ABSTRAK

Pelabelan data merupakan salah satu langkah dalam membuat data latih dan data uji pada analisis sentimen dengan pendekatan *machine learning*. Data yang telah dilabeli dibutuhkan dalam metode *supervised learning* untuk dapat mengklasifikasikan data baru. Pada umumnya, pelabelan data dilakukan secara manual oleh ahli dalam bidang terkait. Namun, pelabelan data secara manual merupakan pekerjaan berat yang membutuhkan waktu lama, apalagi jika datanya berjumlah banyak, dan rawan terjadi *human-error*.

Terdapat beragam metode *lexicon-based* yang dapat digunakan untuk melabeli data secara otomatis pada analisis sentiment, salah satunya adalah VADER (*Valence Aware Dictionary and sEntiment Reasoner*). Metode merupakan metode *lexicon-based* yang sensitif dalam konteks *micro-blog* dan mampu mendeteksi intensitas kekuatan emosional yang tersirat dari sebuah teks. Penelitian ini menggabungkan metode pelabelan VADER dan algoritma *Support Vector Machine* (SVM) kemudian membandingkannya dengan model gabungan pelabelan manual dan algoritma SVM.

Data yang digunakan pada penelitian ini diambil dari Twitter tentang ujaran kebencian bertajuk *anti-Asian* yang muncul selama pandemi COVID-19. Terdapat 1530 data, yang kemudian dibagi menjadi 1224 data latih dan 306 data uji. Berdasarkan hasil pengujian, model dengan gabungan pelabelan VADER dan algoritma SVM memperoleh nilai akurasi, presisi, *recall*, dan *f1-score* yang lebih baik daripada gabungan pelabelan manual dan algoritma SVM. Gabungan pelabelan VADER dan algoritma SVM memperoleh nilai akurasi, presisi, *recall*, dan *f1-score* sebesar 90%. Sedangkan gabungan pelabelan manual dan algoritma SVM memperoleh nilai akurasi, presisi, *recall*, dan *f1-score* sebesar 85%. Selain itu, dilakukan pengujian dengan data uji hasil pelabelan VADER pada model dengan gabungan pelabelan manual dan algoritma SVM yang memperoleh nilai akurasi, presisi, *recall*, dan *f1-score* sebesar 88%.

Kata kunci: analisis sentimen, ujaran kebencian, Twitter, VADER, *Support Vector Machine*

ABSTRACT

Data labeling is one of the steps in creating training data and data testing in sentiment analysis with a machine learning approach. Data that has been labeled is needed in the supervised learning method to be able to classify new data. In general, data labeling is done manually by experts in related fields. However, manually labeling data is hard work that takes a long time, especially if there is a large amount of data, and it is prone to human-error.

There are various lexicon-based methods that can be used to automatically label data in sentiment analysis, one of which is VADER (Valence Aware Dictionary and Sentiment Reasoner). The method is a lexicon-based method that is sensitive in the context of micro-blogs and is able to detect the intensity of emotional strength implied from a text. This research combines the VADER labeling method and the Support Vector Machine (SVM) algorithm and then compares it with the combined model of manual labeling and the SVM algorithm.

The data used in this study was taken from Twitter regarding anti-Asian hate speech that appeared during the COVID-19 pandemic. There are 1530 data, which is then divided into 1224 training data and 306 test data. Based on the test results, the model with a combination of VADER labeling and the SVM algorithm obtained better accuracy, precision, recall, and f1-score than the combination of manual labeling and the SVM algorithm. The combination of VADER labeling and the SVM algorithm obtains an accuracy, precision, recall, and f1-score of 90%. Meanwhile, the combination of manual labeling and the SVM algorithm obtained an accuracy, precision, recall, and f1-score of 85%. In addition, testing was carried out by testing the data resulting from VADER labeling on a model with a combination of manual labeling and the SVM algorithm which obtained an accuracy, precision, recall, and f1-score of 88%.

Keywords: sentiment analysis, hate speech, Twitter, VADER, Support Vector Machine