PENGEMBANGAN MODEL TRANSFORMER MENGGUNAKAN PENDEKATAN SYNONYM REPLACMENT DAN CLASS WEIGHTING UNTUK MENDETEKSI UJARAN KEBENCIAN
Muhammad Salam, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

PENGEMBANGAN MODEL TRANSFORMER MENGGUNAKAN PENDEKATAN *SYNONYM REPLACMENT* DAN *CLASS WEIGHTING* UNTUK MENDETEKSI UJARAN KEBENCIAN

Oleh

Muhammad Salam

23/513107/PPA/06497

Ujaran kebencian di media sosial yang menargetkan individu maupun kelompok berdasarkan etnis, agama, gender, dan kebangsaan kian meningkat serta berpotensi mengganggu kerukunan sosial di Indonesia. Tantangan utama dalam deteksi *hate speech* mencakup keterbatasan dataset, distribusi kelas yang tidak seimbang, serta variasi definisi *hate speech* yang memunculkan inkonsistensi hasil. Kondisi ini menghambat pengembangan model yang akurat dan mampu mengakomodasi kompleksitas klasifikasi *multi-label* dan *multi-class* secara efektif.

Penelitian ini menawarkan pendekatan augmentasi data yang mengintegrasikan dua teknik, yakni *synonym replacement (SR)* dan *class weighting (CW),* untuk deteksi multi-label dan multi-class ujaran kebencian berbahasa Indonesia. *SR* memperkaya variasi leksikal dengan mengganti kata menggunakan sinonim yang relevan, sedangkan *CW* menyeimbangkan kontribusi kelas minoritas melalui pemberian bobot. Model berbasis Transformer dilatih pada korpus berbahasa Indonesia yang telah melalui tahapan *preprocessing* (pembersihan teks, normalisasi, penghapusan *stopwords*, dan tokenisasi) serta dievaluasi menggunakan metrik akurasi dan *F1- Score* pada beberapa konfigurasi.

Hasil eksperimen menunjukkan 4 skenario pembanding: *baseline* tanpa *SR* dan *CW* mencatat akurasi 76,11% dan *F1- Score* 67,55%, sementara kombinasi *SR* + *CW* mencapai akurasi 76,99% dan *F1- Score* tertinggi hingga 83,57%. Temuan ini menegaskan sinergi positif antara augmentasi leksikal dan penyeimbangan bobot kelas dalam meningkatkan ketepatan sekaligus keadilan deteksi pada skema multi-label dan multi-class. Dengan demikian, konfigurasi SR + CW berpotensi besar mendukung otomatisasi moderasi konten berbahaya di platform media sosial, membantu pemangku kepentingan menjaga ruang daring yang lebih aman dan harmonis.

Kata Kunci: Ujaran kebencian, *Transformer*, Data Augmentasi, *Synonym Replacment, Class Weighting, Multi-Label, Multi-Class*, Bahasa Indonesia

UNIVERSITAS
GADJAH MADA
PENGEMBANGAN MODEL TRANSFORMER MENGGUNAKAN PENDEKATAN SYNONYM REPLACMENT DAN CLASS WEIGHTING UNTUK MENDETEKSI UJARAN KEBENCIAN
Muhammad Salam, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

DEVELOPMENT OF A TRANSFORMER MODEL USING SYNONYM REPLACMENT AND CLASS WEIGTHING APRROACHES FOR HATE SPEECH DETECTION

Oleh

Muhammad Salam

23/513107/PPA/06497

Hate speech on social media that targets individuals and groups based on ethnicity, religion, gender, and nationality is increasingly prevalent and threatens social cohesion in Indonesia. The main challenges in hate-speech detection include limited datasets, class imbalance, and divergent definitions of hate speech that lead to inconsistent results. These issues hinder the development of accurate models capable of effectively handling the complexity of multi-label and multi-class classification.

This study proposes a data-augmentation approach that integrates two techniques synonym replacement (SR) and class weighting (CW) for multi-label and multi-class hatespeech detection in Indonesian. SR enriches lexical variation by substituting words with appropriate synonyms, while CW balances minority classes by assigning higher weights. A Transformer-based model is trained on an Indonesian corpus that has undergone preprocessing (text cleaning, normalization, stop-word removal, and tokenization) and is evaluated using accuracy and F1-score across several configurations.

Experimental results report four comparative scenarios: the baseline without SR and CW achieves 76.11% accuracy and 67.55% F1- Score, whereas the SR + CW combination attains 76.99% accuracy and the highest F1- Score of 83.57%. These findings confirm a positive synergy between lexical augmentation and class-weight balancing, improving both accuracy and fairness in multi-label, multi-class detection. Consequently, the proposed SR + CW configuration holds strong potential to support automated moderation of harmful content on social media platforms, helping stakeholders maintain a safer and more harmonious online environment.

Keywords: Hatespeech, Transformer, Data Augmentation, Synonym Replacment, Class weighting, Multi-Label, Multi-Class, Indonesian Language