

ABSTRAK

Perkembangan media sosial X telah meningkatkan penyebaran kasus *cyberbullying* yang menjadi salah satu permasalahan sosial di dunia digital. Karakteristik data berupa teks yang tidak terstruktur, mengandung derau (*noise*), serta penggunaan bahasa tidak baku menyebabkan proses identifikasi isu *cyberbullying* menjadi lebih kompleks. Penelitian ini bertujuan mengintegrasikan metode *Named Entity Recognition* (NER) dan *Topic Modeling Latent Dirichlet Allocation* (LDA) untuk mengekstraksi entitas penting serta mengidentifikasi topik utama yang berkaitan dengan *cyberbullying* pada *platform* X. Data penelitian diperoleh melalui proses *web scraping* sebanyak 2.000 *tweet* atau komentar yang kemudian diproses melalui tahapan *text preprocessing*, ekstraksi entitas menggunakan NER, pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF), pemodelan topik menggunakan LDA, serta proses Labeling Kategori untuk memetakan topik ke dalam kategori *body shaming*, rasis, gender, dan netral. Hasil pengujian menunjukkan bahwa sistem memiliki tingkat akurasi sebesar 80,25%, dengan nilai *F1-Score* masing-masing sebesar 94% pada kategori rasis, 93% pada kategori gender, 70% pada kategori *body shaming*, dan 63% pada kategori netral. Hasil penelitian menunjukkan bahwa integrasi NER dan LDA mampu mengidentifikasi entitas penting, menemukan topik utama, serta memetakan isu *cyberbullying* ke dalam kategori yang telah ditentukan. Penelitian ini memberikan kontribusi berupa pendekatan analisis yang mengintegrasikan informasi entitas dan topik sehingga menghasilkan representasi isu *cyberbullying* yang lebih bagus dibandingkan penggunaan satu metode saja. Sistem yang dikembangkan diharapkan dapat mendukung penelitian lanjutan serta menjadi sarana dalam memahami pola *cyberbullying* di media sosial.

Kata Kunci: *Cyberbullying, Latent Dirichlet Allocation, Named Entity Recognition, Platform X, TF-IDF.*

ABSTRACT

The rapid growth of the X social media platform has contributed to the increasing prevalence of cyberbullying, making it a significant social issue in the digital era. The characteristics of textual data, including its unstructured nature, the presence of noise, and the use of informal language, make the identification of cyberbullying issues more challenging. This study aims to integrate Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) Topic Modeling to extract important entities and identify the main topics related to cyberbullying on the X platform. The dataset consisted of 2,000 tweets and comments collected through a web scraping process. The collected data were processed through several stages, including text preprocessing, entity extraction using NER, Term Frequency–Inverse Document Frequency (TF-IDF) weighting, topic modeling using LDA, and a Category Labeling process to map the extracted topics into four categories: body shaming, racism, gender, and neutral. The test results show that the system achieved an accuracy of 80.25%, with F1-scores of 94% for the racism category, 93% for the gender category, 70% for the body shaming category, and 63% for the neutral category. The findings demonstrate that the integration of NER and LDA is capable of identifying important entities, discovering the main discussion topics, and mapping cyberbullying issues into predefined categories. This study contributes an analytical approach that integrates entity and topic information, resulting in a better representation of cyberbullying issues compared to using a single method alone. Furthermore, the developed system has the potential to support future research and serve as a tool for understanding cyberbullying patterns on social media.

Keywords: Cyberbullying, Latent Dirichlet Allocation, Named Entity Recognition, Platform X, TF-IDF.