Int. J. Sci. R. Tech., 2026 3(11)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

MindCare AI: Multi-Label Emotion Detection from Text for Mental-Wellbeing Awareness with English, Hindi and Marathi Support

Pratiksha Mogal*, Vishakha Lonkar, Disha Nimbalkar, Vedansh Soni, Dr. Rama Bansode

Department of Master of Computer Applications Progressive Education Society's Modern College of Engineering, Pune, Maharashtra, India.

ABSTRACT

Emotional distress often goes unnoticed, and text tools that turn what people write into emotion signals could support self-awareness, yet many such tools are English-only, give a single score, and say little about their errors, privacy or limits. This paper presents MindCare AI, an end-to-end prototype that predicts multi-label emotions (27 emotions plus neutral) from short text. A TF-IDF (1–2-gram) representation feeds 28 one-vs-rest logistic regressions trained on the official GoEmotions splits (43,410 / 5,426 / 5,427 comments), with a single decision threshold (0.45) chosen on validation data only. On the official test split the model reaches macro-F1 0.445, micro-F1 0.514, micro precision 0.430, micro recall 0.638, subset accuracy 0.276 and Hamming loss 0.0503; the top-scoring label is one of the true labels for 56.9% of comments. Per-label F1 ranges from 0.888 (gratitude) to 0.129 (relief), and eight of 28 labels stay below 0.30. The model over-predicts (1.73 labels per comment against 1.17 annotated) and assigns neutral to 31.6% of comments whose annotation is not neutral. The classifier is embedded in a Flask web application with registration, opt-in text storage, per-user history, a JSON API, an English / Hindi / Marathi interface with a small phrase-and-emoji normalisation lexicon, rule-based suggestions, a 10-item self-check questionnaire and a keyword-based crisis notice. These last four components have not been validated, and transformer baselines were not run. This is a single-seed study of a prototype; it is an awareness aid, not a diagnostic tool.

Keywords: emotion detection, multi-label classification, GoEmotions, TF-IDF, logistic regression, mental wellbeing, Hindi, Marathi, Flask

INTRODUCTION

Mental-health conditions are common: the World Health Organization estimated that almost one billion people were living with a mental disorder in 2019 [1], and India's National Mental Health Survey (2015–16) found that roughly one adult in ten had a current mental morbidity, with a wide gap between need and treatment [2]. People usually put their feelings into words, in messages, notes and posts, long before they meet a professional. Software that turns such text into emotion signals could help a person reflect on how they feel, but only if it is honest about what it can and cannot do.

Early emotion datasets and tools used a handful of basic emotions [3, 4]. GoEmotions [5] moved to 27 fine-grained emotions plus neutral, annotated on about 58,000 English Reddit comments, and allows several labels per comment. Fine-grained labels are hard to predict: many are close in meaning and annotators disagree, so even a fine-tuned BERT model is reported at a macro-F1 of only about 0.46 [5]. Practical tools add further problems: most work only in English, they seldom show the threshold or model version behind a prediction, and they rarely state clearly that an emotion label is not a diagnosis.

This paper documents a complete prototype, MindCare AI, and reports its evaluation on the official GoEmotions test split. Our contributions are modest and stated as such:

- a multi-label emotion classifier (TF-IDF 1–2-grams with 28 one-vs-rest logistic regressions) whose single threshold is chosen on validation data only, with a full per-label evaluation, a

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

threshold-sensitivity analysis and an audit of duplicate text across splits;

- a privacy-conscious Flask application: opt-in text storage, per-user isolation of results, deletion of single analyses, history or account, CSRF protection and a JSON API;
- a lightweight English / Hindi / Marathi layer (interface, phrase and emoji normalisation, self-check questionnaire, suggestions) and a keyword crisis notice, described as design components whose accuracy has not been measured; and
- an account of observed failure modes, including a defect we found in the order of the normalisation lexicon.

We do not claim state-of-the-art accuracy. The model is deliberately simple so that it is fast, inspectable and runs on a CPU, and transformer models were not evaluated (Section VI-G). Our official-split protocol matches that of the GoEmotions paper, but we did not re-run its models, so we make no head-to-head comparison.

II. PROBLEM STATEMENT AND OBJECTIVES

People find it hard to track their own emotional state over time, and services cannot read free text at scale. Automated emotion tools often (i) support English only, (ii) output one label or a polarity score, (iii) hide thresholds and model versions, (iv) keep user text by default, and (v) blur the line between emotional awareness and clinical diagnosis. The research therefore targets multi-label emotion prediction, transparent evaluation, privacy-conscious deployment, multilingual access and clear safety boundaries.

Objectives. The project has the following objectives, each described with the purpose it serves:

- **1. Multi-label emotion prediction.** Predict any subset of the 27 GoEmotions emotions plus neutral for a short text, so that mixed feelings are not forced into one class.
- **2. Transparent, leakage-aware evaluation.** Use the official train / validation / test splits, choose the threshold on validation data only, report per-label metrics and audit duplicate text between splits, so that the reported scores can be checked.
- **3. Class-imbalance handling.** Use balanced class weights so that rare emotions (for example grief, pride, relief) can be predicted at all.
- **4. Transparent, privacy-conscious output.** Show model scores, the threshold and the model version, store text only when the user opts in, and let users delete their data.
- **5. Usable, accessible workflow.** Provide a web application with English, Hindi and Marathi interface text, a self-check questionnaire, suggestions, a crisis notice and a printable report.
- **6. Planned baselines and ablations.** Specify the comparisons (Section VI-G) needed to show which design choices matter.

Objectives 1–5 are met at the implementation level. The held-out evaluation (Section IX) shows that objectives 1 and 3 are only partly achieved: the classifier recovers true labels often (recall 0.638) but with modest precision (0.430) and a macro-F1 of 0.445. The baselines and ablations of objective 6 have not yet been run.

III. LITERATURE SURVEY

A. Emotion taxonomies and datasets

Basic-emotion theories [3] and wheel models [4] underlie most emotion datasets. SemEval-2007 Task 14 [6] annotated news headlines, and CARER [7] labelled tweets with six emotions. GoEmotions [5] is larger and finer: 27 emotions plus neutral on Reddit comments, with multiple labels per comment, which is the dataset used here.

B. Bag-of-words and linear models

TF-IDF weighting [8] with a linear classifier is a classic, strong baseline for text. Lexicon methods such as the NRC word–emotion lexicon [9] and VADER [10] need no training but score words in isolation. Linear models over n-grams remain competitive for many text tasks and train quickly [11]. Multi-label problems are often handled by training one binary classifier per label (binary relevance, or one-vs-rest) [12], here with the liblinear solver [13].

C. Transformer models

Pre-trained transformers such as BERT [14], RoBERTa [15] and the smaller DistilBERT [16] are the usual route to higher accuracy on text classification. The GoEmotions authors fine-tuned BERT and

reported a macro-F1 of about 0.46 on the test split [5]. These models need more memory and compute than a linear model, which matters for a small CPU-only deployment.

D. Mental-health tools and screening instruments

Systematic reviews of mental-health chatbots find promising but uncertain evidence and stress safety [17]; Woebot is a well-known example of a conversational agent evaluated in a trial [18]. Validated screening questionnaires such as PHQ-9 [19] and GAD-7 [20] are the clinical reference for depression and anxiety symptoms. Work on predicting mental-health status from social-media text warns that validity, bias and ethics are often under-examined [21]. These warnings are why this project presents emotion labels as awareness signals and not as diagnoses.

E. Indian languages and code-mixing

Multilingual encoders such as XLM-R [22] and MuRIL [23] cover Hindi and Marathi, benchmarks such as GLUECoS [24] address code-switched text, and L3CubeMahaSent [25] provides Marathi sentiment data. We did not use any of these; our Hindi and Marathi support is a small lexicon in front of an English-trained model (Section VII-B).

F. Comparison of the surveyed approaches

Table 1 compares the surveyed approaches with the proposed system.

Table 1: ***Comparison of surveyed approaches with the proposed system.***

Approach	Task	Relation to this work
Lexicons (NRC, VADER) [9, 10]	Word-level emotion or polarity scores	Simplest alternative; not evaluated here
TF-IDF + linear models [8, 11]	Text classification	Method family of this work
Fine-tuned BERT-type models [14, 16]	Fine-grained emotion on GoEmotions	Likely stronger, heavier; not evaluated here
MuRIL, XLM-R [22, 23]	Multilingual and Indic text	Possible replacement for the lexicon layer
Mental-health chatbots [17, 18]	Conversational support	Different goal; this work is not conversational
This work	**GoEmotions, 28 labels, multi-label; web app**	**Single seed; wellbeing modules unvalidated**

IV. RESEARCH GAP

Fine-grained emotion classification on GoEmotions is not new, and this work does not claim a new task or model. It addresses several gaps that remain around such classifiers, each tackled here only at the system level.

The first gap is transparency of evaluation. Emotion tools are often described by one headline score. We use the official splits, choose the threshold on validation data only, report all 28 per-label results, show how sensitive the scores are to the threshold, and audit duplicate text across splits. The audit also corrected a figure in our own project notes (Section V-E).

The second gap is access and safety in an Indian setting. Most emotion tools take English input only. We add English, Hindi and Marathi interface text, a self-check questionnaire, suggestions and a crisis notice, with the openly stated caveat that the Hindi and Marathi handling is a small phrase lexicon and that none of these components has been validated.

The third gap is baselines and deployment evidence. Controlled comparisons against transformer models and per-label thresholds are specified in Section VI-G but have not been performed, and end-to-end latency has not been benchmarked, so the contribution of individual design choices remains unmeasured.

V. DATASET AND PREPROCESSING

A. GoEmotions

GoEmotions [5] contains English Reddit comments labelled with 27 emotions or neutral; a comment can carry several labels. We use the simplified release with its official train, dev and test files, each a tab-separated file of text, comma-separated label ids and comment id. The 28 labels are admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise and neutral. Each comment becomes a 28-dimensional multi-hot vector. Comments are short (mean 12.7 words in the

test split). No Hindi or Marathi data are used for training or evaluation. The pipeline is shown in Fig. 1.

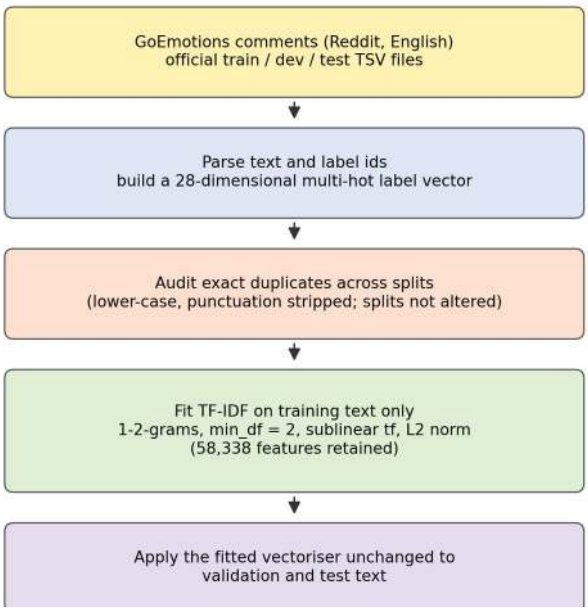


Figure 1: ***Data preparation pipeline for GoEmotions.***

B. Text representation

No stemming or stop-word removal is applied. The scikit-learn vectoriser lower-cases the text and uses its default token pattern, which keeps words of two or more characters and ignores punctuation and emoji. Consequently an emoji or an unmatched Devanagari word contributes nothing to the TF-IDF vector (Section IX-D shows the effect).

C. Official partition

Table 2: ***Official GoEmotions partition used in this study.***

Split	Comments	Role
Train	43,410	Fit TF-IDF and classifiers
Validation (dev)	5,426	Select the threshold
Test	5,427	Report final results once

Because the official splits are random samples of the same Reddit data, the test split measures performance on comments like the training comments. It does not measure performance on diary-style text, other platforms or other languages.

D. Class distribution

Labels are strongly imbalanced (Table 3). Neutral appears in about a third of the comments, whereas grief appears in 0.18% of training comments (77 of 43,410) and 6 test comments. Because a comment can have several labels, the percentages do not sum to 100. The three splits have very similar label profiles.

Table 3: ***Share of comments carrying each label (selected labels) and label cardinality.***

Label	Train (%)	Val. (%)	Test (%)
neutral	32.76	32.55	32.93
admiration	9.51	8.99	9.29
approval	6.77	7.32	6.47
annoyance	5.69	5.58	5.90
embarrassment	0.70	0.65	0.68
nervousness	0.38	0.39	0.42
pride	0.26	0.28	0.29
relief	0.35	0.33	0.20
grief	0.18	0.24	0.11

Label	Train (%)	Val. (%)	Test (%)
Labels per comment	1.18	1.18	1.17
Multi-label comments (%)	16.4	16.2	15.4

E. Duplicate-text audit

We compared lower-cased, punctuation-stripped comment text between splits. There are 98 distinct strings shared by train and validation, 90 by train and test, and 25 by validation and test. Counted per comment, 109 of the 5,427 test comments (2.0%) and 113 of the 5,426 validation comments (2.1%) have text that also occurs in the training split. Our project notes had described this overlap as about 0.2% of test examples, which is inconsistent with these counts (90 shared strings against 5,427 test comments is already 1.7%); the per-comment figure is 2.0%. The official splits were not altered. To gauge the effect, we re-scored the test comments whose text does not occur in training (n = 5,318): macro-F1 is 0.4426 and micro-F1 0.5118, against 0.4451 and 0.5140 on the full split, so the overlap changes the scores only slightly. A model re-trained on de-duplicated data is future work.

VI. PROPOSED METHODOLOGY

A. TF-IDF representation

Each comment is mapped to a sparse vector of 1–2-gram features. With sublinear term frequency, the weight of term t in document d is

$$w_{t,d} = (1 + \ln tf_{t,d}) \cdot idf_t,\ idf_t = \ln((1 + n)/(1 + df_t)) + 1 \quad (1)$$

and every vector is L2-normalised [8, 26]. Terms must occur in at least two training comments (min_df = 2); at most 200,000 features are kept, and 58,338 were retained. The vectoriser is fitted on training text only.

B. One-vs-rest logistic regression

For each label c a separate L2-regularised logistic regression gives a score

$$p_c(x) = 1 / (1 + \exp(-(w_c \cdot x + b_c))) \quad (2)$$

with C = 4.0, the liblinear solver and balanced class weights, so that within each binary problem errors on the rarer class are weighted more heavily. The model therefore has 28 × (58,338 + 1) = 1,633,492 parameters. Because of the class weights, the scores are not calibrated probabilities, and the application shows them as "model scores".

C. Decision threshold

A label is predicted when its score reaches a single global threshold τ:

$$\hat{y}_c = 1\,[p_c(x) \geq \tau] \quad (3)$$

τ was searched over 0.10, 0.15, …, 0.90 and chosen to maximise validation macro-F1, giving τ = 0.45. The test split played no part in this choice. The application also lists the top five scores for every comment, so a user still sees ranked evidence when no score reaches τ (this happens for 2.7% of test comments).

D. Evaluation metrics

Metrics are computed from the 28-column prediction matrix. For each label we report precision P = TP/(TP+FP), recall R = TP/(TP+FN) and F1 = 2PR/(P+R). Macro-F1 is the unweighted mean of the 28 per-label F1 values (neutral included) and is the headline figure because it weights rare labels equally; micro-F1 pools all decisions. Subset accuracy is the share of comments whose full label set is exactly right, Hamming loss is the share of wrong label decisions, and the top-1 hit rate is the share of comments whose highest-scoring label is one of the true labels.

E. Training configuration

Table 4: ***Training configuration.***

Item	Configuration
Libraries	scikit-learn 1.8.0, NumPy 2.4.4, pandas 3.0.2, joblib 1.5.3
Vectoriser	TfidfVectorizer: 1–2-grams, min_df 2, sublinear tf, max 200,000 features (58,338 kept)
Classifier	OneVsRest(LogisticRegression): C 4.0, liblinear, class_weight balanced, max_iter 300
Parameters	28 × (58,338 + 1) = 1,633,492
Threshold	Global τ, grid 0.10–0.90 (step 0.05), maximising validation macro-F1: τ = 0.45
Tuning	Only τ is tuned in the training script; no search over C, n-grams or weights is documented
Seed	42; single run, no confidence intervals
Artefact	Bundle (vectoriser, classifier, labels, τ, version) saved with joblib, 13.1 MB

F. Comparability

Our protocol (official splits, 28 labels, macro-F1) is the one used by the GoEmotions authors [5], who report a macro-F1 of about 0.46 for fine-tuned BERT. We did not re-run their model, and details such as the thresholding rule may differ, so the two numbers should not be ranked against each other. An earlier configuration of ours (different hyper-parameters, $\tau = 0.55$), as recorded in the project notes and not re-run for this paper, had test macro-F1 0.4544, micro-F1 0.5157 and subset accuracy 0.2898. Differences of about 0.01 between settings are of the same size as the validation–test gap (Section IX-C) and should not be over-interpreted.

G. Planned baselines and ablations

The most important next step is a stronger evaluation: several seeds or bootstrap intervals, and the comparisons in Table 5 under the same protocol. Only experiments actually performed will be reported.

Table 5: ***Planned baselines and ablations (not yet run).***

Configuration	Question answered
A. Proposed (TF-IDF 1–2-grams, balanced LR, global τ)	Reference
B. Same without class weights	Value of balanced weights
C. Unigrams only	Value of bigrams
D. Per-label thresholds tuned on validation	Value of label-specific τ
E. Linear SVM or Naive Bayes	Choice of linear learner
F. DistilBERT / RoBERTa fine-tuned on GoEmotions	Gap to transformers
G. Multilingual encoder on labelled Hindi / Marathi text	Accuracy of the lexicon layer
H. Training on de-duplicated data	Effect of duplicate text

VII. WELLBEING AND SAFETY COMPONENTS

A. Transparent output

For every analysis the application shows the predicted labels, the top five scores, the threshold and the model version, and states that scores are model outputs and not probabilities of having a mental-health condition. We did not implement any attribution method. Because the classifier is linear, the n-grams that drive a prediction could be read from its weights; this is future work and is not claimed here.

B. Hindi, Marathi and emoji normalisation

Before classification, 36 Hindi, Marathi and emoji entries (for example words for “happy”, “sad”, “worried”, “thank you” and several emoji) are replaced by English words by plain substring replacement, and the English-trained model sees the result. This is not translation. Entries are applied in dictionary order, and we found that seven of the 36 keys can never match because an earlier, shorter key is a substring of them and replaces it first. For example, in the Hindi phrase meaning “I am happy”, the short key for “khush” fires first and inserts the English word “happy” but leaves a stray vowel sign behind. The prediction in that example is still sensible, but the defect is real; applying longer keys first would remove it.

C. Wellbeing self-check

The questionnaire has 10 items answered on a four-point frequency scale (0–3) over the last two weeks, in English, Hindi and Marathi. Eight items follow the wording style of PHQ-9 [19] items 1–8 (interest, mood, sleep, energy, appetite, self-view, concentration, restlessness) and two follow GAD-7 [20] items 1–2 (nervousness, uncontrollable worry); the PHQ-9 item on thoughts of self-harm is not included. The total (0–30) is placed in one of five bands (0–4, 5–9, 10–14, 15–19, 20–30) that reuse the PHQ-9 severity cut-points. Because the item set and the 30-point total are new, these bands have no validated clinical meaning, and the page says that it is not a diagnosis.

D. Suggestions

Suggestions are rule-based text in three languages for seven emotions (sadness, anger, fear, nervousness, joy, gratitude, neutral), chosen in a fixed priority order from the predicted labels; other labels, such as grief or disappointment, fall back to generic text unless a covered emotion is also predicted. The wording has not been reviewed by a clinician.

E. Crisis notice

A crisis notice is shown when the raw text contains one of nine English phrases (for example “want to die” or “self-harm”). The classifier has no self-harm label, since GoEmotions has none, and the check does not cover Hindi or Marathi, paraphrases or mis-spellings. It will also fire on text that merely discusses the topic. It is a minimal safeguard and not a risk assessment.

VIII. SYSTEM DEPLOYMENT

The application is written with Flask [27] and SQLAlchemy over SQLite. The trained bundle is loaded once and reused. Fig. 2 shows the emotion pipeline and Fig. 3 the application components.

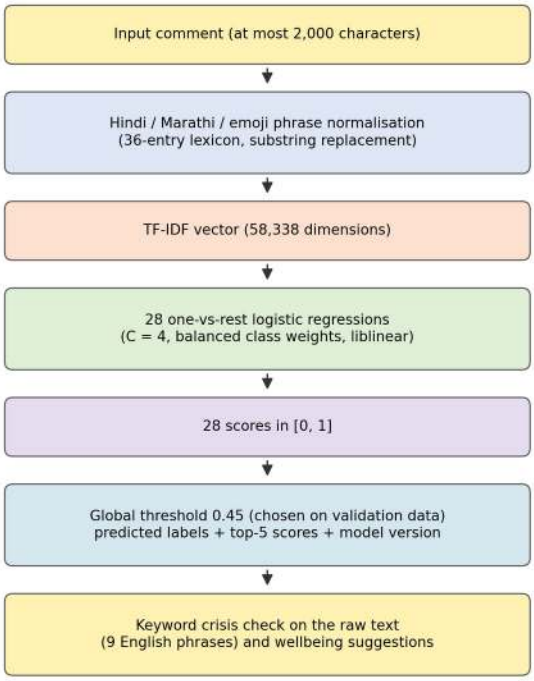


Figure 2: ***Emotion pipeline: normalisation, TF-IDF, 28 one-vs-rest classifiers and a global threshold.***

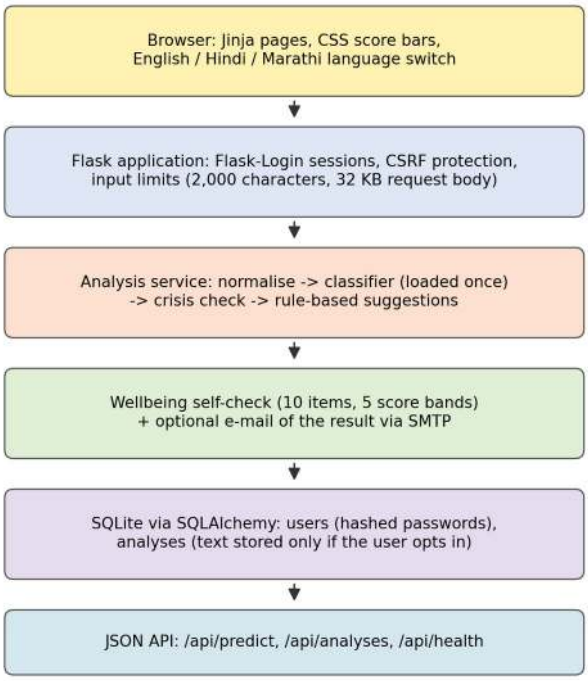


Figure 3: ***System architecture of MindCare AI.***

Accounts and data. Users register with a username (3–40 letters, digits or underscores), an e-mail address and a password of at least eight characters; passwords are stored as Werkzeug hashes. Input is limited to 2,000 characters and 32 KB per request. A saved analysis stores the predicted labels, the scores, the top emotion, the model version and an optional note of up to 300 characters; the text itself is stored only if the user ticks a box. Results are visible only to their owner, and a user can delete one analysis, clear the history or delete the account. State-changing forms use CSRF tokens, and the stateless prediction endpoint is exempt from them. A report page can be printed to PDF from the browser. The self-check result can be e-mailed to the account address when SMTP settings are present in the environment; otherwise nothing is sent.

Interface. Pages cover home, analysis, dashboard, history, detail, report, settings, about and the self-check, with English, Hindi and Marathi interface text selectable per session. The JSON API offers /api/predict, /api/analyses and /api/health. The about page lists the limits of the tool and points to the national Tele-MANAS service [28]; the contact details must be verified before release.

Testing and speed. The project includes 16 automated tests covering the label file and training-set size, model output format and edge inputs (unseen tokens, very long text), the crisis check, language selection, API validation (status codes 422, 400 and 413), authentication, password hashing, per-user isolation, deletion, CSRF enforcement and an end-to-end form flow. In a micro-benchmark of the classifier alone (200 repeated single-comment predictions, one process), a prediction took about 1.8 ms and loading the bundle about 0.45 s; these figures depend on the machine and exclude web-server overhead, so end-to-end latency and memory use have not been benchmarked.

Privacy. All inference runs locally; no text is sent to a third-party AI service. Even so, storing emotional text is sensitive, and any real deployment must follow applicable data-protection law, such as India's Digital Personal Data Protection Act, 2023 [29].

IX. EXPERIMENTAL RESULTS

A. Official test split

Table 6 lists per-label results on the 5,427 test comments at $\tau = 0.45$. Macro-F1 is 0.445 (macro precision 0.391, macro recall 0.534); micro-F1 is 0.514 with micro precision 0.430 and micro recall 0.638. Subset accuracy is 0.276, Hamming loss 0.0503, and the top-scoring label is a true label for 56.9% of comments. Re-scoring the saved model on the test split reproduced these values exactly.

Table 6: ***Per-label results on the official test split (sorted by test F1; supp. = test comments with the label; last column = validation F1).***

Label	P	R	F1	Supp.	Val. F1
gratitude	0.857	0.920	0.888	352	0.889
amusement	0.709	0.879	0.785	264	0.746
love	0.685	0.849	0.758	238	0.753
neutral	0.545	0.770	0.638	1787	0.621
remorse	0.511	0.839	0.635	56	0.747
fear	0.581	0.692	0.632	78	0.591
admiration	0.538	0.736	0.622	504	0.661
joy	0.419	0.689	0.521	161	0.533
optimism	0.437	0.629	0.515	186	0.551
sadness	0.397	0.571	0.468	156	0.453
grief	0.429	0.500	0.462	6	0.400
pride	0.600	0.375	0.462	16	0.533
disgust	0.414	0.512	0.458	123	0.372
surprise	0.361	0.518	0.426	141	0.465
desire	0.379	0.470	0.419	83	0.475
anger	0.343	0.520	0.414	198	0.420

Label	P	R	F1	Supp.	Val. F1
curiosity	0.317	0.518	0.394	284	0.393
confusion	0.262	0.536	0.352	153	0.345
disapproval	0.242	0.457	0.316	267	0.345
excitement	0.236	0.456	0.311	103	0.285
annoyance	0.236	0.409	0.300	320	0.251
embarrassment	0.289	0.297	0.293	37	0.515
realization	0.232	0.372	0.286	145	0.213
approval	0.211	0.390	0.274	351	0.314
caring	0.203	0.363	0.261	135	0.355
disappointment	0.189	0.272	0.223	151	0.241
nervousness	0.227	0.217	0.222	23	0.316
relief	0.100	0.182	0.129	11	0.176
Macro (28 labels)	**0.391**	**0.534**	**0.445**	–	**0.463**

Labels with explicit cue words are recognised well: gratitude (F1 0.888), amusement (0.785) and love (0.758), followed by neutral (0.638), remorse (0.635) and fear (0.632). The weakest labels are relief (0.129, 11 test comments), nervousness (0.222), disappointment (0.223), caring (0.261), approval (0.274), realization (0.286) and annoyance (0.300). Labels that depend on context or on tone, such as approval and realization, are hard for a bag-of-words model. Rare labels have very few test comments (grief 6, relief 11, pride 16), so their F1 values carry large uncertainty. Fig. 4 compares validation and test F1 for every label.

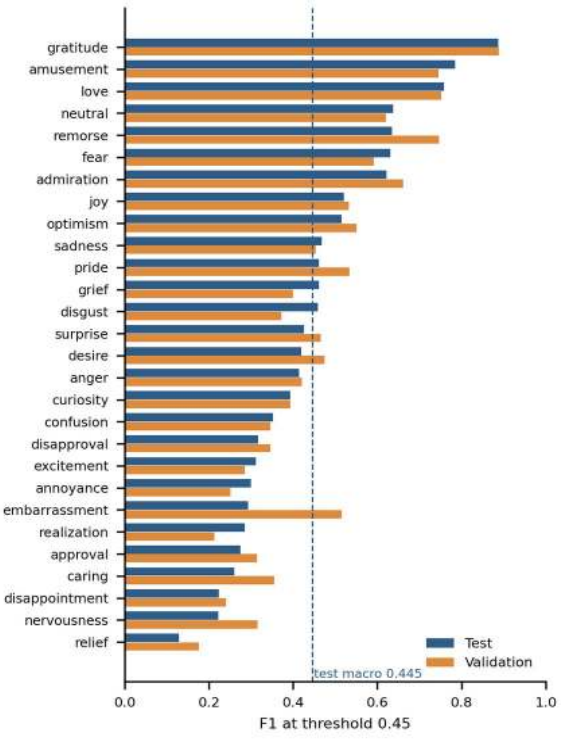


Figure 4: ***Per-label F1 on the validation and test splits at threshold 0.45.***

B. Error structure

The model recalls many true labels (0.638) but at low precision (0.430): it predicts 1.73 labels per comment on average against 1.17 annotated, and macro recall exceeds macro precision (0.534 against 0.391). Neutral is predicted for 46.6% of comments although it is annotated for 32.9%, and 31.6% of comments whose annotation is not neutral receive a neutral prediction, which is the main way in which real emotion can be missed. The balanced class weights push scores of rare labels up, which lifts recall and costs precision. Fig. 5 shows only a weak, non-significant relation between a label's test support and its F1 (Spearman $\rho = 0.32$, $p = 0.10$), which suggests that how distinctive a label's wording is matters more than its frequency; this is a hypothesis, not a result.

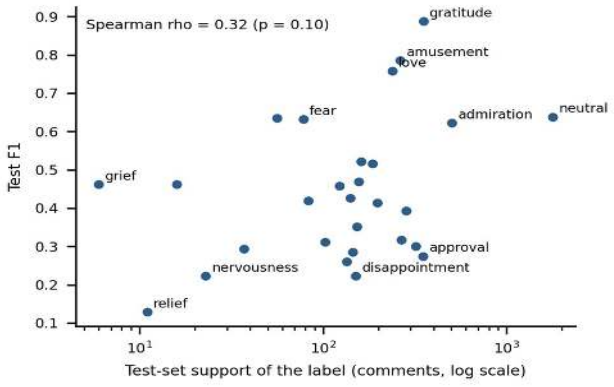


Figure 5: ***Test F1 against test support for the 28 labels.***

C. Validation versus test, and threshold sensitivity

Table 7: ***Summary on the validation and test splits ($\tau = 0.45$).***

Metric	Validation	Test
Macro-F1	0.4628	0.4451
Macro precision / recall	0.409 / 0.545	0.391 / 0.534
Micro-F1	0.5189	0.5140
Micro precision / recall	0.441 / 0.629	0.430 / 0.638
Subset accuracy	0.2879	0.2762
Hamming loss	0.0490	0.0503

Validation and test scores are close (macro-F1 0.463 against 0.445), as expected when both splits are random samples of the same Reddit data and τ was chosen on validation data. This agreement says nothing about other kinds of text. Fig. 6 shows that macro-F1 changes little for τ between 0.40 and 0.65 (validation 0.455–0.463, test 0.440–0.445), whereas micro-F1 rises from 0.502 at τ = 0.40 to 0.526 at 0.55 as precision improves; the choice of τ therefore trades emotion coverage against false alarms, and a single global τ is a simplification.

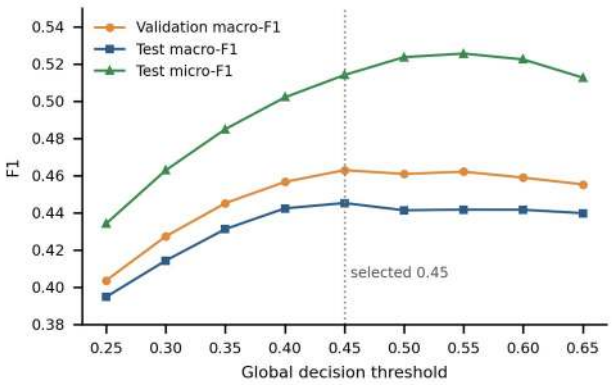


Figure 6: ***Effect of the global threshold on validation and test F1.***

D. Qualitative probes

Table 8 shows hand-written probes passed through the deployed pipeline. They are illustrations and not an evaluation; the Hindi and Marathi inputs were written in Devanagari and are given here in transliteration.

Table 8: ***Qualitative probes (scores rounded to two decimals).***

Input	Top scores (model)	Predicted
Thank you so much, this really helped me!	gratitude 1.00, admiration 0.25, approval 0.12	gratitude
I am so nervous about my exam tomorrow	nervousness 0.99, confusion 0.25, excitement 0.23	nervousness
My grandmother passed away last week and I miss her	sadness 0.97, grief 0.68, neutral 0.50	sadness, grief, neutral
I can't believe he lied to me again, I am furious	surprise 0.91, anger 0.31, confusion 0.21	surprise (anger missed)
Not bad at all, honestly	disapproval 0.97, disgust 0.86, annoyance 0.71	three negative labels (negation missed)
Hindi: "mujhe khushi hai" (I am happy)	without lexicon: neutral 0.74; with lexicon: joy 1.00, excitement 0.99, gratitude 0.54	joy, excitement, gratitude
Marathi: "mala vait vatatay" (I feel bad)	with lexicon: disappointment 1.00, disgust 0.97, sadness 0.96	six labels, incl. disgust and annoyance
Crying-face emoji alone	without lexicon: neutral 0.74; with lexicon: sadness 1.00, disappointment 0.46	sadness, disappointment
I want to die	desire 0.97, sadness 0.91, anger 0.85	desire, sadness, anger, disappointment; crisis notice shown

The probes confirm the pattern seen in the metrics: explicit emotion words work, whereas negation ("not bad"), ambiguous wording and mixed feelings do not. Without the lexicon, every unmatched Hindi, Marathi or emoji input yields the same output (neutral 0.74), so the model has no knowledge of those inputs at all. The last row shows why a separate crisis check is

needed: the classifier reads the sentence as desire, sadness and anger.

DISCUSSION

The classifier recognises emotions that have explicit cue words and is weak on labels that need context. It over-predicts, and its largest practical error is calling emotional text neutral. The macro-F1 is not far below the roughly 0.46 reported for BERT on this dataset, but this closeness should not be read as equivalence, since we did not re-run BERT and our figures come from a single run. Plausible contributors to the weak labels are label noise and annotator disagreement, the lack of word order beyond bigrams, and a single global threshold; we have not isolated any of them.

The larger concern is transfer. The model was trained on public Reddit comments, whereas MindCare AI is meant for personal reflection, possibly in Hindi or Marathi, possibly mixed with English. Predicting a person's emotion from a sentence is also not the same as assessing their mental health [21]. For these reasons the interface frames every output as an awareness signal, and the self-check, suggestions and crisis notice are described as unvalidated aids. Measuring how the system behaves on its intended text and languages is the main open question.

XI. LIMITATIONS

- **Evidence level.** One run on one random split, no confidence intervals, and no baselines or ablations. The scores describe Reddit comments in English.
- **Model choice.** Only the threshold is tuned. Transformer models and per-label thresholds were not evaluated, and the scores are uncalibrated.
- **Duplicates.** About 2% of test comments share their text with training comments. The effect on the score is small (macro-F1 0.4426 without them) but the splits were not cleaned.
- **Multilingual layer.** A 36-entry substring lexicon in front of an English model, with seven keys that never match. Hindi and Marathi accuracy, code-mixed text and emoji were not measured.
- **Self-check.** Items and totals are not validated and the cut-points are borrowed from PHQ-9. The breakdown bars in the interface use a fixed maximum of six although four of the six bars rest on a single item with maximum three, and appetite and restlessness are not shown in the breakdown.
- **Crisis notice and suggestions.** A nine-phrase English keyword list, with no coverage of Hindi or Marathi, and suggestions for seven of the 28 labels that no clinician has reviewed.
- **Security and deployment.** No rate limiting, account lock-out, password reset or encryption of the database at rest is implemented; the prediction endpoint is open and CSRF-exempt; a development fallback secret key is used when none is set; and the application is started with the Flask built-in server. These must be addressed before any public use.
- **Scope.** The system is not a diagnostic or emergency tool and must not be used for clinical decisions. User studies and expert review have not been done.

XII. FUTURE WORK

The most important next step is a stronger evaluation: several seeds or bootstrap intervals, the baselines and ablations in Table 5, and training on de-duplicated data.

- **Better models.** Fine-tune DistilBERT or RoBERTa on GoEmotions and compare accuracy, size and speed against the linear model; add per-label thresholds and probability calibration.
- **Real multilingual support.** Replace the lexicon with a multilingual encoder such as MuRIL or XLM-R and evaluate it on labelled Hindi, Marathi and code-mixed text.
- **Safety.** Use a validated risk-screening approach with multilingual coverage and expert review, and replace the keyword list.
- **Validated self-check.** Use PHQ-9 and GAD-7 in their validated forms, or validate the 10-item set against them with ethical approval.
- **Explanations.** Show the n-grams that drive a prediction and test whether users find them helpful.
- **Engineering.** Fix the lexicon ordering, add security hardening and a production server, and measure end-to-end latency.
- **User study.** Evaluate usability and perceived usefulness with students and counsellors.

CONCLUSION

We presented MindCare AI, a documented pipeline and web application for multi-label emotion detection from short text, and evaluated its classifier on the official GoEmotions test split. The model reaches a macro-F1 of 0.445 and a micro-F1 of 0.514, recognises explicit emotions such as gratitude, amusement and love well, over-predicts, and fails on relief, nervousness, disappointment and context-dependent labels. The multilingual layer, self-check, suggestions and crisis notice are working design components that have not been validated. The results are a starting point and not evidence of generalisation: they come from one run on one dataset without baselines. Until stronger models, multilingual evaluation, security hardening and expert review are in place, the system should be used only as a research prototype and an awareness aid.

DECLARATIONS

Data availability. GoEmotions is publicly available from its authors [5]. The evaluation uses its official train, dev and test splits; the project code, trained bundle and the per-label results file (reports/evaluation_v2.json) are in the project archive.

Use of AI tools. The application does not call any external AI service at run time; all inference runs locally with the bundled scikit-learn model.

REFERENCES

1. World Health Organization, *World Mental Health Report: Transforming Mental Health for All*. Geneva, 2022.
2. National Institute of Mental Health and Neuro Sciences, *National Mental Health Survey of India, 2015–16: Summary*. Bengaluru, 2016.
3. P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169 –200, 1992.
4. R. Plutchik, "A general psychoevolutionary theory of emotion," in *Emotion: Theory, Research, and Experience*, vol. 1. Academic Press, 1980, pp. 3–33.
5. D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proc. ACL*, 2020, pp. 4040–4054.
6. C. Strapparava and R. Mihalcea, "SemEval-2007 Task 14: Affective text," in *Proc. SemEval*, 2007, pp. 70–74.
7. E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. EMNLP*, 2018, pp. 3687–3697.
8. G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.
9. S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.
10. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. ICWSM*, 2014.
11. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. EACL*, 2017, pp. 427–431.
12. M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014.
13. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.
14. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
15. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
16. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.
17. A. A. Abd-Alrazaq, M. Alajlani, A. A. Ali, and M. Househ, "Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis," *J. Med. Internet Res.*, vol. 22, no. 7, art. e16021, 2020.

18. K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot)," *JMIR Ment. Health*, vol. 4, no. 2, art. e19, 2017.
19. K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001.
20. R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," *Arch. Intern. Med.*, vol. 166, no. 10, pp. 1092–1097, 2006.
21. S. Chancellor and M. De Choudhury, "Methods in predicting mental health status in social media: A scoping review of methods and datasets," *npj Digit. Med.*, vol. 3, art. 43, 2020.
22. A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. ACL*, 2020, pp. 8440–8451.
23. S. Khanuja et al., "MuRIL: Multilingual representations for Indian languages," arXiv:2103.10730, 2021.
24. S. Khanuja et al., "GLUECoS: An evaluation benchmark for code-switched NLP," in *Proc. ACL*, 2020, pp. 3575–3585.
25. A. Kulkarni, M. Mandhane, M. Likhitkar, G. Kshirsagar, and R. Joshi, "L3CubeMahaSent: A Marathi tweet-based sentiment analysis dataset," in *Proc. WASSA*, 2021, pp. 213–220.
26. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
27. M. Grinberg, *Flask Web Development*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2018.
28. Ministry of Health and Family Welfare, Government of India, *Tele-MANAS: National Tele Mental Health Programme*, launched 2022.
29. Government of India, *The Digital Personal Data Protection Act, 2023*.

HOW TO CITE: Pratiksha Mogal, Vishakha Lonkar, Disha Nimbalkar, Vedansh Soni, Dr. Rama Bansode, MindCare AI: Multi-Label Emotion Detection from Text for Mental-Wellbeing Awareness with English, Hindi and Marathi Support, Int. J. Sci. R. Tech., 2026, 3 (11), 196-208. https://doi.org/10.5281/zenodo.23942283