Text Emotion Classification in Mental Health Counseling Conversations Using Soft Voting Ensemble of BERT and BiLSTM

Sabrina Ella Zafira[1], Fitra Abdurrachman Bachtiar[1]*, and Fajar Pradana[2]

[1] Department of Informatics Engineering, Faculty of Computer Science, Universitas Brawijaya, Malang 65145, Indonesia
sabrinaella@student.ub.ac.id, fitra.bachtiar@ub.ac.id

[2] Department of Information System, Faculty of Computer Science, Universitas Brawijaya, Malang 65145, Indonesia
fajar.p@ub.ac.id

Abstract. Text-based mental health counseling is increasingly adopted due to its accessibility and anonymity; however, the absence of nonverbal cues makes emotion recognition challenging, particularly under high conversation volume and class imbalance. This study evaluates BERT, BiLSTM with GloVe embeddings, and their soft-voting ensemble for text emotion classification in counseling conversations, using only the client/help-seeker side of the Amod and MentalChat datasets. A semi-supervised pseudo-labeling approach (pre-trained emotion model, confidence threshold 0.7) was used, with low-confidence samples manually verified by a single non-expert annotator; this single-annotator setup, without inter-annotator agreement, is treated as a limitation. After cleaning and deduplication, 6,941 labeled samples across seven emotion classes were obtained. BERT was fully fine-tuned; BiLSTM was trained with a class-weighted loss to address imbalance. Simple and macro-F1-weighted soft voting were both evaluated. BERT consistently outperforms BiLSTM, achieving its best result on MentalChat (accuracy 0.804, macro F1 0.643). The ensembles do not consistently surpass BERT, reflecting the performance gap between base learners, though they substantially outperform standalone BiLSTM. Cross-dataset evaluation shows BERT has the most moderate degradation (≈12.3%). Per-class analysis shows the minority disgust class was correctly recognized in only 1 of 16 model–scenario combinations (BERT, Amod-trained and Amod-tested; recall 14.3%), and 0% elsewhere, highlighting the limits of class weighting under severe imbalance. These results suggest BERT is the more promising candidate among the models compared here for downstream emotion-support tools, though clinical deployment would require expert-validated labels and prospective clinical evaluation beyond this study's scope.

Keywords: text emotion classification · mental health counseling · BERT · BiLSTM · soft voting ensemble · cross-dataset evaluation · pseudo-labeling.

* Corresponding author.

© The Author(s) 2026

M. Murtadha et al. (eds.), *Proceedings of the 2026 11th International Conference on Sustainable Information Engineering and Technology (SIET 2026)*, Atlantis Highlights in Computer Sciences 26,
https://doi.org/10.2991/978-94-6239-828-3_10

1 Introduction

Text-based communication increasingly dominates human interaction, but the loss of nonverbal cues (facial expression, tone, body language) often causes ambiguity in interpreting emotion from text, risking misunderstanding [9]. This is especially critical in mental health counseling: text-based services are popular for their anonymity and accessibility, but high message volume makes it easy for counselors to miss critical emotional signals such as depression or acute anxiety. The WHO reports approximately 720,000 suicide deaths annually, the third leading cause of death among people aged 15–29 [19], and longer mental health treatment waiting times have been linked to increased mortality [2], underlining the practical stakes of reliable emotion classification in counseling text.

A major obstacle is the lack of labeled data: counseling datasets such as Amod and MentalChat provide no emotion annotations, and manual labeling is costly [17]. Pseudo-labeling within a semi-supervised framework offers an alternative: a pre-trained model assigns labels automatically, with only low-confidence cases manually verified [14], a strategy shown to improve performance by 5–7% over purely supervised learning on limited data [12]. However, reusing a model's own predictions as training labels risks confirmation bias, where early prediction errors are reinforced rather than corrected [1]; this risk is addressed explicitly in Section 4.6.

Prior work has combined BERT, BiLSTM, and ensembling: Gou and Li [5] combined BERT and BiLSTM (F1 59.81% on MELD); Nimmi et al. [15] used average voting across BERT, DistilBERT, and RoBERTa (macro F1 85.20% on COVID-19 text); Zhou and Wu [21] improved macro F1 by 12.5% with soft voting on Twitter. These studies target general dialogue or social media text rather than mental health counseling, and the specific BERT–BiLSTM soft-voting combination with systematic cross-dataset evaluation remains rarely explored in this domain.

This study therefore (1) implements a BERT–BiLSTM soft-voting ensemble, (2) applies confidence-threshold pseudo-labeling with manual validation, and (3) evaluates both in-dataset and cross-dataset robustness on Amod and MentalChat. The individual components used here (BERT, BiLSTM, GloVe, soft voting, pseudo-labeling) are each established techniques; the contribution of this study is a systematic empirical evaluation of these techniques in the underexplored mental health counseling domain, including a cross-dataset robustness assessment and an honest characterization of when BERT–BiLSTM soft voting does, and does not, help over a single strong model.

2 Theoretical Background

2.1 Basic Emotions and Mental Health

Ekman's six basic emotions, anger, disgust, fear, happiness, sadness, and surprise, are considered universal and consistently recognizable across cultures [4]. This study extends them with a neutral class, yielding seven target classes. Persistent sadness may indicate depressive symptoms, while heightened fear is often associated with anxiety, consistent with NLP-based reviews linking sadness- and hopelessness-related language to depressive symptomatology [18].

2.2 Text Classification and Pre-processing

Text classification assigns categories to textual documents [9]; modern deep learning approaches such as BERT and BiLSTM learn feature representations automatically and substantially improve accuracy [13]. Pre-processing includes cleaning, whitespace normalization, and duplicate removal, tailored per model: BERT uses WordPiece tokenization, while BiLSTM uses word-level tokenization with padding/truncation.

2.3 GloVe Word Embedding

GloVe combines global matrix factorization with local context-window methods for word vector learning [16]. This study uses 100-dimensional GloVe 6B embeddings as static inputs for BiLSTM.

2.4 BERT

BERT learns bidirectional contextual representations via self-attention [3]. This study fine-tunes bert-base-uncased (12 layers, hidden size 768, 12 heads) on seven emotion classes with batch size 16, learning rate 2e-5, warmup ratio 0.1, up to 10 epochs, and early stopping (patience 3).

2.5 BiLSTM

BiLSTM processes sequences in both directions to capture long-term dependencies [8]. Here it uses hidden size 128, 2 layers, dropout 0.5, Adam optimizer (lr 1e-3), up to 20 epochs, and patience 5.

2.6 Pseudo-Labeling

The pre-trained model j-hartmann/emotion-english-distilroberta-base [7] generates initial labels, samples with confidence ≥ 0.7 are used directly, following evidence that thresholds of 0.7–0.8 balance pseudo-label quality and quantity [20], while lower-confidence samples are manually verified. Because both the automatic labeling and the low-confidence verification were done by a single source (one model, one non-expert annotator) rather than multiple independent raters, no inter-annotator agreement statistic such as Krippendorff's Alpha [10] was computed, and confirmation bias in the sense described by Arazo et al. [1] cannot be fully ruled out (Section 4.5).

2.7 Soft Voting Ensemble

Soft voting aggregates class probabilities from each base model [21]. This study implements (1) simple soft voting (equal-weighted average) and (2) weighted soft voting, where BERT's weight is $w_{BERT} = F_{BERT}/(F_{BERT} + F_{BiLSTM})$ based on in-dataset macro F1, and $w_{BiLSTM} = 1 - w_{BERT}$. Using the macro-F1 values in Section 4.1, this gives approximately $w_{BERT} \approx 0.61$ for the Amod-trained scenario and $w_{BERT} \approx 0.64$ for the MentalChat-trained scenario, i.e., BERT is weighted more heavily due to its stronger performance [15].

Table 1: Dataset Characteristics

Attribute	Amod	MentalChat	Raw Combined
Samples	3,512	6,338	9,850
Type	Q&A	Interview	–
Language	EN	EN	EN

2.8 Evaluation Methods

Performance is evaluated with accuracy, precision, recall, and F1 (macro and micro). Macro-average weights all classes equally, which matters given the severe class imbalance in this dataset; micro-average reflects overall performance under the actual class distribution. Confusion matrices are used to identify misclassification patterns, particularly for minority classes.

3 Methodology

3.1 Research Overview

The workflow comprises: (1) collection and preprocessing of two counseling datasets, (2) pseudo-labeling with manual validation, (3) separate BERT and BiLSTM training, (4) simple and weighted soft-voting ensembles, and (5) in-dataset and cross-dataset evaluation across four train–test combinations.

3.2 Datasets

Two English mental health counseling datasets from Hugging Face are used. Amod/mental health counseling conversations contains ≈3,512 single-turn Q&A pairs between anonymous clients and counselors; only the "Context" (client) field is used as input. ShenLab/MentalChat16K contains 6,338 Q&A pairs from 378 PISCES clinical trial transcripts; only the caregiver (help-seeker) input is used. Overview of the dataset can be seen in Table 1.

The two datasets were chosen because both represent authentic counseling text but differ in register: Amod is short, public, single-turn Q&A, while MentalChat is longer, structured clinical-interview transcript data. This lets the cross-dataset evaluation test generalization across two meaningfully different manifestations of the same domain, rather than across unrelated domains.

"Raw Combined" is the total before cleaning. After removing empty rows, within-source duplicates, and cross-source duplicates, 6,941 unique samples remained; all confidence statistics below refer to this cleaned set, not the raw 9,850.

3.3 Pre-processing and Pseudo-Labeling

Each raw dataset is cleaned separately (empty-row removal, deduplication, normalization), then merged and deduplicated again across sources, reducing 9,850 raw samples

Table 2: Label Distribution in the Full Labeled Pool (6,941 samples, pre-split)

Emotion	Count	%
Sadness	3,592	51.76
Fear	2,110	30.40
Neutral	310	4.47
Surprise	296	4.26
Anger	295	4.25
Happiness	290	4.18
Disgust	48	0.69

to 6,941 unique cleaned samples. These are pseudo-labeled with j-hartmann/emotion-english-distilroberta-base; 4,732 samples (68.18%) with confidence ≥ 0.7 are used directly, while the remaining 2,209 (31.82%) are manually validated by one author acting as a single non-expert annotator. This 6,941-sample pool is the full labeled corpus prior to the per-source 80:20 stratified train/test split, not the final training set alone.

3.4 Emotion Label Distribution

Table 2 presents the emotion label distribution across the 6,941-sample dataset. The distribution is highly imbalanced, with Sadness (3,592; 51.76%) and Fear (2,110; 30.40%) jointly accounting for over 82% of samples, reflecting the distress- and anxiety-laden nature of mental health-related text. The remaining categories, Neutral (4.47%), Surprise (4.26%), Anger (4.25%), Happiness (4.18%), and Disgust (0.69%), are each severely underrepresented, with Disgust forming an extreme minority class.

3.5 Experimental Design

Two training scenarios are used: Scenario A (Amod train) and Scenario B (MentalChat train), each producing a BERT and a BiLSTM model. Each model is evaluated in-dataset and cross-dataset, giving four train–test combinations (Amod→Amod, Amod→MentalChat, MentalChat→MentalChat, MentalChat→Amod), across which single models and both ensemble variants are compared.

4 Results and Discussion

4.1 BERT vs. BiLSTM, and Ensemble Impact

As seen in Table 3, BERT consistently outperforms BiLSTM, peaking on MentalChat→MentalChat (accuracy 0.804, macro F1 0.643) versus BiLSTM's 0.457/0.364 in the same scenario. This reflects BERT's bidirectional self-attention capturing implicit emotional nuance and long-range context more effectively than BiLSTM's static-embedding, sequential representation [3,6].

Soft voting improves macro F1 (class balance) more consistently than accuracy. In the Amod→Amod scenario, the weighted ensemble ($w_{BERT} \approx 0.61$) raises macro F1

Table 3: Summary of Model Evaluation Results (Accuracy / Macro F1)

Model	Source	Target	Accuracy	Macro F1
BERT	Amod	Amod	0.659	0.369
BERT	Amod	MentalChat	0.625	0.326
BERT	MentalChat	MentalChat	0.804	0.643
BERT	MentalChat	Amod	0.665	0.548
BiLSTM	Amod	Amod	0.323	0.239
BiLSTM	Amod	MentalChat	0.296	0.116
BiLSTM	MentalChat	MentalChat	0.457	0.364
BiLSTM	MentalChat	Amod	0.234	0.119
Ens. simple	Amod	Amod	0.575	0.394
Ens. simple	Amod	MentalChat	0.520	0.245
Ens. simple	MentalChat	MentalChat	0.775	0.614
Ens. simple	MentalChat	Amod	0.623	0.542
Ens. weighted	Amod	Amod	0.611	0.415
Ens. weighted	Amod	MentalChat	0.554	0.259
Ens. weighted	MentalChat	MentalChat	0.796	0.634
Ens. weighted	MentalChat	Amod	0.641	0.525

from 0.369 to 0.415 but lowers accuracy from 0.659 to 0.611. On MentalChat→MentalChat, both ensembles (simple: 0.775/0.614; weighted, $w_{BERT} \approx 0.64$: 0.796/0.634) trail single BERT (0.804/0.643) slightly. Across all cross-dataset scenarios, neither ensemble surpasses BERT, though both clearly outperform standalone BiLSTM. This is because BERT's strong probabilities are consistently "pulled down" by the much weaker BiLSTM; soft voting benefits most when base learners have comparable strength, as in [15], where BERT, DistilBERT, and RoBERTa (comparable performance) were combined successfully. Soft voting is therefore best reserved for base learners of similar strength rather than applied by default.

4.2 Cross-Dataset Robustness

All models degrade cross-dataset, but BERT degrades least: macro F1 falls from 0.369 to 0.326 (Amod→MentalChat) and from 0.643 to 0.548 (MentalChat→Amod), a decline of ≈12.3%, still acceptable for imbalanced multiclass tasks. BiLSTM degrades far more sharply (macro F1 down to 0.116–0.119 cross-dataset), reflecting the sensitivity of static GloVe embeddings to linguistic and distributional shift; ensembles fall in between. This matches prior findings that transformer-based models generalize more robustly across datasets than static-embedding approaches [11].

4.3 Per-Class Performance

Confusion matrices were produced for all four models across all four train–test combinations (16 total). Table 4 summarizes recall per class, averaged over the four train–test combinations for each model, to keep the analysis compact while still surfacing minority-class behavior.

Table 4: Average Recall (%) per Emotion Class by Model, Averaged Over All Four Train–Test Combinations

Model	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
BERT	52.3	3.6	67.2	48.7	85.9	71.9	27.5
BiLSTM	32.4	0.0	44.2	27.7	50.8	26.8	21.9
Ens. simple	50.9	0.0	69.2	58.9	74.7	63.4	29.3
Ens. weighted	51.6	0.0	69.6	58.9	75.6	68.6	28.4

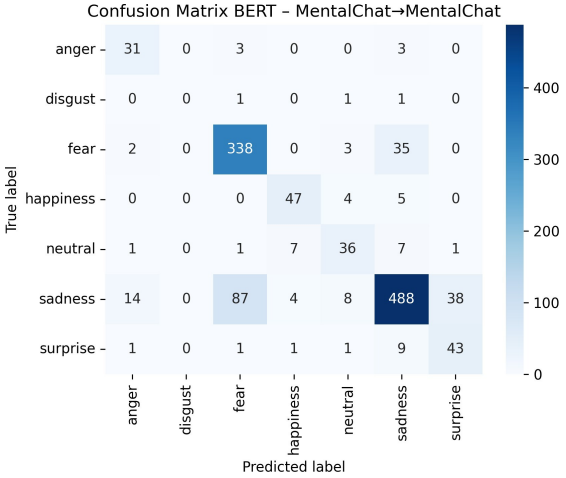


Fig. 1: Confusion Matrix BERT – MentalChat→MentalChat (in-dataset). This is BERT's best-performing scenario overall (accuracy 0.804, macro F1 0.643).

Disgust recall is 0% for every model except BERT, and even for BERT it is driven entirely by a single case: BERT correctly identified 1 of 7 disgust samples (14.3%) only in the Amod-trained, Amod-tested scenario; in the other 15 of 16 model–scenario combinations examined, no disgust sample was ever correctly classified. Figures 1 and 2 illustrate two representative cases, BERT's best-performing scenario and the only scenario with a correctly classified disgust sample.

Anger and surprise also show weak, inconsistent recall, especially cross-dataset, while fear and sadness, the two majority classes, are frequently confused with each other across all models, most pronounced for BiLSTM and least pronounced for BERT. These results support macro F1 over accuracy as the primary metric in this domain, since accuracy alone would obscure the near-total failure to detect disgust, and point to targeted data augmentation or disgust-specific annotation as necessary future work rather than an incidental detail.

4.4 Pseudo-Labeling Effectiveness

Of the 6,941 cleaned samples, 68.18% received high-confidence pseudo-labels directly and 31.82% were manually validated by a single non-expert annotator to reduce la-

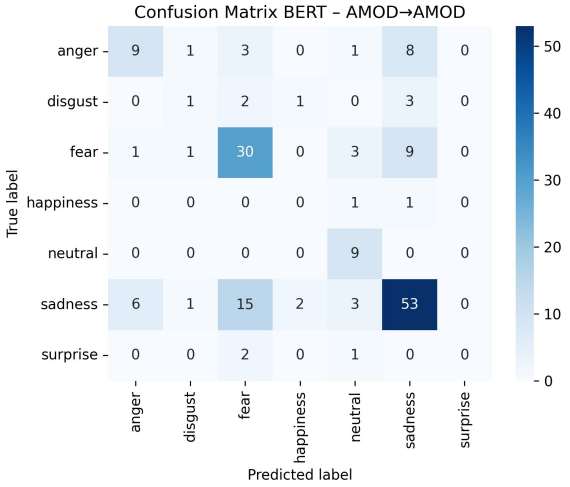


Fig. 2: Confusion Matrix BERT – Amod→Amod (in-dataset). This is the only model–scenario combination in which a disgust sample was correctly classified.

bel noise before entering the training pool; MentalChat contributes 88.03% and Amod 11.97% of this pool. This validation was not cross-checked by a second rater, so no inter-annotator agreement could be computed, and the high-confidence pseudo-labels were not independently checked against expert ground truth (Section 4.6).

4.5 Implications and Limitations

Among the models compared, BERT alone is a more promising foundation than the BERT–BiLSTM ensemble for downstream emotion-detection tooling, given its stronger accuracy and cross-dataset robustness; ensembling is worthwhile mainly when base learners have comparable strength, not as a default strategy. These findings describe relative model performance under the tested conditions only, not clinical readiness: deployment in an actual counseling context would require expert-annotated evaluation data, prospective validation with licensed professionals, and attention to the privacy, bias, and interpretability concerns raised for NLP-based mental health tools more broadly [18].

Key limitations: (1) manual validation was performed by a single non-expert annotator without a second rater or inter-annotator agreement statistic [10]; (2) high-confidence pseudo-labels were not separately checked against expert ground truth, so confirmation bias [1] cannot be fully ruled out; (3) the study is restricted to English text from two datasets, untested for other languages or platforms; (4) the weighted-voting scheme uses a simple macro-F1-proportional heuristic rather than a tuned scheme, and its reported weights are reconstructed from logged metrics rather than recorded directly; (5) no clinical outcome validation was performed.

5 Conclusion

BERT consistently outperforms BiLSTM in-dataset and cross-dataset, peaking at accuracy 0.804 / macro F1 0.643 on MentalChat→MentalChat versus BiLSTM's 0.457/0.364. Soft voting (simple or weighted) does not consistently outperform single BERT, improving macro F1 at the cost of accuracy in-dataset and trailing BERT cross-dataset. BERT is also the most robust to dataset shift (≈12.3% decline vs. BiLSTM's collapse to 0.116–0.119 macro F1 cross-dataset). Pseudo-labeling with a 0.7 confidence threshold and manual validation yielded 6,941 labeled samples reflecting the domain's characteristic imbalance, though the single-annotator process and lack of expert ground-truth checks limit how strongly this label quality can be claimed. Finally, disgust recall was 0% in 15 of 16 model–scenario combinations, showing that class weighting alone cannot overcome severe imbalance.

Future work should: explore ensembles of comparably strong base learners (e.g., RoBERTa, DistilBERT, MentalBERT variants); involve multiple annotators, ideally licensed counselors, with Krippendorff's Alpha [10] reported; separately validate pseudo-label accuracy against expert ground truth on a held-out sample to bound confirmation-bias risk; extend to Indonesian counseling corpora; and apply targeted augmentation or specialized loss functions for minority classes (disgust, surprise, anger). A prospective, professionally validated prototype counselor-support tool integrating BERT is a natural next step beyond this study's scope.

Acknowledgments. Part of this research was funded by the Artificial Intelligence Applied Research Funding Program at Brawijaya University for the 2025 academic year under contract No. 02340.14/UN10.A0501/B/PT.01.03/2025.

Disclosure of Interests. The authors declare that they have no competing interests that are relevant to the content of this article.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
2. Costantini, S.: How do mental health treatment delays impact long-term mortality? American Economic Review **115**(5), 1672–1707 (2025)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186. Association for Computational Linguistics, Stroudsburg, PA, USA (2019)
4. Ekman, P.: An argument for basic emotions. Cognition and Emotion **6**(3–4), 169–200 (1992)
5. Gou, Z., Li, Y.: Integrating bert embeddings and bilstm for emotion analysis of dialogue. Computational Intelligence and Neuroscience **2023**(1) (2023)
6. Hadikhah Mozhdehi, M., Eftekhari Moghadam, A.M.: Textual emotion detection utilizing a transfer learning approach. Journal of Supercomputing **79**(12), 13075–13089 (2023)
7. Hartmann, J.: Emotion english distilroberta-base. Hugging Face (2022), https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
9. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall (2017)
10. Krippendorff, K.: Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research **30**(3), 411–433 (2004)
11. Li, T., Chen, X., Zhang, S., Dong, Z., Keutzer, K.: Cross-domain sentiment classification with contrastive learning and mutual information maximization (2020), http://arxiv.org/abs/2010.16088, arXiv preprint
12. Liu, C., Cui, J., Shang, R., Xiao, Y., Jia, Q., Gehringer, E.: Improving problem detection in peer assessment through pseudo-labeling using semi-supervised learning. In: Proceedings of the 15th International Conference on Educational Data Mining (EDM). International Educational Data Mining Society, Durham, United Kingdom (2022)
13. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M.N., Gao, J.: Deep learning–based text classification. ACM Computing Surveys **54**(3), 1–40 (2022)
14. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(8), 1979–1993 (2019)
15. Nimmi, K., Janet, B., Selvan, A.K., Sivakumaran, N.: Pre-trained ensemble model for identification of emotion during covid-19 based on emergency response support system dataset. Applied Soft Computing **122**, 108842 (2022)
16. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Stroudsburg, PA, USA (2014)
17. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 254–263. Association for Computational Linguistics, Honolulu, HI, USA (2008)
18. Teferra, B.G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., Bhat, V.: Screening for depression using natural language processing: Literature review. Interactive Journal of Medical Research **13**, e55067 (2024)
19. World Health Organization: Suicide worldwide in 2019: Global health estimates. Tech. rep., World Health Organization (2021)
20. Yu, Z., Li, Y., Lee, Y.J.: Inpl: Pseudo-labeling the inliers first for imbalanced semi-supervised learning (2023), http://arxiv.org/abs/2303.07269, arXiv preprint
21. Zhou, Q., Wu, H.: Nlp at iest 2018: Bilstm-attention and lstm-attention via soft voting in emotion classification. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 189–194. Association for Computational Linguistics, Stroudsburg, PA, USA (2018)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

