ABSTRAK

DETEKSI *HATE SPEECH* PADA KOLOM KOMENTAR TIKTOK DENGAN MENGGUNAKAN ALGORITMA *NAÏVE BAYES* DENGAN *SMOOTHING*

Oleh

ROY RAFLES MATORANG PASARIBU

Hate speech adalah ekspresi yang bersifat bias, bermusuhan, dan diskriminatif yang ditujukan kepada individu atau kelompok berdasarkan karakteristik mereka. Di era digital, *platform* media sosial seperti TikTok sangat rentan terhadap penyebaran *hate speech*, terutama melalui kolom komentar. Metode tradisional untuk mendeteksi *hate speech* sering kali tidak efektif dalam menangani volume data yang besar dan beragamnya ekspresi pengguna. Oleh karena itu, penelitian ini mengembangkan model deteksi *hate speech* berbasis *Multinomial Naïve Bayes* dengan *smoothing*. Penerapan *smoothing* bertujuan untuk mengatasi permasalahan *zero probability* pada kata-kata yang jarang muncul, sehingga meningkatkan akurasi prediksi model. *Dataset* dibagi dengan rasio 80:20, di mana 80% digunakan untuk pelatihan dan 20% untuk pengujian. Model yang dikembangkan mencapai akurasi sebesar 88,41%, dengan nilai presisi, *recall*, dan *F1-score* yang serupa. Evaluasi lebih lanjut melalui pengujian pengguna dengan 35 partisipan yang menganalisis 7.415 komentar menunjukkan akurasi deteksi sebesar 68,6%. Sebagai implementasi, model ini diintegrasikan ke dalam *plugin* Google Chrome yang mendeteksi *hate speech* pada komentar TikTok secara *real time*, dilengkapi dengan visualisasi probabilitas dan mekanisme validasi pengguna.

Kata kunci : Deteksi *hate speech*, *Naïve Bayes*, *smoothing*, TikTok, analisis sentimen, *machine learning*, *plugin* Google Chrome.

ABSTRACT

HATE SPEECH DETECTION IN TIKTOK COMMENT COLUMN USING NAÏVE BAYES ALGORITHM WITH SMOOTHING

By

ROY RAFLES MATORANG PASARIBU

Hate speech is a biased, hostile, and discriminatory expression directed at individuals or groups based on their inherent characteristics. In the digital era, social media platforms such as TikTok have become highly susceptible to the spread of hate speech, particularly through comment sections. Traditional methods for detecting hate speech are often ineffective in handling large volumes of data and diverse user expressions. Therefore, this study develops a hate speech detection model based on Multinomial Naïve Bayes with smoothing. The application of smoothing aims to address the issue of zero probability for rarely occurring words, thereby enhancing the model's predictive accuracy. The dataset is split into an 80:20 ratio, with 80% used for training and 20% for testing. The developed model achieves an accuracy of 88.41%, with precision, recall, and F1-score at similar levels. Further evaluation through user testing with 35 participants analyzing 7,415 comments indicates a detection accuracy of 68.6%. As an implementation, this model is integrated into a Google Chrome plugin that detects hate speech in TikTok comments in real time, featuring probability visualization and user validation mechanisms.

Keywords : *Hate speech detection, Naïve Bayes, TikTok, sentiment analysis, machine learning, Google Chrome plugin.*