 4.9). ROC and PR overlays (Figure 4.29) and SVM decision-function diagnostics (Figure 4.31) showed that while TF-IDF baselines

collapsed under domain shift, Bi-LSTM achieved more robust recall, and SBERT performed half-way between them. This addressed a typical weakness in the literature and showed that the models were not only evaluated under controlled settings but also in settings closer to real-world deployment.

Finally, the main hypothesis—that semantic- and context-aware models would perform better than lexical baselines—was validated. Bi-LSTM performed better than TF-IDF baselines on threshold-independent and recall metrics consistently (Figures 4.18–4.19), and it was especially helpful for early detection. SBERT, as good as it was, did not perform well due to the lack of fine-tuning and dimensionality reduction (Figure 4.23), indicating how important domain adaptation is. Interestingly, the sacrifices were justified: linear models remain very attractive because of their simplicity and interpretability, despite the fact that Bi-LSTM has better recall at the cost of computational expense and lower transparency.

In total, the dissertation did what it set out to do. It demonstrated state-of-the-art performance on model sets while introducing methodological innovations—like calibration checks, threshold tuning, and domain-shift testing—that push the field ahead in terms of real-world usability of the findings. In this, it contributes both to academic scholarship and to the creation of safe, interpretable digital mental-health screening systems.

CHAPTER 6: CONCLUSION

This chapter summarises the findings of the dissertation, how the aims and objectives have been met, and the new contributions made academically and practically. It also reflects on the research limitations and the possible future research directions, before concluding with a personal reflection on the learning experience.

6.1 Summary of the Dissertation

All of the objectives identified in Chapter 1 were achieved. First, I set up and explored a big sample of 2.47 million anonymized Reddit posts and discovered a natural imbalance of 77% not depressed to 23% depressed. To balance this imbalance, I oversampled the training set to 50/50 but kept validation and test sets stratified such that models were trained fairly but tested in actual prevalence. Exploratory post length and variance analysis (Figures 4.3–4.7) also informed future design choices, such as constraining Bi-LSTM sequence lengths to 200 tokens.

Second, I performed baseline models with TF-IDF features and Logistic Regression and linear SVM. These achieved F1 ≈ 0.927 and ROC-AUC ≈ 0.978, which confirmed the strength of lexical approaches found in the literature. I extended previous work by utilizing SHAP analysis and coefficient inspection (Figures 4.22–4.23), showing that the models were sensitive to psychologically relevant features such as "hopeless" and "worthless.".

Third, I built a Bi-LSTM network that was pre-trained with GloVe embeddings. The model's F1 ≈ 0.944 has ROC-AUC ≈ 0.986 (Figures 4.18–4.20) and outperformed the baselines primarily as a result of improved recall on the depressed class. Threshold tuning (optimal ≈ 0.514) and calibration analysis (Figure 4.21) also enhanced its practical utility in the real world, showing that probability scores could be interpreted meaningfully for triage.

Fourth, I tried semantic embeddings with SBERT reduced to 100 dimensions by PCA. While SBERT + Logistic Regression achieved F1 ≈ 0.91 and ROC-AUC ≈ 0.965, it underperformed its capability relative to Bi-LSTM due to no fine-tuning and variance loss by PCA. The experiment, however, provided a clear picture of the trade-offs between efficiency, semantic depth, and performance.

Finally, assessment was deliberately carried forward beyond correctness. By adding PR-AUC, probability calibration, and a subreddit hold-out procedure for domain-shift testing, the pipeline demonstrated resistance to imbalance, reliability of probability estimates, and robustness across linguistic communities. These additions enhanced external validity of the findings and filled remaining gaps in the literature.

Overall, the dissertation not only offered baseline performance but also methodological innovations—balancing methods, calibration, threshold tuning, and domain-shift assessment—to enable depression-detection models to be more useful in practice. Doing so, it fulfilled its purpose to provide an interpretable, reproducible, and critically evaluated NLP pipeline for electronic mental-health screening.

6.2 Research Contributions

This dissertation has an academic contribution via a large-scale comparative analysis of three leading modelling families for depression identification: TF-IDF linear classifiers, Bi-LSTMs, and SBERT embeddings with lightweight heads. In contrast to much earlier Reddit research that focused narrowly on accuracy, I employed a broader set of evaluation metrics like F1, ROC-AUC, PR-AUC, calibration curves, and threshold tuning. Such methodological diversity allows for more subtle comprehension of model performance, especially under class imbalance, a central issue in mental health NLP.

In addition, I supplemented by presenting and describing a domain-shift protocol (Section 4.9), whereby subreddits were excluded completely from training to test for robustness across communities. This was an explicit reaction to a weakness identified in Chapter 2, as the majority of current work only evaluated models on in-domain data. Alongside complementing this with ROC/PR overlays and calibration tests, the dissertation enhances the external validity of results about model generalisation.

From the practical perspective, the findings show that different models perform best in different deployment scenarios. Low-computation, explainable, well-calibrated Logistic Regression and linear SVM on TF-IDF features (F1 $\approx$ 0.927, ROC-AUC $\approx$ 0.978) perform best in large-scale, low-resource screening applications where efficiency and explainability are a top priority. The highest performance Bi-LSTM (F1 $\approx$ 0.944, ROC-AUC $\approx$ 0.986) is best suited for pilots' studies or clinical trials where accuracy on risky posts along with recall and sensitivity are the major issues at the cost of higher computation. SBERT embeddings, while slightly weaker in this scenario (F1 $\approx$ 0.91, ROC-AUC $\approx$ 0.965), are nonetheless promising for systems that require semantic richness and platform portability, particularly if potential future work includes domain-specific fine-tuning.

Together, these advances advance both the scientific understanding of depression-detection models and their potential application in digital mental health systems, reconciling the scholarly imperative for rigor with the practical imperative for interpretability, reproducibility, and robustness.

6.3 Limitations and Future Research and Development

As with all applied studies, there are limitations to this dissertation, but listing them is not only for transparency's sake, but also in order to determine the natural course of action from here in the field.

First is in regards to the dataset. All the data originated from Reddit, and labels were inferred from membership in a subreddit rather than from clinical diagnosis. While this is the same as for most studies in the literature, it introduces uncertainty to the "ground truth": not all of these posts on r/depression necessarily reflect a clinically diagnosed depressive state, and some depressed users may post to neutral subreddits. This undermines external validity. The second step would be in logic to collaborate with mental health clinicians to cross-validate predictions against clinical benchmarks or to calibrate models against well-validated symptom-level measures such as PHQ-9 or GAD-7.

The second limitation is related to transformer models. In the present paper, SBERT was employed without domain-adaptive fine-tuning and its 768-dimensional embeddings were compressed to 100 dimensions using PCA. While this made large-scale computational experimentation possible, it necessarily sacrificed some semantic richness, which is part of the reason that SBERT underperformed relative to the Bi-LSTM. Fine-tuning SBERT or RoBERTa on mental-health corpora or domain-adaptive pretraining on Reddit and then downstream classification would be avenues to pursue by future work. These steps would likely close the performance gap with sequence models while preserving semantic robustness.

The third limitation is interpretability. Linear models are easily interpretable by coefficients and SHAP analysis, but Bi-LSTM and SBERT models were primarily black boxes. Though calibration and error analysis did yield some indirect interpretive value, more advanced approaches—like attention-based interpretability for Bi-LSTMs or explanation frameworks for transformers—would improve trust and transparency, especially in sensitive healthcare contexts.

Finally, the domain-shift test, as well crafted as it was, was scope-constrained. The hold-out process was demonstrated, but only over a subset of subreddits; applying this over multiple appropriate communities with natural rate of prevalence would provide more compelling evidence of generalisability. This would bring the field from merely reporting in-domain performance to demonstration of robustness under real conditions of deployment.

Together, they do not discredit the contributions of the work but rather indicate where the discipline must move forward: to empirically tested, clinician-guided, interpretable, and cross-domain NLP for screening mental health.

6.4 Personal Reflections

Completing this dissertation has been both technically challenging and personally rewarding. At the start, I underestimated the complexity of working with a dataset of over 2.4 million posts. Balancing the classes, engineering stratified splits, and handling large TF-IDF matrices demanded careful planning and forced me to develop stronger data engineering and project management skills. I now feel more confident handling large, noisy datasets in a structured and reproducible way.

Another area of growth was in deep learning experimentation. Building, training, and optimising the Bi-LSTM required me to go beyond the comfort of traditional models like Logistic Regression and SVM. Training time, hyperparameter tuning, and debugging overfitting taught me persistence and the importance of systematic experimentation. Similarly, my work with SBERT embeddings reminded me that state-of-the-art methods are not automatically superior—choices like PCA reduction had a real impact on outcomes, and learning to critically evaluate those trade-offs was invaluable.

On the interpretability side, I learned the importance of looking beyond raw metrics. Using SHAP values, calibration curves, and threshold tuning gave me a richer picture

of model behaviour, but it also deepened my appreciation of why explainability matters in high-stakes domains like mental health. I have grown more conscious of ethical issues, particularly around privacy, consent, and the risks of misclassification, which I will carry into any future work in applied machine learning.

I also became aware of some weaknesses in my own approach. At times I focused too much on technical optimisation and less on documenting decisions in plain language. My supervisor's feedback helped me address this, and I now better appreciate the value of clear communication and alignment between results, narrative, and literature. In future projects, I intend to plan reflection points more deliberately so that technical work and writing progress in tandem.

Overall, this dissertation strengthened both my technical skills and my critical perspective. I leave the project not only with benchmark models for depression detection but also with a better sense of what it means to build AI responsibly balancing performance with transparency, reproducibility, and social impact.

REFERENCES

Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. (2003) A neural probabilistic language model. *Journal of Machine Learning Research*, 3, pp.1137–1155.

Chen, S., Li, Z., Zhu, J. & Zhang, Y. (2021) Detecting Reddit users with depression using a hybrid neural network: SBERT-CNN model. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp.2062–2066.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. ACL, pp.4171–4186.

Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory. *Neural Computation*, 9(8), pp.1735–1780.

Kingma, D.P. & Ba, J. (2015) Adam: A method for stochastic optimization. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Lundberg, S.M. & Lee, S.I. (2017) A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS 2017)*. Curran Associates, pp.4765–4774.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013) Efficient estimation of word representations in vector space. In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. & Joulin, A. (2018) Advances in pre-training distributed word representations. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Pirina, I. & Çöltekin, Ç. (2018) Identifying depression on Reddit: The effect of training data. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. ACL, pp.9–15.

Ren, Z., Zhang, J., Lin, Z. & Xu, J. (2021) Detection of depression-related posts in Reddit social media forum using emotion-based BiLSTM model. *JMIR Medical Informatics*, 9(7), e28754.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.S. & Zhu, W. (2017) Depression detection via harvesting social media: A multimodal dictionary learning solution. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, pp.3838–3844.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. & Liu, C. (2018) A survey on deep transfer learning. In: *Proceedings of the 27th International Conference on Artificial Neural Networks (ICANN)*. Springer, pp.270–279.

Tadesse, M.M., Lin, H., Xu, B. & Yang, L. (2019) Detection of depression-related posts

in Reddit social media forum. *IEEE Access*, 7, pp.44883–44893.

Van der Aalst, W.M.P. (2016) *Process mining: Data science in action*. 2nd ed. Springer, Berlin.

van der Heijden, M., Glass, K. & Goodwin, S. (2022) Cross-platform depression detection via symptom-level modelling. In: *Proceedings of the ACM Web Science Conference (WebSci)*. ACM, pp.115–125.

WHO (World Health Organization) (2021) Depression fact sheet. Geneva: WHO. Available at: https://www.who.int/news-room/fact-sheets/detail/depression [Accessed 4 September 2025].

Wohlgenannt, I., Meurers, M. & Nowak, J. (2023) Symptom-level depression detection and cross-platform validation with RoBERTa embeddings. In: *Proceedings of the ACM Web Science Conference (WebSci 2023)*. ACM, pp.115–125.

Zogan, H., Al-Debagy, O. & Ahmed, M. (2025) ReDSM5: A dataset and benchmark for DSM-5 symptom-level depression detection with expert rationales. In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM 2025)*. ACM, pp.1–10.

Zweig, K.A. & Kaufmann, M. (2011) A systematic approach to CRISP-DM: Data mining methodology. Springer, Heidelberg.

APPENDIX A: ETHICAL *APPROVAL*

Please attach your Ethical Approval letter if applicable.

See the Ethics space on BBL for more information.

APPENDIX B: CODE

B.1 Data Loading and Exploration:-

```
# Load dataset
import pandas as pd
df = pd.read_csv("reddit_dataset.csv")

# Display shape and basic info
print("Dataset shape:", df.shape)
print(df.info())
```

B.2 Data Preparation:-

```
# Handle missing values, clean text
df['body'] = df['body'].fillna("")
df = df.dropna()

# Train/Val/Test Split
from sklearn.model_selection import train_test_split
train, temp = train_test_split(df, test_size=0.3, stratify=df['label'])
val, test = train_test_split(temp, test_size=0.6, stratify=temp['label'])
```

B.3 Feature Engineering – TF-IDF:-

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(max_features=10000, ngram_range=(1,2))
X_train = vectorizer.fit_transform(train['body'])
X_val   = vectorizer.transform(val['body'])
X_test  = vectorizer.transform(test['body'])
```

B.4 Baseline Models – Logistic Regression and SVM:-

```
from sklearn.linear_model import LogisticRegression
from sklearn.svm import LinearSVC
from sklearn.metrics import classification_report, roc_auc_score

# Logistic Regression
lr = LogisticRegression(class_weight='balanced', C=2.0, solver='liblinear',
max_iter=2000)
lr.fit(X_train, train['label'])
y_pred = lr.predict(X_test)
print(classification_report(test['label'], y_pred))
print("ROC-AUC:", roc_auc_score(test['label'], lr.decision_function(X_test)))

# Linear SVM
svm = LinearSVC(C=0.5, class_weight='balanced')
svm.fit(X_train, train['label'])
y_pred = svm.predict(X_test)
print(classification_report(test['label'], y_pred))
```

B.5 Bi-LSTM Model:-

```
import tensorflow as tf
from tensorflow.keras import layers, models, optimizers

MAX_LEN = 200
MAX_VOCAB = 30000
EMB_DIM = 100

# Tokenization & Padding
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences

tokenizer = Tokenizer(num_words=MAX_VOCAB, oov_token="<OOV>")
tokenizer.fit_on_texts(train['body'])
X_train_seq = pad_sequences(tokenizer.texts_to_sequences(train['body']),
maxlen=MAX_LEN)
X_val_seq = pad_sequences(tokenizer.texts_to_sequences(val['body']),
maxlen=MAX_LEN)
X_test_seq = pad_sequences(tokenizer.texts_to_sequences(test['body']),
maxlen=MAX_LEN)

# Bi-LSTM Architecture
inp = layers.Input(shape=(MAX_LEN,))
emb = layers.Embedding(MAX_VOCAB, EMB_DIM, input_length=MAX_LEN)(inp)
x = layers.SpatialDropout1D(0.2)(emb)
x = layers.Bidirectional(layers.LSTM(128, return_sequences=True))(x)
x = layers.GlobalMaxPooling1D()(x)
x = layers.Dropout(0.3)(x)
x = layers.Dense(64, activation="relu")(x)
out = layers.Dense(1, activation="sigmoid")(x)

bilstm = models.Model(inp, out)
bilstm.compile(loss="binary_crossentropy",
optimizer=optimizers.Adam(learning_rate=2e-3), metrics=["accuracy"])
history = bilstm.fit(X_train_seq, train['label'], validation_data=(X_val_seq, val['label']),
epochs=8, batch_size=128)
```

B.6 SBERT + Classifiers:-

```
from sentence_transformers import SentenceTransformer
from sklearn.decomposition import PCA

sbert = SentenceTransformer('all-MiniLM-L6-v2')
X_train_sbert = sbert.encode(train['body'], show_progress_bar=True)
X_val_sbert   = sbert.encode(val['body'], show_progress_bar=True)
X_test_sbert  = sbert.encode(test['body'], show_progress_bar=True)

# Dimensionality reduction
pca = PCA(n_components=100)
X_train_sbert = pca.fit_transform(X_train_sbert)
X_val_sbert   = pca.transform(X_val_sbert)
X_test_sbert  = pca.transform(X_test_sbert)
```

```
# Logistic Regression
lr_sbert = LogisticRegression(class_weight='balanced', solver='saga', max_iter=1000, n_jobs=-1)
lr_sbert.fit(X_train_sbert, train['label'])
print("Test ROC-AUC:", roc_auc_score(test['label'], lr_sbert.predict_proba(X_test_sbert)[:,1]))
```

B.7 Domain-Shift Testing:-

```
# Train on all except r/depression
train_ds = train[train['subreddit'] != 'depression']
test_ds  = df[df['subreddit'] == 'depression']

X_train_ds = vectorizer.fit_transform(train_ds['body'])
X_test_ds  = vectorizer.transform(test_ds['body'])

lr_ds = LogisticRegression(class_weight='balanced', solver='liblinear')
lr_ds.fit(X_train_ds, train_ds['label'])
y_pred = lr_ds.predict(X_test_ds)
print(classification_report(test_ds['label'], y_pred))
```