How does the accuracy gap between traditional machine learning classifiers and BiLSTM models for spam detectio

Assignee Research

June 10, 2026

Abstract

Sentiment analysis of product reviews on e-commerce platforms plays a critical role in automatically understanding customer satisfaction and providing actionable insights for sellers seeking to improve product quality. This paper presents a comprehensive benchmarking study comparing a Machine Learning (ML) approach via the PyCaret AutoML framework against a Deep Learning (DL) approach based on a Bidirectional Long Short-Term Memory (BiLSTM) architecture with an Attention mechanism for binary sentiment classification on Indonesian product reviews. The dataset comprises 19,728 samples balanced e

1 Introduction

This paper examines: Benchmarking Logistic Regression, SVM, and LightGBM Against BiLSTM with Attention for Sentiment Analysis on Indonesian Product Reviews. Research question: How does the accuracy gap between traditional machine learning classifiers and BiLSTM models for spam detection vary across low-resource languages in the Indo-European family?.

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 4.4/10.

3 Results

12 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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
PyCaret setup() automatically handled missing value imputation and MinMax feature normalization prior to 10-fold cross-v	×	0.06
Hyperparameter optimization for the Deep Learning model was conducted over multiple trials; the best configuration (Tria	×	0.04
The Deep Learning model was trained for up to 10 epochs with early stopping based on validation loss, reaching the optim	×	0.05
All experiments were executed using Google Colab with GPU acceleration.	×	0.02
Both ML and DL models are evaluated using Accuracy, Precision, Recall, and F1-Score (macro-averaged).	×	0.10
For the ML models, metrics are averaged across 10 cross-validation folds.	×	0.05
For the DL model, metrics are computed on the held-out test set of 3,946 samples.	×	0.08
Logistic Regression emerged as the best-performing model overall in the 10-fold cross-validation results.	×	0.11
The PyCaret classification module was used to orchestrate the ML experimental pipeline with train_size=0.8 and automatic	×	0.04
Three algorithms were benchmarked: Logistic Regression (LR), Support Vector Machine (SVM) with a linear kernel, and Ligh	✓	0.28
All ML models were evaluated using 10-fold Stratified Cross-Validation on the training set, ensuring each fold maintains	×	0.09
The Deep Learning architecture was developed within the PyTorch framework, trained from scratch without pre-trained embe	×	0.06
The Deep Learning architecture consists of an Embedding Layer with a vocabulary size of 13,600 and dense vectors of dime	×	0.04

References

- http://arxiv.org/abs/2604.25452v1

- http://arxiv.org/abs/2605.01322v1
- http://arxiv.org/abs/2108.03739v2