ABSTRACT

IMPLEMENTATION OF ROBUSTLY OPTIMIZED BERT PRETRAINING APPROACH (ROBERTA) MODEL FOR CYBERBULLYING TWEET CLASSIFICATION WITH BACK TRANSLATION DATA AUGMENTATION

By

Fathan Alhindami

The rapid development of internet technology has led to a significant increase in the number of social media users worldwide. Despite its benefits, social media can also be a space for negative behavior such as cyberbullying. This condition requires an automatic classification system that can accurately detect and identify types of cyberbullying in text. This research aims to build a RoBERTa-based text classification model and evaluate the effect of data augmentation techniques on improving model performance. The data used is a collection of English tweets that have been labeled according to the type of cyberbullying. The back translation data augmentation technique is applied to overcome imbalanced data and increase the variety of training data. The RoBERTa model that has gone through the fine-tuning process is implemented to perform classification and its performance is evaluated using accuracy, precision, recall, F1 score, and ROC-AUC metrics. The results showed that the method without data augmentation produced an accuracy of 88%, while the method with data augmentation achieved the highest accuracy of 93%, with precision, recall, and F1-score of 92% each, and an average AUC value of 95.19%. These results show that data augmentation techniques can significantly improve the model's performance in classifying cyberbullying tweets.

Keywords: Cyberbullying, RoBERTa, Data Augmentation, Text Classification.

ABSTRAK

IMPLEMENTASI MODEL *ROBUSTLY OPTIMIZED* BERT *PRETRAINING APPROACH* (ROBERTA) UNTUK KLASIFIKASI *TWEET CYBERBULLYING* DENGAN AUGMENTASI DATA *BACK TRANSLATION*

Oleh

Fathan Alhindami

Perkembangan teknologi internet yang pesat telah memberikan peningkatan yang signifikan dalam jumlah pengguna media sosial di seluruh dunia. Terlepas dari manfaatnya, media sosial juga bisa menjadi ruang munculnya perilaku negatif seperti *cyberbullying*. Kondisi ini memerlukan adanya sistem klasifikasi otomatis yang dapat mendeteksi dan mengidentifikasi jenis *cyberbullying* dalam teks secara akurat. Penelitian ini bertujuan untuk membangun model klasifikasi teks berbasis RoBERTa dan mengevaluasi pengaruh teknik augmentasi data terhadap peningkatan kinerja model. Data yang digunakan merupakan kumpulan *tweet* berbahasa Inggris yang telah diberi label sesuai jenis *cyberbullying*. Teknik augmentasi data *back translation* diterapkan untuk mengatasi data yang tidak seimbang dan menambah variasi data pelatihan. Model RoBERTa yang telah melalui proses *fine-tuning* diimplementasikan untuk melakukan klasifikasi dan kinerjanya dievaluasi menggunakan metrik *accuracy, precision, recall, F1-score*, dan ROC-AUC. Hasil penelitian menunjukkan bahwa metode tanpa augmentasi data menghasilkan *accuracy* sebesar 88%, sedangkan metode dengan augmentasi data mencapai *accuracy* tertinggi sebesar 93%, dengan *precision, recall*, dan *F1-score* masing-masing sebesar 92%, serta rata-rata nilai AUC sebesar 95,19%. Hasil ini menunjukan bahwa teknik augmentasi data secara signifikan mampu meningkatkan kinerja model dalam mengklasifikasikan *tweet cyberbullying*.

Kata kunci: *Cyberbullying*, RoBERTa, Augmentasi Data, Klasifikasi Teks.