Klasifikasi Toxic Comment Pada Twitter Menggunakan Metode SVM danWord Embeddings

Gede Agus Hendra C.[1], Prof. Dr. Adiwijaya, S.Si., M.Si. [2], Mahendra Dwifeberi P., S.Kom., M.Kom.[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]hendrahdr@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]mahendradp@telkomuniversity.ac.id

Abstrak

Perkembangan teknologi yang pesat pada era modern ini, sosial media merupakan sebuah sarana untuk melakukan interasi dan mengutarakan pendapat secara online. Meskipun memiliki dampak positif, sosial media juga memiliki dampak negatif dari penggunaan sosial media salah satunya adalah *cyberbullying* dalam bentuk *toxic comment. Toxic comment* dapat dibagi menjadi banyak jenis, maka penelitian ini dilakukan untuk mengklasifikasikan *toxic comment* tersebut dengan menggunakan metode *machine learning.* Dalam penelitian ini dilakukan pengumpulan data melalui *data crawling* pada twitter dengan jumlah data 1.200 *records.* Keseluruhan data tersebut lalu diberi label secara manual dengan menggunakan 4 label, antara lain *non-toxic*, SARA, katakasar, dan ujaran kebencian. Algoritma yang digunakan untuk pengklasifikasian menggunakan *word embeddings* dengan pendekatan *word2vec* sebagai *feature extraction* dan *support vector machine* (SVM) sebagai *classifier.* Pada penelitian ini didapatkan hasil f1-score paling tinggi 96% dengan menggunakan SVM dengan kernel *polynomial* pada label 'Toxic' dan *word2vec* dengan dimensi 100. Hasil analisis pada penelitian ini menunjukan algoritma *word embeddings* dengan pendekatan *word2vec* dapat meningkatkan hasil f1-score dari *classifier* SVM.

Kata kunci: svm, word2vec, *toxic comment*, sosial media.

Abstract

Rapid technological developments in this modern era, social media is a means to interact and express opinions online. Although it has a positive impact, social media also has a negative impact from the use of social media, one of which is cyberbullying in the form of toxic comments. Toxic comments can be divided into many types, so this study was conducted to classify these toxic comments using machine learning methods. In this study, data was collected through data crawling on Twitter with a total of 1,200 records. The entire data is then manually labeled using 4 labels, including non-toxic, SARA, abusive words, and hate speech. The algorithm used for classification uses word embeddings with a word2vec approach as feature extraction and support vector machine (SVM) as a classifier. In this study, the highest accuracy results were 96% using SVM polynomial on 'Toxic' label and vector dimensions of 00. The results of the analysis in this study showed that the word embeddings algorithm with the word2vec approach could improve the accuracy of the SVM classifier.

Keywords: svm, word2vec, toxic comment, social media.

1. Pendahuluan

1.1 Latar Belakang

Perkembangan teknologi yang pesat pada era modern ini, sosial media merupakan salah satu media yang sangat popular di kalangan masyarakat. Berdasarkan hasil survey yang diambil dari tahun 2005-2015 di Amerika Serikat, jumlah pengguna sosial media meningkat secara signifikan setiap tahunnya [1]. Per Januari 2021 pengguna internet di Indonesia sudah mencapai 202.6 juta pengguna dan pengguna sosial media sudah mencapai 170 juta pengguna dengan peningkatan 10 juta (6.3%) pengguna dari tahun 2020 [2].

Akibat semakin populernya sosial media di kalangan pengguna internet, terjadi satu hal yang marak terjadi yaitu cyberbullying. Cyberbullying adalah kekerasan verbal yang secara disengaja yang dilakukan berulang – ulang melalui teks elektronik [3]. Dari beberapa studi yang dilakukan menemukan bahwa cyberbullying sering ditemukan melalui sosial media yang berbasis pesan teks seperti Facebook dan Twitter, di sosial media cyberbullying sering kali ditemukan pada kolom komentar karena anonimitas yang dirasakan oleh korban cyberbullying [4]. Cyberbullying yang terjadi pada kolom komentar tersebut dapat dikatakan sebagai toxic comment. Toxic comment merupakan komentar kasar berupa ancaman, pelecehan seksual, maupun diskriminasi yang dapat mempengaruhi psikologi seseorang sehingga mereka akan meninggalkan