Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 8, No.4, October 2026, pp. 02604033-01 ~ 02604033-012
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v8i4.2741

Enhanced Self-Esteem Classification: Leveraging Data Augmentation and Transformer-Based Sentence Embeddings

Reza Ahmadiansah[1], Mukti Ali[1], Rasimin[2], Achmad Maimun[3], Kastolani[1], Imam Subqi[1], Embun Bening Di Moravia[4], Andi Bahtiar Semma[5*]

[1]Faculty of Da'wa, Universitas Islam Negeri Salatiga, Jl. Ring Road Salatiga Km. 2 Salatiga 50241, Central Java, Indonesia

[2]Faculty of Teacher Training and Education, Universitas Islam Negeri Salatiga, Jl. Ring Road Salatiga Km. 2 Salatiga 50241, Central Java, Indonesia

[3]Postgraduate Program, Universitas Islam Negeri Salatiga, Jl. Ring Road Salatiga Km. 2 Salatiga 50241, Central Java, Indonesia

[4]Faculty of Psychology, Universitas Islam Negeri Syarif Hidayatullah, Jl. Ir. H. Djuanda No. 95, Tangerang 15412, Indonesia

[5]Faculty of Science and Technology, Universitas Islam Negeri Salatiga, Jl. Ring Road Salatiga Km. 2 Salatiga 50241, Central Java, Indonesia

*andisemma@gmail.com

Abstract. This study investigates automated self-esteem assessment from self-descriptive text using transformer-based sentence embeddings. Although prior research has explored text, behavioral, and multimodal signals, the combined effects of data augmentation, embedding choice, and classifier complexity in text-only self-esteem classification remain insufficiently understood. Accordingly, this study aims to systematically evaluate embedding–classifier combinations under both low-resource and augmented data conditions. Textual self-descriptions were collected from 298 undergraduate students at UIN Salatiga and labeled using the Indonesian version of the Rosenberg Self-Esteem Scale, yielding three self-esteem categories. To address data scarcity, a controlled translation-based augmentation pipeline with expert psychological validation was applied exclusively to the training set. Seven multilingual sentence embedding models were paired with eight classification algorithms, and performance was evaluated using macro-averaged metrics, along with training and inference time. Results reveal a two-regime pattern: (1) in limited-data settings, strong embeddings with simple classifiers perform best, (2) whereas in augmented settings, representation quality dominates and classifier choice has a marginal effect. The findings suggest that prioritizing high-quality embeddings and carefully validated data augmentation enables accurate, scalable, and cost-effective text-based self-esteem assessment for real-world psychological applications.

Keywords: Self-esteem classification, machine learning, data augmentation, computational psychology

(Received 2025-05-05, Revised 2026-07-01, Accepted 2026-07-04, Available Online by 2026-08-29)

1. Introduction

The advancement of Machine Learning (ML) has fundamentally transformed data processing and analysis methodologies across numerous disciplines [1]. This rapidly evolving field of artificial intelligence has transcended theoretical frameworks to deliver practical applications that address complex problems through algorithmic learning. Contemporary ML systems demonstrate the capacity to generate predictions and decisions with minimal human intervention by identifying patterns within extensive datasets [2]. The exponential growth in computational power, combined with the availability of vast data repositories, has accelerated ML development, enabling sophisticated applications in natural language processing, computer vision, and predictive analytics that were previously inconceivable [3]. The versatility of machine learning has facilitated its integration across diverse fields, spanning healthcare, finance, education, and psychology [4]. In psychology, ML algorithms analyze behavioral patterns, emotional responses, and cognitive processes with unprecedented precision [5], [6]. Researchers now employ these tools to identify subtle correlations between psychological variables, predict mental health outcomes, and personalize therapeutic interventions. ML models can process multidimensional psychological assessment data, recognize patterns imperceptible to human observation, and contribute to more objective diagnostic processes [7]. This technological synergy with psychology represents a significant advancement in understanding human behavior and mental processes beyond traditional statistical approaches [8].

Self-esteem plays a critical role in university students' academic performance, mental well-being, and social integration [9]. As a psychological foundation, healthy self-esteem enables students to approach challenges with confidence, persist through academic difficulties, and maintain resilience during stressful periods [10]. Research consistently demonstrates that students with positive self-regard tend to establish higher educational goals, engage more actively in learning processes, and demonstrate greater adaptability when confronting setbacks [11]. Conversely, low self-esteem correlates with increased vulnerability to anxiety, depression, and maladaptive coping strategies, potentially undermining academic achievement and the overall university experience [12], [13]. As institutions increasingly recognize this connection, support systems targeting self-esteem enhancement have become essential components of comprehensive student success initiatives [14]. The exploration of self-esteem through various methodological approaches has been examined by multiple studies utilizing different datasets and analytical frameworks. [15] utilized smartphone passive sensing data to predict self-esteem among 51 students over five weeks. The study incorporated physical activity, conversation, and communication patterns and employed linear regression, gradient boosted regression, and deep learning-based multilayer perceptron models, with gradient boosted regression yielding optimal results in predicting performance, social, and appearance self-esteem (SMAPE values of 8.61%, 5.64%, and 7.53%, respectively). Despite significant findings, the limited sample size and lack of consideration for digital footprints pose limitations, raising questions about the generalizability of the results.

[16] Focusing on frontostriatal white matter integrity and its impact on self-esteem, leveraged probabilistic tractography and ridge regression models to analyze neuroimaging data from two datasets. Their multivariate frontostriatal model exhibited modest predictive accuracy ($r = 0.237$), despite challenges with sample differences and limited standardization of outcomes. Although the univariate model showed a significant positive relationship between white matter integrity and self-esteem, the generalizability of these findings depends on the similarity of the test samples. In a context more grounded in psychometric variables, [17] examined the interplay between depression, internet addiction, and self-esteem among 461 Dhaka-based students. Using logistic regression, their prediction models for self-esteem achieved

impressive accuracies (86% from depression, 82% from internet addiction). Despite promising results, the study's voluntary sampling method introduces potential biases that question its applicability beyond Dhaka's undergraduate population. [18] further explored adolescent internet addiction and its interaction with self-esteem and resilience using a comprehensive dataset of over 8,000 junior high school students. Employing SHapley Additive exPlanations (SHAP) for interpretability, their XGBoost model attained the highest predictive scores (AUC of 0.790). The study, however, remains constrained by its reliance on self-reported questionnaires and a cross-sectional design, raising questions about how incorporating longitudinal data could enhance causal inferences.

[19] developed an innovative Owl Search Optimized Dynamic Deep Neural Network (OSO-DDNN) for predicting elementary students' academic performance linked to self-esteem and individualism. This model excelled compared to conventional methods, achieving 92% accuracy. Nevertheless, the model's reliance on clean data underscores the need for future work considering broader socio-economic factors. [20] introduced a unique approach by analyzing web search data to detect low self-esteem among college students, achieving high F1 scores when using search category distributions (average F1 = 0.86). However, the small sample size and reliance on college students limit the study's scope. [21] investigated electroencephalogram (EEG) data for classifying high and low self-esteem. Utilizing a Random Forest classifier, they achieved notable accuracy (79.38%) but acknowledged limitations such as the lack of middle-range data and a culturally specific dataset.

Despite substantial findings from prior studies in understanding self-esteem through diverse datasets and methodologies, significant research gaps persist, particularly concerning the need for data augmentation and advanced model building to enhance predictive accuracy and generalizability. Existing studies predominantly rely on limited datasets, such as smartphone passive sensing [15], neuroimaging data [16], or self-reported questionnaires [18]. These datasets often lack diversity, represent small sample sizes, or fail to account for broader socio-economic and cultural contexts, which are crucial for capturing the multifaceted nature of self-esteem. For instance, Bin Morshed et al.'s [15] reliance on physical activity and communication patterns excludes digital footprints, while [20] use of web search data is constrained by a small and specific population. The voluntary sampling method employed by [17] introduces selection bias, further limiting the applicability of findings beyond particular demographics. Moreover, the models utilized in previous studies have demonstrated varying levels of success but often fall short due to dataset-related limitations. For example, the OSO-DDNN model [19] achieves high accuracy but depends heavily on clean, curated data, neglecting the complexities of real-world scenarios where data may be noisy or incomplete. Similarly, the EEG-based Random Forest classifier [21] lacks middle-range data, which could provide deeper insights into nuanced self-esteem levels. To address these limitations, this research focuses on incorporating data augmentation methods, such as synthetic data generation, which could enhance the quality and diversity of training data, thereby improving model performance and generalizability. Furthermore, exploring diverse model-building approaches holds promise for better predictive capabilities. Such advancements would bridge existing gaps in self-esteem prediction and pave the way for practical applications in mental health assessment and intervention.

This research aims to address gaps in predicting and understanding self-esteem by integrating advanced data augmentation techniques and diverse model-building approaches. Specifically, it seeks to enhance the quality and diversity of training datasets through synthetic data generation and other augmentation methods while exploring a range of machine learning models that can better capture the multifaceted nature of self-esteem. The study will focus on improving predictive accuracy and generalizability across broader populations, addressing limitations such as small sample sizes, selection biases, and reliance on specific data types (e.g., physical activity, neuroimaging, or self-reported questionnaires). Additionally, the research aims to contribute to practical applications in mental health assessment and intervention for university students, leveraging machine learning's capacity to identify patterns and predict outcomes related to self-esteem.

2. Method

2.1. Data acquisition

A stratified random sample of 298 undergraduate students was recruited from UIN Salatiga across various academic programs and year levels to ensure demographic representativeness. Participants were students from humanities, social sciences, and engineering programs, with 31.2% male and 68.8% female students. The inclusion criteria required participants to be native Indonesian speakers currently enrolled as full-time students, while the exclusion criteria eliminated those with self-reported severe mental health conditions or language comprehension difficulties that might affect response quality. Self-esteem was measured using the Indonesian version of the Rosenberg Self-Esteem Scale (RSE) *[22], [23]*, a validated 10-item instrument with established psychometric properties in Indonesian populations. The scale employs a 4-point Likert response format ranging from strongly agree to strongly disagree, with higher scores indicating greater self-esteem. Self-descriptions were collected through structured prompts asking participants to *"describe yourself in 3-5 sentences, focusing on your personal qualities, abilities, and characteristics that you believe define who you are."*

Data collection occurred in controlled classroom environments during regular class periods over a four-week period. Each session was administered by trained research assistants following standardized protocols to ensure consistency. Participants first completed informed consent procedures, including explicit information about data anonymization and research purposes. The RSE was administered first, followed by the self-description task, with participants given unlimited time to complete both measures. All responses were collected using anonymous identification codes, with no personally identifiable information recorded. RSE total scores (range: 10-40) were categorized using tertile splits on the base dataset: Low self-esteem (10-23, 29.3%), Medium (24-30, 41.9%), High (31-40, 28.2%). This distribution-based approach ensured balanced representation while maintaining psychological meaningfulness. Post-augmentation, class distributions were balanced to 33.3±2% per class through stratified sampling from the augmented pool.

2.1.1 Data Augmentation and Validation

Due to the limited size of the original dataset, a structured computational data augmentation protocol was implemented to expand the training data while preserving semantic integrity and class balance. The augmentation pipeline consisted of three sequential stages: (1) stratified data splitting, (2) controlled generative augmentation through translation-based transformation, and (3) expert validation. First, a stratified train–test split (80:20) was performed to ensure proportional class distribution across subsets. Augmentation procedures were applied exclusively to the training set, while the test set (20%) contained only original, non-augmented samples. This design prevented data leakage and ensured unbiased performance evaluation. For the generative augmentation stage, translation-based paraphrasing was conducted using Ollama with the pre-trained model zongwei/gemma3-translator:4b**.** The base Indonesian data were translated into English and subsequently back-translated into Indonesian. In addition, cross-lingual transformation was performed through Indonesian–Malay translation and back-translation, leveraging the linguistic proximity between the two languages to introduce controlled lexical and syntactic variation while maintaining semantic equivalence. This approach generated diverse paraphrastic variants without altering the underlying psychological meaning.

The final dataset composition was controlled to ensure that no single class exceeded 35% of the total samples. Both expert reviewers (master's-level psychologists with 5+ years of experience) independently evaluated 500 augmented descriptions using a structured protocol. Reviewers rated each sample on two dimensions: (1) semantic preservation, whether the augmented text maintained the core meaning and linguistic naturalness of the original samples, and (2) psychological coherence, whether the content appropriately reflected the assigned self-esteem category. Disagreements were resolved through discussion to reach a consensus. Of the 500 reviewed samples, 47 (9.4%) were rejected, resulting in a 90.6% acceptance rate applied to filter the complete augmented dataset. While this validation process ensured

quality control, formal inter-rater reliability statistics were not calculated, representing a limitation for future replication studies.

2.2. Classification Process

Feature extraction was performed using seven pre-trained multilingual embedding models to obtain dense semantic representations of the text. The evaluated models included indolem/indobert-base-uncased (IndoBERT Base, 768 dimensions), firqaaa/indo-sentence-bert-base (IndoSentence-BERT, 768 dimensions), paraphrase-multilingual-MiniLM-L12-v2 (MiniLM-L12, 384 dimensions), paraphrase-multilingual-mpnet-base-v2 (Multilingual MPNet, 768 dimensions), sentence-transformers/LaBSE (LaBSE, 768 dimensions), and distiluse-base-multilingual-cased-v2 (DistilUSE, 512 dimensions). These models were selected to represent a diverse range of transformer-based architectures, including Indonesian-specific, multilingual, and lightweight sentence embedding models. Each model generated fixed-length vector representations that served as input features for the classification stage.

Following feature extraction, eight classification algorithms were evaluated. The gradient boosting family included XGBoost (max_depth=8, eta=0.1), CatBoost (depth=8, iterations=500), and LightGBM (num_leaves=31, learning_rate=0.05). Ensemble learning was further represented by Random Forest (n_estimators=200, max_depth=15). Linear modeling was implemented using Logistic Regression (C=1.0, max_iter=1000), while kernel-based classification employed Support Vector Machines with an RBF kernel (C=1.0). Instance-based learning was conducted using k-Nearest Neighbors (n_neighbors=5, cosine distance metric). Additionally, a feedforward neural network implemented in PyTorch (hidden layer dimensions: 256, 128, and 64) was included to assess the performance of a deep neural architecture on fixed embedding inputs. All experiments used Python 3.9.7 with scikit-learn 1.0.2, TensorFlow 2.8.0, and transformers 4.18.0 on an Apple Silicon M1 with the extreme programming paradigm [24].

Each embedding model was systematically combined with each classifier, resulting in a comprehensive evaluation of embedding–classifier pairings. For every combination, six performance indicators were computed: Accuracy, macro-averaged Precision, macro-averaged Recall, macro-averaged F1-score, training time, and inference time. The use of macro-averaged metrics ensured balanced evaluation across all classes, particularly in the presence of potential class imbalance. Training and inference time measurements were included to assess computational efficiency in addition to predictive performance. The research design is illustrated in Figure 1.

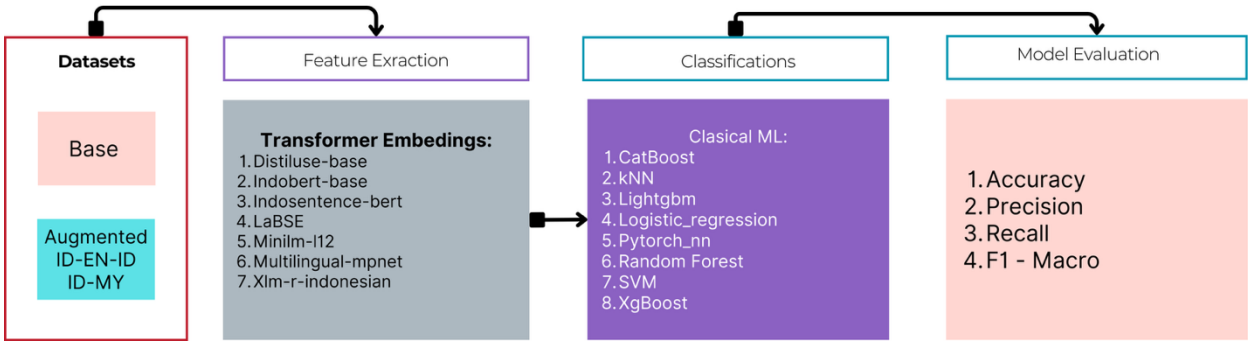

Figure 1. Research Design

3. Result and Discussion

3.1. Base dataset performance

The results in Table 1 summarize the 10 best-performing model configurations by accuracy. Overall, the combination of the distiluse-base embedding with logistic regression achieved the strongest performance, with the highest accuracy (0.45) and macro F1-score (0.4288), and was also the most computationally efficient in both training and inference. This indicates that simpler linear classifiers can be highly effective

when paired with strong sentence embeddings. Among configurations using the multilingual-mpnet embedding, logistic regression and the PyTorch neural network both reached an accuracy of 0.4333. However, these results came with trade-offs: the neural network required significantly more training time, while logistic regression showed a lower macro F1-score. Tree-based methods such as LightGBM and CatBoost generally produced competitive but slightly weaker macro-level metrics, often at the cost of substantially longer training times. More complex models, particularly CatBoost paired with indosentence-bert or multilingual-mpnet, did not consistently outperform simpler approaches despite training times exceeding 100 seconds. The weakest performance among the top ten was observed for the minilm-l12 embedding with a PyTorch neural network, which achieved the lowest accuracy and macro F1-score in this subset.

Table 1. 10 Best - Base datasets

Embedding Model	Classifier Model	Accuracy	F1 Macro
distiluse-base	logistic_regression	0.4500	0.4288
distiluse-base	lightgbm	0.4167	0.4127
multilingual-mpnet	lightgbm	0.4167	0.4120
multilingual-mpnet	pytorch_nn	0.4333	0.4049
distiluse-base	xgboost	0.4000	0.3912
multilingual-mpnet	logistic_regression	0.4333	0.3842
indosentence-bert	catboost	0.4167	0.3837
distiluse-base	kNN	0.4167	0.3787
multilingual-mpnet	catboost	0.4333	0.3737
minilm-l12	pytorch_nn	0.3667	0.3710

3.2. Data Augmentation Impacts

The augmented datasets show particularly strong performance when paired with LaBSE embeddings. The best results are achieved by combining LaBSE with both kNN and LightGBM classifiers, with each achieving an accuracy of 0.9665 and an F1-macro score of 0.9657. These configurations clearly outperform the others, indicating that LaBSE representations, when coupled with either distance-based or gradient-boosting methods, provide highly discriminative features for the task. Other LaBSE-based models also perform competitively. The PyTorch neural network and CatBoost classifiers both achieve an accuracy of 0.9553, with F1-macro scores of 0.9563 and 0.9545, respectively. These results confirm the robustness of LaBSE embeddings across diverse learning paradigms, from neural networks to ensemble methods, while maintaining consistently high predictive performance.

Beyond LaBSE, multilingual-mpnet embeddings demonstrate solid performance, particularly with a PyTorch neural network, reaching an accuracy of 0.9553 and an F1-macro of 0.9537. Slightly lower but still strong results are observed when multilingual-mpnet is combined with LightGBM, achieving an accuracy of 0.9274 and an F1-macro of 0.9252, suggesting that this embedding model remains effective but somewhat more sensitive to classifier choice. The remaining top performers include distiluse-base with kNN and minilm-l12 with a PyTorch neural network, both achieving accuracies above 0.93 and F1-macro scores closely aligned with their accuracy values. LaBSE paired with Random Forest and XGBoost also appears in the top ten, reinforcing the overall conclusion that high-quality multilingual embeddings, especially LaBSE, dominate performance on augmented datasets, while classifier selection fine-tunes but

does not fundamentally alter the ranking of the best approaches. The detailed 10 augmented dataset performer is illustrated in Table 2.

Table 2. 10 Best Performance - Augmented Dataset

Embedding Model	Classifier Model	Accuracy	F1 Macro
LaBSE	kNN	0.9665	0.9657
LaBSE	lightgbm	0.9665	0.9657
LaBSE	pytorch_nn	0.9553	0.9563
LaBSE	catboost	0.9553	0.9545
multilingual-mpnet	pytorch_nn	0.9553	0.9537
distiluse-base	kNN	0.9441	0.9438
LaBSE	random_forest	0.9441	0.9424
LaBSE	xgboost	0.9330	0.9335
minilm-l12	pytorch_nn	0.9330	0.9315
multilingual-mpnet	lightgbm	0.9274	0.9252

3.2.1. Embedding model Analysis

The results show that embedding model choice plays a dominant role in downstream classification performance, often outweighing the effect of the classifier itself. LaBSE consistently achieves the strongest results across nearly all classifiers, reaching the highest accuracy and macro F1 scores overall. Its performance remains robust even with simpler models such as kNN and LightGBM, which achieve identical top scores, indicating that LaBSE embeddings provide highly separable, well-structured representations that generalize effectively across different learning paradigms.

Among mid-tier embedding models, multilingual-mpnet, MiniLM-L12, IndoSentence-BERT, and distiluse-base demonstrate strong but more varied performance patterns. These embeddings perform best when paired with neural networks or distance-based methods such as kNN, suggesting that their representation spaces benefit from either non-linear decision boundaries or local similarity exploitation. Tree-based ensemble methods generally yield competitive but slightly lower results, while linear classifiers consistently underperform, indicating that these embeddings encode class information in ways that are not linearly separable.

IndoBERT-base shows comparatively weaker performance across classifiers, with a noticeable drop in both accuracy and macro F1. Even with neural networks, its results lag behind those of sentence-level embedding models, suggesting that its token-level pretraining objective is less well aligned with sentence-level semantic discrimination. Similarly, XLM-R Indonesian and MiniLM-L12 show stable but moderate performance, reinforcing the trend that sentence-transformer-style models are better suited for embedding-based classification tasks.

3.2.1. Classifier model analysis

Across embedding settings, kNN, neural networks, and gradient-boosting classifiers consistently delivered the strongest performance, while linear models showed clear limitations. kNN emerged as a top-performing classifier in many configurations, achieving very high accuracy and macro F1 when paired with strong embeddings, and demonstrating robustness across different representation spaces. Neural networks implemented with PyTorch also showed consistently strong results, often ranking among the top two classifiers for a given embedding, indicating their ability to effectively leverage dense semantic representations.

Tree-based ensemble methods such as LightGBM, CatBoost, and Random Forest generally formed a solid middle tier. LightGBM in particular showed competitive and stable performance, sometimes matching or closely trailing the best results, especially when combined with high-quality embeddings. CatBoost and Random Forest followed a similar pattern but with slightly lower macro F1 scores, suggesting reduced effectiveness in handling class balance compared to kNN and neural networks. XGBoost performed reliably but lagged behind other boosting methods, suggesting weaker alignment with the embedding features.

Support Vector Machines consistently underperformed relative to the leading classifiers, with noticeable drops in both accuracy and macro F1 across all embeddings. This suggests limited scalability or suboptimal margin separation in the high-dimensional embedding space. Logistic regression was the weakest classifier overall, producing the lowest scores in every embedding scenario. Its poor performance highlights the inadequacy of linear decision boundaries for capturing the complex semantic structure encoded in modern sentence embeddings.

3.3. Timing analysis

Across all embedding models, logistic regression consistently delivers the lowest average inference time, often well below 0.002 seconds. This pattern is evident for distiluse-base, LaBSE, multilingual-mpnet, and xlm-r-indonesian, where logistic regression achieves the fastest or near-fastest performance among all classifiers. XGBoost and CatBoost also show relatively low inference times, generally remaining under 0.004 seconds across most embeddings, indicating good efficiency for tree-based boosting methods.

kNN and SVM exhibit the highest inference times across nearly all embedding models. kNN inference times frequently exceed 0.02 seconds and reach 0.035 seconds with indosentence-bert, while SVM shows a similar pattern, peaking at approximately 0.039 seconds. Random forest inference times are consistently high, clustering around 0.026–0.027 seconds across embedding models, suggesting limited sensitivity to embedding choice but higher overall computational cost.

Neural network classifiers exhibit greater variability across embedding models. For most embeddings, PyTorch-based neural networks achieve moderate inference times of 0.002-0.006 seconds. However, a notable exception occurs with indobert-base, where inference time rises sharply to 0.053 seconds, making it the slowest combination in the entire set. When comparing embedding models, no single embedding consistently dominates in terms of inference speed across all classifiers. Instead, classifier choice has a stronger influence on inference time than the embedding itself. Lightweight classifiers such as logistic regression and gradient boosting methods consistently outperform distance-based and kernel-based approaches, regardless of whether the embeddings are derived from distiluse-base, IndoBERT variants, LaBSE, MiniLM, multilingual-mpnet, or xlm-r-indonesian.

Discussion

On the base dataset, the strongest performance was achieved by combining the distiluse-base embedding with logistic regression, yielding the highest accuracy (0.45) and macro F1-score (0.4288) among the top configurations. This result highlights that, in low-resource or non-augmented settings, high-quality sentence embeddings can compensate for simpler linear decision boundaries. While more complex classifiers such as neural networks and gradient-boosting methods were competitive, they did not consistently outperform logistic regression, despite incurring substantially higher training and inference costs.

In contrast, the augmented dataset produced a dramatic performance improvement across nearly all models. The combination of LaBSE embeddings with kNN or LightGBM achieved near-ceiling performance, with accuracies and macro F1-scores exceeding 0.96. These findings clearly demonstrate that data augmentation fundamentally alters the learning landscape, enabling embedding models. The LaBSE particularly produces highly separable semantic representations that are easily exploited by both distance-based and ensemble classifiers. Taken together, these results suggest a two-regime behavior: in limited-data scenarios, simpler classifiers paired with strong embeddings are sufficient and efficient, whereas in enriched data settings, representation quality dominates so strongly that classifier choice mainly fine-tunes, rather than determines, overall performance.

Our findings align strongly with and extend previous research on automated self-esteem assessment and psychological trait prediction. Earlier work by [17] reported an accuracy of approximately 86% using logistic regression with traditional feature extraction techniques. In comparison, our best-performing kNN-based configuration on the augmented dataset achieved around 95–97% accuracy and macro F1-score. This substantial improvement is likely attributable to the use of transformer-based sentence embeddings, which capture contextual and semantic nuances far beyond surface-level lexical features. Similarly, [20] achieved an F1-score of 89% using search behavior data, demonstrating the predictive power of behavioral signals. Our results show that textual content alone, when represented using modern multilingual embeddings, can reach or exceed this level of performance. This suggests that language-based signals are not merely complementary but can serve as a primary modality for self-esteem inference when adequately modeled.

Our findings also complement multimodal approaches reported by [25] and [21], who relied on smartphone sensing and EEG data, respectively. Rather than contradicting these approaches, our results reinforce the notion that self-esteem is a multifaceted construct. Text-based classification, as demonstrated here, is a scalable, cost-effective component within a broader multimodal assessment framework. Finally, the strong performance observed in this study provides empirical support for the theoretical framework proposed by [18], who identified self-esteem as a critical predictor in their XGBoost-based model for internet addiction. Their emphasis on the interaction between self-esteem and psychological resilience underscores the importance of accurate and reliable self-esteem measurement, which is an objective that our proposed classification pipeline directly supports.

The observed differences between base and augmented datasets can be explained primarily by data diversity and class boundary clarity. In the base dataset, limited samples constrain the model's ability to learn robust decision surfaces, favoring simpler classifiers that generalize better under scarcity. This explains why logistic regression performs surprisingly well despite its linear nature.

In the augmented setting, however, richer and more balanced data allow embeddings such as LaBSE to fully express their representational capacity. The resulting embedding spaces exhibit strong local clustering and global separability, which explains why kNN performs exceptionally well and why tree-based methods such as LightGBM closely match its performance. The reduced performance gap among classifiers in this regime suggests that once representations are sufficiently informative, classifier expressiveness becomes less critical. The LaBSE embedding appears to encode sentence-level semantics in a largely classifier-agnostic manner, enabling a wide range of learning algorithms to achieve similarly high performance.

3.4 Implications and limitations

From a practical standpoint, these findings have several important implications. First, they suggest that high-quality multilingual sentence embeddings should be prioritized over classifier complexity when designing real-world self-esteem assessment systems. Second, data augmentation emerges as a critical strategy for achieving reliable and clinically meaningful performance, especially in psychological domains where labeled data are often scarce. In applied settings such as mental health screening, educational assessment, or social media analysis, combining LaBSE embeddings with lightweight classifiers like kNN or LightGBM offers an attractive balance between accuracy, robustness, and interpretability. At the same time, timing analysis indicates that classifier choice has a stronger impact on inference speed than embedding choice, underscoring the need to account for computational constraints when deploying such systems at scale.

Despite strong, consistent performance across multiple embedding–classifier combinations, several limitations should be acknowledged. First, although data augmentation substantially improved classification performance, augmented samples may not fully reflect the linguistic and psychological variability present in real-world self-esteem expressions. Synthetic or transformed data can inadvertently reinforce patterns already present in the original dataset, potentially inflating performance estimates. As a result, the high F1 score achieved, particularly with LaBSE-based embeddings, should be interpreted with caution when considering generalization to unseen data. Second, the study focuses exclusively on textual data and does not incorporate complementary behavioral, physiological, or contextual signals. Prior

research has demonstrated that self-esteem is influenced by multimodal factors, including smartphone usage patterns and neurophysiological indicators. While our results show that text-based approaches can achieve very high accuracy, they capture only one dimension of a complex psychological construct and may miss latent factors better captured through multimodal fusion.

4. Conclusion

This study demonstrates that the effectiveness of automated self-esteem classification is primarily driven by the quality of sentence embeddings and the availability of data rather than classifier complexity. The researchers identified a "two-regime" pattern: in data-scarce environments, simpler linear classifiers paired with high-quality embeddings generalize best, whereas in data-augmented settings, the richness of the semantic representation becomes so dominant that the specific choice of classifier matters less. By showing that text-only models using modern multilingual embeddings can rival or surpass complex multimodal approaches, the work suggests that scalable assessment tools should prioritize representation learning over intricate decision boundaries. While the study is limited by its reliance on synthetic data and a purely textual focus, it provides a strong theoretical foundation for reducing classifier reliance through superior embedding spaces, paving the way for future validation on larger, naturally occurring datasets.

Declaration of AI and AI assisted technologies in the writing process

During the preparation of this work, the author(s) used Grammarly in order to do a grammar check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of Competing Interest

We have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not Applicable

Reference

[1] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Machine learning towards intelligent systems: applications, challenges, and opportunities," *Artif. Intell. Rev.*, vol. 54, pp. 3299–3348, 2021, doi: https://doi.org/10.1007/s10462-020-09948-w.

[2] M. Marsden *et al.*, "Early clinical evaluation of a machine-learning system for risk prediction of trauma-induced coagulopathy in the prehospital setting," *Emerg. Med. J.*, 2025, doi: https://doi.org/10.1136/emermed-2024-214396.

[3] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *Int. J. Intell. Netw.*, vol. 3, pp. 58–73, Jan. 2022, doi: https://doi.org/10.1016/j.ijin.2022.05.002.

[4] M. I. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, pp. 255–260, 2015, doi: https://doi.org/10.1126/science.aaa8415.

[5] C. Du and W. Wang, "Leveraging cognitive computing for advanced behavioral and emotional data insights," *Int. J. Cogn. Comput. Eng.*, vol. 6, pp. 183–190, Dec. 2025, doi: https://doi.org/10.1016/j.ijcce.2024.12.005.

[6] A. Maimun and A. B. Semma, "Emotional Responses to Religious Conversion: Insights from Machine Learning," *Islam. Guid. Couns. J.*, vol. 6, no. 2, Art. no. 2, Oct. 2023, doi: https://doi.org/10.25217/0020236395500.

[7] G. Orrù, M. Monaro, C. Conversano, A. Gemignani, and G. Sartori, "Machine Learning in Psychometrics and Psychological Research," *Front. Psychol.*, vol. 10, Jan. 2020, doi: https://doi.org/10.3389/fpsyg.2019.02970.

[8] R. N. Landers and S. Marin, "Theory and Technology in Organizational Psychology: A Review of Technology Integration Paradigms and Their Effects on the Validity of Theory," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 8, no. Volume 8, 2021, pp. 235–258, Jan. 2021, doi: https://doi.org/10.1146/annurev-orgpsych-012420-060843.

[9] E. Acosta-Gonzaga, "The Effects of Self-Esteem and Academic Engagement on University Students' Performance," *Behav. Sci.*, vol. 13, 2023, doi: https://doi.org/10.3390/bs13040348.

[10] I. A. P. J. Wulandari, "The effect of self-esteem and self-efficacy on the academic resilience of undergraduate students in Jakarta," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 729, 2021, doi: https://doi.org/10.1088/1755-1315/729/1/012094.

[11] Y. Zhao, Z. Zheng, C. Pan, and L. Zhou, "Self-Esteem and Academic Engagement Among Adolescents: A Moderated Mediation Model," *Front. Psychol.*, vol. 12, 2021, doi: https://doi.org/10.3389/fpsyg.2021.690828.

[12] E. Asici and H. I. Sari, "Depression, Anxiety, and Stress in University Students: Effects of Dysfunctional Attitudes, Self-Esteem, and Age," *Acta Educ. Gen.*, vol. 12, no. 1, pp. 109–126, Feb. 2022, doi: https://doi.org/10.2478/atd-2022-0006.

[13] A. Derakhshan and J. Fathi, "The Interplay between Perceived Teacher Support, Self-regulation, and Psychological Well-being among EFL Students," *Iran. J. Lang. Teach. Res.*, vol. 12, no. 3 (Special Issue), pp. 113–138, Dec. 2024, doi: https://doi.org/10.30466/ijltr.2024.121579.

[14] E. Arredondo, L. Moreau, and J. G. Linn, "THE QUALITY OF LIFE IN CHILE: Historic definitions, current challenges and future possibilities," 2024.

[15] M. Bin Morshed, K. Saha, M. De Choudhury, G. D. Abowd, and T. Plötz, "Measuring Self-Esteem with Passive Sensing," in *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*, in PervasiveHealth '20. New York, NY, USA: Association for Computing Machinery, Feb. 2021, pp. 363–366. doi: https://doi.org/10.1145/3421937.3421952.

[16] R. S. Chavez, D. T. Tovar, M. Stendel, and T. D. Guthrie, "Generalizing Effects of Frontostriatal Structural Connectivity on Self-Esteem Using Predictive Modeling," 2021, doi: https://doi.org/10.31234/osf.io/p982u.

[17] N. Rubaiyat *et al.*, "Classification of depression, internet addiction and prediction of self-esteem among university students," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2019, pp. 1–6. Accessed: Apr. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9038211/, doi: https://doi.org/10.1109/ICCIT48885.2019.9038211.

[18] R. Liu *et al.*, "Building Machine Learning Predictive Models for Adolescent Internet Addiction: Key Findings on Self-Esteem and Resilience Interaction," Jan. 2025, doi: https://doi.org/10.21203/rs.3.rs-5606509/v1.

[19] J. Deng, X. Huang, and X. Ren, "A multidimensional analysis of self-esteem and individualism: A deep learning-based model for predicting elementary school students' academic performance," *Meas. Sens.*, vol. 33, p. 101147, Jun. 2024, doi: https://doi.org/10.1016/j.measen.2024.101147.

[20] A. Zaman, R. Acharyya, H. Kautz, and V. Silenzio, "Detecting Low Self-Esteem in Youths from Web Search Data," in *The World Wide Web Conference*, in WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 2270–2280. doi: https://doi.org/10.1145/3308558.3313557.

[21] R. Buettner, D. Sauter, I. Eckert, and H. Baumgartl, "Classifying High and Low Self-Esteem using a Novel Machine Learning Method based on EEG Data," *PACIS 2021*, 2021.

[22] J. Martín-Albo, J. L. Núñez, J. G. Navarro, and F. Grijalvo, "The Rosenberg Self-Esteem Scale: translation and validation in university students," *Span. J. Psychol.*, vol. 10, no. 2, pp. 458–467, 2007, doi: https://doi.org/10.1017/S1138741600006727.

[23] M. Rosenberg, “Rosenberg self-esteem scale (RSE),” *Accept. Commit. Ther. Meas. Package*, vol. 61, no. 52, p. 18, 1965.
[24] A. B. Semma, M. Saerozi, K. Kusrini, A. Syukur, and A. Maimun, “An Extreme Programming Approach to Streamlining Thesis Writing,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 13, no. 6, Art. no. 6, Dec. 2023, doi: https://doi.org/10.18517/ijaseit.13.6.18701.
[25] M. Bin Morshed, K. Saha, M. De Choudhury, G. D. Abowd, and T. Plötz, “Measuring Self-Esteem with Passive Sensing,” in *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*, Atlanta GA USA: ACM, May 2020, pp. 363–366. doi: https://doi.org/10.1145/3421937.3421952.