ABSTRAK

Penelitian ini berfokus pada klasifikasi ujaran kebencian *(hate speech)* pada komentar TikTok berbahasa Indonesia. Platform TikTok sebagai media sosial dengan intensitas interaksi tinggi menghasilkan volume komentar yang besar dengan karakteristik bahasa yang beragam, termasuk penggunaan bahasa formal dan non-formal. Variasi linguistik tersebut menimbulkan tantangan dalam proses moderasi konten, khususnya dalam mengidentifikasi ujaran kebencian secara otomatis. Oleh karena itu, penelitian ini diarahkan untuk membangun model klasifikasi teks yang mampu mengenali pola ujaran kebencian pada komentar dengan karakteristik bahasa non-standar. Algoritma *Multinomial Naïve Bayes* digunakan sebagai metode klasifikasi karena kemampuannya dalam memodelkan distribusi kata pada data teks, sementara pembobotan fitur dilakukan menggunakan *Term Frequency–Inverse Document Frequency (TF-IDF)* untuk meningkatkan representasi numerik teks. *Dataset* penelitian merupakan data sekunder yang diperoleh melalui penggabungan *dataset* publik dan komentar TikTok hasil scraping dengan total awal sebanyak 5.698 komentar. Data yang dikumpulkan merepresentasikan komentar pengguna secara umum dengan variasi bahasa formal dan non-formal. Untuk meningkatkan kualitas data, dilakukan tahapan prapemrosesan yang meliputi pembersihan teks, tokenisasi, normalisasi, penghapusan *stopword*, dan *stemming*. Setelah *preprocessing*, diperoleh 4.542 komentar yang layak digunakan dalam proses pemodelan. Pelabelan data dilakukan menggunakan pendekatan *keyword-based* labeling untuk mengidentifikasi komentar yang mengandung ujaran kebencian. *Dataset* kemudian dibagi menggunakan rasio 80:20 menjadi data latih dan data uji. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa model *Multinomial Naïve Bayes* dengan pembobotan TF-IDF mampu mengklasifikasikan ujaran kebencian dengan performa yang tinggi. Akurasi model mencapai 93% sebelum optimasi parameter dan meningkat menjadi 95% setelah dilakukan hyperparameter tuning dengan nilai alpha sebesar 0,5. Hasil *confusion matrix* menunjukkan tingkat kesalahan klasifikasi yang relatif rendah, meskipun distribusi kelas pada *dataset* masih menunjukkan ketidakseimbangan. Temuan penelitian ini mengindikasikan bahwa pendekatan *Multinomial Naïve Bayes* efektif dalam mengenali pola linguistik ujaran kebencian pada komentar TikTok berbahasa Indonesia, termasuk pada teks dengan karakteristik bahasa non-formal.

Kata Kunci: Ujaran kebencian, TikTok, *Multinomial Naive Bayes*, TF-IDF, klasifikasi teks.

ABSTRACT

This study focuses on the classification of hate speech in Indonesian-language TikTok comments. As a social media platform with a high level of user interaction, TikTok generates a large volume of comments characterized by diverse linguistic patterns, including both formal and informal language. Such linguistic variations present significant challenges for content moderation, particularly in the automatic detection of hate speech. Therefore, this research aims to develop a text classification model capable of identifying hate speech patterns within comments that exhibit non-standard language characteristics. The Multinomial Naïve Bayes algorithm is employed as the primary classification method due to its effectiveness in modeling word distributions in textual data, while feature weighting is performed using the Term Frequency–Inverse Document Frequency (TF-IDF) method to enhance numerical text representation. The dataset used in this study consists of secondary data obtained by combining a public dataset and TikTok comments collected through scraping, resulting in an initial total of 5,698 comments. The collected data represent general user comments containing variations of formal and informal language. To improve data quality, several preprocessing steps were applied, including text cleaning, tokenization, normalization, stopword removal, and stemming. After preprocessing, 4,542 comments were deemed suitable for modeling. Data labeling was conducted using a keyword-based labeling approach to identify hate speech instances. The dataset was then divided into training and testing sets using an 80:20 ratio. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The experimental results indicate that the Multinomial Naïve Bayes model with TF-IDF weighting achieves high classification performance. The model obtained an accuracy of 93% prior to parameter optimization, which improved to 95% after hyperparameter tuning with an alpha value of 0.5. The confusion matrix results demonstrate a relatively low classification error rate, although class imbalance remains present within the dataset. These findings suggest that the Multinomial Naïve Bayes approach is effective in recognizing linguistic patterns of hate speech in Indonesian TikTok comments, including texts characterized by informal language usage.

Keywords: Hate speech, TikTok, Multinomial Naïve Bayes, TF-IDF, text classification.