UNIVERSITAS
GADJAH MADA
**IDENTIFIKASI DAN GENERALISASI UJARAN KEBENCIAN TERSIRAT MENGGUNAKAN TRIPLET NETWORK DAN
COSINE-ANGULAR BASED SEMI-HARD NEGATIVE MINING**
wicaksono leksono muhamad, Yunita Sari, S.Kom., M.Sc., Ph.D., Dr. Sri Mulyana, M.Kom.
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

IDENTIFIKASI DAN GENERALISASI UJARAN KEBENCIAN TERSIRAT MENGGUNAKAN TRIPLET NETWORK DAN *COSINE-ANGULAR BASED SEMI-HARD NEGATIVE MINING*

Oleh

Wicaksono Leksono Muhamad
23/528898/PPA/06691

Penyebaran konten provokatif di media sosial meningkat, namun deteksi ujaran kebencian tersirat tetap menantang. Karakteristik ujaran kebencian tersirat menyerupai teks netral dan menunjukkan variansi intra-kelas tinggi sehingga menyulitkan model konvensional.

Penelitian sebelumnya melakukan *fine-tuning* pada model pra-latih berbasis BERT dengan *Supervised Contrastive Learning* (SCL), menggunakan data augmentasi dan *implied statement* sebagai pasangan positif. Namun, SCL dapat menyebabkan *overclustering* karena menganggap selain pasangan tersebut sebagai negatif, sehingga model cenderung mengelompokkan teks mirip secara berlebihan dan lemah dalam generalisasi.

Penelitian ini mengintegrasikan *triplet loss* dengan *semi-hard negative mining* untuk mengurangi *overclustering*, dengan mempertimbangkan kelas pada pasangan serta memilih negatif yang sepadan agar ruang *embedding* tetap stabil. Rancangan menggunakan metrik jarak kosinus–angular dengan margin aditif pada domain radian maupun kosinus, serta fungsi *reducer* statis (*SmoothMax*) dan adaptif (*attentiveReducer*) guna memaksimalkan informasi dalam satu *batch*. Pendekatan ini meningkatkan kinerja secara konsisten dengan kenaikan *F1-Score* 1.13% pada IHC dan 1.15% pada SBIC, serta uji silang hingga 6.32%. Temuan ini menegaskan keunggulan *triplet loss* berbasis angular dalam optimasi deteksi ujaran kebencian tersirat.

Kata Kunci: **identifikasi Ujaran Kebencian tersirat**, ***Cosine based Semi-hard Negative Mining*,*triplet loss***

UNIVERSITAS
GADJAH MADA
**IDENTIFIKASI DAN GENERALISASI UJARAN KEBENCIAN TERSIRAT MENGGUNAKAN TRIPLET NETWORK DAN
COSINE-ANGULAR BASED SEMI-HARD NEGATIVE MINING**
wicaksono leksono muhamad, Yunita Sari, S.Kom., M.Sc., Ph.D., Dr. Sri Mulyana, M.Kom.
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

IDENTIFICATION AND GENERALIZATION OF IMPLICIT HATE SPEECH USING TRIPLET NETWORK AND COSINE–ANGULAR BASED WITH SEMI-HARD NEGATIVE MINING

By

Wicaksono Leksono Muhamad
23/528898/PPA/06691

The spread of provocative content on social media is increasing, yet detecting implicit hate speech remains a major challenge. The characteristics of implicit hate speech often resemble neutral text and exhibit high intra-class variance, making it difficult for conventional models to handle.

Previous studies have applied fine-tuning on pre-trained BERT-based models using Supervised Contrastive Learning (SCL), with augmented data and implied statements serving as positive pairs. However, SCL may lead to overclustering, since it treats all other instances as negatives. This causes the model to excessively cluster similar texts together, resulting in weaker generalization.

This study integrates triplet loss with semi-hard negative mining to mitigate overclustering by considering class relationships within pairs and selecting appropriate negatives so that the embedding space remains stable. The design employs cosine–angular distance metrics with additive margins in both cosine and radian domains, along with static (SmoothMax) and adaptive (attentiveReducer) reducer functions to maximize information within each batch. This approach consistently improves performance, yielding an F1-score increase of 1.13% on IHC and 1.15% on SBIC, with cross-dataset improvements of up to 6.32%. These findings highlight the effectiveness of angular-based triplet loss in optimizing implicit hate speech detection.

Keywords implicit hate speech identification, ***Cosine based Semi-hard Negative Mining***, ***triplet loss***