ABSTRACT

The rise of toxic speech on social media brings negative impacts on individuals and social dynamics, making automated detection based on Natural Language Processing increasingly critical. The primary challenges in this task are the severe class imbalance between toxic and non-toxic samples, and the majority of models still being monolingual and thus unable to generalize toxic detection across languages. This study proposes an algorithm-level method as a strategy to address class imbalance in the pretrained cross-lingual language model XLM-RoBERTa for toxic comment detection in English and Indonesian without relying on translated datasets. The methods applied include a comparative study of various adaptive fusion loss configurations, two-stage fine-tuning and threshold tuning, with multi-source training data. Experimental results show that Adaptive Fusion Binary Cross-Entropy (BCE) with Focal Loss achieved the best minority-class f1-score of 0,7952, a 0,15% improvement over the Static Fusion BCE with Focal Loss baseline 0,7937. Two-stage fine-tuning further improved the minority-class f1-score to 0,8052 a 1,15% increase over the baseline on the validation set, and threshold tuning provided 2,04% minority class f1-score gain from 0,7932 to 0,8136 on the test set, with an f1-macro of 0,8947 representing a 5,58% improvement over the prior study 0,8389. Future research is recommended to explore multiclass classification, code-mixed language handling, entity and lexical bias mitigation, and real-time system deployment.

Keywords: *Adaptive Fusion Loss, Cross-lingual, Imbalance Data Handling, Toxic Comment Detection, XLM-RoBERTa*

ABSTRAK

Meningkatnya ujaran *toxic* di media sosial membawa dampak negatif terhadap individu dan dinamika sosial, sehingga deteksi otomatis berbasis *Natural Language Processing* menjadi semakin krusial. Tantangan utama dalam tugas ini adalah distribusi data yang tidak seimbang antara kelas *toxic* dan *non-toxic,* serta mayoritas model yang masih bersifat *monolingual* sehingga tidak mampu menggeneralisasi deteksi *toxic* ke bahasa lain. Penelitian ini mengusulkan pendekatan *algorithm-level method* sebagai strategi *imbalance data handling* pada *pretrained cross-lingual language model* XLM-RoBERTa untuk deteksi komentar *toxic* dalam bahasa Inggris dan bahasa Indonesia tanpa memanfaatkan dataset hasil translasi. Metode yang diterapkan mencakup studi komparatif berbagai konfigurasi *adaptive fusion loss, two-stage fine-tuning* dan *threshold tuning* dengan data latih yang bersifat *multi-source.* Hasil eksperimen menunjukkan bahwa *Adaptive Fusion Binary Cross-Entropy (BCE) with Focal Loss* menghasilkan *f1-score* kelas minoritas terbaik sebesar 0,7952, meningkat 0,15% dari *baseline Static Fusion BCE with Focal Loss 0,7937.* Penerapan *two-stage fine-tuning* meningkatkan *f1-score* kelas minoritas menjadi 0,8052 dengan peningkatan sebesar 1,15% dibandingkan *baseline* pada *validation set,* dan *threshold tuning* memberikan peningkatan *f1-score* kelas minoritas sebesar 2,04% dari 0,7932 menjadi 0,8136 pada *test set* dengan *f1-macro* 0,8947 meningkat 5,58% dibandingkan penelitian terdahulu 0,8389. Penelitian selanjutnya disarankan untuk mengeksplorasi klasifikasi multikelas, penanganan *code-mixed*, mitigasi bias entitas dan leksikal, serta implementasi model pada sistem *real-time*.

Kata Kunci: *Adaptive Fusion Loss, Cross-lingual, Imbalance Data Handling, Deteksi Komentar Toxic, XLM-RoBERTa*