XGLM-564M and RoBERTa Performance in Low-Resource Educational Dialogue Classification

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the performance of XGLM-564M on low-resource educational dialogue act classification compare to RoBERTa in Indonesian and English, measured by F1 scores on imbalanced datasets. 10 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: Robust Educational Dialogue Act Classifiers with Low-Resource and Imbalanced Datasets. Research question: How does the performance of XGLM-564M on low-resource educational dialogue act classification compare to RoBERTa in Indonesian and English, measured by F1 scores on imbalanced datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 papers retrieved. 10 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
The averaged F1 score of the AUC approaches generally outperformed the CE approach when the training set size was small	×	0.05
The gap between AUC and CE approaches achieved the most significant difference at 100 sentences.	×	0.04
When the training set size increased to 800 sentences, COMAUC outperformed both DAM and CE on average and demonstrated m	×	0.02
The F1 score of the CE approach was lower than those of AUC approaches (i.e., DAM and COMAUC) when the DA classifier was	×	0.11
The F1 scores of the AUC approaches also decreased under conditions of 80% FP ratio, but the decrease was less pronounce	×	0.02
The DA classifier optimized by CE approach demonstrated vulnerable performance when the FP ratios were 60% and 80% in th	×	0.10
The dataset included records of tutoring sessions where tutors and students worked together to solve problems in various	×	0.05
Our study adopted 50 tutorial dialogue sessions, which contained 3,626 utterances (2,156 tutor utterances and 1,470 stud	×	0.04
The average number of utterances per tutorial session was 72.52 (min = 11, max = 325) where tutors made an average of 43	×	0.01
The DA scheme [25] was originally designed in a two-level structure and included 31 DAs.	×	0.05

References

- http://arxiv.org/abs/2311.00958v1
- http://arxiv.org/abs/2304.07499v1
- http://arxiv.org/abs/2308.02966v1