HATE SPEECH DETECTION PADA VIDEO MENGGUNAKAN METODE *KNN* DAN *NAIVE BAYES*

Christopher Kelvin Pintoro Kwan

Teknik Informatika – Data Science and Artificial Intelligence

Pembimbing:

Vincentius Riandaru Prasetyo, S.Kom., M.Cs., Fitri Dwi Kartikasari, S.Si., M.Si.

ABSTRAK

Hate speech atau ujaran kebencian sudah memberikan banyak dampak yang negatif di Indonesia seperti kerusuhan, pertengkaran fisik maupun verbal, perpecahan di masyarakat, dan masih banyak lagi. Sosial media menjadi tempat untuk menyebarkan *hate speech* paling cepat. Tidak hanya melalui postingan teks, cukup sering juga ditemukan *hate speech* berbentuk video. Dalam penelitian ini, peneliti akan membuat model yang menerapkan model *machine learning* untuk mendeteksi adanya *hate speech* dalam video dimana saat ini kebanyakan model *machine learning* digunakan untuk mendeteksi *hate speech* dalam bentuk teks saja. Dalam penerapannya, model akan mengubah video yang diinput menjadi teks menggunakan *Google API*. Kemudian klasifikasi akan dilakukan menggunakan *Naive Bayes* untuk mengklasifikasikan apakah video *hate speech* atau bukan, dan *KNN* untuk mengklasifikasikan konteks dari video. Pada *dataset* yang tidak seimbang hasil klasifikasi yang didapatkan pada klasifikasi *hate speech* adalah 74% dan klasifikasi konteks video didapatkan akurasi sebesar 45%. Pada *dataset* yang seimbang namun terjadi *overfitting* akurasi yang didapatkan pada klasifikasi *hate speech* adalah 93% dan pada klasifikasi konteks video didapatkan akurasi 55%. Berdasarkan hasil uji coba didapatkan bahwa model yang digunakan dapat memiliki akurasi yang baik apabila *dataset* yang digunakan seimbang antar label dan tidak ada *overfitting* pada label.

Kata kunci : *Hate speech, Machine learning, KNN, Naive Bayes*

HATE SPEECH DETECTION IN VIDEO USING KNN AND NAIVE BAYES METHOD

Christopher Kelvin Pintoro Kwan

Informatics – Data Science and Artificial Intelligence

Contributors:

Vincentius Riandaru Prasetyo, S.Kom., M.Cs., Fitri Dwi Kartikasari, S.Si., M.Si.

ABSTRACT

Hate speech has had many negative impacts in Indonesia, such as riots, physical and verbal altercations, divisions in society, and many more. Social media is the place to spread hate speech most quickly. Not only through text posts, it is quite common to find hate speech in the form of videos. In this research, researchers will create a model that applies machine learning models to detect hate speech in videos, where currently most machine learning models are used to detect hate speech in text form only. In its application, the model will convert the input video into text using Google API. Then classification will be carried out using Naive Bayes to classify whether the video is hate speech or not, and KNN to classify the context of the video. In an unbalanced dataset, the classification results obtained for hate speech classification were 74% and for video context classification the accuracy was 45%. In a balanced dataset but overfitting occurs, the accuracy obtained in hate speech classification is 93% and in video context classification the accuracy is 55%. Based on the test results, it was found that the model used can have good accuracy if the dataset used is balanced between labels and there is no overfitting on the labels.

Kata kunci : ***Hate speech, Machine learning, KNN, Naive Bayes***