AutoML and BiLSTM Accuracy Gaps in Non-English Emotion Classification

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: To what extent does the accuracy gap between PyCaret AutoML and BiLSTM for fine-grained emotion classification vary when applied to non-English datasets, such as those in Spanish or Mandarin. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking PyCaret AutoML Against BiLSTM for Fine-Grained Emotion Classification: A Comparative Study on 20-Class Emotion Detection. Research question: To what extent does the accuracy gap between PyCaret AutoML and BiLSTM for fine-grained emotion classification vary when applied to non-English datasets, such as those in Spanish or Mandarin?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

12 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 4.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
All experiments are conducted on Google Colab with GPU acceleration (NVIDIA Tesla T4).	×	0.01
ML models are trained on CPU using scikit-learn [9], while DL models utilize GPU training via PyTorch [8].	×	0.05
We report the following metrics for all models: Accuracy, Macro F1-Score, Weighted F1-Score, Training Time.	×	0.09
BiLSTM model achieves the highest overall performance with an accuracy of 89%.	×	0.11
BiLSTM model uses two stacked LSTM layers (hidden size = 256) processing input sequences in both forward and backward di	×	0.03
GRU model uses two stacked GRU layers (hidden size = 256).	×	0.01
Transformer model uses 2 layers and 4 attention heads.	×	0.02
All DL models use an embedding dimension of 128, vocabulary size of 30,000, maximum sequence length of 128 tokens, dropo	×	0.02
Logistic Regression (LR) is a linear classifier optimized using the L-BFGS solver with L2 regularization.	×	0.05
Multinomial Naive Bayes (MNB) is a probabilistic classifier based on Bayes' theorem, applied with Laplace smoothing (α =	×	0.07
Support Vector Machine (LinearSVC) is a linear SVM optimized using the squared hinge loss with L2 regularization.	×	0.08

References

- http://arxiv.org/abs/1903.04717v2
- http://arxiv.org/abs/2604.26310v1
- http://arxiv.org/abs/2605.04885v1