Precision versus Efficiency: A Quantized DistilBERT Framework for Automated Mental Health Text Triage

Ayaan Goswami
Carnegie Vanguard High School
Houston, Texas, USA
ayaangoswami17@gmail.com

Clayton Greenberg
Saarland University
Saarbrücken, Germany

Abstract

This study addresses the challenge of automatically triaging online mental health data by engineering a text-analysis system capable of accurately and efficiently categorizing human psychological distress in a lightweight, web-deployable format. As more people turn to social media and other online platforms when experiencing psychological distress, for both interaction with others and to seek support, the volume of such disclosure increasingly outpaces the capacity of moderators and clinicians to review it, forming a bottleneck to accurate and efficient triaging. To overcome the bottleneck, this study proposes an automated classification system, deployed as a web application where users submit text directly. Within this approach, an optimization framework was utilized, targeting the classification head, encoder, inference runtime, and numeric precision of a bidirectional deep learning transformer (BERT). Because inference cost is dominated by the encoder rather than the classifier, optimizing the classification head yielded no speedup, so the encoder itself was replaced with a distilled model and quantized. The final framework was able to successfully categorize raw user text into seven different psychological states, at an overall accuracy of about 82% on unseen test data. Additionally, the distilled and quantized model measurably reduced response time, lowering latency from 186.4 to 43.8 ms and model size from 438.7 to 68.3 MB, with per-class performance reported for every configuration. The result quantifies what each optimization costs: the runtime change is numerically lossless but contributes under 10% of the speedup, distillation and quantization supply the rest, and quantization damages the smallest classes severely for one encoder while leaving the other's per-class F1 intact. The deployed system is offered as a research prototype for prioritizing human review of written text, not as a diagnostic or standalone screening tool.

Keywords

Natural Language Processing, Text Classification, Mental Health Triage, Machine Learning, BERT, Optimization, Sentiment Analysis

IJSCAR VOL. 3, ISSUE 3

ACM Reference Format:
Ayaan Goswami and Clayton Greenberg. 2026. Precision versus Efficiency: A Quantized DistilBERT Framework for Automated Mental Health Text Triage. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 3).* ACM, New York, NY, USA, 15 pages. https://doi.org/10.67149/yhjs2024.5/5wcbbc4m

1 Introduction

1.1 Context and Background

In recent years, the overlap between digital communication and mental health has become a critical focal point for both psychological research and computational linguistics. Individuals who experience psychological distress increasingly turn to online platforms and social media to facilitate social interaction, access peer support, and articulate their struggles to seek mental health help [13]. Dizavandi and colleagues define mental health triage as the process of "assessing whether the patient... needs mental health care [and assessing] what level of urgency [they] fall into" [2]. However, the large amount of digital mental health data has created a severe bottleneck in mental health triage, as the volume of online disclosure far outpaces the capacity of moderators and clinicians to review it. Because of this bottleneck, "rapidly and accurately assessing clinical characteristics to determine the severity of psychiatric emergencies [has become] one of the most critical initial steps to providing appropriate treatment" [19].

To meet the demand for increasingly rapid and accurate assessments, efforts have been made to develop automated tools, such as artificial intelligence (AI) models, for mental health analysis and triage. However, rapid automated tools that can accurately filter and prioritize massive volumes of unstructured text to identify who needs immediate help remain scarce, and there is a further gap in balancing the accuracy and efficiency of such tools, since the most accurate tools are the most computationally expensive to run, and the cheapest tools have subpar accuracy, making both types unsatisfactory for deployment. These gaps lead to the core research question driving this study: under a limited computational budget, how can an automated text-analysis system be optimized to categorize human psychological distress accurately and efficiently in a web-deployable format, and what does each optimization cost?

1.2 Service and Target Audience

To address the research question, this project presents a web-based automated triage application designed to analyze user-inputted text and classify its underlying psychological state. The core users of this application include platform moderators who desire automated safety filters to flag high-risk posts, clinical organizations looking for preliminary evaluation tools, and general users looking for a quick reflection of how an automated classifier reads their own text. Users submit text to the application directly. The system does not monitor or collect posts from social media, and is engineered to minimize the computational cost of each request so that it remains responsive on low-cost hosting.

1.3 Problem Formulation

This study is structured as a supervised classification problem, wherein the system is trained on over 42,000 samples of complex language data from dedicated online mental health support communities [18]. This input consists of raw, unstructured human text, with the output being a set of categorical labels predicting one of seven discrete mental health statuses. The system is intended for risk screening rather than diagnosis, and the labels it predicts are inherited from the categories under which the training data was originally collected rather than from clinical assessment.

2 Dataset

2.1 Dataset Source

This study utilized the public "Sentiment Analysis for Mental Health" dataset compiled by Suchintika Sarkar on Kaggle [18], containing 53,043 samples of raw textual data compiled from social media platforms and public forums such as Reddit and Twitter. The dataset has two primary variables: "statement" (the independent feature), the raw user-generated text to be parsed by the pipeline, and "status" (the dependent feature), a categorical label mapping each statement into one of seven mutually exclusive classes: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality disorder.

The corpus is an aggregation of nine previously published datasets: 3k Conversations Dataset for Chatbot, Depression Reddit Cleaned, Human Stress Prediction, Predicting Anxiety in Mental Health Data, Mental Health Dataset Bipolar, Reddit Mental Health Data, Students Anxiety and Depression Dataset, Suicidal Mental Health Dataset, and Suicidal Tweet Detection Dataset. Each was scraped and labeled independently under its own scheme. The compiler applied no unified annotation pass across them and does not document how each source maps onto the seven classes. The seven statuses were therefore inherited from the source files rather than assigned by clinicians, and the dataset carries no diagnostic assessment, annotation codebook, or inter-annotator agreement statistic. The labels indicate the category under which a statement was originally collected, which correlates with psychological state but does not measure it.

2.2 Preprocessing

Missing data (NaN values) were removed, eliminating 362 rows and leaving 52,681 usable samples. Next, the "status" labels were converted to numerical IDs using label encoding. The dataset was then divided into a 20% testing set and an 80% training set, producing 42,144 training samples and 10,537 testing samples. As seen in Figure 1, the dataset exhibits a clear class imbalance, with Normal (≈31.0%) and Depression (≈29.2%) comprising the majority of entries, followed by Suicidal (≈20.2%), while the specialized clinical conditions of Anxiety, Bipolar, Stress, and Personality disorder together account for under 20%. Because of the extreme class imbalance, overall accuracy is weakly informative on this task, and all results are additionally reported per class. To account for the imbalance, stratified sampling was utilized to maintain the original proportional distribution of each status across both the train and test sets. Finally, a pre-trained Hugging Face tokenizer was applied to the "statement" column, truncating long text to 128 tokens, padding shorter sequences to a uniform length, and generating the input and attention-mask tensors the model requires [20]. The 128-token cap was selected as a deployment tradeoff, since sequence length drives inference cost quadratically in the attention mechanism. Under this cap, 33.5% of statements are truncated, and no ablation at longer sequence lengths was run. Section 3 returns to what this choice may cost in accuracy.

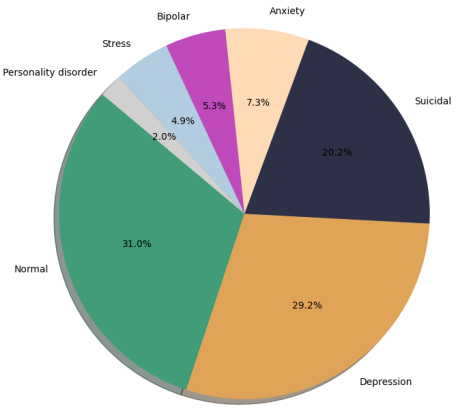

Figure 1: Distribution of Mental Health Conditions in the Sarkar (2024) Dataset.

2.3 Corpus Quality

Because the corpus is an aggregation of independently collected sources, it was audited for duplication and label consistency before any model was trained. Statements were normalized by lowercasing, removing non-alphanumeric characters, and collapsing whitespace. Under this normalization, the 52,681 usable rows contain 50,960 unique statements, so 1,721 rows (3.27%) are duplicates of another row. In addition, 1,453 distinct statements appear more than once after

Table 1: Label pairs in conflict among repeated statements. Depression/Suicidal is the most frequent disagreement in the corpus.

Conflicting labels	Statements	Rows
Depression / Suicidal	17	35
Anxiety / Depression	9	24
Normal / Stress	2	4
Depression / Normal	1	2
Depression / Normal / Stress / Suicidal	1	13
Total	30	78

normalization. A further 494 statements normalize to fewer than ten characters, and three normalize to the empty string.

A portion of the corpus is not mental health disclosure at all: the most common repeated entries include "what do you mean" (22 occurrences), "why not" (16), and a study recruitment advertisement appearing 10 times under the Stress label, none containing clinical content.

Of the 1,453 repeated statements, 30 (2.06%) carry conflicting labels, affecting 78 rows. The most extreme case is a statement normalizing to "name," which appears 13 times under four different statuses, including one instance labeled Suicidal. This is a lower bound on label noise rather than an estimate of it, since disagreement is only observable where a statement happens to repeat, and repeats are unlikely to be representative. What the audit establishes is that the mapping from text to label is not deterministic.

The distribution of conflicting labels, however, is more informative than their frequency. As shown in Table 1, 17 of the 30 conflicting statements (57%), covering 35 of the 78 affected rows, are disagreements between Depression and Suicidal, likely due to their high level of comorbidity [8]. This double-labeling of statements could be indicative that some Depression/Suicidal confusion when modeling may reflect inconsistency in the supervision rather than error by the model.

2.4 Train/Test Leakage

Because duplicates are distributed across the corpus, stratified splitting can place copies of the same text in both partitions, inflating measured performance. The test set was therefore audited for overlap with the training set. A test row counted as leaked if its normalized text exactly matched a training row, or if its maximum character 3–5-gram TF-IDF cosine similarity against the training set reached 0.90. Cosine similarity over-fires on very short texts, where character overlap is high but content unrelated ("very beautiful" and "is it beautiful" score highly while sharing little substance), so near-duplicate matches were confirmed only when the pair additionally shared at least half of its words and the test statement exceeded 30 normalized characters.

Under this rule, 517 test rows are exact duplicates of training rows and a further 101 are confirmed near-duplicates, giving 618 leaked rows, or 5.87% of the test set. Omitting the word-overlap confirmation raises the count to 705 rows (6.69%), so that step removes 87 short-text collisions. The estimate is insensitive to the threshold: varying the cosine cutoff from 0.80 to 0.99 moves the leakage rate only between 6.24% and 5.22%.

Table 2: Train/test leakage by class. Contamination concentrates in the smallest classes, and the Suicidal class is least affected.

Class	Test rows	Leaked	Leaked %
Personality disorder	215	54	25.12
Stress	517	93	17.99
Bipolar	556	97	17.45
Anxiety	768	76	9.90
Depression	3081	134	4.35
Normal	3269	125	3.82
Suicidal	2131	39	1.83
Overall	10537	618	5.87

Leakage is distributed unevenly across classes, as depicted in Table 2. The three smallest classes are most contaminated, with roughly a quarter of Personality disorder test items appearing in training. Suicidal, the class of greatest consequence for triage, is least contaminated at 1.83%, so performance on the safety-critical class is not materially inflated by memorization.

2.5 Consequences for Evaluation

Because leakage concentrates in the smallest classes, it distorts macro-averaged metrics far more than overall accuracy. Across the configurations evaluated in Section 5, removing leaked rows lowers accuracy by 0.30 to 0.46 percentage points but lowers macro-F1 by 1.48 to 1.96 points, an effect roughly four to five times larger. Overall accuracy is therefore doubly misleading on this corpus, as it under-weights the minority classes and conceals most of the inflation that duplication introduces.

All accuracy, macro-F1, and per-class results reported in this paper are computed on the leakage-filtered test set of 9,919 samples, with full test set figures given alongside for comparability with prior work. The calibration analysis of Section 5.8 is the one exception, and is reported on the full test set for the reason given there. The training set was left intact, as deduplicating it would alter both the size and class distribution of the training data, yielding models different from the ones deployed, whereas filtering the test set corrects the measurement without changing what is measured. Duplication internal to the training set may still encourage memorization that a test-side filter cannot detect, so this correction is partial.

3 Related Work

3.1 Classification on the Sarkar Corpus

The dataset has been analyzed in several prior publications, allowing this work to be positioned against reported

results on identical data. Kallstenius et al. compared three approaches on 52,681 statements: a traditional pipeline combining TF-IDF features with a support vector machine, a prompt-engineered GPT-4o-mini, and a fine-tuned GPT-4o-mini [10]. The TF-IDF pipeline achieved the highest overall accuracy at 95%, the fine-tuned model reached 91%, and prompting alone reached 65%. Rehman et al. paired several embedding schemes with recurrent and transformer architectures on 51,074 entries of the same corpus, reporting 95.71% for BERT with Word2Vec embeddings and 94.47% for RoBERTa with FastText, against 81.08% for a bidirectional LSTM [16].

The accuracy reported in this study is substantially lower than the strongest of these figures, a difference that warrants explanation. Several methodological differences prevent direct comparison. Kallstenius et al. augmented the corpus by back-translating statements through French without stating whether augmentation preceded the train/test split; if it did, augmented variants of the same statement could appear on both sides of it. That study also describes the corpus as 52,681 unique statements, whereas the normalization in Section 2.3 finds 50,960 distinct statements among those rows. Rehman et al. removed exact duplicates before modeling but report no split ratio and do not examine near-duplicates. Neither study reports the sequence-length limit used, which materially affects both accuracy and inference cost. A further difference is internal to this work: the 128-token cap truncates the longest statements in the corpus, and no ablation at longer sequence lengths was run, so some part of the gap may be attributable to this deployment-motivated choice rather than to methodology in the compared studies.

This work therefore does not claim to improve on the results of the literature. What the comparison shows is that neither study reports a near-duplicate audit or any latency, throughput, or model size figure, and only Kallstenius et al. report per-class metrics. The two axes central to this paper, per-class behavior under compression and measured inference cost, are therefore absent from prior work on this corpus, as is any quantification of the leakage it contains.

3.2 Data Quality in Social Media Corpora

The audit reported in Section 2.4 follows an established line of work rather than introducing a new method. Elangovan et al. showed that overlap between training and test partitions in public NLP datasets inflates reported results, so measurements interpreted as generalization partly reflect memorization [3]. Mu et al. examined twenty datasets widely used in computational social science, found most contain duplicate and near-duplicate samples producing label inconsistencies and leakage, and concluded that duplication affects existing claims of state-of-the-art performance [12]. Their finding that few dataset papers document any deduplication step is consistent with what is observed here. This study applies those methods to a corpus actively used for mental health classification, while establishing baselines that account for inference cost and compression.

3.3 Efficient Transformer Inference

Reducing the cost of transformer inference is a well-developed area, as seen in the literature surrounding this topic. Knowledge distillation is a method that trains a smaller student model to reproduce the behavior of a larger teacher. In the case of BERT models, DistilBERT retains most of BERT's performance with roughly 40% fewer parameters, and is accordingly the distilled encoder used in this study [17]. Quantization reduces the numeric precision of weights and activations, typically from 32-bit floating point to 8-bit integers, which shrinks the model and allows integer arithmetic at inference [9]. Zafrir et al. applied 8-bit quantization to BERT specifically, reporting a fourfold reduction in model size with limited accuracy loss [21].

Prior work on this corpus has appealed to efficiency without measuring it: Kallstenius et al. argue their pipeline is preferable partly because it is smaller and faster than a large language model, but report no latency, throughput, or model size [10]. This study treats these as measurements rather than assumptions, and reports them for every configuration.

3.4 Per-Class Effects of Compression

Compression can damage some classes far more than aggregate metrics suggest. Hooker et al. found that pruned and quantized image classifiers retain comparable overall accuracy while diverging sharply on a narrow subset of inputs, with the cost falling disproportionately on underrepresented parts of the distribution [5]. They report quantization as considerably less damaging than pruning, but the risk remains that 8-bit quantization alone can cost a minority class substantial performance while overall accuracy barely moves.

4 System

4.1 Text Representation

Since the data were raw textual statements, an algorithm was needed to translate them into a mathematical format a computer can process. Several were tested during development, ranging from word-frequency methods such as Bag-of-Words and TF-IDF to neural models such as BERT [1, 11].

The first approaches tested were frequency-based. A Bag-of-Words (BoW) model counts how often each unique word in the training set appears and classifies new inputs on those frequencies. TF-IDF (Term Frequency-Inverse Document Frequency) refines this by weighting each term frequency by the logarithm of its inverse commonness, reducing the impact of filler words such as "and" and "the." Both are computationally inexpensive, but are rigid and context-free, unable to account for word order, syntax, or negation.

To overcome these limitations this study used neural network-based techniques, specifically embeddings and BERT. Where TF-IDF assigns static values to words, embeddings represent them as high-dimensional vectors that capture conceptual relationships rather than frequencies. BERT (Bidirectional Encoder Representations

from Transformers) was chosen for its bidirectionality and contextual awareness: unlike unidirectional models it reads the words before and after a token to interpret its use, and its attention mechanism weights each word by those around it. These features make it far more nuanced than BoW or TF-IDF, but also much less efficient, which constrains its deployment on the low-cost hardware available to this project.

4.2 Optimization Axes

BERT's efficiency limitation was therefore treated as an optimization problem rather than an architectural one. A fine-tuned BERT model with its native classification head serves as the reference configuration, and four independent axes were identified along which its inference cost might be reduced: the classification head, the encoder, the inference runtime, and the numeric precision of the weights. Each axis was varied separately so its contribution could be attributed, and every configuration was evaluated under the protocol in Section 4.4.

The first axis is the classification head. An earlier version of this system replaced BERT's linear output layer with an optimized Random Forest classifier, using BERT solely as a static feature extractor over 768-dimensional embeddings. Optimizing this axis cannot reduce inference cost, because every input must still pass through the full encoder to produce that embedding, and substituting 300 decision trees for a single linear layer adds work rather than removing it. Section 5 reports the measured outcome. The native head was retained and optimization was directed at the encoder instead.

The second axis is the encoder itself. Two options were tested. The first was to skip fine-tuning and extract embeddings from a pre-trained encoder, which removes the cost of task-specific training. However, as Section 5 shows, this also substantially degrades the model's accuracy. The second was to replace the encoder with a smaller one, namely the released DistilBERT checkpoint described in Section 3 [17]. Rather than distilling from the BERT model directly, this study fine-tuned that checkpoint on the mental health corpus.

The third and fourth axes concern how the encoder is executed rather than what it contains. Both encoders were exported to the ONNX interchange format and served through ONNX Runtime, replacing the PyTorch execution path without altering any weights. Each was then quantized dynamically to 8-bit integers, a standard technique that reduces model size and permits integer arithmetic at inference in place of floating-point operations [9, 21]. Dynamic quantization computes activation scales at run time, so its behavior depends on the composition of each batch. This property is examined in Section 5.

4.3 Model Selection Criteria

Compressing a model trades quality for cost, so a criterion was required to determine which configurations remained acceptable. To avoid selecting a threshold that favored a preferred outcome, this study adopted the tolerance used by the MLPerf Inference benchmark, which requires that an optimized submission achieve at least 99% of the accuracy of the unmodified fp32 reference model, with quantization the permitted form of optimization [15]. The same 99% relative tolerance is applied here in two places: to overall accuracy, and to recall on the Suicidal class. The second constraint reflects a failure asymmetry, in which failing to flag a user at risk is a more consequential error than raising a false alarm, and it prevents a configuration from qualifying by trading away performance on the class that matters most for triage.

Among the configurations satisfying both constraints, the one with the lowest per-request latency was selected for deployment. These criteria were fixed before the configurations were compared.

4.4 Measurement Protocol

All configurations were measured under a single fixed protocol, since latency figures obtained under differing thread counts, batch sizes, or hardware are not comparable. Every measurement was taken on the CPU, matching the deployment environment of Section 6, on an Intel Core i7-14650HX with the thread budget of both PyTorch and ONNX Runtime pinned to two threads to approximate the allocation available to the hosted application. Inputs were truncated to 128 tokens. Accuracy was computed over the complete hold-out test set in batches of 32, while latency was measured one request at a time, matching the way the deployed application serves traffic. Batching of 32 was adopted for tractability; Section 5.9 shows the fp32 figures are insensitive to this choice, while quantized BERT varies by 2.55 accuracy points across batch sizes and the deployed model by 0.28. Each configuration was timed over 300 single requests drawn from the test set, discarding the first 10 as warm-up, and the reported figure is the mean, in ms. Throughput was instead measured under batched inference at batch size 32, since a deployed service amortizes cost across concurrent requests, so it is not the reciprocal of the single-request latency and the two columns are not directly comparable. The small reversals between the fp32 PyTorch and ONNX rows of Table 4 fall within the run-to-run variation of that measurement and should not be read as an ordering. Model size is that of the deployed artifact, comprising weights, configuration, and tokenizer, excluding optimizer state and training checkpoints.

Absolute latencies are specific to this processor, which combines performance and efficiency cores and is subject to thermal limits under sustained load, but comparisons between configurations hold because all were measured on the same machine under identical settings.

Table 3: Per-class performance of the deployed quantized DistilBERT model on the leakage-filtered test set ($n = 9{,}919$).

Class	Precision	Recall	F1	Support
Anxiety	0.8342	0.8873	0.8599	692
Bipolar	0.8416	0.8105	0.8257	459
Depression	0.7766	0.7665	0.7715	2947
Normal	0.9479	0.9612	0.9545	3144
Personality disorder	0.7000	0.6522	0.6752	161
Stress	0.6986	0.7217	0.7100	424
Suicidal	0.7228	0.7103	0.7165	2092
Macro average	0.7888	0.7871	0.7876	9919
Accuracy			0.8231	9919

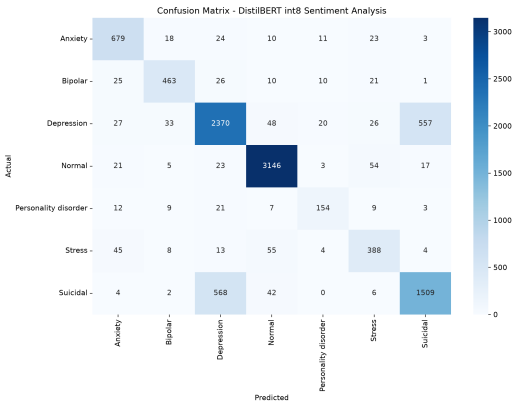

Figure 2: Confusion Matrix of the Deployed Quantized DistilBERT Model.

5 Results and Discussion

All results in this section are reported on the leakage-filtered test set of 9,919 samples described in Section 2.4, with full test set figures given alongside where comparability with prior work matters. The calibration analysis of Section 5.8 is the single exception and is reported on the full test set. Unless stated otherwise, every configuration was measured under the protocol in Section 4.4.

5.1 Model Performance

The deployed quantized DistilBERT model achieved an overall accuracy of 82.31% on the leakage-filtered evaluation set and 82.65% on the full test set, where accuracy is the proportion of correctly predicted statuses out of all predictions. Because overall accuracy is weakly informative under the class imbalance of Section 2.2, precision, recall, and F1 are reported for every class in Table 3, giving a macro-averaged F1 of 0.7876. Figure 2 shows the confusion matrix, with the principal diagonal representing correct predictions.

The model was reasonably accurate across all statuses, with the most prominent confusion coming from mixing up depression and suicidal. The encoder relies on contextual embeddings, and highly similar contexts mapping to distinct but related psychological states blur its decision boundaries. As reported in Section 2.3, this same pair accounts for 57% of the labeling conflicts in the corpus itself, so part of the confusion likely reflects inconsistency in the supervision rather than model error.

Of the 2,092 suicidal statements, 553 were classified as depression and 41 as normal. Section 7.3 takes up the difference between those two failures.

5.2 Hyperparameter Selection

The model was tuned using the Hugging Face "TrainingArguments" module over 3 epochs to prevent overfitting on the 42,144 training samples. A learning rate of 2e-5 ensured stable gradient descent, training and evaluation batch sizes were set to 16, a weight decay of 0.01 penalized overly-large weights, and 16-bit (FP16) mixed precision training accelerated training. The same settings were used for DistilBERT, so any difference between the encoders is attributable to the encoder rather than its training schedule.

5.3 Comparison with Frequency Baselines

To establish what a transformer actually purchases, the Bag-of-Words and TF-IDF approaches of Section 4 were evaluated under the same protocol, and appear in the frequency block of Table 4. The strongest frequency baseline, TF-IDF with logistic regression, reaches 77.58% accuracy and 0.6883 macro-F1 against 82.31% and 0.7876 for the deployed model. The transformer buys 4.73 percentage points of accuracy and 9.93 of macro-F1, at roughly 45 times the latency and 15 times the size.

The gap is wider on the minority classes than the overall figures suggest: on Personality disorder the best baseline reaches 0.5135 F1 against 0.6752, and on Stress 0.4384 against 0.7100, so the advantage of contextual embeddings concentrates in the classes with the least training data.

5.4 Optimization Results

Table 4 reports every configuration retained for deployment consideration, and the two abandoned axes are reported in Section 5.5. Exporting either encoder to ONNX Runtime is numerically lossless, reproducing the PyTorch predictions exactly, and reduces latency by only 5-9%. Quantizing to 8-bit integers reduces it by a further factor of 2.02 for both encoders. The speedup is therefore attributable almost entirely to the encoder and the numeric precision, not the runtime.

Applying the selection criteria of Section 4.3, the reference is fine-tuned BERT in fp32, at 0.8312 accuracy and 0.7132 Suicidal recall. Quantized BERT reaches only 96.33% of reference accuracy and is rejected, while DistilBERT in fp32 reaches 98.73% and is also rejected. Quantized DistilBERT reaches 99.02% of reference accuracy and 99.60% of reference Suicidal recall, satisfying both constraints, and is the fastest to do so. It was therefore deployed, reducing latency from 186.4 ms to 43.8 ms and size from 438.7 MB to 68.3 MB.

Table 4: All configurations retained for deployment consideration. The two abandoned optimizations are reported in Section 5.5. Accuracy, macro-F1, and Suicidal recall are reported on the leakage-filtered test set. Latency, throughput, and size are measured under the protocol of Section 4.4. The deployed configuration is highlighted in bold.

	Configuration	Acc.	Macro-F1	Suic. R	ms/req	req/s	MB
freq.	BoW + logistic regression	0.7608	0.6841	0.6711	0.84	9694.9	4.2
	TF-IDF + logistic regression	0.7758	0.6883	0.6788	0.98	9055.9	4.6
	TF-IDF + multinomial naive Bayes	0.6509	0.4209	0.4990	1.03	9684.0	7.4
BERT	fp32, PyTorch	0.8312	0.7964	0.7132	186.4	3.8	438.7
	fp32, ONNX Runtime	0.8312	0.7964	0.7132	170.3	3.6	439.2
	int8, ONNX Runtime	0.8007	0.7406	0.7787	84.4	7.0	111.4
DistilBERT	fp32, PyTorch	0.8206	0.7839	0.7424	93.2	7.5	268.8
	fp32, ONNX Runtime	0.8206	0.7839	0.7424	88.4	7.0	268.9
	int8, ONNX Runtime	**0.8231**	**0.7876**	**0.7103**	**43.8**	**13.1**	**68.3**

The margin is narrow. The deployed configuration clears the 99% threshold by 0.02 percentage points on the filtered test set and by 0.08 on the full one, but a paired bootstrap with 5,000 resamples places the 95% confidence interval on that ratio at [98.31%, 99.74%] and the probability that it exceeds the threshold at 0.51. The criterion is therefore met at the point estimate but not robustly, and the separation between the quantized and unquantized DistilBERT configurations should not be treated as established.

One consequence of this criterion deserves comment. The 99% tolerance is applied to overall accuracy, a metric this paper elsewhere argues is weakly informative under class imbalance. It is retained as the primary gate because it is externally specified rather than chosen here, and because a criterion defined on a single class is trivially gameable. The cost is visible: quantized BERT attains the highest Suicidal recall of any configuration evaluated, 0.7787, yet is rejected for falling to 96.33% of reference accuracy, and its expected triage cost is the worst of the four transformer configurations at 0.3843, because that recall gain is purchased by damage to Personality disorder and Stress. A rule keyed to Suicidal recall alone would therefore have deployed the model with the worst aggregate cost-sensitive behavior. This tension between a class-specific safety metric and an aggregate one is not resolved here, and a principled multi-objective criterion is left to future work.

5.5 Optimizations That Did Not Succeed

Two of the four axes were investigated and abandoned, both evaluated under the canonical configuration and on the leakage-filtered test set. Replacing the native head with an optimized Random Forest trained on the fixed 768-dimensional embeddings from the fine-tuned model's [CLS] token produced essentially the same accuracy, 0.8300 against 0.8312, and 1.67 points lower Suicidal recall. It did not reduce inference cost: with the embedding already computed, evaluating the 300 class-balanced decision trees of depth 25 is more expensive than a single 768-to-7 linear layer, and both are negligible beside the encoder forward pass preceding them [14].

The second abandoned optimization was to skip fine-tuning and extract embeddings from a pre-trained encoder, removing the cost of task-specific training. This proved far more damaging than any compression step, reducing accuracy by 18.24 points, macro-F1 by 35.36, and Suicidal recall by 33.41. Together these results locate both the cost and the capability in the encoder: the head is too cheap for its replacement to matter, and the fine-tuned representations too important to discard. The only remaining way to reduce cost without destroying accuracy is a smaller encoder.

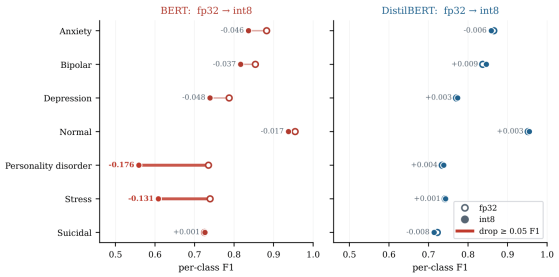

Figure 3: Change in per-class F1 under 8-bit quantization for both encoders, on the leakage-filtered test set. Hollow markers are fp32 and filled markers are int8. Drops of at least 0.05 F1 are highlighted. Quantization is near-lossless for DistilBERT but costs BERT heavily on the two smallest classes.

5.6 Per-Class Effects of Compression

Aggregate accuracy conceals the most important consequence of quantization. Figure 3 shows per-class F1 for both encoders before and after quantization.

Quantizing BERT costs 3.05 points of overall accuracy, which appears tolerable, but the loss is not evenly distributed: Personality disorder falls by 0.129 F1 and Stress by 0.122, while the largest class loses 0.015. The redistribution

is not uniformly negative: recall on Suicidal rises from 0.7132 to 0.7787, apparently because quantization noise shifts the Depression/Suicidal boundary toward the higher-risk class. The gain is paid for by the two smallest classes and by a 3.05-point accuracy loss, which is why the configuration is nonetheless rejected. The identical procedure applied to DistilBERT leaves per-class F1 essentially unchanged and in several classes marginally improves it, so two configurations separated by 2.24 points of overall accuracy differ by 12.9 points of F1 on the smallest class. The exception is Suicidal recall, which falls from 0.7424 to 0.7103, but is offset in F1 as precision rose to compensate. However, the recall loss is still the safety-relevant quantity and is taken up in Sections 5.7 and 7.3.

This extends the findings reviewed in Section 3: 8-bit quantization alone can remove a substantial fraction of a minority class's performance, and the damage depends on which architecture is quantized [5, 6].

5.7 Cost-Sensitive Evaluation

Overall accuracy also treats every error as equally serious, which is inappropriate for triage. An expected triage cost (ETC) was therefore computed by assigning a cost to each type of misclassification and averaging over the test set: classifying a suicidal statement as normal raises no flag at all, and is assigned 10; classifying it as another form of distress still routes the user to an elevated-risk category and is assigned 3; missing any other distress is assigned 3; confusing one distress category for another is assigned 1; and a false alarm on a normal statement is assigned 1. These weights encode an ordering over error types rather than a calibrated clinical cost, and were fixed before configurations were compared.

The deployed configuration scores 0.3533, against 0.3441 for the fp32 reference and 0.3323 for unquantized DistilBERT, the best transformer configuration. Quantized BERT is worst at 0.3843. The deployed model is therefore not what this metric would select, and the 41 suicidal statements it assigns to the normal class compare with 31 for unquantized DistilBERT. This is a tradeoff of the deployment decision: the criteria prioritize the latency benefits that make the system practical, subject to the accuracy and recall constraints, at a cost of 0.0210 ETC.

5.8 Calibration

Because the interface displays prediction probabilities to users, their reliability was evaluated on the full test set using expected calibration error (ECE), Brier score, and negative log-likelihood. This is the one analysis reported on the unfiltered set. Because the leaked rows are ones the model has effectively already seen, retaining them can only make stated confidence appear better matched to accuracy than it is, so the overconfidence reported below is a lower bound. Every configuration is overconfident, in that mean confidence exceeds accuracy, which Figure 4 shows as bars sitting below the diagonal across most of the confidence

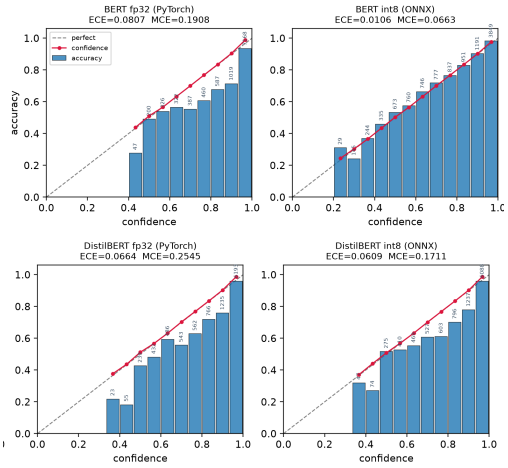

Figure 4: Reliability diagrams for each configuration on the full test set, 15 equal-width bins. Bars give observed accuracy, the line gives mean confidence, and the dashed diagonal is perfect calibration. Bars falling below the diagonal indicate overconfidence.

range. The deployed model assigns confidence of at least 0.99 to 38.2% of predictions while being correct on 82% of them, though its maximum confidence is 0.9992, so the interface never displays a probability of 100%.

Quantization leaves DistilBERT's calibration essentially unchanged, moving ECE from 0.0664 to 0.0609 and Brier score from 0.2510 to 0.2521. Its effect on BERT is far larger, reducing ECE from 0.0807 to 0.0106, but this does not indicate improvement, as ECE measures only whether stated confidence matches observed accuracy, and quantization noise flattens the confidence distribution of a previously overconfident model. Brier score and negative log-likelihood, which reward correct predictions as well as calibrated ones, rank quantized BERT last at 0.2763 and 0.5200. Selecting on calibration error alone would therefore have chosen the least accurate of the transformer configurations.

5.9 Robustness Checks

Three properties of the evaluation were examined to determine how far the reported figures depend on measurement choices. Dynamic quantization computes activation scales at run time from tensors whose padding depends on batch composition: across batch sizes of 1, 32, and 64, quantized DistilBERT varies by 0.28 accuracy points while quantized BERT varies by 2.55, so evaluation batching should match the serving configuration.

The same bootstrap distinguishes which reported differences exceed evaluation noise. The deployed model's accuracy deficit against the fp32 reference is reliable at -0.0082 (95% CI $[-0.0141, -0.0021]$), whereas its deficits on macro-F1 (-0.0088, CI $[-0.0182, +0.0006]$) and on Suicidal recall (-0.0029, CI $[-0.0192, +0.0132]$) are not distinguishable from zero. Compression is therefore measurably costly

in aggregate accuracy while leaving the safety-critical metric statistically indistinguishable from the reference.

The effect of the leakage documented in Section 2.4 was also examined, and is quantified there. The single exception is quantized BERT, which scores marginally lower on leaked rows than on unseen ones.

5.10 Analysis of Edge-Case Classifications

To evaluate the boundary limitations and pragmatic capabilities of the deployed system beyond static test metrics, a qualitative edge-case analysis was performed. Because a probe set written and interpreted by the same authors risks selecting for known weaknesses, the suite was authored externally. Two collaborators who were not involved in developing the system wrote fifty probes, ten for each of five linguistic phenomena, with the label an external human reader would assign. The suite was frozen before any item was run. All results were recorded, and are visible in Appendix A, which lists every probe.

The deployed model classified 24 of the 50 probes correctly, with accuracy varying sharply by phenomenon: 80% on hyperbole, 60% on minimized distress, 40% on figurative idiom, and 30% on both sarcasm and temporal blindness. Idiom, sarcasm, and temporal hedging all require the model to override the literal sentiment of the tokens present, whereas hyperbole is handled comparatively well because its exaggeration is lexically overt.

These failures share a direction, as of the 26 misclassified probes, 20 were assigned to the Normal class and none to Suicidal. The model therefore does not over-escalate on figurative language, it under-reacts to distress expressed indirectly. Nine of the ten sarcastic probes were classified as normal, seven of them incorrectly and at a mean confidence of 0.985, and every idiom failure went to normal. Temporal blindness fails in both directions, with four ongoing disclosures classified as normal and three resolved narratives still assigned a clinical label, indicating that the model keys on clinical vocabulary rather than on tense or trajectory. Minimized distress is handled better, at 60%, but two of its four failures still resolve to normal.

The safety consequence appears in the three probes whose intended label was Suicidal. All three were missed. A sarcastic disclosure of suicidal ideation was classified as normal at 0.919, a temporally hedged one as normal at 0.968, and a minimized one as depression, with 0.305 assigned to the Suicidal class, a distribution whose consequence for the crisis-resource threshold is traced in Section 7. The model is also confidently wrong rather than uncertain, as mean confidence on the twenty misclassifications that resolved to normal was 0.902, so these errors would not be caught by thresholding on the displayed probability.

These results identify a cost the aggregate metrics do not expose. The deployed configuration sits within 0.8 accuracy points of the fp32 reference and is statistically indistinguishable from it on Suicidal recall, yet answers fewer than half of a pragmatics suite correctly and fails in precisely the direction that matters for triage. With ten probes per category these rates carry wide confidence intervals and indicate which phenomena are difficult rather than precisely estimating performance. Pragmatic robustness is nonetheless a further axis, alongside per-class F1 and calibration, on which overall accuracy is uninformative.

5.11 Limitations

Several limitations must also be acknowledged. The training data was sourced from public social media platforms like Reddit and Twitter, biasing the model toward the demographic makeup of those platforms, often younger and internet-literate populations, and lowering its ability to generalize to older populations or to non-digital communication such as spoken transcripts. It is consequently suited to prioritizing which texts a human reviews first rather than to standalone screening. Section 7 sets out the intended human-in-the-loop configuration and the uses the system does not support. The model also classifies text into seven mutually exclusive categories, whereas clinical conditions are often comorbid. Displaying probabilities for all statuses partly addresses this, but the singular classification framework remains a limitation, as do the edge-case failures reported above. A further limitation concerns conversational context. Every item in the corpus is a single statement stripped of the thread it appeared in, and the deployed interface likewise classifies isolated text. A sentence such as “if that passes I’m going to throw myself off a bridge” is a disclosure of intent in one context and political hyperbole in another, and neither the training labels nor the classifier has access to the distinction. The system therefore cannot resolve context-dependent risk, a further argument for using it as a prioritization signal for human review rather than as a filter.

Three limitations concern the evaluation rather than the model. All figures come from a single training run on a single split, without cross-validation or repeated seeds. The bootstrap reported above quantifies the uncertainty arising from the finite test set, but it cannot capture the variance that retraining under a different seed would introduce, so the selection margin in particular should be read as provisional. The leakage audit removed contaminated rows from the test set but not from the training set, so duplication within training may still encourage memorization this study does not quantify. Finally, no concurrency or load testing was performed, so the latency figures describe single-request behavior on one processor.

6 Site Features

The web application is deployed as a two-tier system. The first tier is a static frontend hosted on GitHub Pages at https://jaukg9.github.io/mental-health-sentiment-analysis, while the second tier is a Python inference backend hosted on Hugging Face Spaces [4, 7]. The backend serves the quantized

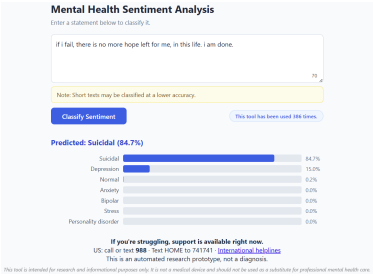

Figure 5: Visualization of a High Risk Classification Using the Website.

DistilBERT model of Section 5 through ONNX Runtime, requiring neither PyTorch nor Scikit-Learn at inference, keeping the deployed image small enough for the CPU-only tier. Every result is accompanied by a notice stating that the tool is a research prototype and not a diagnosis, and results identified as high risk additionally display crisis resources, as described in Section 7. Below are primary use cases that illustrate the platform's core functionality.

6.1 Self Reflection

The first use case supports individuals who wish to see how an automated classifier reads their own writing. A user is presented with a text-input module where they can submit unstructured personal logs, such as a journal entry, for example: "I'm confused, I'm not feeling good lately. Every time I want to sleep, I always feel restless." Upon submission, the application routes the text to the inference model. The interface then updates to display the predicted classification as a probability distribution over the seven categories, accompanied by a notice that it is a research prototype and not a diagnosis. As Section 7.5 sets out, the output is not intended as an assessment of the user's condition, and the self-reflection case is supported only in that weaker sense.

6.2 Digital Content Triage and Moderation

The application also functions as a triage tool for community moderators managing online social forums. A moderator who identifies a concerning post and needs a severity assessment inputs its raw text into the site, and if the model identifies severe emotional distress it outputs a high-risk categorization such as "Suicidal," letting the moderator prioritize the case and respond with emergency hotline numbers or platform-specific support protocols. An example of a high-risk classification can be visualized in Figure 5, which also shows the crisis resources the interface displays alongside such a result.

6.3 Clinical Pre-Screening

Finally, the platform can be used within a telehealth or clinical pre-appointment process. A patient may be prompted before a virtual session to write a brief summary of their current mental state, which the application processes into a preliminary risk category, allowing the provider to tailor their initial line of questioning. That category is a screening signal for the clinician, not a diagnosis, and is not shown to the patient as an assessment of their condition.

7 Ethics and Safety

A publicly deployed classifier that assigns mental health categories to text raises questions that performance metrics do not address. This section documents data provenance, the review status of the study, the asymmetry between kinds of error, the safeguards in the deployed application, and the uses for which the system is and is not intended.

7.1 Data Provenance and Consent

The corpus was not collected by the authors. It is a publicly released aggregation of nine previously published datasets, each scraped from public posts and forum comments by its original compiler, and no interaction with any participant took place at any stage of this work.

The people whose statements appear in the corpus wrote them for other readers, not for research. Public visibility is not consent to research use, and the individuals concerned were not asked and cannot practically be asked. Neither the compiler of the aggregated dataset nor those of its constituents document whether platform terms of service permitted automated collection and redistribution, or whether any opt-out was offered, so this study cannot verify the conditions under which its training data was obtained. No attempt was made to identify any author, link statements across posts, or recover metadata beyond the text and its label, none of which is present in the released dataset. The code, audit scripts, and edge-case suite supporting this study are available at https://github.com/JaukG9/mental-health-sentiment-analysis/.

7.2 Ethical Review

This study was not submitted for institutional review board approval. It involves no intervention, recruitment, or interaction with human participants, and uses only previously published public data, which ordinarily falls outside human-subjects review. That determination was made by the authors rather than a board, and the deployed application is a component a board might reasonably have examined, since it accepts text from the public.

The deployed application maintains a count of the number of classifications performed, which is not associated with any user. Submitted text is processed to produce a classification and is not retained or logged.

7.3 Failure Asymmetry

The errors this system can make are not equivalent: failing to flag a person at risk removes the possibility of a response, whereas flagging one who is not merely consumes reviewer attention. Overall accuracy treats the two as interchangeable, which is why per-class metrics are reported alongside it throughout Section 5. Of the 2,092 suicidal statements

in the filtered test set, the deployed model assigns 553 to depression and 41 to normal. The first group is misclassified but still receives an elevated-risk category and would be routed for review, but the second receives no flag at all, and represents 1.96% of suicidal statements. This is the failure mode that matters most, and the one the expected triage cost of Section 5.7 is weighted to penalize.

As Sections 5.4 and 5.7 establish, the deployed configuration is not the safest available: unquantized DistilBERT and quantized BERT each miss fewer suicidal statements, but neither satisfies the accuracy criterion of Section 4.3 at comparable cost. The deployed model was selected because it meets that criterion at less than half the inference cost, a trade defensible for a prototype on donated compute but one that must be acknowledged.

7.4 Deployment Safeguards

Every classification is presented with a notice stating that the tool is an automated research prototype and not a diagnosis, and a footer stating it is for research purposes, is not a medical device, and is not a substitute for professional care. High-risk classifications additionally display crisis resources beneath the result: the 988 Suicide and Crisis Lifeline, the Crisis Text Line (text HOME to 741741), and a link to international helpline directories, as shown in Figure 5.

Crisis resources are displayed whenever the model assigns a probability of at least 0.25 to the Suicidal class, rather than only when Suicidal is the top-ranked label. Because recall on that class is 0.7103, a rule keyed to the top-ranked label alone would leave roughly 29% of suicidal statements without crisis information. The probability threshold extends coverage to cases ranked as depression that still carry substantial suicidal probability.

The threshold narrows the gap without closing it. Of the three probes in the edge-case suite of Section 5 whose intended label was Suicidal, one clears 0.25: the minimized disclosure, at 0.305, would display crisis resources despite being ranked as depression. The other two do not, having been assigned suicidal probabilities below 0.01. Indirect disclosure expressed sarcastically or with temporal hedging produces confident normal predictions rather than uncertain ones, and no threshold on a probability the model does not assign can recover those cases. This is the clearest limitation of the deployed system, and argues for human review of flagged and unflagged text alike rather than reliance on the classifier as a filter.

7.5 Intended Use and Human Oversight

The system is intended to prioritize human attention, not replace it. Its outputs are screening signals indicating which texts a moderator, clinician, or support worker should read first, and every workflow in Section 6 places a person between the classification and any action affecting the individual concerned. It produces no diagnosis, and its categories are inherited from the collection conditions of the training corpus rather than from clinical assessment.

Several uses are outside what this work supports. The system should not inform consequential decisions about a person, including those concerning care not voluntarily sought, employment, education, or insurance; should not screen individuals without their knowledge; and its outputs should not be presented to a person as an assessment of their condition. It has not been validated against clinical outcomes, has not been evaluated beyond the English-language social media users in its training data, and its probabilities are systematically overconfident.

Finally, a tool that identifies distress at low cost can be turned toward ends its authors did not intend, including identifying vulnerable users for exploitation rather than support. It is released as a research artifact with that risk acknowledged, and the reporting here is intended in part to make its error characteristics legible to anyone evaluating it for deployment.

8 Conclusions

This study developed an automated natural language processing framework to detect and categorize mental health distress from unstructured text, treating deployment as an optimization problem over four axes: the classification head, the encoder, the inference runtime, and numeric precision. A BERT model fine-tuned on 42,144 social media samples served as the reference, and the deployed system classifies inputs into seven psychological states at 82.31% accuracy and 0.7876 macro-F1 on a leakage-filtered hold-out set.

Two axes yielded no benefit. Replacing the classification head with a Random Forest left accuracy unchanged while adding inference cost, since every input must still traverse the full encoder, and skipping fine-tuning cost 18.24 accuracy points. Replacing the encoder and quantizing it cut per-request latency from 186.4 ms to 43.8 ms and the deployable artifact from 438.7 MB to 68.3 MB, a 4.3× and 6.4× improvement over the fp32 BERT reference, while retaining 99.02% of the reference accuracy and 99.60% of its Suicidal recall. Distillation and quantization contribute comparably to that speedup, the distilled encoder halving latency and quantization halving it again (2.02×), while the runtime change alone accounts for only 5-9%.

Overall, this study demonstrates that a mental health triage system can be lightweight and affordable without fixating on a single accuracy metric. By prioritizing measurable efficiency, this framework produces a fast, deployable model, still with transparent strengths and limitations.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[2] Fatemeh Rajab Dizavandi, Razieh Froutan, Hossein Karimi Moonaghi, Abbas Ebadi, and Mohammad Reza. 2023. Mental Health Triage from the Viewpoint of Psychiatric Emergency Department Nurses; a Qualitative Study. *PubMed* 11 (01 2023), e70–e70. https://doi.org/10.22037/aaem.v11i1.2080

[3] Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 1325–1335. https://aclanthology.org/2021.eacl-main.113/

[4] GitHub. 2024. GitHub Pages Documentation. https://docs.github.com/en/pages.

[5] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2019. What Do Compressed Deep Neural Networks Forget? *arXiv preprint arXiv:1911.05248* (2019). https://arxiv.org/abs/1911.05248

[6] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising Bias in Compressed Models. *arXiv preprint arXiv:2010.03058* (2020). https://arxiv.org/abs/2010.03058

[7] Hugging Face. 2024. Hugging Face Spaces Documentation. https://huggingface.co/docs/hub/spaces-overview.

[8] Erkki Isometsä. 2014. Suicidal Behaviour in Mood Disorders—Who, When, and Why? *The Canadian Journal of Psychiatry* 59 (03 2014), 120–130. https://doi.org/10.1177/070674371405900303

[9] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2704–2713.

[10] Thomas Kallstenius, Andrea Johansson Capusan, Gerhard Andersson, and Adam Williamson. 2025. Comparing traditional natural language processing and large language models for mental health status classification: a multi-model evaluation. *Scientific Reports* 15, 1 (2025), 24102. https://doi.org/10.1038/s41598-025-08031-0

[11] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[12] Yida Mu, Mali Jin, Xingyi Song, and Nikolaos Aletras. 2024. Enhancing Data Quality through Simple De-duplication: Navigating Responsible Computational Social Science Research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 12477–12492. https://aclanthology.org/2024.emnlp-main.694/

[13] John A. Naslund, Ameya Bondre, John Torous, and Kelly A. Aschbrenner. 2020. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. *Journal of Technology in Behavioral Science* 5 (04 2020), 245–257. https://doi.org/10.1007/s41347-020-00134-x

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[15] Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, Brian Anderson, Maximilien Breughe, Mark Charlebois, William Chou, et al. 2020. MLPerf Inference Benchmark. In *ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. 446–459. https://doi.org/10.1109/ISCA45697.2020.00045

[16] Abdul Rehman, Muzamil Ahmed, Hikmat Ullah Khan, Amal Bukhari, Ali Daud, and Hussain Dawood. 2025. Mental Health Sentiment Analysis: Exploring an Optimized BERT with Deep Encodings. *Engineering, Technology & Applied Science Research* 15, 5 (2025), 26242–26248. https://doi.org/10.48084/etasr.10469

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019). https://arxiv.org/abs/1910.01108

[18] Suchintika Sarkar. 2024. Sentiment Analysis for Mental Health. https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health/data

[19] Samanvith Thotapalli, Musa Yilanli, Ian McKay, William Leever, Eric Youngstrom, Karah Harvey-Nuckles, Kimberly Lowder, Steffanie Schweitzer, Erin Sunderland, Daniel I Jackson, and Emre Sezgin. 2025. Potential of ChatGPT in youth mental health emergency triage: Comparative analysis with clinicians. *PubMed* 4 (09 2025), e70159–e70159. https://doi.org/10.1002/pcn5.70159

[20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[21] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8BERT: Quantized 8Bit BERT. In *Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2), NeurIPS Edition*. https://arxiv.org/abs/1910.06188

A Externally Authored Edge-Case Suite

This appendix lists the complete edge-case suite discussed in Section 5, together with a sample of ordinary hold-out items for comparison. The suite consists of fifty probes, ten for each of five linguistic phenomena, written by collaborators not involved in developing the system and frozen before any item was run. Probabilities are the softmax outputs of the deployed quantized model, computed one request at a time to match the serving configuration of Section 4.4. For each probe, P(int.) is the probability the model assigns to the label a human reader intends and P(pred.) the probability it assigns to the label it actually returns. The two coincide wherever the model is correct. The suite and the code that produced these results are available at https://github.com/JaukG9/mental-health-sentiment-analysis/.

Table 5: Full prediction probabilities on a sample of hold-out test items. The Suicidal item, classified as Depression, is the only misclassification in the sample and reflects the Depression/Suicidal confusion discussed in Section 5.

Input Text	Actual	Predicted	Prediction Probabilities	
If you really have to go to the ICU, it's okay, the important thing is to get better quickly and be healthy again	Normal	Normal	Normal	97.4%
			Anxiety	1.9%
			Stress	0.3%
			Bipolar	0.1%
			Depression	0.1%
			Suicidal	0.1%
			Personality disorder	0.0%
Housing crisis Global warming Lack of employment Coronavirus No future Despair Hopeless	Suicidal	Depression	Depression	90.3%
			Suicidal	8.2%
			Normal	1.1%
			Anxiety	0.2%
			Stress	0.1%
			Personality disorder	0.0%
			Bipolar	0.0%
Medication I haven't taken my medication in months because I lacked health insurance and could not find someone to sign off on prescriptions. I have insurance now and can afford my medication. Has any of you started medication after being off it for some time? What was it like? Should I be worried I'll have an episode?	Bipolar	Bipolar	Bipolar	89.9%
			Anxiety	9.0%
			Depression	0.6%
			Stress	0.2%
			Normal	0.1%
			Personality disorder	0.1%
			Suicidal	0.1%
If I worry about people, my chest always hurts. Please don't worry :(	Anxiety	Anxiety	Anxiety	98.8%
			Normal	0.4%
			Stress	0.4%
			Depression	0.2%
			Bipolar	0.1%
			Suicidal	0.1%
			Personality disorder	0.0%

Table 6: Externally authored edge-case suite, probes 1-25, with the deployed model's predictions. A tick marks agreement with the label a human reader assigns. Where the model is wrong, the gap between P(int.) and P(pred.) shows how confidently it prefers the wrong class. For probes whose intended label is Suicidal, P(int.) is the quantity governing the crisis-resource display described in Section 7.

Phenomenon		Probe	Intended	P(int.)	Predicted	P(pred.)
Hyperbole	✓	I'm so hungry I could eat a horse right now.	Normal	99.9%	Normal	99.9%
	×	This deadline is actually going to kill me, I've got way too much on my plate this week.	Stress	12.7%	Normal	68.5%
	✓	I laughed so hard I nearly died, that video was hilarious.	Normal	99.3%	Normal	99.3%
	✓	My heart is pounding out of my chest, I'm so nervous about this exam I could throw up.	Anxiety	99.2%	Anxiety	99.2%
	✓	I feel like the weight of the entire world is crushing me, I can't shake this sadness no matter what I do.	Depression	94.4%	Depression	94.4%
	×	I have a thousand things to do and only twenty-four hours in the day, I'm losing it.	Stress	13.6%	Depression	38.2%
	✓	That concert was insane, best night of my entire life, I'm still buzzing.	Normal	98.1%	Normal	98.1%
	✓	I'm so anxious about the flight tomorrow I feel like I'm going to explode.	Anxiety	99.3%	Anxiety	99.3%
	✓	I feel like I'm carrying around a two hundred pound weight of sadness that never goes away no matter how much I try to shake it off.	Depression	96.7%	Depression	96.7%
	✓	I have so much on my plate right now I feel like I'm about to combust, this week has been absolute chaos.	Stress	56.3%	Stress	56.3%
Idiom	✓	I killed two birds with one stone today, finished my report and picked up groceries.	Normal	99.0%	Normal	99.0%
	✓	That joke had me dying, so funny.	Normal	99.8%	Normal	99.8%
	×	I'm swamped at work this week, drowning in emails and back-to-back meetings.	Stress	0.4%	Normal	98.7%
	×	My stomach's in knots waiting for the test results to come back.	Anxiety	0.1%	Normal	99.8%
	×	I've been feeling down in the dumps for weeks now, nothing seems to lift my mood.	Depression	0.2%	Normal	98.7%
	✓	Break a leg tonight, you're going to do great.	Normal	99.8%	Normal	99.8%
	×	I'm running on fumes trying to juggle work and the kids this week.	Stress	0.1%	Normal	99.8%
	✓	I've been walking on eggshells all week worried about the layoffs at work.	Anxiety	98.6%	Anxiety	98.6%
	×	One week I'm on top of the world starting five new projects and the next I can't even get off the couch, it's like living with two different people in my head.	Bipolar	13.5%	Normal	65.5%
	×	I burn bridges the second I feel someone pulling away, better to torch it myself than wait around to get left, that's just how I'm wired I guess.	Personality disorder	0.5%	Normal	89.4%
Sarcasm	×	Oh great, another all-nighter, exactly what I needed with three deadlines this week.	Stress	0.1%	Normal	99.6%
	✓	Wonderful, it's raining again the one day I decided to walk to work.	Normal	99.9%	Normal	99.9%
	×	Fantastic, my hands are shaking again right before the presentation, love that for me.	Anxiety	0.1%	Normal	99.8%
	×	Great, another day where I can't find the motivation to get out of bed, really thriving over here.	Depression	0.2%	Normal	98.9%
	✓	Perfect, spilled coffee all over my shirt right before the meeting, off to a great start.	Normal	99.8%	Normal	99.8%

Table 7: Externally authored edge-case suite, probes 26-50.

Phenomenon		Probe	Intended	P(int.)	Predicted	P(pred.)
Sarcasm	×	Love how my inbox somehow doubled overnight, definitely going to be a relaxing week.	Stress	0.0%	Normal	99.7%
	×	Oh good, my mind is racing through worst case scenarios again at 2am, exactly what I needed.	Anxiety	0.1%	Normal	99.6%
	×	Sure, I'm 'doing great,' that's why I've cancelled plans with everyone three weekends in a row.	Depression	0.0%	Normal	99.8%
	×	Oh I'm SO fine, definitely not lying awake calculating how much easier it would be for everyone if I just wasn't here anymore, nothing to see here.	Suicidal	0.9%	Normal	91.9%
	✓	Great, another 'friend' who disappears the second things get hard, must be something about me right? Awesome pattern I've got going, really love collecting people just to lose them.	Personality disorder	49.6%	Personality disorder	49.6%
Temporal blindness	×	I used to struggle with anxiety attacks in college, but I haven't had one in years and feel completely fine now.	Normal	6.8%	Anxiety	86.8%
	×	I went through a rough depressive episode after my divorce, but that was five years ago and I'm honestly in a really good place now.	Normal	9.4%	Depression	46.3%
	✓	I've been anxious about this move for weeks now and it's still eating at me every single day.	Anxiety	98.7%	Anxiety	98.7%
	×	Ever since the new project started last month, I've been stressed out non-stop and it hasn't let up.	Stress	17.3%	Normal	81.1%
	×	I've felt this heavy sadness since the breakup two months ago and it still hasn't gone away.	Depression	1.9%	Normal	68.2%
	×	I had panic attacks daily back in high school, but that stopped once I graduated and I've been panic-free for almost a decade.	Normal	24.5%	Anxiety	41.7%
	✓	I had a manic episode last year and I'm still cycling between highs and crashes most months.	Bipolar	97.8%	Bipolar	97.8%
	✓	I used to have really bad social anxiety, but after a few years of practice I actually enjoy parties now.	Normal	98.9%	Normal	98.9%
	×	I told my therapist I 'used to' think about not waking up, but honestly some nights it's still there, I just don't say it out loud anymore.	Suicidal	0.1%	Normal	96.8%
	×	I used to tell people my relationships 'just didn't work out,' but looking back, every single one ended the same way, me either clinging so hard they ran or pushing them away before they could leave first, and honestly that hasn't really changed.	Personality disorder	9.2%	Normal	55.1%
Minimized distress	×	I've been swamped with work lately, staying late most nights and barely keeping up, but I think I just need a lighter week.	Stress	3.3%	Normal	95.9%
	×	I'm okay I guess, just haven't felt like myself in weeks and everything feels kind of pointless.	Depression	28.9%	Anxiety	39.1%
	✓	It's nothing, I just get really panicky before flights, my heart races and I can't sit still.	Anxiety	89.6%	Anxiety	89.6%
	×	No big deal, just been crying more than usual and struggling to get out of bed most mornings.	Depression	1.3%	Normal	96.7%
	✓	It's fine honestly, just thinking too much right now and feeling constantly worried about stuff.	Anxiety	98.9%	Anxiety	98.9%
	✓	Not a big thing, I just can't stop worrying about everything lately, my mind won't slow down.	Anxiety	98.8%	Anxiety	98.8%
	✓	It's nothing serious, I just go through these weeks where I barely sleep and work way too much, then crash hard after.	Stress	71.7%	Stress	71.7%
	✓	I'm managing, just feel numb most days and nothing really excites me like it used to.	Depression	57.5%	Depression	57.5%
	×	It's nothing really, I've just been having thoughts about wanting to end my life, but don't make a big deal out of it, I'll be fine.	Suicidal	30.5%	Depression	46.6%
	✓	I know this probably isn't a big deal compared to what others are going through, and I hate to take up anyone's time, but I've been feeling pretty overwhelmed lately.	Stress	78.3%	Stress	78.3%