UNIVERSITAS
GADJAH MADA
Studi Komparatif Teknik Augmentasi pada Klasifikasi Topik Cerita Berbahasa Jawa Sebagai Bahasa dengan Sumber Daya Terbatas
Intan Dwi Febryanti, Moh. Edi Wibowo, S.Kom.,M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

STUDI KOMPARATIF TEKNIK AUGMENTASI PADA KLASIFIKASI TOPIK CERITA BERBAHASA JAWA SEBAGAI BAHASA DENGAN SUMBER DAYA TERBATAS

Oleh

Intan Dwi Febryanti
22/494760/PA/21285

Bahasa Jawa merupakan salah satu bahasa daerah di Indonesia yang termasuk dalam kategori *low-resource language*, sehingga ketersediaan data berlabel untuk pengembangan model *Natural Language Processing* (NLP) masih terbatas. Keterbatasan tersebut dapat memengaruhi performa model klasifikasi teks. Penelitian ini bertujuan untuk membandingkan efektivitas berbagai teknik *data augmentation* dalam meningkatkan performa klasifikasi topik cerita berbahasa Jawa.

Penelitian menggunakan tiga dataset, yaitu dataset yang bersumber dari *Large Language Model* (LLM), dataset yang ditulis oleh penutur asli bahasa Jawa, serta dataset gabungan. Teknik *data augmentation* yang diterapkan meliputi *Easy Data Augmentation* (EDA), *Back Translation*, *Paraphrasing* berbasis GPT, dan *Mixed Augmentation*. Model klasifikasi yang digunakan adalah XLM-RoBERTa yang dilakukan *fine-tuning* pada setiap skenario eksperimen. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Sebagai validasi tambahan, dilakukan eksperimen menggunakan model *Linear Support Vector Machine* (Linear SVM) dengan metode *Repeated Stratified Cross Validation*.

Hasil penelitian menunjukkan bahwa seluruh teknik *data augmentation* mampu meningkatkan performa model dibandingkan *baseline*. Namun, tidak terdapat satu teknik augmentasi yang secara konsisten memberikan performa terbaik pada seluruh dataset. Metode *Synonym Replacement* (SR) memberikan hasil terbaik pada dataset LLM, *Random Swap* (RS) pada dataset penutur asli, dan *Random Insertion* (RI) pada dataset gabungan. Selain itu, validasi menggunakan Linear SVM menunjukkan bahwa metode berbasis EDA memberikan peningkatan performa yang konsisten dengan variasi hasil yang rendah.

Penelitian ini menunjukkan bahwa *data augmentation*, khususnya metode berbasis EDA, merupakan pendekatan yang efektif untuk meningkatkan performa klasifikasi topik cerita berbahasa Jawa sebagai bahasa dengan sumber daya terbatas.

Kata kunci: *bahasa Jawa, klasifikasi topik, augmentasi data, XLM-RoBERTa*

UNIVERSITAS
GADJAH MADA
Studi Komparatif Teknik Augmentasi pada Klasifikasi Topik Cerita Berbahasa Jawa Sebagai Bahasa dengan Sumber Daya Terbatas
Intan Dwi Febryanti, Moh. Edi Wibowo, S.Kom.,M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

A COMPARATIVE STUDY OF DATA AUGMENTATION TECHNIQUES FOR JAVANESE STORY TOPIC CLASSIFICATION AS A LOW-RESOURCE LANGUAGE

By

Intan Dwi Febryanti
22/494760/PA/21285

Javanese is one of the regional languages in Indonesia that is categorized as a *low-resource language*, resulting in limited availability of labeled data for developing *Natural Language Processing* (NLP) models. This limitation may affect the performance of text classification models. This study aims to compare the effectiveness of various *data augmentation* techniques in improving the performance of topic classification for Javanese stories.

This study employs three datasets, namely an LLM-generated dataset, a dataset written by native Javanese speakers, and a combined dataset. The applied *data augmentation* techniques include *Easy Data Augmentation* (EDA), *Back Translation*, GPT-based *Paraphrasing*, and *Mixed Augmentation*. XLM-RoBERTa was used as the classification model and fine-tuned for each experimental scenario. Model performance was evaluated using *accuracy*, *precision*, *recall*, and *F1-score*. As an additional validation, experiments were conducted using a *Linear Support Vector Machine* (Linear SVM) with the *Repeated Stratified Cross Validation* method.

The results show that all *data augmentation* techniques improve classification performance compared to the baseline model. However, no single augmentation technique consistently achieved the best performance across all datasets. *Synonym Replacement* (SR) produced the best performance on the LLM-generated dataset, *Random Swap* (RS) on the native-speaker dataset, and *Random Insertion* (RI) on the combined dataset. Furthermore, validation using Linear SVM demonstrated that EDA-based methods consistently improved model performance with low performance variation.

This study concludes that *data augmentation*, particularly EDA-based methods, is an effective approach for improving Javanese topic classification as a *low-resource language*.

Keywords: *Javanese language, topic classification, data augmentation, XLM-RoBERTa*