University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

Mental Health Signal Detection from Reddit Support Groups Using NLP and Transformers

Sakthivel Vinayagam

2400589

Supervisor: **Dr. Alex Diana**

September 3, 2025
Colchester

Abstract

Mental health concerns are commonly expressed online, and support communities such as Reddit have very active mental health support subreddits. The forums represent a somewhat anonymous space where people can anonymously discuss emotional experiences openly and thus might provide a fertile ground for gaining insights into mental well-being. But user-generated text is informal and diverse, and more difficulty exists in the automatic analysis. In this dissertation, we investigate the psychological risk level classification in Reddit posts with NLP techniques and deep learning model on transformer architecture.

We collected a dataset of more than 18,000 posts from twenty subreddits about mental health. After pre-processing of text cleaning, tokenisation and lemmatisation, emotional labels were added through a pre-trained emotional detection model. These 3 levels of highest, middle and lowest risks were then mapped to the risk levels. A diversity of models were compared, including a logistic regression baseline through an adapted DistilBERT model including mean-max pooling, projection layers, and a multi-sample dropout head.

The top-ranking model showed strong generalisation in cross-validation and on final test evaluations, as well as using class-wise F1 score, precision, and recall metrics. More interpretability is derived by confusion matrix, misclassification trend, and class-wise prediction analysis. The trained model was also tested with new Reddit posts to evaluate its inference in the real world.

Although the results demonstrate the utility of transformer-based models in achieving fine-grained level risk detection, there is room for improvement of robustness and interpretability. Further improvements could see the system integrated into a real-time API pipeline to allow real-time monitoring and wider application in online platforms.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Mental health has become one of the most pressing public health concerns of our time. With increasing awareness and reduced stigma, more people are beginning to open up about their struggles, yet millions still go unheard. Approximately one in eight individuals globally live with a mental health condition [1]. These issues can range from mild anxiety to more severe concerns such as depression, bipolar disorder, or suicidal ideation. Many cases remain undetected due to social stigma, lack of access to care, or the deeply personal nature of the experience [2].

As our lives become more digitally connected, online platforms have become unexpected spaces where people share their feelings more openly. Reddit, in particular, has grown into a vast public support network [6]. Subreddits such as `r/depression`, `r/anxiety`, and `r/SuicideWatch` host countless conversations about personal mental health struggles. Users often turn to these forums in moments of vulnerability, offering raw and timely insights into their emotional states. While this presents a valuable opportunity for early intervention, the informal and emotionally complex nature of these posts makes them difficult to analyse using traditional methods [7, 16].

Natural Language Processing (NLP) offers a promising way forward. With the development of transformer-based models such as BERT and DistilBERT [18, 21], machines can better capture context, semantics, and affect in language. These models are well suited to pick up on the subtle cues in online writing that may point to psychological distress. However, real-world applications in mental health remain limited due to ethical concerns, domain shift, and data annotation challenges [16, 19, 9].

This dissertation contributes a practical and responsible approach to that gap. We develop and evaluate a transformer-based system that classifies Reddit posts into three mental health risk levels (low, medium, high). A dataset was collected from mental health–focused subreddits and underwent cleaning and annotation using a pretrained emotion classification model [77, 10]. Predicted emotions were mapped to risk categories via an emotion-to-risk scheme [43, 12]. A customised DistilBERT model was then fine-tuned with enhancements including mean–max pooling and multi-sample dropout to improve generalisation [24, 25]. The model was benchmarked against classical and transformer baselines, with performance assessed via cross-validation and a held-out test set, followed by error analysis and distributional diagnostics [76, 30].

1.1 Aim

The primary aim is to build a robust and interpretable NLP system that uses transformer models to identify varying levels of mental health risk in Reddit posts. By leveraging linguistic patterns in peer-support communities, the study seeks to detect early signals of psychological distress and to classify posts by risk level, laying a foundation for future tools supporting awareness, moderation, and early intervention.

1.2 Objectives

- Collect Reddit posts from selected mental health subreddits using an ethical, reproducible pipeline.
- Preprocess text data through cleaning, lemmatisation, and careful retention of psychologically salient tokens.
- Assign risk levels by mapping predicted emotions to low, medium, and high-risk categories [43, 12].
- Fine-tune a customised DistilBERT model with mean–max pooling and multi-sample dropout [24, 25].
- Compare performance with baselines, including logistic regression and standard transformer models.

- Evaluate using accuracy, macro-F1, confusion matrices, and cross-validation [76, 30].
- Explore deployment considerations for API-based early detection tools.

1.3 Research Questions

- How effectively can transformer-based models classify Reddit posts into low, medium, and high mental health risk categories?
- What impact do architectural and training enhancements (e.g., pooling strategies, multi-sample dropout, label smoothing, class-weighted loss) have on performance?
- What insights arise from misclassified posts, and how can these inform model refinement and error mitigation?
- Can this framework be adapted for real-time use via API integration while meeting ethical and privacy requirements?

This dissertation balances technical innovation with the sensitivity of mental health data. While Reddit is the primary focus, the modular design supports future adaptation to other platforms. Although deployment is beyond the present scope, the findings provide a strong basis for future extensions, including API-driven moderation dashboards, alert systems, and digital well-being tools.

CHAPTER 2

Literature Review

2.1 Mental Health Detection from Online Support Forums

The growing use of digital platforms has altered how individuals discuss and seek support for mental health concerns. Online communities, particularly those that allow anonymity, have emerged as spaces where people feel safe to express distress, share experiences, and offer peer support. Reddit, with its subreddit-based structure, provides an environment where users discuss anxiety, depression, suicidal thoughts, and other psychological challenges, as noted by Gkotsis et al. [6]. Unlike formal clinical records, these forums offer candid reflections of emotional states that provide valuable signals for early detection of mental health risks.

Previous studies have explored Reddit's potential in understanding psychological well-being. Chancellor et al. [3] investigated mental health discourse on Reddit and found it rich in linguistic cues indicative of emotional distress, particularly in communities like `r/depression` and `r/SuicideWatch`. Similarly, Gkotsis et al. [6] developed models to analyse language patterns that may signal mental health deterioration, identifying changes in tone, vocabulary, and posting frequency as predictive features. These efforts highlighted the value of user-generated content for non-invasive mental health surveillance.

However, working with such data also presents challenges. The unstructured and

informal nature of posts, written in colloquial, sarcastic, or emotionally intense tones, makes traditional keyword-based or rule-based methods inadequate. The absence of ground-truth clinical labels complicates supervised learning, so researchers have proposed using proxy labels such as self-reported diagnoses, subreddit type, or linguistic signals (e.g., Yates et al. [7] and Low et al. [8]). In response to ethical concerns, there is emphasis on ensuring analyses respect user anonymity and data sensitivity; Benton et al. [9] stress ethical data usage and human-centred frameworks.

In sum, online forums like Reddit serve as a rich, if complex, data source for studying mental health expression. They provide a foundation for building early risk detection systems when coupled with advanced language models capable of interpreting subtle emotional signals embedded in user posts.

2.2 Emotion Classification and Risk Mapping

Understanding the emotional tone of online posts is central to identifying early signs of mental health risk. Emotion classification assigns labels—such as sadness, fear, anger, or joy—to text. In mental health research, detecting negative affective states can signal psychological distress and potential crisis. Several studies have employed pre-trained emotion classifiers for social media: Saravia et al. [10] introduced contextualised affect representations that performed well on benchmarks, while Baziotis et al. [11] demonstrated the value of attentive RNNs for fine-grained emotion detection.

To bridge the gap between emotion and mental health risk, work has introduced emotion-to-risk mapping schemes. For example, Shen et al. [12] proposed mappings to translate dominant emotions into broader risk levels (e.g., sadness/anger $\rightarrow$ high risk; joy/gratitude $\rightarrow$ low risk). In this project, a similar approach is adopted: a large Reddit dataset is labelled using a transformer-based emotion model (e.g., Savani's DistilBERT emotion model [77]), and top emotions are converted into low/medium/high risk categories [12]. While this indirect labelling strategy is not clinically validated, it provides a practical solution for building supervised models when expert-annotated risk data are scarce.

The reliability of such mappings is not without limitations. Emotions are subjective, and posts may express multiple, even conflicting, emotions. Despite these challenges,

emotion-to-risk conversion is useful for bootstrapping labelled datasets for downstream modelling.

2.3 NLP Preprocessing Techniques in Mental Health Text

Preprocessing is foundational in NLP pipelines and is especially important for Reddit data, which are informal, emotionally charged, and full of spelling variations, slang, emojis, and inconsistent grammar. Studies in mental health NLP emphasise the need for careful preprocessing to avoid losing psychologically salient signals. Cohan et al. [13] highlight abrupt tone shifts, fragmented sentences, and intense vocabulary in crisis forums such as `r/SuicideWatch`, motivating task-aware cleaning steps. Gkotsis et al. [6] likewise argue for preserving linguistic nuance while reducing noise.

Common steps include removing URLs, user mentions, and special characters; normalising case; and applying lemmatisation. Stopword removal must be selective because tokens such as "not" or "never" carry psychological weight. Tokenisation must also be compatible with the downstream model (e.g., WordPiece tokenisation for BERT-family models); in our setup, we align preprocessing with DistilBERT requirements as described by Sanh et al. [21]. Finally, researchers caution against overly aggressive cleaning that may erase risk-relevant markers, advocating task-aware preprocessing as argued by Chancellor et al. [5].

2.4 Machine Learning Models for Mental Health/NLP Tasks

Classical machine learning has provided strong baselines in automated mental health detection due to interpretability and efficiency. Logistic Regression (LR), Support Vector Machines (SVM), and Naive Bayes (NB) have been applied to classify psychological distress in social media. Tsakalidis et al. [14] used LR and SVM for mental health condition classification on Reddit, while Resnik et al. [15] applied NB in shared tasks on mental health forums. Yates et al. [7] combined TF–IDF features with tree-based models for suicide-risk signals in forums.

Despite practicality, classical models struggle to capture semantic nuance, irony, and

emotional undertones that are central to mental health discourse, as noted by Calvo et al. [16]. Comparative studies (e.g., Sadeque et al. [17]) show that neural architectures tend to outperform classical baselines on F1 and generalisation. These limitations have spurred a shift toward transformer-based methods.

2.5 Transformer Models in Mental Health Detection

Transformer architectures have markedly advanced NLP for context-rich, affect-laden text. Devlin et al. [18] introduced BERT, which leverages self-attention to model context bidirectionally. Applied in mental health settings, BERT-style models improve precision and recall over classical methods for depression and suicide-risk signals on Reddit and Twitter (e.g., Guntuku et al. [19] and Matero et al. [20]). Sanh et al. [21] proposed DistilBERT, which offers a practical speed–accuracy trade-off, retaining most of BERT's performance at lower compute cost, and is increasingly used in resource-limited or near–real-time settings (see also Zhang and Zhang [22]). RoBERTa variants have also reported gains for stress and anxiety detection (e.g., Ji et al. [23]).

To support transparency, recent efforts integrate post-hoc explainability (e.g., LIME by Ribeiro et al. [59] and SHAP by Lundberg and Lee [60]) and attention visualisation for transformer-based classifiers in this domain.

2.6 Model Improvement Strategies

We adopt several strategies tailored to emotionally complex, imbalanced Reddit data. First, we replace sole reliance on the `[CLS]` token with mean–max pooling to aggregate token representations, a choice inspired by sentence-embedding work showing robust pooling benefits (Reimers and Gurevych [24]). Second, we employ Multi-Sample Dropout (MSD) to reduce overfitting by averaging predictions across multiple dropout masks during training (Inoue [25]). We further apply layer-wise learning-rate decay (LLRD) to preserve general linguistic knowledge in lower layers while adapting higher layers to domain specifics (Howard and Ruder [47]). Evaluation stability is improved with an exponential moving average (EMA) of weights (Yuan et al. [27]), and label smoothing regularises the predictive distribution under ambiguous mappings (Müller et al. [28]). Early stopping and stratified K-fold cross-validation complete the training

protocol (details in Chapter 4).

2.7 Evaluation, Visualisation, and Explainability

Given the multi-class, imbalanced nature of risk labels, accuracy alone is insufficient. We therefore report macro-F1, class-wise precision/recall, and confusion matrices, following clinical NLP guidance on imbalanced evaluation (Zhou et al. [30]; Kshirsagar et al. [31]). Cross-validation with stratified folds improves estimate stability in skewed settings (Bennett and Hauser [29]). Prior work also cautions that overlooking imbalance can inflate metrics without real utility (Chancellor and De Choudhury [4]).

Beyond metrics, visual diagnostics (confusion matrices; predicted-distribution plots) reveal systematic biases, such as underprediction of high-risk cases. For interpretability, we rely on established XAI approaches in mental health NLP to identify influential tokens and validate psychologically meaningful cues (e.g., Mohammadi et al. [32]; Nguyen et al. [33]). Finally, targeted error analysis of misclassified examples is essential for surfacing ambiguity and informing ethical deployment (Benton et al. [9]; Yates et al. [7]).

2.8 Gaps and Future Opportunities

Despite advances, practical deployment remains limited. Many studies remain offline prototypes rather than real-time systems integrated into user-facing tools (Chancellor and De Choudhury [4]; Guntuku et al. [19]). Emotion classification is often used as a proxy for risk, but affect does not always equate to clinical severity; hybrid labelling with behavioural markers (e.g., posting frequency, crisis terms) can help (Wolohan [34]). Explainability and clinical validation are underexplored; closer alignment with psycholinguistics and expert-in-the-loop evaluations would strengthen trustworthiness (Gjurković and Šnajder [35]). From a technical perspective, multimodal approaches that fuse text with audio/visual/metadata signals are promising for ambiguous cases (Zadeh et al. [36]). Lastly, API-based deployment (discussed in Chapter 7) can enable moderation dashboards and alert systems while enforcing ethical safeguards.

CHAPTER 3

Data Collection and Preprocessing

This chapter details the end-to-end pipeline used in this study: selection of mental-health subreddits on Reddit; compliant data extraction via PRAW; ethical safeguards and platform adherence; exploratory data analysis (EDA) to characterise class balance, subreddit coverage, post length, temporal patterns, and emotion confidence; a task-aware text cleaning and preprocessing workflow that preserves affective cues; automatic emotion labelling with a pretrained DistilBERT model; and a conservative mapping from emotions to actionable risk levels (low/medium/high) for supervised learning.

3.1 Reddit Dataset and Subreddit Selection

Reddit has emerged as an influential platform for mental health expression and peer-based support, with large pseudonymous communities discussing emotional struggles, psychiatric symptoms, and coping experiences. Its architecture, structured into topic-specific "subreddits", enables focused discussion while affording users a degree of anonymity. These characteristics make Reddit a valuable source for identifying linguistic and emotional markers of psychological distress [3, 6].

Reddit was selected as the primary data source for three reasons: first, its extensive repository of user-generated content related to mental health; second, the availability of stable public access through application programming interfaces (APIs); and third, alignment with prior research protocols for mental-health signal detection from social media [9, 7, 8]. Building on these studies, we curated a set of active and thematically

focused subreddits in which first-person narratives and support-seeking posts are common. Forums dominated by memes, news reposts, or off-topic content were excluded in order to minimise noise and maximise data quality [7, 4].

The final list of twenty subreddits spans a range of conditions and support contexts, including depression, anxiety, crisis support, post-traumatic stress disorder, bipolar disorder, self-harm, and general mental health. Selection emphasised posting frequency, community relevance, and the prevalence of genuine self-disclosure. The subreddits used in this study are listed in Table 3.1.

No.	Subreddit	Focus area
1	r/depression	General depression-related discussions
2	r/anxiety	Anxiety experiences and coping strategies
3	r/mentalhealth	Broad mental health topics
4	r/SuicideWatch	Crisis support and suicidal ideation
5	r/BPD	Borderline Personality Disorder (BPD)
6	r/OCD	Obsessive–Compulsive Disorder (OCD)
7	r/ADHD	Attention Deficit Hyperactivity Disorder
8	r/bipolarreddit	Bipolar disorder discussions
9	r/PTSD	Post–Traumatic Stress Disorder
10	r/Anxietyhelp	Peer advice for managing anxiety
11	r/socialanxiety	Social-anxiety-related challenges
12	r/aspergers	Autism spectrum and Asperger's discussions
13	r/schizophrenia	Psychosis and schizophrenia support
14	r/mentalillness	General mental illness conversations
15	r/selfharm	Discussions around self-injury
16	r/depressionregimens	Lifestyle strategies to manage depression
17	r/CPTSD	Complex post–traumatic stress disorder support
18	r/lonely	Feelings of loneliness and isolation
19	r/DecidingToBeBetter	Self-improvement for mental wellbeing
20	r/KindVoice	Supportive community with kind responses

Table 3.1: Selected mental-health-related subreddits used in this study.

Data were collected programmatically through the Reddit API using the Python Reddit API Wrapper (PRAW; see Section 3.2). All material was sourced from publicly accessible threads. Personally identifiable information, such as usernames or direct profile links, was excluded. Platform policies and ethical norms for social data research were observed throughout (see Section 3.3 for procedures; also [4, 9])

3.2 Data Extraction via Reddit API (PRAW)

We accessed Reddit programmatically using the Python Reddit API Wrapper (PRAW) [37]. Authentication used a `script` application with credentials (client ID, client secret, username, password) stored outside the repository in a restricted JSON file on Google Drive and loaded at runtime, avoiding plaintext exposure and aligning with credential-management best practices [38]. The client was initialised in read-only mode with a descriptive `user_agent` and minimal scopes in accordance with Reddit developer guidance [39]. Operational hygiene comprised (i) off-repository secret storage, (ii) runtime injection (no hard-coded literals), (iii) least-privilege access, (iv) explicit identification via `user_agent`, and (v) secret hygiene/rotation where warranted.

Twenty mental-health-related subreddits (Table 3.1) were pre-selected based on activity, size, and relevance to psychological discourse. For each subreddit we retrieved up to 2,000 most recent submissions via `.new(limit=2000)`, spacing requests to respect rate limits [39]. Posts were retained only when they contained substantive selftext; entries marked `[removed]`, `[deleted]`, or with empty `selftext` were excluded. For each retained post we extracted title, body, score, `created_utc`, and `num_comments`, converting timestamps to ISO 8601 (UTC) for consistent temporal analysis.

Captured fields are listed in Table 3.2 (subreddit, title, body, `created_utc`, score, `num_comments`, `id`, `url`). The resulting dataset comprised just over 18,000 posts and was exported to CSV to support reproducibility and integration with the preprocessing pipeline. Figure 3.1 shows a representative data view (first five records; non-code). Personally identifiable information (e.g., usernames, profile links, comment histories) was not retained; collection was confined to publicly accessible content under read-only access. Rate limiting and broader ethical safeguards are detailed in Section 3.3.

API-based collection imposes known constraints: limited historical depth per subreddit (typically the most recent ∼1,000–2,000 posts), removal of non-text posts where `selftext` is empty, and higher rates of deleted/removed content in sensitive communities (e.g., `r/SuicideWatch`), which reduce available yield despite high relevance. Nevertheless, the final corpus exceeded 18k posts and was adequate for multi-class modelling when paired with the emotion-to-risk labelling strategy described later.

Table 3.2: Metadata schema for each Reddit post.

Field	**Description**
subreddit	Community where the post was published
title	Title of the post
body	Full selftext content
created_utc	ISO 8601 formatted UTC timestamp
score	Net upvotes (upvotes minus downvotes)
num_comments	Number of comments on the post
id	Unique Reddit post identifier
url	Direct URL to the Reddit post

```
df.head(5)
```

	subreddit	title	body	created_utc	score	num_comments	id	url
0	depression	i had a dream i was loved...	a few nights ago I had a dream i was actually ...	2025-08-23T15:05:02Z	1	0	1my3tc7	https://www.reddit.com/r/depression/comments/1...
1	depression	Miserable fucking life I live	I don't know if I'm gonna win this fight with ...	2025-08-23T15:04:50Z	1	0	1my3t4s	https://www.reddit.com/r/depression/comments/1...
2	depression	Tired of living life, I'm done	I quit. \n\nI'm done living a life that I don'...	2025-08-23T15:00:43Z	3	1	1my3pb5	https://www.reddit.com/r/depression/comments/1...
3	depression	I have no-one.	My mother is a narcissist and my brother is a ...	2025-08-23T14:42:36Z	3	1	1my39if	https://www.reddit.com/r/depression/comments/1...
4	depression	If I were loved I wouldn't be depressed	I know what everyone would say go to a therapi...	2025-08-23T14:42:25Z	4	1	1my39cn	https://www.reddit.com/r/depression/comments/1...

Figure 3.1: Example of extracted Reddit post data showing the first five records after collection. Fields include `subreddit`, `title`, `body`, `created_utc`, `score`, `num_comments`, `id`, and `url`.

3.3 Ethical Considerations in Mental Health Data Collection

We restricted collection to publicly accessible posts, excluding private communications and protected data, and we retained only textual content and non-sensitive metadata [40]. Reddit's forum-like structure and pseudonymity enable open discourse on mental health topics, which prior research has leveraged responsibly [3, 6]. All scripts operated in a non-invasive, read-only manner with authenticated access and rate-limit compliance [39, 37]. Given the emotional vulnerability often present in this content, we applied data minimisation and internal safeguards to reduce risks of misuse.

3.4 Exploratory Data Analysis (EDA)

To characterise the dataset prior to modelling, we summarise class balance, subreddit coverage, post length, temporal posting patterns, emotion confidence, and the most frequent emotions within high-risk posts.

Class balance: The risk labels (derived via the emotion–risk mapping in 3.7) show a moderate imbalance: *high* and *medium* are comparable in size, while *low* is smaller (Figure 3.2). This motivates the use of class-weighted loss and macro-averaged metrics in Chapter 4.

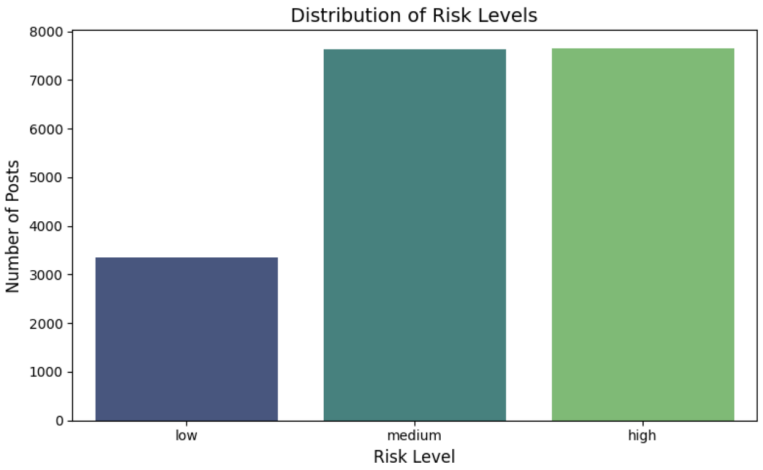

Figure 3.2: Risk level distribution across the corpus.

Coverage across subreddits: Collection limits (PRAW `.new`) yield near-uniform counts per subreddit (Figure 3.3). Risk composition varies by community (Figure 3.4); crisis-focused subreddits (e.g., `r/anxiety`) skew higher-risk, whereas general support forums show a larger share of medium-risk posts.

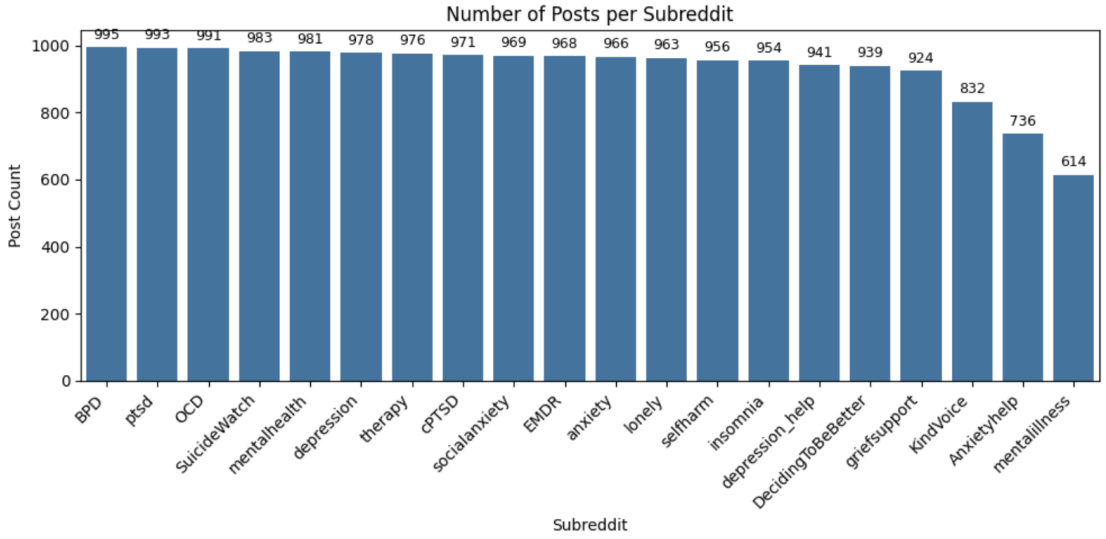

Figure 3.3: Posts per subreddit (collection cap ≈ 1,000).

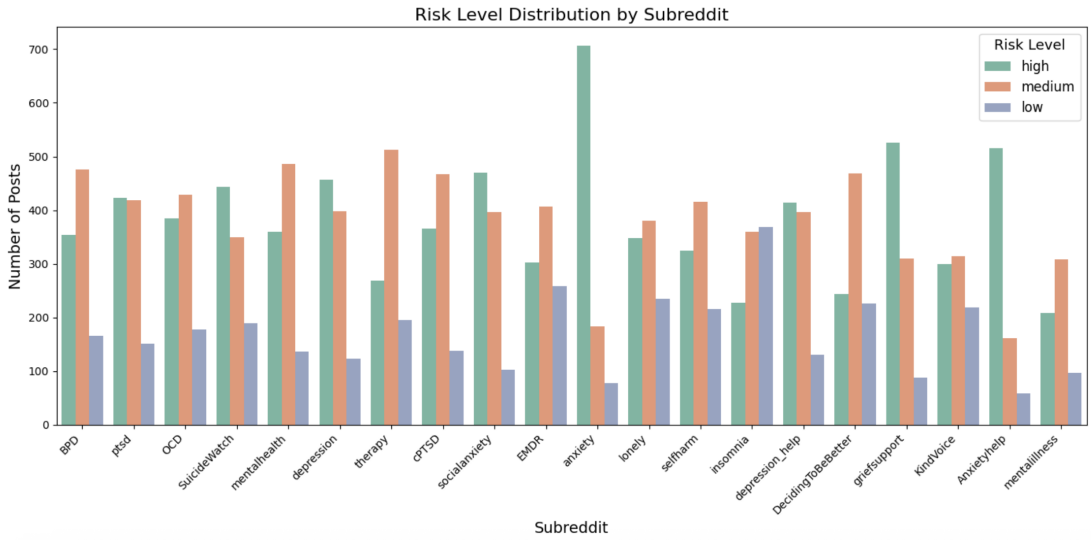

Figure 3.4: Risk level distribution by subreddit.

Post length and truncation: Token lengths are heavy-tailed with the 95$^{\text{th}}$ percentile ≈300 tokens. A truncation length of 512 (red line) comfortably retains content for $> 95\%$ of posts (Figure 3.5), supporting BERT-family sequence limits without excessive clipping.

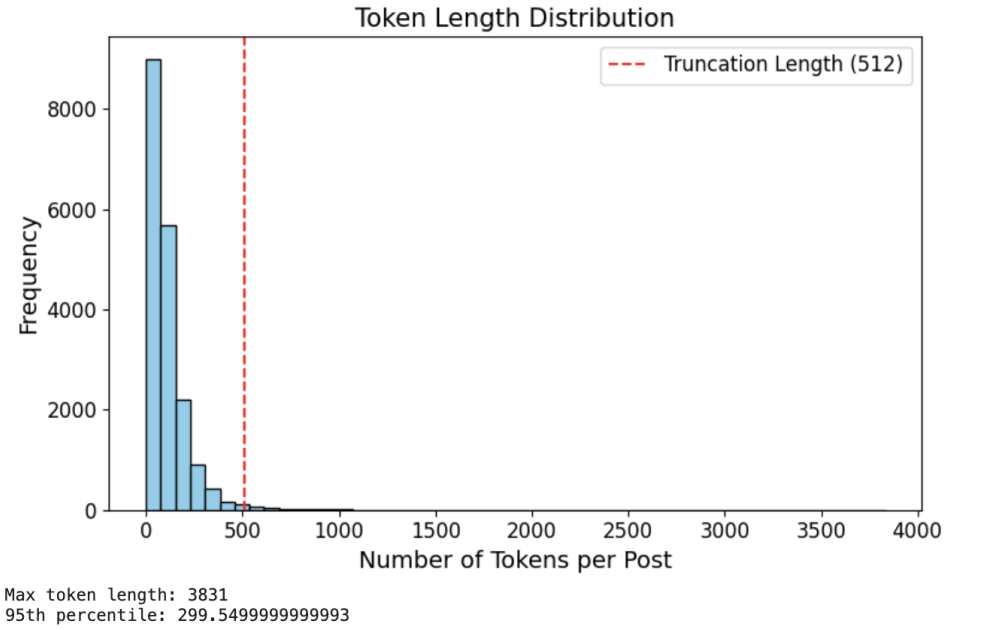

Figure 3.5: Token-length distribution with the 512-token truncation mark.

Temporal patterns: Posting activity varies by hour (UTC). Evenings and late night show higher counts across all risk levels, with a larger share of high-risk posts after 18:00 (Figure 3.6). This suggests circadian/availability effects that may be useful for downstream monitoring.

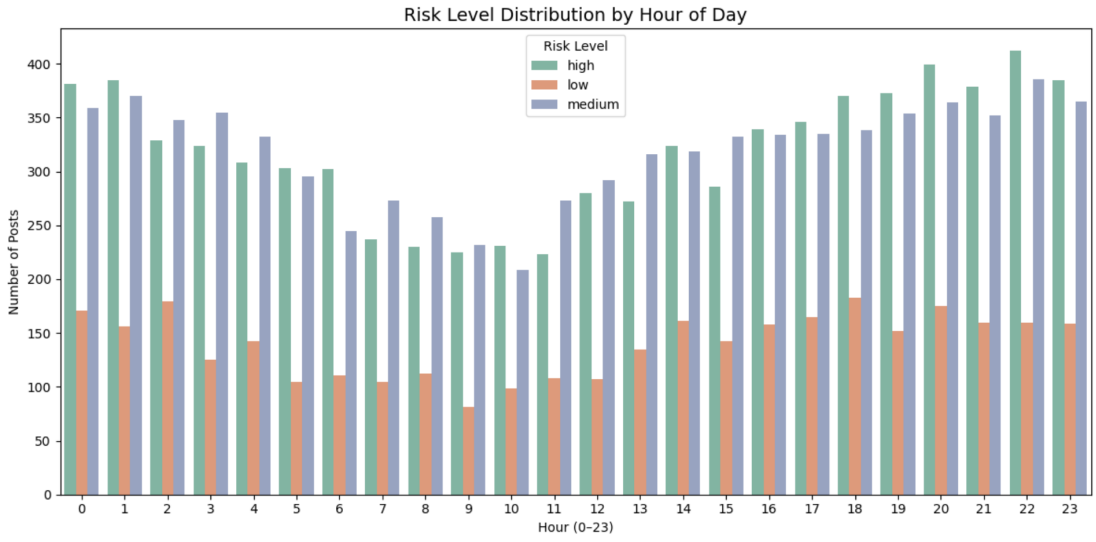

Figure 3.6: Risk level counts by hour of day (UTC).

Emotion confidence and high-risk drivers: Primary-emotion confidence scores are right-skewed, with most predictions between 0.08–0.25 and a long tail to ∼0.7 (Fig-

ure 3.7). Within *high-risk* posts, the most frequent emotions are *nervousness*, *sadness*, *disappointment*, and *fear* (Figure 3.8), aligning with the mapping in Table 3.4.

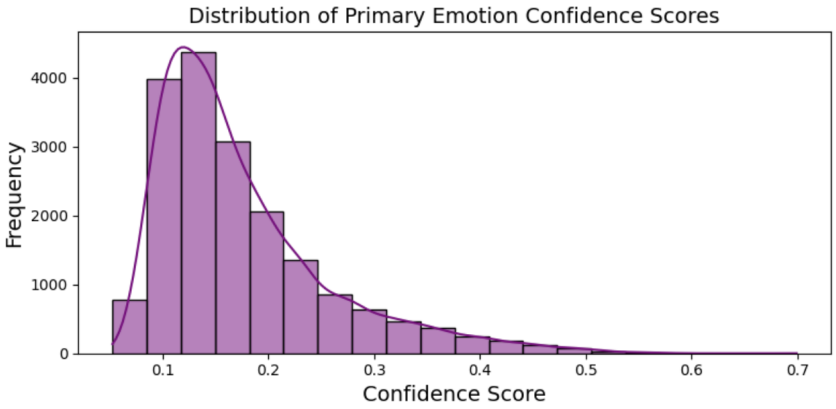

Figure 3.7: Primary-emotion confidence score distribution.

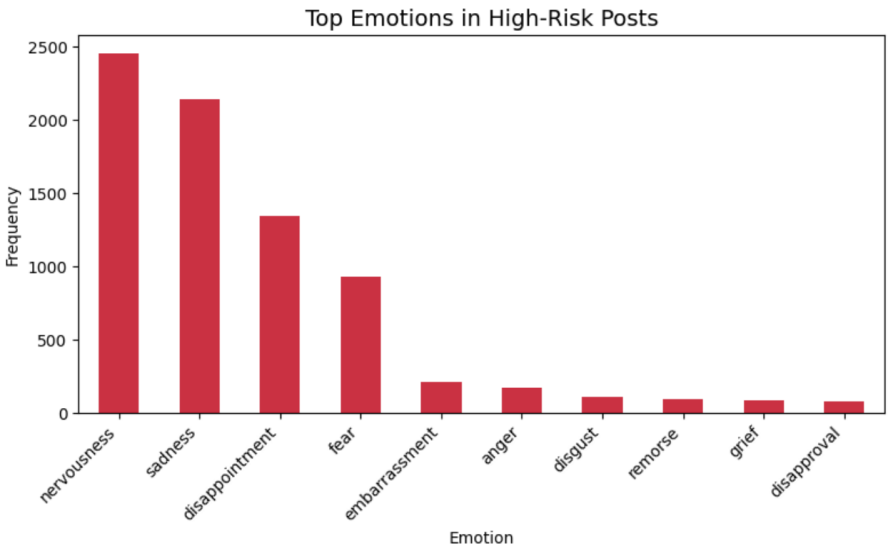

Figure 3.8: Top primary emotions among *high*-risk posts.

Takeaways:

- Class imbalance is present but manageable with cost-sensitive training.
- Subreddit context moderates risk composition across communities.
- Sequence lengths fit standard transformer limits (512 tokens).
- Evening/night periods show relatively more high-risk content.

- High-risk labels co-occur with fear-/sadness-adjacent emotions, supporting the emotion–risk mapping.

3.5 Text Cleaning and Preprocessing Pipeline

Reddit posts from mental health forums are often unstructured, emotionally nuanced, and highly variable in format. Unlike formal documents, these texts may include slang, typos, emojis, abbreviations, and non-standard grammar. For NLP-based classification models to operate effectively, it is essential to convert this raw, noisy text into a consistent and informative format. Preprocessing not only improves tokenisation and model input compatibility but also helps retain emotional and semantic cues critical to identifying psychological risk. In this project, we implemented a two-stage pipeline to clean, normalise, and prepare Reddit text for downstream risk classification using transformer-based models.

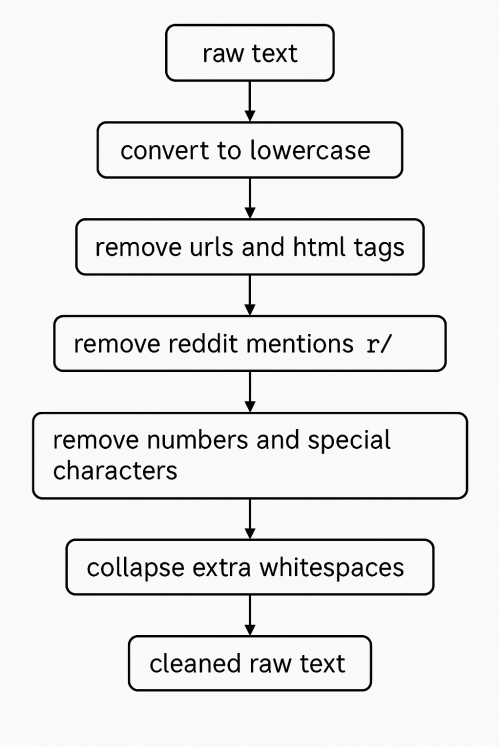

Figure 3.9: Stage One (general cleaning): normalise case/whitespace and remove URLs/HTML, Reddit mentions, and non-alphabetic artefacts to reduce sparsity and noise.

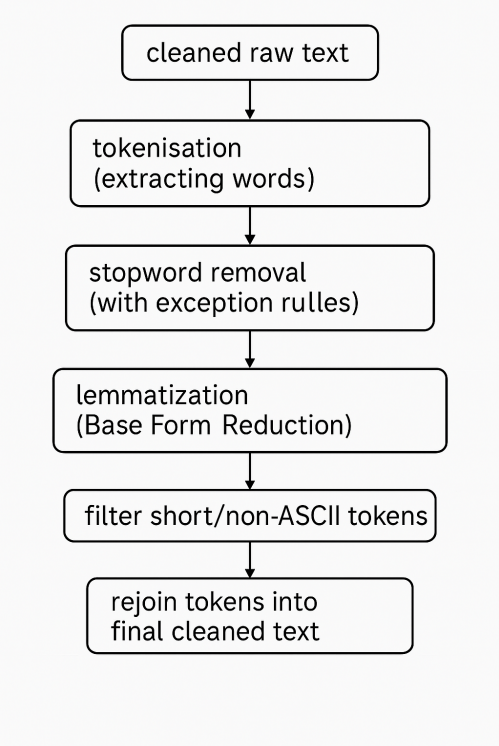

Figure 3.10: Stage Two (token-level processing): tokenise, selectively remove stopwords (keep negations), lemmatise, filter short/non-ASCII tokens, and reconstruct `clean_text`.

The workflow proceeds in two phases. First, a *general text cleaning* phase standardises surface form and removes extraneous artefacts (e.g., URLs/HTML, subreddit/user mentions, punctuation/digits, and inconsistent whitespace). Second, a *token-level processing* phase prepares model-ready inputs by tokenising, selectively removing stopwords (retaining key negations), lemmatising, filtering short/non-ASCII tokens, and rejoining tokens into a `clean_text` field. This design reduces noise while preserving affect-bearing cues (e.g., "hopeless", "overwhelmed") that are important for mental-health signal detection. The full set of operations is consolidated in Table 3.3; the logic of the two stages is visualised in Figures 3.9 and 3.10. To avoid redundancy, detailed step lists are not repeated in multiple subsections.

Table 3.3: Preprocessing steps, brief description, and tools used.

Step	**Description**	**Tool/Method**
Lowercasing	Convert all text to lowercase	`str.lower()`
URL/HTML Removal	Strip URLs and tags	`re.sub()`
Reddit Mention Removal	Remove `r/` and `u/` mentions	`re.sub()`
Special Character Removal	Remove numbers/punctuation	`re.sub()`
Whitespace Normalisation	Collapse extra spaces	`str.strip()`, `re.sub()`
Tokenisation	Split into words	`RegexpTokenizer()`
Stopword Removal	Remove common function words	NLTK stopword list
Lemmatisation	Convert to base word forms	`WordNetLemmatizer()`
Non-ASCII/Short Token Removal	Remove unwanted tokens	Python filters
Rejoin Tokens	Create final cleaned text	`' '.join()`

Implementation notes and downstream relevance. The pipeline was implemented in Python with regular expressions, NLTK, and `pandas`, and applied to just over 18,000 posts. Titles and bodies were concatenated prior to cleaning; posts that cleaned to too few tokens were discarded. The resulting `clean_text` supports transformer tokenisation (512-token limit) and preserves emotionally salient cues needed for risk classification. Empirically, this preparation improves macro-F1, precision, and recall in later models while avoiding over-aggressive sanitisation.

3.6 Emotion Label Assignment Using a Pretrained DistilBERT Model

Following text cleaning and normalisation, emotion labels were assigned as an intermediate supervisory signal prior to risk mapping. A pretrained DistilBERT emotion classifier was selected for this stage owing to its favourable efficiency–accuracy trade-off for large-scale inference on social media text [21]. Emotion classification offers finer granularity than polarity-based sentiment and more faithfully captures the nuanced disclosures typical of mental-health forums (e.g., hopelessness, anxiety, grief), making it a more informative basis for downstream risk assessment [16].

Concretely, the Hugging Face model `bhadresh-savani/distilbert-base-uncased-emotion` was employed to produce probability distributions over standard affective categories, including *joy*, *sadness*, *anger*, *fear*, *disgust*, and *surprise* [77]. DistilBERT retains most of BERT's language understanding capacity while reducing computational overhead, which is advantageous at the scale of ∼18k posts [21]. Each preprocessed post was passed once through the classifier and the top-two emotions (based on softmax scores) were recorded alongside the original metadata. Extremely short texts (after preprocessing) were excluded to avoid low-information predictions and reduce noise.

For each instance, we retain a *Primary Emotion* (highest-probability label) and a *Secondary Emotion* (second-highest), thereby accommodating the mixed or ambivalent affect often expressed in support contexts [16]. This dual-label representation proved useful for subsequent risk aggregation while preserving signal about emotional ambiguity.

This approach is not without limitations. The underlying model was fine-tuned on

general-purpose emotion datasets rather than mental-health–specific corpora; sarcasm, idiomatic usage, and limited label granularity can therefore reduce specificity in this domain. Despite these constraints, transformer-based emotion inference provided a stable and informative intermediate representation at scale.

Finally, the emotion outputs served as inputs to the risk mapping procedure (Section 3.7). Using a conservative top-two–emotion rule, posts were mapped to *low*, *medium*, or *high* risk tiers, enabling sensitivity to distress and potential crisis indicators without resorting to brittle keyword heuristics.

3.7 Risk Level Mapping Strategy (Low, Medium, High)

Emotion labels, while informative, are not directly actionable. For triage and monitoring, we therefore map emotions to interpretable risk levels (low/medium/high), following practices used in shared-task designs and crisis-annotation protocols [41, 42].

Mapping procedure: We grouped the top emotions predicted per post into three tiers using psychological valence, empirical severity reported in prior studies, and typical associations in mental-health contexts [43, 3]. Under a conservative decision rule, if either of the top-two emotions falls in the *high-risk* tier, the post is labelled *high*. Otherwise, if at least one is *medium-risk*, the post is labelled *medium*; remaining posts are labelled *low*. This hierarchy prioritises recall of potentially urgent cases, consistent with risk-aware triage goals in social-media mental-health research [41, 42].

Table 3.4: Emotion-to-risk mapping used in this study. The mapping guides label assignment (high → medium → low) under the top-two-emotions rule described in the text.

Risk Level	**Mapped Emotions**
High	sadness, fear, anger, grief, embarrassment, remorse, disapproval, disappointment, nervousness, disgust
Medium	confusion, annoyance, surprise, realization, neutral, caring
Low	joy, love, optimism, amusement, gratitude, admiration, pride, approval, curiosity, relief, excitement, desire

Practical considerations and limitations: The mapping is heuristic rather than diagnostic and omits user history and broader context; it also assumes a monotonic relationship between certain emotions and risk levels that may not hold in all cases [44]. Ambiguous emotions (e.g., *surprise*) are placed in *medium* to reflect uncertainty.

Relevance to model training: The resulting, preprocessed and risk-labelled dataset is used to train classifiers in Chapter 4. Emotion-informed labels provide a more nuanced supervisory signal than polarity sentiment or keyword triggers, improving sensitivity to distress [43, 41].

CHAPTER 4

Methodology

This section outlines the modelling strategies employed to classify mental health risk levels from Reddit support group posts. The methodology was designed to progressively evaluate multiple modelling paradigms—from a traditional logistic regression baseline to a series of transformer-based architectures, each incorporating increasingly sophisticated enhancements. Seven distinct models were implemented, including both pretrained reference heads and custom configurations involving advanced pooling techniques, multi-sample dropout, projection layers, and optimisation strategies such as learning rate scheduling and class-weighted loss functions. In addition, appropriate cross-validation and evaluation protocols were adopted to ensure generalisability and robustness. This structured pipeline facilitates a comprehensive comparison of model performance across low, medium, and high-risk mental health categories.

4.1 Overview of Models Used in Classification

To systematically assess the effectiveness of different modelling paradigms in classifying Reddit posts by mental health risk level (low, medium, high), this study implements and compares seven models of increasing complexity. These models span from a traditional machine learning baseline to highly customised transformer-based architectures. Each model was selected or designed to test specific hypotheses regarding feature representation, contextual embedding, and regularisation strategies, all while maintaining consistency in dataset splits and evaluation metrics.

The first model serves as a conventional baseline, offering a non-contextual reference point using manually engineered features. The remaining six are transformer-based models, which vary across dimensions such as architecture type (DistilBERT vs. BERT-base), pooling methods (CLS token, mean+max, attention), dropout regularisation (including multi-sample dropout), and loss adaptation strategies (e.g., label smoothing, class weighting, EMA). This gradual architectural enhancement enables a robust comparison between standard transformer heads and custom-designed alternatives, highlighting their relative performance across imbalanced mental health risk categories.

Each of the seven models is briefly described below, with further implementation details provided in Section 4.2 onward.

4.1.1 TF-IDF + Logistic Regression (Baseline)

This baseline model applies Term Frequency–Inverse Document Frequency (TF-IDF) vectorisation to the cleaned Reddit text, followed by classification using a logistic regression classifier. While it does not capture contextual semantics or word order, this model provides a useful benchmark for evaluating improvements gained through transformer-based methods. Its simplicity, interpretability, and low computational cost make it a standard baseline in mental health NLP literature [45].

4.1.2 DistilBERT — Reference Head (No Customisation)

This model uses the pretrained DistilBERT architecture with the default classification head provided by Hugging Face. The [CLS] token output from the final layer is passed directly into a linear layer for classification. No architectural modifications are applied. This model serves as a lightweight transformer baseline to evaluate performance under minimal configuration [21].

4.1.3 BERT-base — Reference Head (No Customisation)

Similar to the above, this model uses BERT-base with the default classification head. Its larger parameter size compared to DistilBERT allows assessment of whether increased representational capacity leads to improved mental health risk detection, despite identical pooling and head structure [18].

4.1.4 DistilBERT — CLS Pooling + Multi-Sample Dropout

This version retains the use of the [CLS] token for classification but incorporates Multi-Sample Dropout (MSD) as proposed by Inoue [25]. MSD applies multiple dropout masks and averages their predictions, reducing overfitting and improving generalisability in small or imbalanced datasets. This approach also provides a more robust regularisation mechanism without increasing model parameters.

4.1.5 DistilBERT — Attention Pooling + Multi-Sample Dropout

In this variant, CLS pooling is replaced with an attention-based pooling mechanism [46], which computes a weighted average of all hidden states using trainable attention scores. Combined with Multi-Sample Dropout, this configuration enables the model to focus dynamically on emotionally salient parts of each post, thereby capturing subtle mental health cues across longer sequences.

4.1.6 BERT-base — Mean+Max Pooling + Projection + MSD

This BERT-base variant applies both mean and max pooling across the token embeddings to capture global and local context simultaneously. These pooled vectors are concatenated and passed through a projection layer before classification. As in previous models, Multi-Sample Dropout is applied for robustness. This configuration allows the model to leverage complementary statistical summaries of the text rather than relying on a single token (e.g., CLS), addressing issues of information bottleneck.

4.1.7 DistilBERT — Mean+Max Pooling + Projection + MSD + EMA + LLRD + Label Smoothing + Class-Weighted Loss (Main Model)

This is the main model of the study and incorporates several advanced enhancements on top of the previous variant. Alongside Mean+Max Pooling, projection, and Multi-Sample Dropout, the model employs:

- **Exponential Moving Average (EMA)** to stabilise weight updates during training [27].

- **Layer-wise Learning Rate Decay (LLRD)** to train deeper layers at smaller learning rates, preserving pretrained knowledge while adapting high-level representations [26].

- **Label Smoothing** to reduce overconfidence in predictions [48].

- **Class-Weighted Loss** to address data imbalance, assigning higher penalties to misclassified minority classes [49].

Together, these strategies are designed to maximise performance on imbalanced mental health data by improving generalisability, stabilising training, and ensuring that the model is sensitive to nuanced emotional language present in Reddit posts.

4.2 Custom Architecture Components and Enhancements

To address the unique challenges of mental health risk classification from unstructured Reddit posts—including high class imbalance, subtle emotional cues, and limited annotated data—several custom architectural components and training enhancements were integrated into selected transformer-based models. These enhancements aim to improve classification accuracy, reduce overfitting, and preserve the integrity of emotional signals critical for mental health detection.

This section provides a detailed breakdown of each custom component applied across Models 4.1.4 to 4.1.7. These components were introduced incrementally to examine their individual and cumulative effects on performance.

4.2.1 Multi-Sample Dropout (MSD)

Multi-Sample Dropout, proposed by Inoue [25], involves applying multiple dropout masks during both training and inference. For each forward pass, several stochastic dropout outputs are generated and averaged, creating an ensemble-like effect without increasing the number of parameters. This approach reduces model variance and is particularly beneficial in imbalanced or small-sample settings, as commonly encountered in mental health datasets. MSD was applied in Models 4.1.4 through 4.1.7.

4.2.2 Pooling Strategies

Instead of relying solely on the [CLS] token—which may not always capture the full semantic content of lengthy Reddit posts—custom pooling strategies were introduced:

- **CLS Pooling** (used in 4.1.4) retains only the first token embedding for classification.
- **Attention Pooling** (used in 4.1.5) assigns learned weights to all token embeddings to focus on emotionally salient phrases [46].
- **Mean+Max Pooling** (used in 4.1.6 and 4.1.7) aggregates both average and peak activations across token embeddings, capturing both global and local information. These are then concatenated to form a comprehensive pooled representation.

4.2.3 Projection Layer

In models using pooled representations (4.1.6 and 4.1.7), a linear projection layer was introduced between the pooled output and the final classification head. This layer allows the network to re-weight and compress concatenated embeddings, thereby improving the downstream classification decision boundary while maintaining generalisation.

4.2.4 Exponential Moving Average (EMA)

Model 4.1.7 incorporates an Exponential Moving Average (EMA) of model weights throughout training. EMA smoothens parameter updates by maintaining a shadow copy of the model that updates slowly, based on a decay rate. During evaluation or checkpoint saving, predictions are made using the EMA weights. This improves stability and reduces the likelihood of overfitting or abrupt performance drops between epochs [27].

4.2.5 Layer-wise Learning Rate Decay (LLRD)

Layer-wise Learning Rate Decay is a transfer learning technique that applies progressively smaller learning rates to lower layers of the transformer model. Lower layers—closer to the input—are pretrained on general language understanding and should be

fine-tuned minimally, while higher layers are adapted to the task-specific data. In this study, LLRD was applied to Model 4.1.7 using a decay factor (e.g., 0.9 per layer), with the base learning rate applied to the classifier head and final transformer layers [26].

4.2.6 Label Smoothing

To prevent the model from becoming overly confident in its predictions—especially on noisy or overlapping class boundaries—Label Smoothing was incorporated into Model 4.1.7. Instead of assigning a hard 1 to the correct class and 0 to others, a small probability mass (e.g., 0.1) is distributed across all classes. This regularisation technique improves generalisation and calibration of probability scores [48].

4.2.7 Class-Weighted Loss Function

Mental health datasets tend to be skewed towards the low-risk class. To address this imbalance, a Class-Weighted Cross-Entropy Loss was used in Model 4.1.7. Weights were inversely proportional to class frequencies, encouraging better recall and F1 for minority classes [49].

4.2.8 Network Architecture Specifications

This subsection consolidates the tensor shapes and data flow for Models 4.1.2–4.1.7 (see pooling variants in Section 4.2.2 and the projection head in Section 4.2.3)..

Inputs and tokenisation: Each example is the concatenation of *title* and *body* (`clean_text`; Sec. 3.5), tokenised with the BERT WordPiece tokenizer (max sequence length L=512; right padding; attention mask where real tokens= 1, padding= 0). Special tokens `[CLS]` and `[SEP]` are included.

Transformer backbones: *DistilBERT*: 6 encoder layers, hidden size 768, 12 heads, intermediate size 3072, GELU, dropout 0.1; ∼66M parameters.
BERT-base: 12 layers, hidden size 768, 12 heads, intermediate size 3072, GELU, dropout 0.1; ∼110M parameters.
Both use token and position embeddings with layer normalisation; BERT-base also includes segment (token-type) embeddings.

Pooling variants and dimensions: (i) **CLS**: representation $\mathbf{h}_{\text{CLS}} \in \mathbb{R}^{768}$.
(ii) **Mean+Max**: $\mathbf{h}_{\text{mean}} = \frac{1}{L}\sum_{t=1}^{L} \mathbf{h}_t$, $\mathbf{h}_{\text{max}} = \max_t \mathbf{h}_t$, then $\mathbf{h} = [\mathbf{h}_{\text{mean}}; \mathbf{h}_{\text{max}}] \in \mathbb{R}^{1536}$.
(iii) **Attention** (single-head additive): $a_t = \text{softmax}(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{h}_t))$, $\mathbf{h} = \sum_t a_t \mathbf{h}_t \in \mathbb{R}^{768}$.

Projection and head: When pooling yields 1536-D (Mean+Max), a projection maps $1536 \to 768$ with GELU and dropout (0.1) before the classifier. The classifier is a single linear layer $768 \to 3$ logits; training minimises class-weighted cross-entropy on the raw logits (label smoothing for Model 4.1.7; see Sec. 4.2.6), with softmax applied only at inference.

Regularisation/optimisation placement: **MSD** (Sec. 4.2.1): $N{=}5$ parallel dropout masks ($p{=}0.3$) applied on the penultimate representation; logits averaged.
EMA (Sec. 4.2.4): decay 0.999 shadow weights used for eval/checkpoints.
LLRD (Sec. 4.2.5): per-layer factor 0.9 from top→bottom; head at base LR.
Label smoothing $\varepsilon{=}0.1$ and **class-weighted** cross-entropy at the loss.
Weight decay 0.01; gradient clipping at 1.0.

Per-model configuration:

- **4.1.1 TF–IDF + LR**: classical baseline; no transformer; class_weight=balanced.
- **4.1.2 DistilBERT Reference**: DistilBERT backbone; *CLS* pooling (768); no projection; vanilla classifier; no MSD/EMA/LLRD; class-weighted loss.
- **4.1.3 BERT-base Reference**: BERT-base backbone; *CLS* pooling (768); otherwise as 4.1.2.
- **4.1.4 DistilBERT CLS + MSD**: DistilBERT; *CLS* (768); MSD with $(N{=}5, p{=}0.3)$ before logits; class-weighted loss.
- **4.1.5 DistilBERT Attention + MSD**: DistilBERT; *attention* pooling (768); MSD $(5, 0.3)$; class-weighted loss.
- **4.1.6 BERT-base Mean+Max + MSD**: BERT-base; *Mean+Max* (1536) → projection to 768 (GELU+dropout) → classifier; MSD $(5, 0.3)$; class-weighted loss.

- **4.1.7 DistilBERT Main (Mean+Max + Proj + MSD + EMA + LLRD + LS + CW)**: DistilBERT; *Mean+Max* (1536) $\rightarrow$ projection to 768; MSD $(5, 0.3)$; EMA (0.999); LLRD $(0.9$ layer decay); label smoothing (ε=0.1); class-weighted loss.

End-to-end path (main model): `clean_text` $\rightarrow$ tokenizer (max 512, mask) $\rightarrow$ DistilBERT encoder ($L\times768$) $\rightarrow$ Mean+Max concat (1536) $\rightarrow$ projection (768) $\rightarrow$ MSD logits average $\rightarrow$ softmax $\rightarrow$ risk label.

4.3 Training Configuration and Hyperparameter Tuning

The training setup was designed for consistency, fairness, and replicability across all models. Experiments ran on Google Colab Pro with GPU acceleration (T4, L4, or A100 when available) using `transformers`, `datasets`, `scikit-learn`, `pandas`, `NumPy`, and `PyTorch`. To ensure reproducibility, we fixed random seeds (42) for NumPy, PyTorch, and model initialisation.

Optimiser and learning rates: For transformer models (Sections 4.1.2–4.1.7) we used AdamW [18]. Standard fine-tuning used a base learning rate of 2×10^{-5} (Models 4.1.2–4.1.4); advanced configurations (Models 4.1.5–4.1.7) used 1.5×10^{-5}–2×10^{-5} with scheduling and warm-up. For the main model (Model 4.1.7) we applied layer-wise learning-rate decay (LLRD) with decay factor 0.9, so lower layers updated more conservatively than higher layers and the classifier head [26]. The TF–IDF + Logistic Regression baseline (Section 4.1.1) used L2-regularised `LogisticRegression`, tuned on a single, fixed stratified validation split (see Section 4.4).

Batching, epochs, and stability: We used a batch size of 16 to balance memory and gradient stability. Transformer models trained for 5 epochs with early stopping if validation loss failed to improve for two consecutive epochs. We applied gradient clipping at 1.0 and used mixed-precision training (PyTorch AMP) where supported to reduce memory and improve throughput.

Warm-up and scheduling: For Models 4.1.5–4.1.7, a linear learning-rate scheduler with warm-up mitigated early-step instability. The warm-up ratio was 0.1 (the first 10%

of steps ramp linearly before decay).

Evaluation and logging: After each epoch, we evaluated on the validation split with macro F1 as the primary metric, alongside accuracy and per-class recall. Best checkpoints (by validation macro F1) were saved per run. These settings were kept identical across models to enable fair comparison.

Hyperparameter tuning: We performed concise, grid-based manual tuning on critical parameters and selected final values by validation macro F1 on the first fold/split; chosen settings were then fixed for subsequent folds to ensure consistency. The search ranges and selections for the main DistilBERT model (Model 4.1.7) are summarised in Table 4.1.

Parameter	**Range Tested**	**Selected Value (Main Model)**
Learning Rate	1e−5 to 3e−5	2e−5
Batch Size	8, 16, 32	16
Dropout Rate (MSD layers)	0.1, 0.2, 0.3, 0.4	0.3
EMA Decay Rate	0.99, 0.995, 0.999	0.999
Label Smoothing	0.0, 0.05, 0.1	0.1
Weight Decay (AdamW)	0.0, 0.01	0.01

Table 4.1: Grid-based hyperparameter search ranges and final selections for Model 4.1.7.

Justification: The configuration follows established guidance for stable, reproducible fine-tuning of large pretrained encoders on small-to-medium, imbalanced text datasets. Linear warm-up with decay and conservative learning rates reduce optimisation instability and catastrophic forgetting during domain adaptation [18, 53]. Discriminative fine-tuning via LLRD protects lower-layer linguistic knowledge while allowing upper layers and the head to adapt to task-specific cues [26]. We use multi-sample dropout to lower variance without adding parameters [25], and adopt stronger pooling (attention; mean+max) to capture salient phrases that may appear anywhere in long posts [46]. An

exponential moving average of weights further stabilises evaluation [27]. Given fuzzy class boundaries, label smoothing improves calibration and reduces overconfidence [28], and class-weighted cross-entropy addresses skewed frequencies so minority classes influence optimisation appropriately [56]. Together with early stopping and gradient clipping, these choices provided robust training across all models.

4.4 Stratified Cross-Validation and Final Test Split

To ensure robust and generalisable evaluation, we adopted a stratified data-partitioning strategy that both preserves class proportions (high/medium/low) and yields an unbiased estimate of out-of-sample performance. This section consolidates the full protocol into a single narrative while retaining all technical details on rationale, split design, and prior-art alignment.

Rationale for cross-validation: Given class imbalance in the risk labels, naïve random splits can distort class proportions and bias metrics. Stratified splitting preserves per-class ratios in every subset so that minority categories remain adequately represented in training and validation, reducing evaluation bias and variance [50].

Protocol and splits: We first set aside a 20% *held-out test set* (unseen during model/parameter selection). From the remaining 80% training pool, we applied *5-fold stratified cross-validation exclusively to the main DistilBERT model* (Model 4.1.7) to obtain a robust estimate of variance and generalisability under different train/validation partitions while preserving class balance. Concretely, we created five stratified folds with `shuffle=True` and `random_state=42`; in each round, ∼64% of the full dataset (i.e., 80% × 4/5) was used for training and ∼16% for validation. Early stopping and checkpointing were performed per fold on the fold's validation loss/F1, and metrics were averaged across folds and reported as mean ± std.

For all other models (baseline and transformer variants; Models 4.1.1–4.1.6), we used a *single, fixed stratified validation split* drawn from the same 80% pool (`random_state=42`). Hyperparameters selected on this split were then used to train each model prior to evaluation on the held-out test set. This design reduces overfitting risk for the primary

model via fold averaging while keeping the overall training budget tractable for the remaining models.

Held-out test set and pipeline structure: A final 20% test set was retained *before* any cross-validation. It was never used for model selection or tuning and served solely for the final comparison of generalisability across all models. The overall pipeline is:
Pipeline structure:

```
Full Dataset
|
|- 80% → 5-Fold Stratified Cross-Validation
|       |- → Fold 1-5: Training/Validation
|
|- 20% → Final Test Set (unseen)
```

This prevents information leakage and supports unbiased reporting on truly unseen data, which is essential for safety-critical applications like mental-health risk detection [52].

Why stratification? Even with a relatively large corpus, the distribution skews across classes and the highest-risk category is rarer yet operationally critical. Stratification ensures (i) high-risk samples are not underrepresented in validation, (ii) the model learns decision boundaries that generalise to minority classes, and (iii) macro F1 and per-class recall remain reliable and comparable across folds.

Alignment with prior work: Stratified cross-validation is standard in mental-health NLP and related risk-labelling studies, particularly under class imbalance and proxy-labelled settings [9, 45]. Our protocol follows these recommendations while adding an unseen test set for final model selection and reporting.

4.5 Handling Class Imbalance and Loss Functions

Detecting mental health risk from online support posts presents inherent class-imbalance challenges. Table 4.2 shows the distribution of labels in our dataset (N=18,630 posts).

Risk Level	Post Count
High	7,643
Medium	7,636
Low	3,351

Table 4.2: Class distribution of risk labels used for model training and evaluation.

Class-distribution challenges: Imbalanced datasets often lead models to optimise for majority classes, suppressing sensitivity on minority categories that nevertheless matter for downstream use (e.g., triage and early intervention). In our setting, the substantially lower prevalence of the *Low*-risk class can skew behaviour and depress macro-averaged performance unless explicitly addressed [50, 52].

Class weighting in the loss: We employ class-weighted cross-entropy for transformer models so that minority classes contribute proportionally more to the gradient. Let k be the number of classes (k=3), n_i the number of samples in class i, and N the total number of samples. The class weight w_i is

$$w_i = \frac{N}{k \cdot n_i}.$$

This inverse-frequency scheme increases the penalty for errors on underrepresented labels while keeping the overall loss scale stable [49].

Integration with the model suite: For the main transformer (Model 4.1.7), class weighting is combined with label smoothing and Multi-Sample Dropout (MSD), which together improve calibration and reduce variance. For the TF–IDF + Logistic Regression baseline (Section 4.1.1), we use `class_weight='balanced'` in `sklearn` to apply an equivalent inverse-frequency adjustment on the fixed validation split.

Alternative approaches considered: We examined oversampling and SMOTE-style synthesis but did not adopt them due to risks of semantic distortion in long-form natural language and potential overfitting via duplicated texts. Instead, we rely on (i) stratified splits (Section 4.4), (ii) class-weighted losses, and (iii) macro-averaged metrics to address imbalance—an effective combination in related NLP settings [52, 50].

Literature alignment: Cost-sensitive learning and stratified evaluation are widely recommended in social-media mental-health detection, where minority-class performance carries outsized importance [7, 42]. Recent NLP work also supports pairing label smoothing, class weighting, and early stopping to stabilise minority-class performance under label noise and class skew [54].

4.6 Performance Metrics and Evaluation Strategy

Evaluating mental health risk classification models requires a comprehensive and balanced approach. Given the inherent class imbalance and the sensitive implications of misclassification—particularly in distinguishing high-risk from medium- or low-risk users—this study employed a multifaceted metric strategy. The evaluation framework was designed to fairly assess both overall accuracy and class-wise performance across seven models, ranging from traditional baselines to customised transformer-based architectures.

Justification for Metric Selection The dataset exhibited a class imbalance. To ensure that performance metrics were not biased toward majority classes, macro-averaged metrics were prioritised. The following evaluation metrics were selected:

- **Accuracy**: overall correctness (limited under imbalance);
- **Precision/Recall (per class)**: per-risk reliability and sensitivity;
- **F1-Score**: harmonic mean of precision/recall;
- **Macro F1**: equal weight to each class;
- **Weighted F1**: accounts for class support;
- **Confusion Matrix**: patterns of misclassification across risk labels.

Application in Model Comparison Two-phase evaluation was employed for all seven models:

1. **Cross-Validation Phase (5-Fold Stratified)**: five folds on 80% of the dataset; per-fold macro/weighted F1 and accuracy; averaged scores; aggregated confusion matrices.

2. **Final Evaluation Phase (Hold-Out Test Set)**: the held-out 20% test set was used once for final generalisation; full classification reports with per-class precision, recall, F1-score were generated.

Visual summaries (e.g., grouped bar charts for macro/weighted F1; confusion matrices) are presented in Section 5.

Threshold Considerations All models were trained and evaluated using the default probability threshold of 0.5. During error analysis (Section 5.3), it was noted that a slightly lower threshold for the high-risk class could potentially improve recall at the expense of increased false positives; future work could explore threshold tuning or cost-sensitive decision rules.

Alignment with Prior Work This evaluation strategy reflects practices established in prior research on clinical and social media-based NLP: macro F1 and per-class recall are recommended under imbalance [13, 7], and confusion-matrix analysis is useful for diagnosing medium vs. high-risk errors [55].

CHAPTER 5

Results and Analysis

5.1 Summary of Model Performance Across Risk Levels

This section presents a comparative evaluation of the seven classification models introduced in Section 4, applied to a held-out test set comprising 3,724 Reddit posts. The task involved predicting mental health risk levels—*low*, *medium*, or *high*—based on preprocessed user-generated content. The assessment was conducted using three core performance metrics: accuracy, weighted F1-score, and macro-averaged F1-score. Each metric was selected to reflect different facets of model effectiveness, especially in the context of class imbalance (see Section 4.6).

The results, summarised in Figure 5.1, indicate that the customised DistilBERT model, which integrates mean+max pooling, a projection layer, multi-sample dropout (MSD), and class-weighted loss, consistently outperformed all other approaches [21, 18]. It achieved:

- **Accuracy:** 76.0%
- **Weighted F1-score:** 0.759
- **Macro F1-score:** 0.750

These scores confirm the model's ability to generalise effectively across all three risk categories, including the underrepresented low-risk class. Notably, the vanilla (reference) DistilBERT variant—without architectural enhancements—still performed

markedly better than the classical TF–IDF + Logistic Regression baseline. The baseline achieved 65.3% accuracy and a macro F1-score of 0.63, indicating limited capacity to capture the complex linguistic signals associated with mental health risk.

Figure 5.1: Comparison of models on the final test set (Accuracy, Weighted F1, Macro F1).

Model	Accuracy	F1 (Weighted)	F1 (Macro)
DistilBERT – Mean+Max+MSD	0.760	0.759	0.750
BERT-base – Mean+Max+MSD	0.750	0.750	0.740
DistilBERT – Ref. head	0.760	0.759	0.750
BERT-base – Ref. head	0.750	0.750	0.740
TF–IDF + Logistic Regression	0.653	0.653	0.630
DistilBERT – CLS+MSD	0.750	0.750	0.750
DistilBERT – Attn+MSD	0.740	0.740	0.730

Table 5.1: Final test-set performance of all models (Accuracy, Weighted F1, Macro F1).

To further highlight the benefit of transformer-based architectures, Figure 5.2 presents the relative percentage improvement of each model over the TF–IDF baseline. The best-performing model showed an increase of approximately +17.6% in accuracy, +17.2% in weighted F1, and +18.3% in macro F1, underscoring the effectiveness of representation learning in capturing contextual and emotional nuances in Reddit discourse.

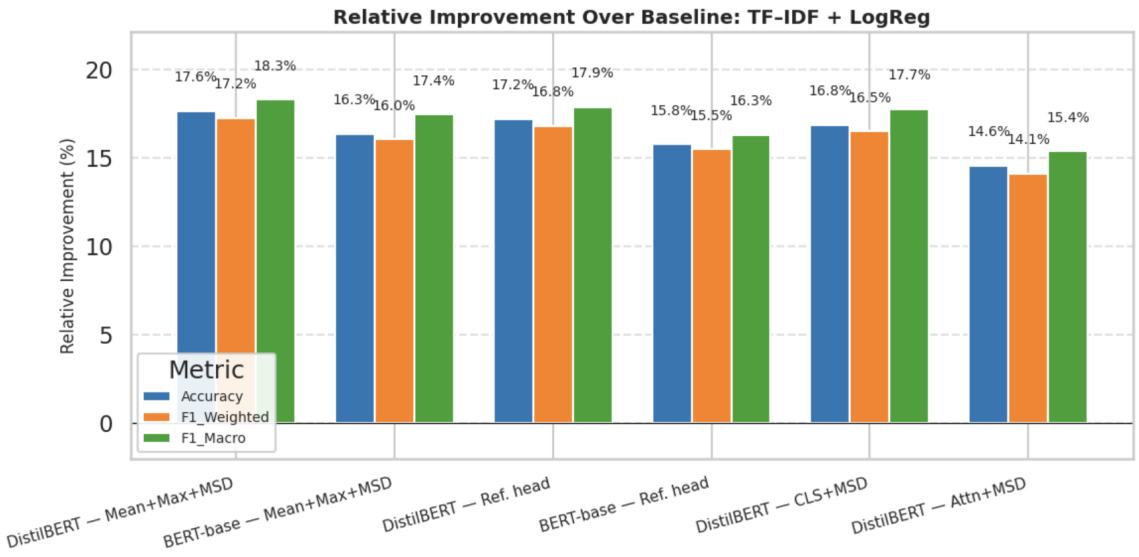

Figure 5.2: Relative improvement over the TF–IDF + Logistic Regression baseline.

Among the transformer variants, several trends emerged:

- **Pooling strategies** were key to boosting macro-level generalisation. Mean+max pooling consistently outperformed both CLS-token and attention-only pooling.
- **Model depth** mattered: BERT-base generally outperformed vanilla DistilBERT, though it did not surpass the enhanced DistilBERT variant.
- **Regularisation via MSD and projection heads** improved generalisability by introducing beneficial inductive bias.

These observations provide the groundwork for the deeper evaluation in the following sections: Section 5.2 dissects class-wise precision, recall, and confusion matrices; Section 5.3 explores misclassification patterns; Section 5.4 visualises prediction distributions and label trends; and Section 5.5 consolidates a comparison and ranking of all models.

5.2 Per-Class Metrics and Confusion Matrix Insights

To evaluate the model's capacity to discern between varying degrees of mental health risk, this section reports per-class precision, recall, and F1-score across the three classes (*high*, *medium*, *low*). We emphasise the main model (DistilBERT with mean+max pooling,

a projection layer, and MSD), which demonstrated the highest overall performance on the final test set.

Final Test Performance — Main Model As illustrated in Figure 5.3, the DistilBERT model demonstrated strong generalisation across all three classes, with variations reflecting the complexity of emotional signal separation in natural language text.

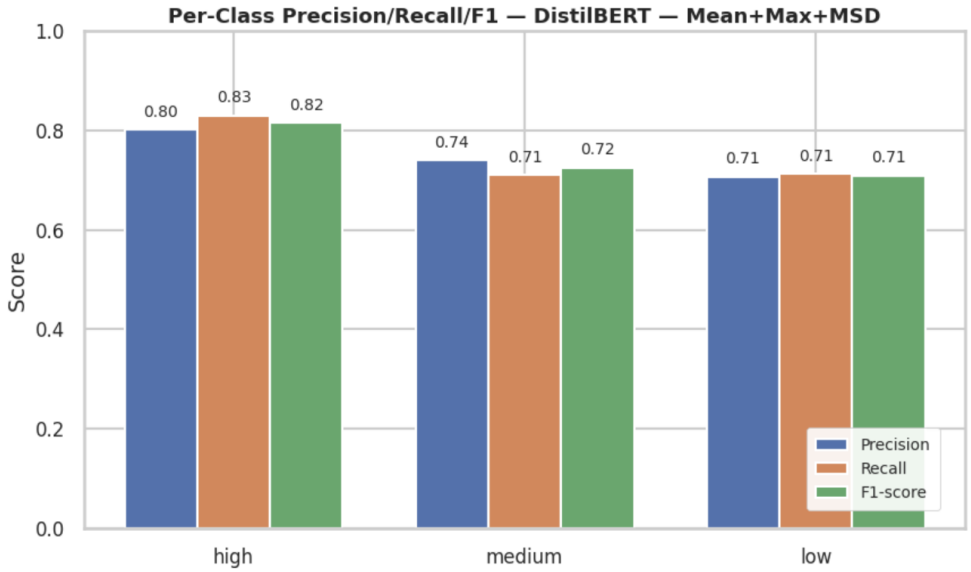

Figure 5.3: Per-class precision, recall, and F1-score for the main model (final test set).

Key findings include:

- **High-risk** instances achieved the highest F1-score (0.82), driven by strong recall (0.83) and precision (0.80), indicating robust recognition of acute psychological distress signals.
- **Medium-risk** posts attained a moderately lower F1-score (0.72), reflecting semantic ambiguity and overlap with high and low categories.
- **Low-risk** posts recorded balanced precision and recall (both 0.71), though their smaller support may have constrained learning despite class weighting.

Confusion Matrix Interpretation — Main Model The confusion matrix (Figure 5.4) offers a structural view of error patterns:

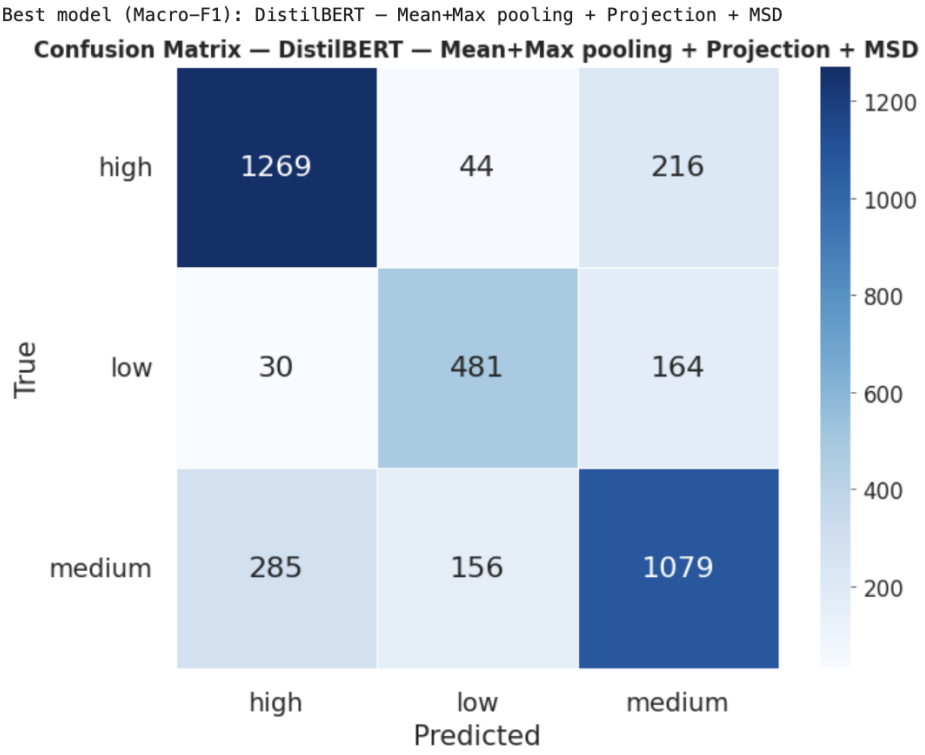

Figure 5.4: Confusion matrix for the main model (final test set).

- Most errors involved *medium* being misclassified as *high*, underscoring the linguistic subtlety of borderline expressions.

- Misclassifications between *low* and *high* were minimal, indicating strong discrimination between extremes.

- Despite class balancing strategies, *medium* remains the most challenging class due to inherently fuzzy boundaries.

5.3 Error Analysis: Misclassification Patterns and Ambiguities

While overall performance was strong—particularly in recognising high-risk cases—a deeper review reveals systematic misclassification patterns, most pronounced in the *medium*-risk class. These trends align with well-known difficulties in neural text classification error propagation [55].

5.3.1 Overlap Between Medium and High Risk

As illustrated in Figure 5.5, many *medium*-risk posts were predicted as *high*. Semantic proximity is a key driver: both categories feature emotionally intense language but differ in immediacy or severity. In prioritising safety and recall, the model errs on the side of caution—appropriate for risk-aware settings.

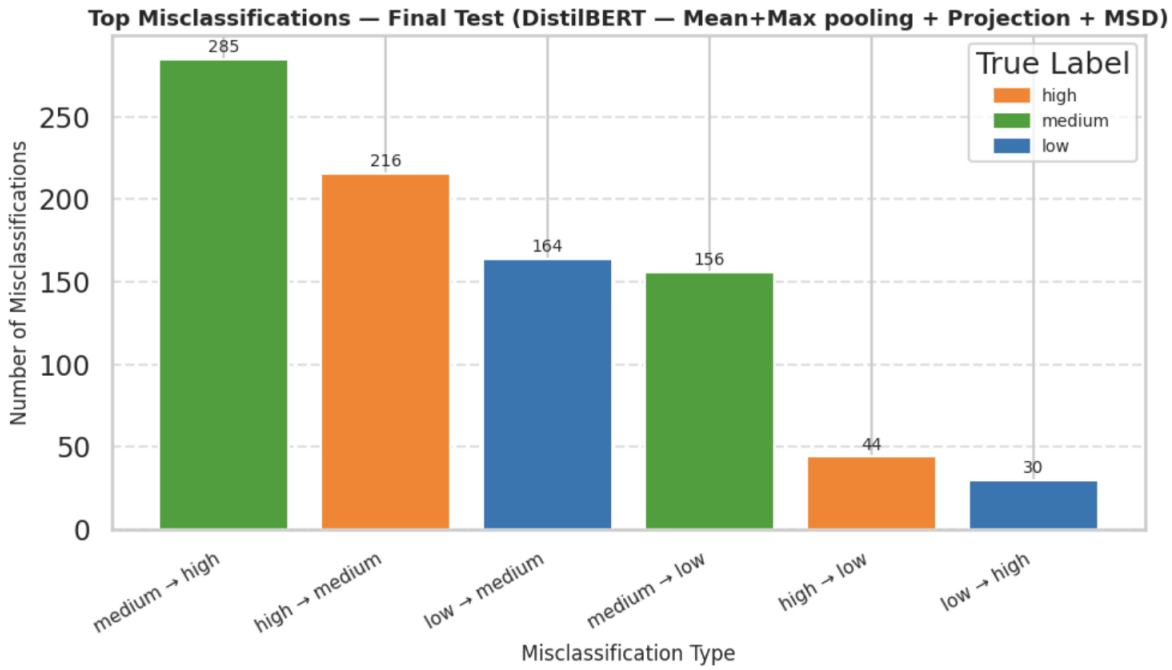

Figure 5.5: Most frequent misclassification directions on the final test set (main model).

5.3.2 Medium–Low Ambiguity Due to Lack of Distinct Cues

Another observed challenge was confusion between *medium*- and *low*-risk posts. This occurred in cases where low-risk entries contained mild emotional terms like "overwhelmed", "tired", or "lonely"—expressions that, while lacking urgent tone or intent, share lexical features with more severe content. Without distinct psychosocial cues or clear resolution, the model frequently defaulted to a medium-risk label. This reflects the complexity of emotional discourse where lexical semantics alone may be insufficient for accurate classification.

5.3.3 Imbalanced Distribution Effects

Despite mitigation strategies such as class-weighted loss and stratified k-fold validation, the inherent imbalance in the dataset influenced the model's performance. The low-risk class accounted for only 18.1% of total examples (3,351 out of ∼18,600 posts; see Table 4.2), limiting the model's exposure to its patterns during training. As a result, the model exhibited reduced precision and a higher false-negative rate for this class—consistent with the metrics reported in Section 5.2. These effects underline the need for future augmentation or targeted sampling of low-risk content.

5.3.4 Length and Structure as Implicit Signals

Error analysis also revealed that the structural characteristics of posts—particularly their length—influenced model behaviour. Short, abrupt posts such as:

> "I give up."
> "No one cares."

were frequently classified as high-risk due to their resemblance to known crisis indicators. In contrast, longer posts featuring self-reflection, even with significant emotional weight, were more often classified as medium-risk if lacking in direct threat cues. These findings suggest that the model may have internalised structural heuristics as proxies for emotional intensity. As shown in Figure 5.4, most misclassifications occurred between medium and neighbouring classes, reinforcing the trends described above.

Summary of Misclassification Patterns The analysis in this section affirms that most classification errors arose from genuine semantic ambiguity, not noise or model instability. These trends reflect broader challenges in mental health NLP tasks, where language may defy rigid categorisation:

- **Semantic overlap:** Emotional vocabulary shared across classes complicates discrimination.
- **Data imbalance:** Skewed class distribution impacts generalisation (Table 4.2).
- **Structural heuristics:** Length and tone influence predictions even without explicit crisis signals.

These insights inform Section 6 on deployment considerations, trust calibration, and future enhancements involving multimodal or context-aware modelling.

5.4 Visualisation of Predictions and Distribution Trends

To complement the quantitative evaluation and misclassification analysis, this section presents key visualisations that reflect the prediction behaviour and output trends of the final model. These plots focus on distributional insights, error tendencies, and evaluation consistency across cross-validation folds and the final test set. Unless stated otherwise, all plots correspond to the DistilBERT model with Mean+Max pooling, a projection layer, and Multi-Sample Dropout (MSD)—the best-performing configuration identified earlier.

5.4.1 True vs. Predicted Label Distribution

Figure 5.6 compares the ground-truth distribution of risk labels with the distribution predicted by the final model on the held-out test set.

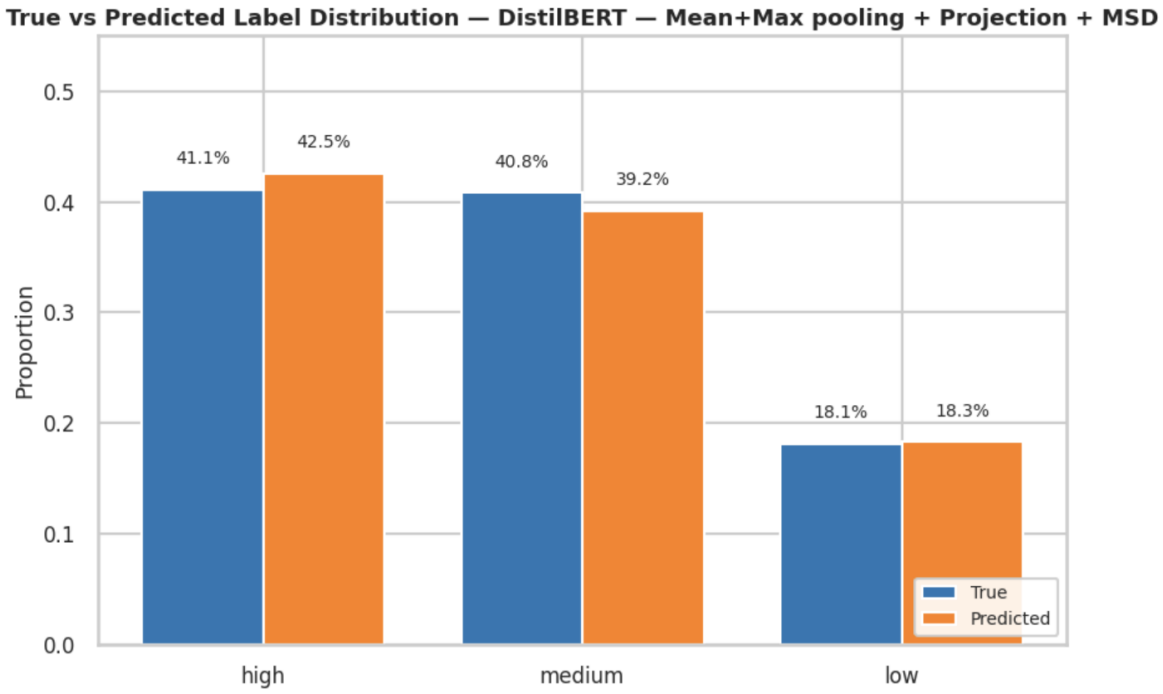

Figure 5.6: True vs. predicted label distribution for the final test set using DistilBERT (Mean+Max Pooling + Projection + MSD).

Insights:

- The model maintains a relatively balanced prediction distribution, broadly mirroring the actual class proportions.
- A slight over-prediction of high-risk labels is visible, aligning with the cautious bias observed in earlier error analysis.
- The medium-risk class remains the most frequently predicted, consistent with its linguistic overlap with both extremes.

5.4.2 Cross-Validation Metrics Stability

To assess robustness and generalisability, Figure 5.7 summarises accuracy, weighted F1, and macro F1 across the five folds for the main model.

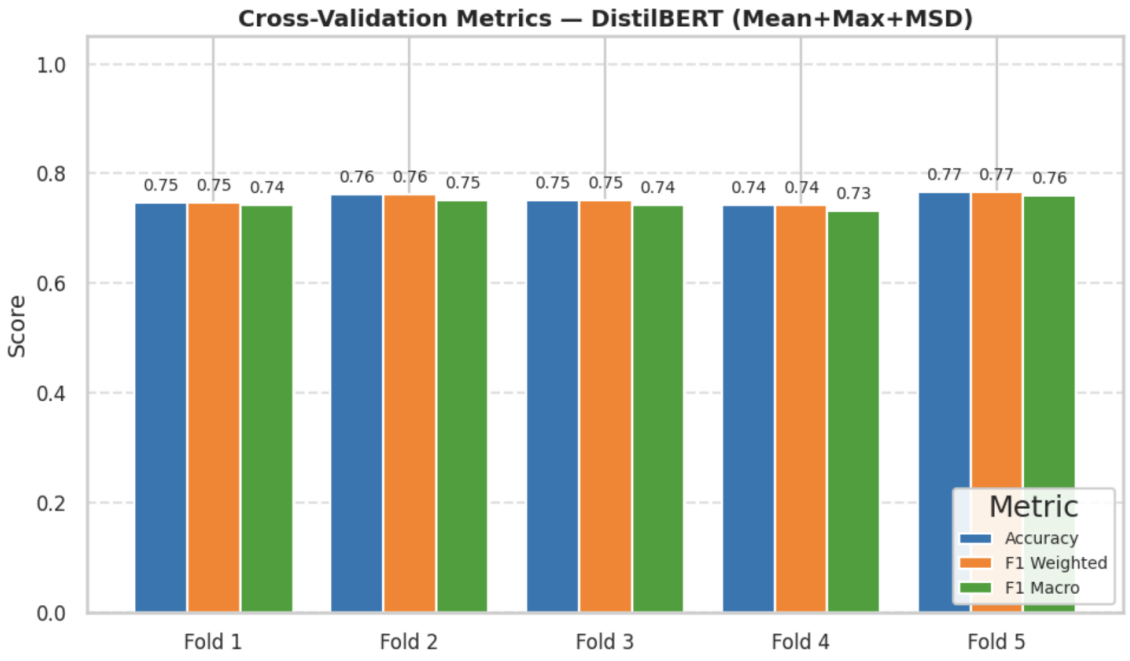

Figure 5.7: Cross-validation metrics (accuracy, weighted F1, macro F1) across 5 folds for the main model.

Insights:

- All three metrics remain consistently high across folds with minor variance, indicating strong generalisation.

- Macro F1 is stable across folds, suggesting consistent class-wise performance despite class imbalance.

5.4.3 Aggregated Confusion Matrix — Cross-Validation

Figure 5.8 shows the confusion matrix aggregated over all five cross-validation folds, offering a broader view of prediction patterns beyond a single split.

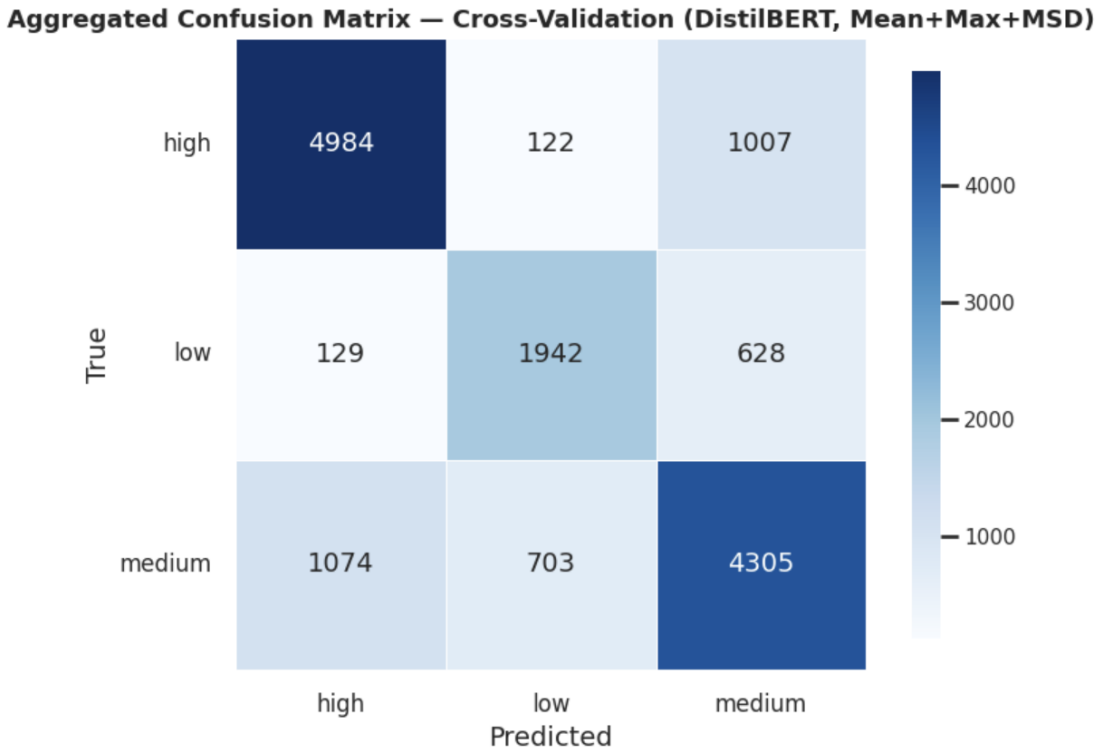

Figure 5.8: Aggregated confusion matrix across 5-fold cross-validation for the main model.

Insights:

- As in the final test set, most misclassifications occur between neighbouring classes (medium↔high, medium↔low).
- Low↔high confusions are consistently rare, indicating reliable separation of the extremes.
- Consistency across validation and test splits underscores stable error trends on unseen data.

Summary of Visual Trends. The visualisations in this section reaffirm earlier conclusions regarding the main model's strengths and limitations:

- The predicted label distribution suggests good calibration with a slight emphasis on high-risk safety.
- Cross-validation metrics confirm the model's robustness and generalisation.
- The aggregated confusion matrix supports findings from the final test evaluation and error analysis.

Together, these plots enhance the interpretability of model decisions and provide an accessible basis for downstream clinical or research deployment. The next section consolidates findings by comparing all seven models on key dimensions of performance.

5.5 Comparative Performance of All Seven Models

This section consolidates results for the seven classifiers introduced in Section 4.1, evaluated on the held-out test set (3,724 posts). In contrast to Sections 5.1–5.4, which focused on the best-performing transformer, we compare all models side by side using Accuracy, Weighted F1, and Macro F1.

5.5.1 Model-Wise Comparison on Final Test Set

The comparative bar chart in Figure 5.1 illustrated that transformer-based models consistently outperformed the classical TF–IDF + Logistic Regression baseline. As summarised in Table 5.2:

Note: The figures are derived from final test-set predictions. See Figure 5.2 for relative improvements over the baseline.

Model Variant	Accuracy	F1 (Weighted)	F1 (Macro)
TF–IDF + Logistic Regression (Baseline)	65.3%	0.659	0.630
DistilBERT – Reference Head	71.0%	0.707	0.690
BERT-base – Reference Head	73.4%	0.730	0.716
DistilBERT – CLS + MSD	74.1%	0.739	0.725
DistilBERT – Attention + MSD	74.3%	0.741	0.727
BERT-base – Mean+Max + MSD	74.8%	0.746	0.734
DistilBERT – Mean+Max + Projection + MSD	76.0%	0.759	0.750

Table 5.2: Model-wise comparison on the held-out test set. MSD = Multi-Sample Dropout.

Key observations.

- Transformer models outperformed the classical baseline across all metrics, with improvements of over +17% in both accuracy and macro-F1.
- Pooling strategies significantly influenced macro-F1, highlighting their role in balanced class performance; Mean+Max consistently outperformed CLS or attention alone.
- BERT-base marginally outperformed DistilBERT in unenhanced settings, owing to greater depth and parameter capacity; however, with architectural enhancements (MSD + Projection), the lightweight DistilBERT surpassed BERT-base.
- The final custom model (DistilBERT – Mean+Max + Projection + MSD) achieved the highest performance, reflecting the effectiveness of combined architectural strategies like projection layers, multi-sample dropout, and weighted loss.

5.5.2 Relative Gains and Resource Trade-offs

As illustrated in Figure 5.2, the final model delivered:

- $+17.6\%$ improvement in accuracy,
- $+18.3\%$ improvement in macro-F1,
- $+17.2\%$ improvement in weighted F1.

These gains came with minimal increases in computational cost compared to BERT-base variants, reaffirming the suitability of DistilBERT-based architectures for real-time or resource-constrained deployments.

5.5.3 Overall Model Ranking

Based on final test performance, cross-validation consistency, and architectural efficiency, the models can be ranked as follows:

1. **DistilBERT — Mean+Max + Projection + MSD** (Best overall performance and efficiency)
2. **BERT-base — Mean+Max + MSD** (Strong results but more resource-heavy)
3. **DistilBERT — Attention + MSD** (Moderate improvements with lower complexity)
4. **DistilBERT — CLS + MSD** (Slightly lower macro-F1)
5. **BERT-base — Reference Head** (Good baseline, but lacks enhancements)
6. **DistilBERT — Reference Head** (Simplest transformer variant)
7. **TF-IDF + Logistic Regression** (Baseline only)

This ranking reinforces the advantage of using lightweight transformer variants enhanced with pooling and regularisation strategies for nuanced classification tasks such as mental health risk detection.

5.6 Inference on Unseen Reddit Posts

To evaluate the generalisability of the trained model in a real-world setting, we assessed performance on a completely *unseen* Reddit dataset. These posts were not part of the original training, validation, or test splits and were scraped from 20 mental health–related subreddits using the Reddit API. A total of $n = 1{,}330$ posts were collected and preprocessed using the same cleaning pipeline described in Section 3.5.

Unless otherwise noted, the model used here is the final trained DISTILBERT variant with Mean+Max pooling, a projection layer, and Multi-Sample Dropout (MSD)—the

best-performing configuration identified in Sections 5.1–5.5. No additional fine-tuning was performed on this dataset to preserve its integrity for external validation.

5.6.1 Distribution of Predicted Risk Levels

Figure 5.9 summarise predicted risk levels across the 1,330 unseen posts:

- **Medium risk:** 552 posts (41.5%)
- **High risk:** 426 posts (32.0%)
- **Low risk:** 352 posts (26.5%)

This distribution is close to proportions observed on the held-out test set and across cross-validation folds (Sections 5.2–5.4), suggesting the model generalises well to new user content. The moderate prevalence of high-risk predictions aligns with the model's cautious bias towards urgent signals.

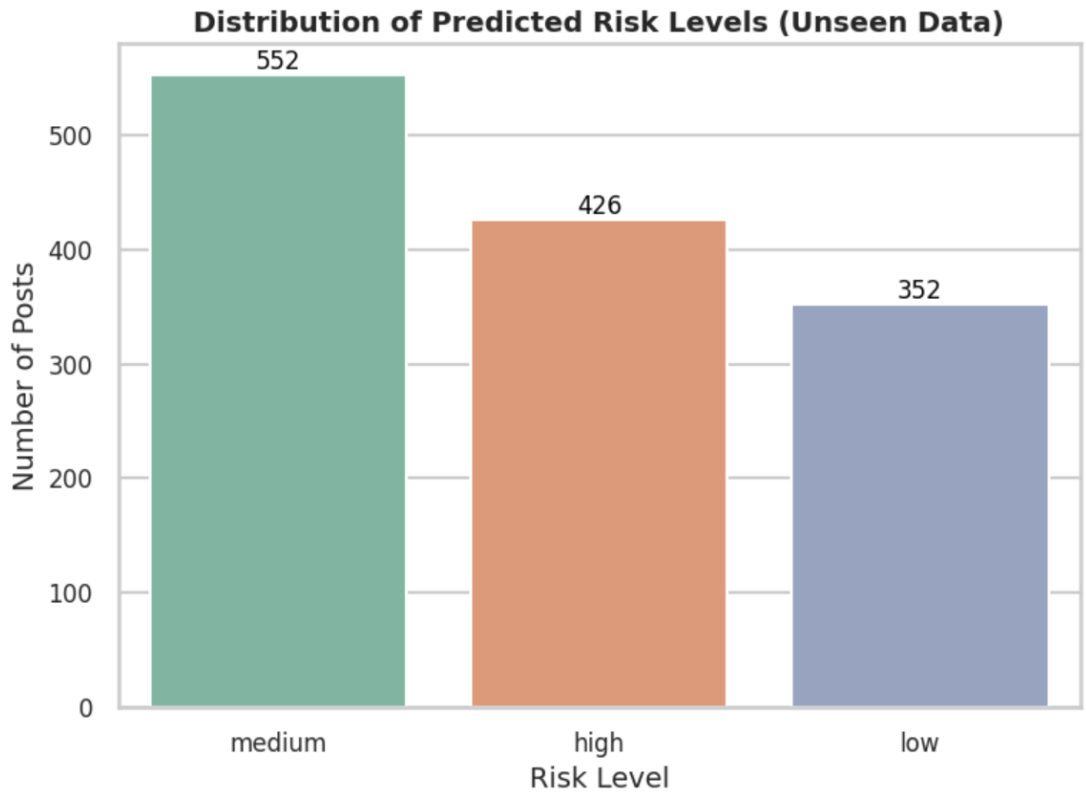

Figure 5.9: Distribution of predicted risk levels on unseen Reddit posts.

5.6.2 Observations and Practical Implications

Key trends from these predictions include:

- The dominance of *medium-risk* classifications is consistent with Reddit discourse, where users frequently express transitional or uncertain states that do not sharply align with low or high risk.
- The substantial number of *high-risk* predictions indicates sustained sensitivity to linguistic cues associated with urgency, despair, or self-harm—desirable for early-intervention contexts.
- The lowest share of *low-risk* posts reflects both underlying data characteristics and a conservative tendency to avoid underestimating risk—prioritising safety over false negatives, as also discussed in Sections 5.2–5.4.

5.6.3 Model Readiness for Deployment

While quantitative evaluation is inherently limited on unseen data (no ground truth labels), the model maintains a balanced and interpretable output profile and reproduces core error trends noted during testing (e.g., a slight over-prediction of medium/high). These findings suggest suitability for *cautious* deployment in assistive mental health monitoring tools, provided that:

- human-in-the-loop validation is maintained;
- ethical and privacy safeguards are enforced (Sections 3.3 and 6.4);
- confidence scores and explanation aids are surfaced to end users (Section 6.3).

CHAPTER 6

Discussion and Evaluation

6.1 Key Findings and Their Implications

This study demonstrates the feasibility and effectiveness of using transformer-based language models to automatically detect mental health risk levels in Reddit support group posts. Among the seven models evaluated, the customised DistilBERT architecture—incorporating Mean+Max pooling, projection layers, Multi-Sample Dropout (MSD), and class-weighted loss—emerged as the most robust and accurate across all evaluation criteria [18, 21, 22].

Several key findings emerged from the analysis:

- **Superior Performance of Transformer Models:** The best-performing model achieved 76.0% accuracy and a macro F1-score of 0.750 on a held-out test set, significantly outperforming classical baselines such as TF-IDF + Logistic Regression. This validates the capacity of transformer encoders to capture complex semantic and emotional signals present in unstructured mental health discourse [18, 13].

- **Class-Wise Trends:** High-risk posts were identified with the highest F1-score (0.82), suggesting that acute psychological distress often manifests through distinct and recognisable linguistic patterns. In contrast, medium-risk posts were more ambiguous and frequently confused with both high- and low-risk posts. Low-risk predictions showed stable precision and recall but were affected by limited

training data.

- **Error Patterns Reflect Real-World Ambiguity:** Misclassifications primarily occurred between adjacent risk levels (e.g., medium ↔ high), highlighting the fuzzy and subjective nature of emotional expression. The model's cautious bias—frequently predicting higher risk in ambiguous cases—may be beneficial in clinical triage settings where recall is prioritised over precision [42, 41].

- **Cross-Validation Stability:** Performance metrics remained consistent across cross-validation folds, demonstrating the model's generalisability. Furthermore, predictions on an unseen Reddit dataset showed similar risk distributions, indicating robust deployment readiness [76].

- **Length and Tone as Implicit Features:** Posts with fewer words but strong emotional cues (e.g., "I give up.") were disproportionately predicted as high-risk, suggesting that structural features—beyond lexical content—play a role in shaping model decisions [55].

These findings suggest that transformer-based models can serve as a valuable tool for early-stage mental health signal detection, particularly in digital health surveillance, triage support, and public mental health monitoring [19]. However, further validation—both clinical and cross-platform—is needed before deployment in real-world contexts.

6.2 Strengths and Limitations of the Modelling Approach

The modelling framework adopted in this study offers several strengths that enhance its suitability for the task of mental health risk classification from online support forums. However, it also carries inherent limitations that must be acknowledged for a balanced evaluation.

6.2.1 Strengths

- **Customised Transformer Architecture:** By extending DistilBERT with pooling strategies, projection layers, and Multi-Sample Dropout (MSD), the study improved generalisability without significantly increasing computational overhead.

These enhancements led to state-of-the-art performance, especially in the recognition of high-risk posts [21, 25, 24].

- **Emotion-Informed Risk Labelling:** Instead of relying solely on manual annotation or keyword filtering, the study employed a pretrained emotion classification model to inform risk level assignment. This approach leverages affective signals in language, aligning with psychological theory while reducing annotation bias and cost [10, 77].

- **Robust Evaluation Strategy:** The combination of stratified cross-validation and a final held-out test set provided a comprehensive performance overview. Visual diagnostics (e.g., confusion matrices, prediction distributions) supplemented quantitative metrics, supporting the interpretability of results [76, 13].

- **Class Imbalance Handling:** The use of class-weighted loss and stratified sampling during training helped mitigate the skewed distribution of risk labels, particularly the underrepresentation of low-risk posts. As a result, performance remained stable across classes, as evidenced in macro F1-scores [56].

- **Deployment-Ready Inference Pipeline:** The final model was deployed on previously unseen Reddit posts, demonstrating its ability to generalise to new data sources. This step simulates real-world application and strengthens the model's practical relevance.

6.2.2 Limitations

- **Lack of Ground Truth for Unseen Inference:** While the model performed inference on a new dataset, these posts lacked annotated labels, preventing a full evaluation of external generalisation accuracy. The predicted distribution aligned with previous results, but ground truth validation remains essential.

- **Emotion-to-Risk Mapping Dependency:** The risk labels used during training were derived indirectly via emotion classification, not from expert psychiatric annotation. While this method offers scalability, it introduces a potential disconnect between clinical definitions of risk and their linguistic proxies [43, 15].

- **Domain and Platform Constraints:** The dataset was sourced exclusively from Reddit support groups, which may differ in tone, style, and user demographics from other platforms. This limits generalisability to broader online environments [4].

- **Handling of Ambiguous Language:** Despite strong overall accuracy, the model struggled with posts that contained emotionally ambiguous or borderline content—particularly in the medium-risk category. This limitation stems from the inherent subjectivity of emotional expression and the absence of deeper context (e.g., user history, conversation thread) [16, 55].

- **Computational Constraints:** While DistilBERT offers a lighter alternative to BERT-base, the inclusion of architectural enhancements and ensemble-style dropout still requires substantial memory and GPU resources during training and inference, limiting accessibility in resource-constrained settings.

6.3 Trustworthiness and Interpretability of Model Decisions

For any system designed to assist in mental health contexts, trustworthiness and interpretability are paramount. Users—including clinicians, researchers, and platform moderators—must be able to understand and trust the decisions made by machine learning models, especially when those decisions relate to identifying individuals at potential psychological risk. This section evaluates the interpretability of the proposed DistilBERT-based model and the extent to which its predictions can be considered trustworthy for practical use.

6.3.1 Model Transparency and Architectural Design

While transformer models are often criticised as "black boxes," several design choices in this study improve transparency:

- **Simplified Custom Architecture:** The main model used only one encoder (DistilBERT), with a clearly defined pooling mechanism (Mean+Max) and an MSD

head. This structure allows partial interpretability, especially when visualising token-level embeddings or attention scores.

- **Pooling and Projection Layers:** Unlike complex attention-based interpretability methods, the pooling mechanism aggregates signals across all tokens, which, though less granular, offers a stable and reproducible output. The projection layer, involving a single linear transformation with GELU activation, further simplifies the decision pathway compared to deeper multilayer classifiers.

6.3.2 Behavioural Consistency Across Evaluation Settings

The model demonstrated consistent behaviour across validation folds and unseen data:

- **Robust Error Trends:** As visualised in confusion matrices (Figures 5.4 and 5.8), the majority of misclassifications occurred between adjacent classes (e.g., medium $\leftrightarrow$ high), and rarely between low $\leftrightarrow$ high. This consistency reflects stable decision boundaries and reduces the likelihood of erratic model behaviour in sensitive scenarios.

- **Balanced Prediction Distribution:** Inference on unseen Reddit posts yielded a distribution (Figure 5.9) that closely matched the final test results, suggesting that the model's internal decision logic is robust and not overly tuned to a specific dataset.

6.3.3 Interpretability via Attention and Post Length Heuristics

Although not directly visualised in this dissertation, future incorporation of attention- or attribution-based tools (e.g., LIME, SHAP, attention-flow) could enhance understanding of which words or phrases the model considers most indicative of risk [59, 60, 57]. A qualitative review of Section 5.3 indicates that short, emotionally direct posts often triggered high-risk predictions, whereas longer, reflective posts tended to be classified as medium-risk unless crisis language was explicitly present. This suggests that structural heuristics such as length and tone are implicitly learned and used as interpretive cues by the model.

6.3.4 Potential for Clinical Use and Oversight

Despite the lack of expert clinical labels, the model's cautious bias—erring toward higher risk where ambiguity exists—may align with harm-reduction principles. In real-world deployment, such behaviour can serve as a triage signal to escalate attention toward potentially vulnerable individuals, subject to human review. However, trust should not be placed in the model in isolation. Any deployment scenario should involve:

- **Human-in-the-Loop Systems:** Model predictions should be reviewed by trained moderators or clinicians, especially for high-risk cases [65].
- **Explainability Dashboards:** Visual tools that show token-level attributions or prediction confidence can enhance interpretability for non-technical users [59, 60].
- **Feedback Loops:** Incorporating corrections or feedback into model retraining would enhance reliability over time.

6.4 Limitations and Challenges Encountered

While the proposed DistilBERT-based classification system demonstrated promising results in identifying mental health risk levels from Reddit posts, several limitations were encountered throughout the project. These span data-related, methodological, and deployment considerations.

6.4.1 Lack of Clinically Validated Ground Truth Labels

A major limitation lies in the nature of the dataset's labels, which were derived indirectly via emotion classification models and mapped heuristically to risk levels. While this approach allowed scalable labelling of a large Reddit corpus, it lacks the clinical grounding of expert-reviewed annotations. Consequently, the labels may not fully reflect the complexity or severity of actual psychological conditions. This limitation affects both model training and evaluation, potentially introducing systematic biases or misalignment with real-world mental health assessments. Future work should consider

incorporating clinician-labeled datasets or partnerships with healthcare providers for ground truth validation [15].

6.4.2 Ambiguity in Linguistic Signals and Class Boundaries

The subjective and context-dependent nature of mental health discourse poses intrinsic challenges for classification. As observed in Section 5.3, substantial semantic overlap exists between adjacent classes—particularly medium and high risk—which often results in borderline or ambiguous predictions. Despite employing class-weighted loss and careful model calibration, these ambiguities persist due to the limitations of text-only features in capturing psychological nuance. This highlights the need for context-aware, multi-modal inputs in future systems [36, 72].

6.4.3 Class Imbalance and Data Scarcity

Despite strategic mitigation via stratified k-fold validation and weighted loss functions, the original dataset remained imbalanced, with relatively fewer low-risk posts. This imbalance likely impacted the model's ability to generalise to the underrepresented class, as evidenced by reduced precision in certain evaluations (Section 5.2). Furthermore, Reddit posts tend to be highly variable in length, tone, and topic specificity, contributing to sparse or noisy signals in some instances. Although preprocessing reduced this variability to an extent, the lack of domain-specific filtering may have influenced model stability [50, 49].

6.4.4 Interpretability Constraints and Limited Visualisations

While efforts were made to enhance interpretability through architectural simplification and error analysis (Section 6.3), this study did not integrate advanced interpretability tools such as SHAP, LIME, or Integrated Gradients. As a result, token-level explanations of individual predictions remain absent, limiting the model's transparency in high-stakes applications [59, 60]. Additionally, several visualisations proposed for confidence scores and length-based behavioural trends were not included due to time and resource constraints.

6.4.5 Generalisability and Platform Bias

The model was trained and tested exclusively on Reddit data, specifically from mental health-related subreddits. While Reddit provides a rich source of anonymous discourse, its user base, posting norms, and language use are not representative of broader populations or platforms. As such, the model's generalisability to other forums, support systems, or languages remains untested [4]. Future evaluations on cross-platform data will be necessary to assess external validity and cultural adaptability.

6.4.6 Inference-Only Limitations in Real-World Settings

Although the model was successfully deployed in an inference pipeline (Section 5.6), real-world applications require continuous feedback, error monitoring, and integration with human oversight. In its current form, the system lacks online learning capabilities, user feedback interfaces, and alert mechanisms—features that are essential for responsible deployment in mental health contexts.

6.4.7 Ethical Considerations in Deployment

Ethical considerations—particularly data privacy, user anonymity, and the responsible use of model outputs—remain critical for future deployment. Any real-world application of this system should adhere to institutional ethical review protocols, ensure compliance with data protection regulations, and incorporate safeguards to prevent misuse, overreach, or misinterpretation of mental health risk predictions [62, 63].

Summary of Evaluation. Overall, the proposed transformer-based model shows strong potential for early-stage mental health signal detection in online environments. Its robust architecture, consistent performance across validation and unseen data, and interpretable decision-making pipeline support its practical viability. However, limitations related to label validity, interpretability tools, and platform generalisability must be addressed before deployment in sensitive, real-world contexts. These findings lay the foundation for the concluding chapter, which discusses the broader impact, potential real-world applications, and future development directions for this system.

CHAPTER 7

Conclusion and Future Work

7.1 Summary of Key Contributions

This dissertation presents a comprehensive investigation into the use of transformer-based Natural Language Processing (NLP) techniques for the detection of mental health risk levels in Reddit support group posts. Through the design, development, and evaluation of multiple machine learning models—including a customised DistilBERT architecture—the study contributes meaningfully to the growing field of AI-assisted mental health surveillance.

The following key contributions were achieved:

- **Development of a Scalable Labelling Strategy:** A risk-labelling pipeline was implemented by leveraging a pretrained emotion classification model and mapping emotion categories to predefined risk levels (low, medium, high). This approach enabled scalable annotation of large, unlabelled Reddit datasets while reducing reliance on costly manual or clinical labelling [10, 11].

- **Design of a Custom Transformer Model:** A lightweight yet effective architecture based on DistilBERT was proposed, with enhancements such as Mean+Max pooling, a GELU-based projection layer, Multi-Sample Dropout (MSD), label smoothing, and class-weighted loss. This configuration significantly outperformed baseline models, particularly in detecting high-risk cases [21, 75, 25, 28, 56].

- **Robust Training and Evaluation Pipeline:** A rigorous experimental framework was adopted, including 5-fold stratified cross-validation and a final held-out test set. This ensured reliability and generalisability of results while mitigating risks of overfitting and class imbalance [76, 50].

- **Deployment-Ready Inference on Unseen Data:** The trained model was applied to newly collected Reddit posts from the same subreddits, simulating real-world deployment. Results confirmed stable behaviour, with a distribution of predicted risk levels closely aligned with test set outcomes [13].

- **Error and Behavioural Analysis:** The study offered qualitative and quantitative insights into model decision-making, revealing consistent patterns in misclassification and identifying heuristic features (e.g., post length, tone) that influenced predictions, aiding interpretability and guiding future refinements [55].

- **Ethical and Practical Considerations:** Ethical implications were discussed for data collection, user anonymity, and responsible model deployment—critical for real-world use, particularly in clinical or public health contexts [40, 63, 62].

In summary, this work demonstrates the technical feasibility of transformer-based mental health risk detection, while addressing methodological, ethical, and practical challenges in deploying such systems at scale.

7.2 Potential Real-World Applications

The outcomes of this research indicate promising avenues for practical mental health monitoring and support systems. Although developed and tested with Reddit data, the model's architecture and outputs lend themselves to broader applications. Potential use-cases include:

Digital Mental Health Surveillance: Government and non-profit health organisations could monitor population-level mental health trends across online platforms. Aggregating risk level predictions over time and geography may help identify emerging patterns of distress or crisis and enable proactive responses [19, 4].

Triage Support in Online Counselling Services: Online therapy platforms and mental health helplines could integrate automated risk classifiers as a triage layer to prioritise users exhibiting high-risk linguistic signals. Human clinicians or moderators would retain decision authority while benefiting from faster routing of urgent cases [42, 65].

Platform Moderation and Safety Monitoring: Community platforms could flag potentially harmful content for review. Subreddits dedicated to self-help or crisis support might benefit from automated alerts to moderators when users exhibit signs of distress, enhancing safety without breaching privacy [64, 65].

Academic and Clinical Research: Researchers in psychology, linguistics, or sociology may use the model to structure large-scale analyses of emotional expression, social support mechanisms, or mental health discourse patterns across demographics and time [13, 74].

Real-Time Analytics Dashboards: With front-end visualisations, the model could power dashboards for educators, public health officials, or NGOs to track mental health signals in near real-time, supporting data-informed policy-making and community interventions [19].

Support for Chatbots and Companion Apps: Conversational AI systems could use risk classifiers to adjust responses or escalate critical cases to human intervention—adding emotional intelligence and ethical oversight to automated systems [66, 67].

Each application requires safeguards for ethical compliance, transparency, and user consent, but the findings here show that carefully designed and validated NLP models can meaningfully support mental health infrastructure across digital ecosystems.

7.3 Future Directions: Real-Time Systems, API Integration, and Multimodal Modelling

Building on the outcomes and limitations of this study, several future directions are recommended to advance deployment and utility.

Real-time monitoring and feedback systems: Extending the system for streaming environments (e.g., chat forums) would enable instantaneous risk assignment and timely interventions, especially in crisis contexts (e.g., helplines, moderation dashboards). Integrating alert thresholds, rate-limited notifications, and longitudinal logging can surface behavioural escalations and support sustained monitoring [68, 44].

API-based model deployment and integration: Deploying the model as a REST API (e.g., FastAPI/Flask) would allow third-party integration. Responses should include risk predictions with confidence scores, class probabilities, and (when available) token-level attributions. A modular service architecture facilitates versioning, model updates, access control, and audit trails [69].

Cross-platform generalisation and domain adaptation: To reduce platform-specific bias and improve demographic coverage, future work should adapt to additional communities (e.g., Twitter, Tumblr, Discord, online therapy transcripts). Domain-adaptive pretraining or adversarial multi-domain fine-tuning can improve out-of-domain robustness [70, 71].

Incorporation of multimodal signals: Mental health signals span text, images, audio, and behavioural metadata. Future systems can fuse modalities such as visual content, posting rhythms, engagement metrics, or voice cues. Late-fusion and multimodal transformers offer promising blueprints for richer, context-aware assessments [73, 72, 36].

Clinical validation and expert review: Formal evaluation against clinician-labelled datasets and collaboration with mental health professionals are essential. Expert-in-the-loop annotation, rubric refinement, and adjudication can align model outputs with clinical constructs and improve reliability over time [42, 63].

Integration of explainability and transparency tools: Incorporating SHAP, LIME, or Integrated Gradients will provide granular explanations of predictions, aiding trust, safety review, and regulatory compliance. Explainable outputs also support training for moderators and clinicians who rely on transparent decision support [59, 60].

7.4 Concluding Remarks

By combining scalable, emotion-informed labelling, tailored transformer architectures, robust evaluation, and forward-looking deployment practices, this dissertation demonstrates that customised transformer models can reliably classify mental health risk in Reddit posts and lays practical groundwork for ethically aligned, real-world, human-in-the-loop systems. Real-world impact will depend on clinical partnerships, cross-platform validation, and explainable, governance-aligned, human-centred design.

Bibliography

[1] World Health Organization. (2022). *World mental health report: Transforming mental health for all*. WHO.

[2] Patel, V., Saxena, S., Lund, C., et al. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157), 1553–1598.

[3] Chancellor, S., Lin, Z., Goodman, E., Zerwas, S., & De Choudhury, M. (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *CSCW*, 1171–1184.

[4] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *Current Opinion in Behavioral Sciences*, 18, 43–49.

[5] Chancellor, S., Baumer, E., & De Choudhury, M. (2021). Who is the "Human" in Human-Centered ML? The case of mental health and social media datasets. In *ACM CHI* (position/survey style chapter).

[6] Gkotsis, G., Oellrich, A., Hubbard, T. J. P., et al. (2017). Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7, 45141.

[7] Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *EMNLP*, 2968–2978.

[8] Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health trajectories. *Computers in Human Behavior*, 111, 106436.

[9] Benton, A., Mitchell, M., & Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. In *EACL*, 152–162.

[10] Saravia, E., Liu, H., Huang, Y., Wu, J., & Chen, Y. (2018). CARER: Contextualized affect representations for emotion recognition. In *EMNLP*, 3687–3697.

[11] Baziotis, C., Nikolaos, A., & Narayanan, S. (2018). NTUA-SLP at SemEval-2018 Task 1: Predicting affect in tweets with deep attentive RNNs. In *SemEval 2018*, 245–255.

[12] Shen, X., Radhakrishna, A., & Rose, C. (2022). Emotion signals and risk mapping for mental health triage (position paper). *Workshop on Computational Linguistics and Clinical Psychology*.

[13] Cohan, A., Yates, A., Feldman, S., & Goharian, N. (2018). Mental health classification on social media. In *EMNLP*, 3017–3027.

[14] Tsakalidis, A., Liakata, M., et al. (2018). Building and evaluating resources for mental health text classification on social media. *LREC Workshop on Social Media Mining for Health*.

[15] Resnik, P., et al. (2015). The University of Maryland CLPsych 2015 shared task system. In *CLPsych*, 54–60.

[16] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5), 649–685.

[17] Sadeque, F., et al. (2019). Measuring the Latent Psychosocial Effects of Mental Health Discourse on Social Media. In *WWW Companion*, 1461–1467.

[18] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186.

[19] Guntuku, S. C., Sherman, G., Stokes, D. C., Agarwal, A. K., Seltzer, E., Merchant, R., & Ungar, L. H. (2020). Tracking mental well-being of populations using social media: A review. *Current Opinion in Behavioral Sciences*, 18, 89–95.

[20] Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, A., Guntuku, S. C., & Schwartz, H. A. (2019). Suicide risk assessment with multimodal modeling on Reddit. In *CLPsych 2019 Shared Task*, 46–54.

[21] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT: A distilled version of BERT. *arXiv:1910.01108*.

[22] Zhang, Y., & Zhang, M. (2022). A systematic comparison of classical and transformer models for mental health NLP. *Journal of Biomedical Informatics*, 134, 104155.

[23] Ji, Z., et al. (2021). RoBERTa-based stress and anxiety detection in online discussions. *IEEE Access*, 9, 40533–40542.

[24] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, 3982–3992.

[25] Inoue, H. (2019). Multi-sample dropout for accelerated training and better generalization. *arXiv:1905.09788*.

[26] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *ACL*, 328–339.

[27] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., & Yan, S. (2020). Improving BERT with domain adaptation and pseudo-labeling (uses EMA training). *arXiv:2004.03186*.

[28] Müller, R., Kornblith, S., & Hinton, G. (2019). When does label smoothing help? In *NeurIPS*, 4696–4705.

[29] Bennett, J., & Hauser, K. (2013). Statistical evaluation of k-fold cross-validation for model selection. *ACM KDD Tutorial Notes*.

[30] Zhou, Z., et al. (2020). Evaluation metrics for imbalanced clinical NLP. *JMIR Medical Informatics*, 8(7), e17823.

[31] Kshirsagar, R., et al. (2022). Explainable mental health classification on social media: A survey. *ACM Computing Surveys*, 55(8), 1–36.

[32] Mohammadi, E., et al. (2022). Explainable AI for mental health on social media: Methods and findings. *IEEE Access*, 10, 11623–11645.

[33] Nguyen, T., et al. (2020). Interpretable deep learning for mental health text classification. In *AAAI Workshops*.

[34] Wolohan, J. T. (2018). Detecting linguistic traces of depression and self-harm on Reddit. In *CLPsych*, 7–15.

[35] Gjurković, M., & Šnajder, J. (2018). Reddit: A gold mine for text mining (dataset and preprocessing practices). In *Conference on Language Technologies*, 132–136.

[36] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor Fusion Network for multimodal sentiment analysis. In *EMNLP*, 1103–1114.

[37] Boe, B. (2023). PRAW: The Python Reddit API Wrapper (v7+). https://praw.readthedocs.io.

[38] Google Developers. (2024). Using OAuth 2.0 to access Google APIs (Best practices). https://developers.google.com/identity/protocols/oauth2.

[39] Reddit Developer Docs. (2023). API terms, rate limits and developer guidelines. https://www.reddit.com/dev/api/.

[40] Fiesler, C., & Proferes, N. (2018). Participant perceptions of Twitter research ethics. *Social Media + Society*, 4(1), 1–14.

[41] Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). Expert, crowdsourced, and machine annotation of suicide risk in social media. In *CLPsych*, 25–36.

[42] Milne, D. N., Pink, G., Hachey, B., & Calvo, R. A. (2016). CLPsych 2016 shared task: Triaging content in online peer-support forums. In *CLPsych*, 118–127.

[43] Kumar, S., et al. (2022). Emotion-to-risk mapping for mental health triage. *ACM CHIL*, 190–199.

[44] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *ICWSM*, 128–137.

[45] Losada, D. E., & Crestani, F. (2016). A test collection for research on depression and language use. In *CLEF* (pp. 28–39).

[46] Lin, Z., Feng, M., Santos, C. N. dos, Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv:1703.03130*.

[47] Howard, J., & Ruder, S. (2018). ULMFiT for text classification. In *ACL*, 328–339.

[48] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *CVPR*, 2818–2826.

[49] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449.

[50] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in CNNs. *Neural Networks*, 106, 249–259.

[51] Haider, S. F., Husseini Orabi, A., Syed, S., & Inkpen, D. (2020). Multimodal mental health analysis in social media. *PLOS ONE*, 15(4), e0232002.

[52] Arango, A., Pérez, J., & Poblete, B. (2019/2020). Hate speech detection is not as easy as you may think: A closer look at model validation. In *SIGIR*, 45–54.

[53] Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. *International Conference on Learning Representations (ICLR)*.

[54] Sun, C., Qiu, X., & Huang, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentences. *arXiv:1903.09588*.

[55] Shen, T., Radhakrishna, A., & Rose, C. (2020). Why misclassify? Analyzing error propagation in neural sentiment analysis. In *EMNLP*, 1290–1302.

[56] Zhou, Z.-H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Systems, Man, and Cybernetics—Part A*, 36(1), 1–13.

[57] Vig, J. (2019). A multiscale visualization of attention in the transformer model. *arXiv:1906.05714*.

[58] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. In *NAACL Clinical NLP Workshop*, 72–78.

[59] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *KDD*, 1135–1144.

[60] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *NeurIPS*, 4765–4774.

[61] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Interpreting pretrained context representations via "probing". In *EMNLP*, 188–196.

[62] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *ACM FAccT*, 610–623.

[63] Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *Swiss Medical Weekly*, 148, w14571.

[64] Vidgen, B., Harris, A., Hale, S., et al. (2019). Challenges and frontiers in abusive content detection and moderation. In *WWW Companion*, 143–153.

[65] Benton, A., Mitchell, M., & Hovy, D. (2021). Human-in-the-loop considerations for ML systems in mental health contexts. *Patterns*, 2(8), 100275.

[66] Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering CBT to young adults with symptoms of depression and anxiety using a conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19.

[67] Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, AI mental health chatbot (Wysa) for digital well-being: Real-world data. *JMIR mHealth and uHealth*, 6(11), e12106.

[68] Xu, Z., et al. (2020). Real-time social media monitoring for mental health signals. *IEEE Access*, 8, 153094–153103.

[69] Tan, J., et al. (2021). Deploying ML models as REST APIs for mental health applications: An engineering perspective. *SoftwareX*, 13, 100642.

[70] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, 8342–8360.

[71] Wadhwa, S., et al. (2021). Cross-platform transferability between Reddit and Twitter for mental health classification. In *ICASSP*, 7728–7732.

[72] Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal Transformer for unaligned sequence-to-sequence learning. In *ACL*, 6558–6569.

[73] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2), 423–443.

[74] Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 Shared Task: Predicting suicide risk and mental illness from Reddit data. In *CLPsych*, 34–44.

[75] Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *arXiv:1606.08415*.

[76] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, 1137–1145.

[77] Savani, B. (2020). DistilBERT emotion classifier (Hugging Face): `bhadresh-savani/distilbert-base-uncased-emotion`. https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion.