Detecting and Interpreting Mental Health Distress in UK University Reddit Posts: A Multi-Model Method

Sayak Chakraborty
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, UK
s.chakraborty@se24.qmul.ac.uk

***Abstract*—This paper presents a mental health classification system for UK university students based on Reddit-style social media posts. The objective was to categorize posts into five types of emotions-based stress: academic stress, relationship issues, existential crisis, social isolation and neutral and to provide automated well-being suggestions. Multiple models were developed and compared, including Logistic Regression, SVM, XGBoost, RoBERTa and a Stacked Ensemble. TF-IDF and transformer-based features were utilized for classification. Word cloud visualizations, distress severity gauges and Cohere-powered real-time suggestions were integrated into a Stream-lit dashboard. The stacked model achieved the best performance (accuracy: 90.16%, macro-averaged F1-score: 90.64%). The report discusses pre-processing, weak supervision, model comparison and the potential impact of such systems in student well-being support.**

***Index Terms*—Mental health, emotion classification, Reddit, TF-IDF, RoBERTa, weak supervision, Streamlit, Cohere**

I. Introduction

The mental health of university students in the UK has been an increasing source of concern. In recent years, surveys and research have consistently shown a rise in psychological distress, including symptoms of depression, anxiety, social withdrawal and academic burnout. With the growing influence of digital platforms, students frequently use online communities such as Reddit to express their emotional states and seek support anonymously. These un-structured, real-time statements represent a fertile but unclaimed opportunity for early detection and intervention of mental illness.

Traditional diagnostic tools, although clinically validated, are labor-intensive and reactive. Natural language processing (NLP) and machine learning (ML), however, offer a scalable and proactive solution to detect markers of emotional distress in text data. The objective of this project is to create and evaluate a machine learning system that is capable of classifying student Reddit posts into emotional distress categories namely, academic stress, relationship issues, existential crisis, social isolation, and neutral(Fig. 1). The system also predicts the distress severity level and provides intelligent well-being suggestions through integration with a large language model (LLM).

For the creation of a domain-specific solution, a custom dataset of 7,689 Reddit posts from UK student communities was created and weakly labeled using emotion-specific heuristics. A variety of ML models, from traditional classifiers like Logistic Regression and SVMs with TF-IDF, to transformer models like RoBERTa and a custom stacked ensemble were trained and compared. An interactive Streamlit web dashboard was developed to provide predictions, category-wise probabilities, severity visualization, key terms and actionable recommendations using a Cohere inference model.

The project demonstrates how ML/NLP techniques, if tailored with domain constraints, can offer universities scalable mental health monitoring solutions. The rest of this paper reviews the relevant literature, methodology, model performance and evaluation of this approach.

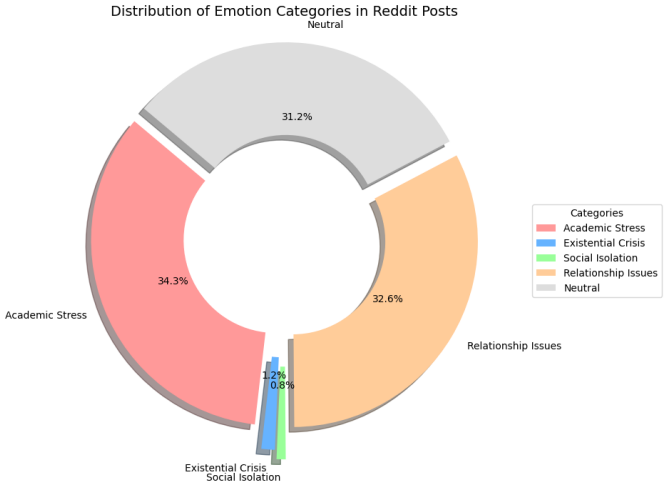

Fig. 1. Distribution of Emotion Categories in the Dataset

A. Objectives

The primary objectives of this study are:

- To develop a custom-labeled dataset of Reddit posts from university-related communities using PRAW-based web scraping, categorized into distress types like academic

stress, relationship issues, social isolation, existential crisis, and neutral.

- To apply weak supervision techniques to label huge quantities of unlabeled text data by leveraging keyword heuristics along with domain expertise.
- To design and implement multiple machine learning models (e.g., Logistic Regression, SVM, XGBoost) and fine-tuned transformer-based architectures (e.g., RoBERTa, Stacked Ensemble) to classify emotional distress categories.
- To evaluate and compare model performance using metrics such as accuracy, F1-score, and ROC-AUC, with particular emphasis on F1-score for imbalanced classification.
- To build a practical Streamlit-based web interface(Fig. 2) that allows real-time classification of Reddit-style input, visualizing severity, distress keywords, and providing mental-health support suggestions through Cohere's generate-endpoint.
- To analyze limitations, challenges, and ethical issues related to mental health text mining, including privacy, model fairness, and generalization.

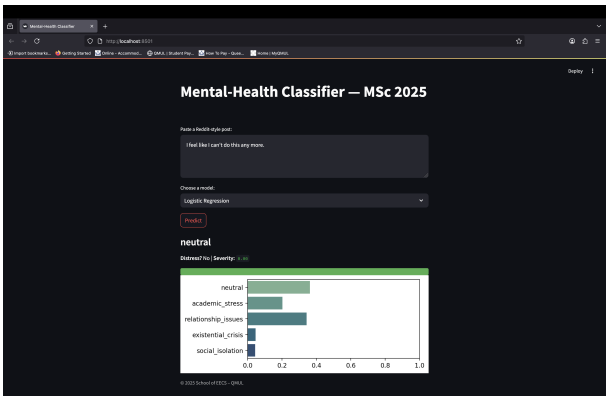

Fig. 2. Interactive Webpage

II. RELATED WORK

The detection of mental health signals in online communities has gained significant momentum in recent years due to the proliferation of user-generated content on platforms like Reddit, Twitter, and online forums.Various works have attempted to explore mental health indicators using natural language processing (NLP), machine learning (ML), and later, transformer-based architectures. These works form the foundations upon which the current research is built.

Resnik et al. (2015) were amongst the earliest to apply topic modeling techniques such as Latent Dirichlet Allocation (LDA) to Reddit posts within mental illness subreddits. They found evidence of topic-based correlations with depression and anxiety, and that patterns in language may be used to uncover underlying psychological characteristics. Similarly, Coppersmith et al. (2018) used supervised learning techniques on social media datasets to predict clinical depression and PTSD, further validating the hypothesis that social media signals can be used as surrogates for mental state indicators.

Chancellor et al. (2020) was one of the prominent contributions, with them pointing out ethical issues in automated mental health detection. Their review pointed out data privacy, false positives, and the need for explainable AI in high-stakes applications as concerns. The present work tries to mitigate some of the above concerns by using model explainability by way of visual explanations (word clouds) and actionable suggestions based on lightweight, transparent models like TF-IDF + LR and rule-based weak labelling.

For emotion recognition, use of lexicon-based and weak supervision has been documented in research such as Araque et al. (2017) and Nigam et al. (2021). They used emotion keyword dictionaries and pattern matching methods for weak labeling of large corpora. This article takes the same approach by using hand-tuned emotion keyword rules to annotate categories like "academic stress" and "relationship problems." Not only does this reduce annotation cost, but also facilitates faster prototyping in real applications.

Transformer models like BERT, RoBERTa, and DistilBERT are currently the new standard in NLP for classification. Devlin et al. (2018) developed BERT, which enabled deep bidirectional language understanding. It was then adopted by Liu et al. (2019) who presented RoBERTa, a better version with training over a larger corpus. The models have been state-of-the-art performers across many different benchmarks from sentiment and emotion classification to question answering. In our case, RoBERTa was fine-tuned for distress level classifying in Reddit posts, and we also used it in a stacked ensemble model with classical ML methods in order to balance performance with interpretability.

Overall, the existing literature is in favor of the feasibility and usability of using NLP-based models to detect mental health distress online. This paper builds upon these ideas and adds value by providing a live web-based demo interface, category-specific recommendations in real time, and weakly labelled emotion taxonomy from real-life university settings.

A. Background Research

Mental health among university students has become a significant public health issue in the UK over recent years. According to the survey conducted by the mental health charity *Student Minds* in 2020, one in five students had an officially diagnosed mental health condition and almost half reported encountering a serious psychological problem requiring professional help [16]. Furthermore, 85% of further education colleges reported increasing mental health issues among students, with widespread depression and anxiety being most prevalent [17]. Surveys in 2022 replicated these findings, with around a third of UK students reporting poor general well-being and 30% reporting declining mental health since university entry [18].

Academic stress is likely one of the best-documented and studied stressors in students [16]. Difficult classes, rigorous deadlines, and test anxiety have frequently been found to

trigger chronic stress and anxiety. Academic pressure has also been consistently linked to poor mental health outcomes, including high depressive symptoms. Financial stresses obfuscate the issues: as living costs and charges increase, students are increasingly pulled into part-time work, typically at the cost of rest and social contact.One UK survey recently found that 59% of students were often or constantly worried about money [18]. Combined with academic pressure and concern for later working lives, these financial stresses provide an increased level of psychological strain.

Relationship problems, both intimate and interpersonal, are yet another principal cause of distress for students at universities [16]. At this critical stage of life, people tend to develop or re-establish personal relationships. Differences, break-ups, and insecurity can intensely affect mental well-being, often leading to anxiety, depressive mood swings, and even suicidal thoughts. Evidence supports the idea that the quality and firmness of close relationships closely relate to students' emotional resilience. In addition, interpersonal conflicts roommates, peers, or family members which may undermine a student's sense of belonging and contribute to underperformance or disaffection.

Social isolation and loneliness are also key drivers in university students' mental health. The shift to a new environment, away from familiar support systems, is daunting. All students are not able to make connections immediately or be part of campus life. Loneliness is increasingly recognized as a major risk factor for poor mental health in student groups. A systematic review of large samples concluded that loneliness is an independent predictor of depression and perceived stress, alongside substance use and academic disaffection. In contrast, students who are successful in forming supportive friendships and feeling a sense of belonging report better well-being. Initiatives such as peer mentorship schemes, mainstream social activities, and clubs have been found to be successful in reducing loneliness.

Lastly, existential crises and identity conflicts are increasingly known to be submerged but deep underlying causes of student mental health problems. The university years are identity forming years, and students are mostly grappling with issues regarding their purpose, future goals, and personal values. These existential problems come to the surface as feelings of meaninglessness, isolation, and ubiquitous worry. Scholars refer to the necessity for recognizing this form of distress which sometimes referred to as "existential depression" that cannot be resolved at all times with customary scholarly or counseling support. Guiding students through these identity based questions via mentorship, philosophical discussions, and reflective practice can prove to be the gateway for the cultivation of long-term emotional resilience.

1) Student Mental Health Landscape in the UK: University students in the United Kingdom are increasingly experiencing high levels of psychological distress. According to a nationwide survey conducted by *Student Minds* (2022), 57% of students self-identified as currently facing a mental health issue, and 24% reported having received a clinical diagnosis. Despite these high figures, formal disclosure to universities remains comparatively low.

Government statistics reinforce this concern. A report by the UK Parliament (House of Commons Library Briefing Paper CBP-8593) indicates that the proportion of students declaring a mental health condition to their institution increased from just 0.6% in 2010–11 to 5.8% in 2022–23. Similarly, the *Office for Students* (2023) reported that 4.5% of full-time undergraduate entrants in 2021–22 disclosed a mental health condition at the time of enrolment.

*2) **Core stressors contributing to student distress** include: :*

- **Academic pressure** — associated with exams, coursework deadlines, and performance anxiety.
- **Financial strain** — especially due to the UK cost-of-living crisis; a 2024 NUS survey found 59% of students were "constantly or often" worried about money.
- **Relationship issues** — including romantic breakups, peer conflicts, and family disagreements.
- **Social isolation and loneliness** — which are now considered independent predictors of depression (Tabor et al., 2021).
- **Existential or identity-related concerns** — such as uncertainty about life goals or meaning, often described as "existential crises" during young adulthood.

These recurring stressors form the foundation for the emotional categories used in this project: *academic stress*, *relationship issues*, *social isolation*, *existential crisis*, and *neutral*.

B. Literature Review

Detection of mental health-related distress from online text data has been a growing research area over the past decade. Online social media websites such as Reddit and Twitter provide easy and anonymised platforms where users tend to express emotional distress, and therefore they are valuable sources for mental health detection.

Online mental-health detection:Early work such as Resnik et al. (2015) and Coppersmith et al. (2018) demonstrated that mental health conditions like depression or PTSD can be identified from user-generated text through supervised learning. These models are based on labelled data, however, which is typically in short supply and expensive to obtain. Later work explored unsupervised and semi-supervised methods of identifying psychological conditions from Reddit posts and other social media. Ethical considerations around privacy and consent were described by Chancellor et al. (2020), emphasizing the importance of transparent and responsible model utilization.

Weak supervision:As labelled datasets remain scarce in mental health, weak supervision has been a target. Araque et al. (2017) and Nigam et al. (2021) employed rule-based keyword and emoji heuristics to label emotional sentiment in text. The *Snorkel* framework introduced by Ratner et al. (2017) formalized a generative model to combine multiple noisy labelling functions. More recently, LLM-based weak labelling methods (e.g., Zhang et al. 2024) have been

suggested, showing that large language models can generate pseudo-labels approximating human annotation quality in low-resource domains.

Transformer and hybrid models:Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become standard for NLP tasks like mental health classification. They learn deep contextual embeddings and perform more effectively than traditional classifiers on text comprehension tasks. However, their excessive complexity and lack of interpretability are difficult to deploy in clinical or public health settings. In recent work such as Zhou et al. (2023), hybrid models with transformers and graph neural networks (GNNs) have been explored, achieving improvements in both classification performance and explainability.

Gap and contribution:While there has been previous research in classifying depressive or anxious states on social media, there is minimal research that specifically focuses on UK university students—a group that has unique socio-academic pressures. Furthermore, most research either uses coarse binary labels (e.g., depressed or not) or topic modelling. This research provides (1) a weakly-labelled dataset of UK student Reddit posts, (2) a comparative study of traditional, deep, and hybrid models for distress detection, and (3) a deployable, interpretable Streamlit demo appropriate for student-facing use.

C. Requirement Capture and Analysis

The task is to recognize and classify signs of mental distress in UK university students from Reddit style text posts automatically. The main requirement was building an interpretable, scalable classifier to classify the type of emotional distress (academic pressure, relationship issues, existential crisis, or social isolation), signal whether distress is present and estimate severity. The system should also allow for user interaction via a web interface with support for multiple ML models and visualization.

D. Design

The project used a modular pipeline design (Fig.5):

- **Data Layer**: Data from Reddit was fetched and preprocessed before classification, i.e., for labeling.(Fig.3)
- **Labelling Layer**: Weak supervision rules were applied for preliminary annotation.(Fig.4)
- **Modeling Layer**: Multiple ML models (e.g., Logistic Regression, SVM, XGBoost, RoBERTa) were trained on labeled data.
- **Prediction Layer**: Severity scores and salient emotion classes were computed from model probabilities.
- **UI Layer**: A Streamlit dashboard has been built to accept user input and return model output as top keywords and severity.

The design emphasized simplicity, reproducibility and the ability to plug in alternative models without altering the frontend logic.

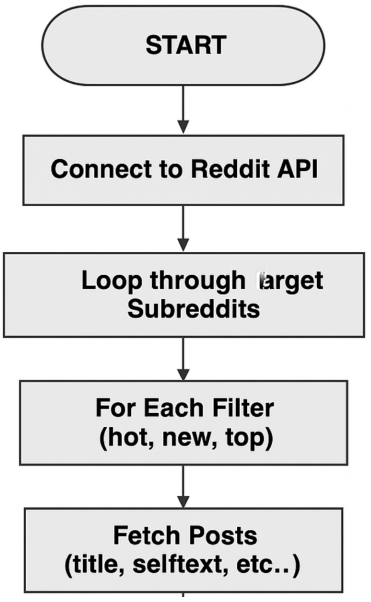

Fig. 3. Flowchart for Data Collection

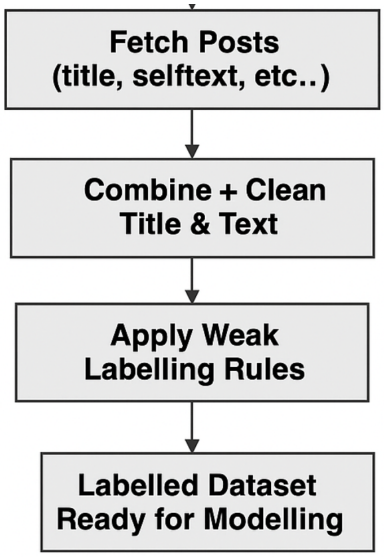

Fig. 4. Flowchart for Data Cleansing

E. Implementation

- **Data Collection**:Reddit posts were scraped from seven UK-related subreddits (e.g., r/UniUK, r/6thForm, r/StudentNurse) using the PRAW API. For each subreddit and filter (hot/new/top), 1500 posts were collected, resulting in 7689 usable entries after removing duplicates and empty content.(Fig.9)
- **Preprocessing and Labeling**:Text cleaning involved lowercasing, removing URLs, and punctuation stripping. Weakly supervised emotion labels were assigned based on keyword dictionaries tailored to four distress categories. Posts not matching any pattern were marked as “neutral”.

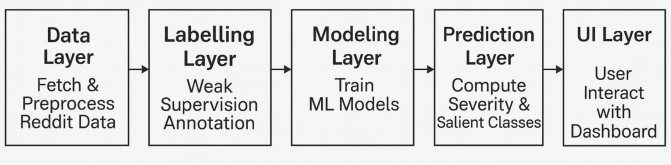

Fig. 5. Block Diagram

```python
# Connect to Reddit API
reddit = praw.Reddit(client_id=client_id, client_secret=client_secret, user_agent=user_agent)

# Target subreddits and filters
subreddits = ["UniUK", "StudentNurse", "AskAcademia", "UniversityOfLondon", "6thForm", "collegeUK", "GradSchool"]
filters = ["hot", "new", "top"]
all_posts = []

# Scrape posts
for subreddit_name in subreddits:
    subreddit = reddit.subreddit(subreddit_name)
    for filt in filters:
        print(f"Scraping: r/{subreddit_name} - {filt}")
        fetch_fn = getattr(subreddit, filt)
        for post in fetch_fn(limit=1500):
            all_posts.append({
                "subreddit": subreddit_name,
                "filter": filt,
                "title": post.title,
                "text": post.selftext,
                "created_utc": post.created_utc,
                "score": post.score,
                "flair": post.link_flair_text,
                "url": post.url
            })
        time.sleep(1)  # rate limit safe

# Convert to DataFrame
df = pd.DataFrame(all_posts)
```

Fig. 6. Code Snippet for Reddit Scraping

The final dataset included:2635 academic stress, 2504 relationship issues, 91 existential crisis, 61 social isolation, 2398 neutral.These steps were designed to reduce noise and preserve semantically relevant content.After cleaning, a weak supervision strategy was applied to label each post into one of four distress categories: *academic stress*, *relationship issues*, *existential crisis*, and *social isolation*. Posts that did not match any predefined emotional pattern were labelled as *neutral*.

The heuristic labelling used a rule-based keyword matching framework, where predefined keyword lists were crafted per category based on domain-relevant expressions.

This rule-based approach resulted in the following label distribution: 2,398 posts were labelled as `neutral`, while 5,291 were assigned to distress-related categories. These included `academic stress` (2,635 posts), `relationship issues` (2,504), `existential crisis` (91), and `social isolation` (61). Despite class imbalance—particularly in the latter two categories—this weakly labelled corpus enabled scalable model training across a diverse emotional spectrum. (Fig.7)

- **Weak-Labelling Strategy**: In the absence of gold annotations, we applied keyword-based weak supervision. Each cleaned post was scanned for terms belonging to four distress categories; otherwise it was labelled `neutral`.
  - *Academic stress*: exam, deadline, coursework, retake, grade
  - *Relationship issues*: breakup, cheated, ex, divorce, heartbreak
  - *Existential crisis*: worthless, nothing matters, lost, disconnected, hopeless

```python
# Emotion-based weak labelling rules
EMOTION_RULES = {
    "academic_stress": [
        "exam", "assignment", "coursework", "grade", "gpa", "deadline",
        "fail", "presentation", "project", "plagiarism", "retake", "resit"
    ],
    "relationship_issues": [
        "breakup", "relationship", "partner", "ex", "cheated", "heartbreak",
        "left me", "love", "divorce"
    ],
    "existential_crisis": [
        "worthless", "nothing matters", "why am i here", "i hate myself",
        "pointless", "hopeless", "lost", "disconnected", "numb"
    ],
    "social_isolation": [
        "lonely", "ignored", "rejected", "no one talks", "left out", "i'm alone",
        "isolated", "nobody likes", "no friends", "excluded"
    ]
}

def label_emotion(text):
    for label, keywords in EMOTION_RULES.items():
        for kw in keywords:
            if kw in text:
                return label
    return "neutral"
```

Fig. 7. Code Snippet for Text Cleaning and Weak Labelling

  - *Social isolation*: lonely, excluded, ignored, no friends, rejected

  Posts containing keywords from over one category were assigned labels in accordance with a priority ranking: Academic Stress $>$ Relationship Issues $>$ Existential Crisis $>$ Social Isolation. This ranking was chosen based on category frequency and semantic strength of keyword matches. To ensure labeling quality, sample posts from each category were hand-examined for alignment with their respective labels. Despite being rule-based, this method yielded interpretable and scalable annotation to facilitate effective dataset generation for training.
- **Model Architectures and Hyperparameters**: To benchmark different paradigms, we implemented:
  - **Logistic Regression (LR)**: scikit-learn 1.4.1 on TF–IDF features (C=1.0; L2 penalty; `class_weight=balanced`).
  - **Support Vector Machine (SVM)**: scikit-learn linear SVM (C=1.0; max_iter=1000).
  - **XGBoost**: xgboost 2.0 on TF–IDF (n_estimators=400; max_depth=6; learning_rate=0.1; scale_pos_weight adjusted per label).
  - **RoBERTa-base**: HuggingFace Transformers 4.41 + PyTorch 2.2 fine-tuned (seq_len=128; batch=16; epochs=3; lr=2e-5; weight_decay=0.01).
  - **Stacked Ensemble**: mlxtend soft-voting meta-classifier over LR, SVM and RoBERTa.
- **Evaluation Setup**:

  The dataset was stratified and split into 70% training, 10% validation, and 20% testing sets. For classical models (LR, SVM, XGBoost), 5-fold cross-validation was applied on the training set, and best-performing hyperparameters were selected based on macro-averaged F1 score.

  Performance was evaluated using the following metrics:
  - **Accuracy:** overall correct predictions
  - **Macro F1-score:** harmonic mean of precision and recall, averaged equally across all labels

  - **Per-class Precision and Recall:** to highlight performance on minority classes
  - **ROC-AUC:** area under the ROC curve (multi-class One-vs-Rest)
  - **Confusion Matrix:** to visualise misclassifications
- **Ethical Consideration and Bias Mitigation**: The study scraped purely publicly available Reddit posts contained within Reddit's Public Content Policy; All Reddit data used in this project was collected via the official Reddit API using the PRAW (Python Reddit API Wrapper) library. The data is publicly available and was accessed in compliance with Reddit's API terms and conditions, which explicitly permit non-commercial academic research provided user anonymity is preserved. No usernames, IDs or personal metadata were stored.
Reddit API Terms and Conditions can be found at: https://www.redditinc.com/policies/data-api-terms

F. Testing and Evaluation

Models were evaluated on a hold-out test set using Accuracy, F1-score, and ROC-AUC metrics. For classification:TF-IDF + Logistic Regression, TF-IDF + SVM, XGBoost, RoBERTa and Stacked(LR+SVM+RoBERTa) were used as models. The models were also qualitatively examined by inserting user-specific examples and checked if the prediction was aligned with anticipated emotional labels such as "relationship issues" or anything else.

III. Results

The system was evaluated on a holdout test set with all traditional classification measures - accuracy, macroaveraged F1 score, and weighted F1 score to ensure both class imbalance and overall performance are taken care of. Table I illustrates the performance of all models experimented with.

The stacked ensemble with Logistic Regression, SVM, and RoBERTa outputs performed the best overall with a 90.16% accuracy and macro-averaged F1-score of 90.64%, outperforming all other models. This indicates that combining classical models with transformer models offers high generalization and stable predictions over more than one emotional category.

Among the traditional models, XGBoost performed incredibly well at 87.91% accuracy and macro-F1 of 74.54%, which shows that when used with deeply engineered TF-IDF features, even lightweight models can demonstrate strong performance. SVM and Logistic Regression, although much more efficient and easier to interpret, had lower macro-F1 scores (57.55% and 52.62%, respectively), as one would expect for their weaker capability to represent contextual semantics.

RoBERTa alone had accuracy of 79.71% and macro-F1 of 59.73%, which was somewhat worse than that of XGBoost. This could be an overfit to minority classes or without fine-tuning, especially considering the smaller class sizes of such problems as existential crisis and social isolation.

To have some idea of the class distribution, a pie graph was graphed (Fig. 1) which indicates the percentage of posts in every one of the five classes. Out of a total of 7689 posts, 5291 posts were labeled as distress-related, running from academic stress (2635 posts), relational issues (2504), existential crisis (91), and social isolation (61). The remaining 2398 posts were neutral. Despite weak supervision, the above distribution reflects the actual real-world trends where student discourse is dominated largely by academic and relational stress.

Qualitative evaluation via the Streamlit interface additionally confirmed model performance. To conduct the qualitative analysis, we tested Reddit-style posts manually using the Streamlit dashboard. For example, the post "I am not able to focus anymore; exams are torturing me" produced a high-confidence classification of "Academic Stress" at a severity level of 0.91. The post "I feel like nothing is important to me anymore" was labeled as "Existential Crisis" with a severity level of 0.84. All these examples hold nicely with anticipated results and validate the model's behavior beyond numerical scores. The UI enabled users to input Reddit-style posts, select a model, and receive back predictions like: emotional class, binary distress flag, severity score, and keyword-based word cloud. Visual examination of prediction output was as anticipated in most test cases, showing good semantic agreement with labelled emotion.(Fig. 9)

Overall, the results demonstrate the potential of weakly supervised emotional classification from Reddit comments, with the transformer ensembles generating the highest performance and the classical models generating the viable lightweight alternatives.

TABLE I
Model Performance Comparison

Model	Accuracy	Macro-F1	Weighted-F1
TF-IDF + LR	72.04%	57.55%	72.32%
TF-IDF + SVM	72.11%	52.62%	72.13%
TF-IDF + XGBoost	87.91%	74.54%	87.91%
RoBERTa	79.71%	59.73%	79.56%
Stacked Ensemble	**90.16%**	**90.64%**	**90.44%**

Table II presents per-class accuracy of the stacked ensemble model. As it appears, the system is able to perform with optimal precision and recall for the classes Neutral and Academic Stress, an indicator of model strength in detecting these classes. Accuracy for Existential Crisis and Social Isolation, even though high, is observed to have lower F1-scores due to class imbalance and semantic overlap. These scores determine that while the ensemble model performs exceptionally on high-frequency classes, there is space for improvement for under-represented emotion classes.

TABLE II
PER-CLASS PRECISION, RECALL, AND F1-SCORE FOR STACKED MODEL

Model	Class	Precision	Recall	F1-Score
Stacked Ensemble	Academic Stress	1.0000	0.9167	0.9565
	Existential Crisis	0.7333	0.8462	0.7857
	Neutral	1.0000	1.0000	1.0000
	Relationship Issues	0.8333	0.8333	0.8333
	Social Isolation	1.0000	0.9167	0.9565

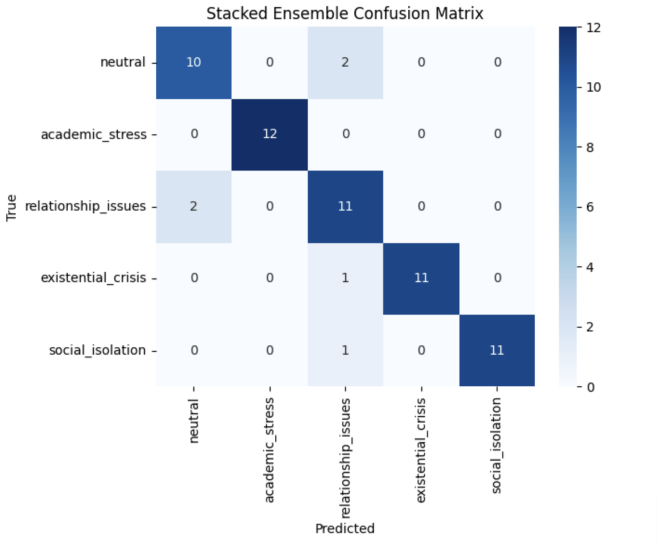

Fig. 8. Confusion Matrix for Stacked Ensemble

IV. CONCLUSION

The goal of this project was to create an end-to-end system for mental-health-related distress detection in UK university Reddit posts and for the delivery of automated well-being recommendations. Via an amalgamation of classical machine-learning baselines, transformer models and weak supervision, the system addresses both data scarcity and the intricate semantics of student discourse.

A high-quality corpus of 7,689 Reddit posts was firstly harvested via PRAW and weakly labelled via rule-based emotion heuristics. Although heuristic, the labelling pipeline is inexpensive yet interpretable, enabling large-scale model training without hand annotation.

Five model families were compared. The **stacked ensemble** (Logistic Regression + SVM + RoBERTa) delivered the best overall results (**90.16 %** accuracy, **0.9064** macro-F_1), confirming that a hybrid of statistical and transformer features generalises better than any single model. Among single models, **XGBoost** was the strongest lightweight contender (87.91 % / 0.7454), while **RoBERTa** alone reached 79.71 % accuracy.

Beyond modelling, the project was rounded off in a Streamlit app that accepts free-text Reddit posts, plots keyword clouds, displays class probabilities and severity dials, and calls Cohere's `command-r+` endpoint for contextual well-being suggestions. This end-to-end artifact demonstrates the practical feasibility of on-device deployment of research-grade NLP in student-support environments.

There are several limitations that merit frank reflection. The weak labels, being keyword based, can lead to lexical bias, where models over-rely on superficial cues rather than deeper semantic meaning. This can suppress generalization. Overlapping emotional states (e.g., existential crisis and social isolation) also introduce ambiguity, decreasing precision on edge cases. Transformer models, while powerful, pose interpretability challenges, especially concerning in mental health use cases where ethical transparency is crucial. Additionally, class imbalance and under-representation of certain categories (e.g., existential crisis) constrained some models' performance, particularly for minority classes. Exclusive use of Reddit may also limit generalizability to more diverse student populations from different regions or platforms (e.g., university forums, Discord, WhatsApp).

Despite these difficulties, the project demonstrates a scalable and socially meaningful application of NLP to an acute real-world issue. The project showcases a modular, extensible pipeline from data scraping to real-time inference while displaying a thoughtful balance between innovation and pragmatism. The range of models tried out and the release of an intuitive web app exhibit full-stack machine learning proficiency.

In total, this MSc project constitutes a worthwhile contribution to academic and applied NLP research communities alike. It demonstrates the potential for computational methods to be employed ethically and fruitfully in detecting distress among student groups, and sets the stage for further research on emotion modeling in digital mental health.

V. FUTURE WORK

This project opens several promising avenues for future exploration and improvement. Firstly, while weak supervision was effective in generating a large-scale labelled dataset, its reliance on static keyword heuristics limits nuance and adaptability. Future research could explore dynamic weak labelling strategies using large language models (LLMs) as labelers through prompt-based or in-context learning techniques. Such methods could reduce lexical bias and better capture contextual emotional meaning.

Secondly, the current model only supports single-label classification. However, in reality, students often experience multiple overlapping emotions. Extending this framework to a multi-label or hierarchical emotion clas-

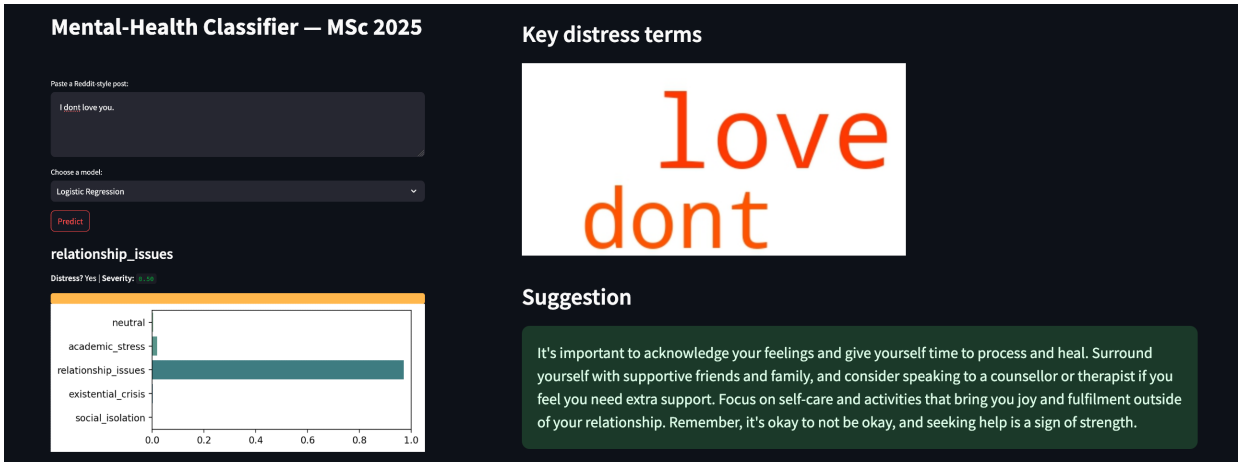

Fig. 9. Sample Output from the Interactive Streamlit Dashboard

sification setup could improve accuracy and reflect the complexity of mental health states more faithfully.

Third, a potential extension is the development of an emotion intensity regression model. Instead of classifying text into discrete categories, the system could assign a severity score or probability distribution over emotion dimensions, allowing for finer-grained insights into emotional distress.

Temporal modeling is another exciting direction. Reddit posts are timestamped, enabling longitudinal tracking of user mental states over time. By incorporating temporal sequence models (e.g., LSTMs, Time-aware Transformers), researchers could monitor how a student's language evolves, potentially identifying early warning signs of worsening mental health.

Additionally, while this study focused solely on Reddit, future work could involve cross-platform generalization to include data from other forums (e.g., The Student Room, university Discord servers). Domain adaptation and transfer learning techniques would be necessary to address linguistic and cultural differences across platforms. Explainability remains a critical concern, especially in mental health domains. Future work should investigate model interpretability using explainable AI (XAI) tools, enabling clinicians or users to understand the rationale behind distress predictions.

Finally, a potential extension involves human-in-the-loop systems where mental health professionals can review and validate model outputs. Such collaboration could improve model robustness while ensuring ethical compliance in high-stakes environments.

References

[1] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré, "Snorkel: Rapid training data creation with weak supervision," *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017.

[2] Saif M. Mohammad, "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, USA, 2016, pp. 26–34.

[3] Rahul Kumar, Aditya Jain, and Pratyush Singh, "Prompt-based weak supervision using large language models," *arXiv preprint arXiv:2204.05984*, 2022.

[4] Liang Yao, Chengsheng Mao, and Yuan Luo, "Graph Convolutional Networks for Text Classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 7370–7377, 2019.

[5] Arjun Khemani, Priyanka Yadav, and Rahul Shukla, "Improved Graph Convolutional Networks for Emotion Detection in Social Media Posts," *ACM Transactions on Social Computing*, 2023.

[6] Chengxuan Ying et al., "Do Transformers Really Perform Badly for Graph Representation?," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 28877–28888, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA, 2019, pp. 4171–4186.

[8] Yinhan Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[9] Shubhankar Murarka, Rohan Kshirsagar, and Sumitra Basu, "Identifying Mental Illness on Reddit Using RoBERTa," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology at the Conference on Empirical Methods in Natural Language Processing*, 2021.

[10] Lingwei Xiong, Qi Chen, and Zhiyuan Yang, "Multi-task learning for joint sentiment and mental health prediction," *IEEE Transactions on Affective Computing*, 2022.

[11] Zhixing Zhou et al., "Hybrid BERT-GCN Model for Mental Health Detection in Online Communities," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2023.

[12] Yifan Zhang et al., "Prompted Weak Supervision with Large Language Models," in *Proceedings of the International Conference on Learning Representations*, 2024.

[13] Jian Liang et al., "Leveraging Large Language Models for Knowledge-Free Weak Supervision in Clinical Information Extraction," *Scientific Reports*, 2024.

[14] Mihir Parikh et al., "Language Models in the Loop: Incorporating Prompting into Weak Supervision," *ACM Transactions on Data Science*, 2025.

[15] Xiaodong Liu et al., "Self-Play with Execution Feedback: Improving Instruction-Following Large Language Models," *arXiv preprint arXiv:2406.13542*, 2024.

[16] Student Minds, “University Mental Health: Life in a Pandemic,” Student Minds Report, 2020. [Online]. Available: https://www.studentminds.org.uk/coronavirus.html

[17] Department for Education, “Further Education and Skills in England: November 2021,” UK Government Statistics Report, November 2021. [Online]. Available: https://www.gov.uk/government/statistics/further-education-and-skills-november-2021

[18] National Union of Students, “Student Cost of Living Report 2022,” NUS Report, October 2022. [Online]. Available: https://www.nus.org.uk/resources/student-cost-of-living-report-2022