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

kolom diskusi [5].

Dengan maraknya cyberbullying yang terjadi, maka akan dilakukan klasifikasi teks menggunakan metode machine learning untuk mengidentifikasi komentar – komentar yang ada pada sosial media. Dalam penelitian ini metode classifier yang dipilih adalah Support Vector Machine (SVM) dan Word Embedding digunakan untuk feature extraction. Metode SVM ini telah dibandingkan dengan berbagai metode lain, salah satunya adalah Decision tree. Dari penelitian tersebut menunjukan bahwa SVM memiliki hasil yang lebih baik daripada metode Decision Tree [6]. Dalam studi yang telah dilakukan juga menunjukan bahwa Word Embedding dapat meningkatkan akurasi dari classifier klasifikasi teks [7]. Data diambil dari Twitter menggunakan *library tweepy* dengan memanfaatkan fitur *hashtag* pada Twitter .

1.2 Topik dan Batasannya

Topik yang dibahas dalam tugas akhir ini adalah bagaimana mengimplementasikan metode *SVM* dan *word embeddings* untuk melakukan klasifikasi teks dan bagaimana mengukur performansi dari metode yang digunakan.

Batasan masalah pada tugas akhir ini adalah sebagai berikut: Pertama, sistem hanya mampu untuk menerima masukan berupa *metadata* dari *test case* dan memberi luaran berupa hasil f1-score dari pengimplementasian dari algoritma yang dipakai. Kedua pengklasifikasian *toxic comment* untuk *labelling* dibagi menjadi 4, antara lain non-toxic, SARA, kata kasar, ujaran kebencian.

1.3 Tujuan

Tujuan dari tugas akhir ini adalah untuk menguji performasi *classifier* SVM dengan *Word* embeddings sebagai *feature extraction*-nya pada kasus klasifikasi *toxic comment*.

1.4 Organisasi Tulisan

Struktur penulisan dari tugas akhir ini disusun sebagai berikut: Bagian pertama berisi pendahuluan terkait tugas akhir ini. Bagian kedua menjelaskan studi yang terkait dengan tugas akhir ini. Bagian ketiga akan menjelaskan pemodelan dan performansi dari sistem yang dibangun. Bagian keempat menjelaskan hasil dan evaluasi hasil pengujian yang telah dilakukan pada bagian ketiga. Kemudian, pada bagian terakhir menjelaskan kesimpulan dan saran berdasarkan hasil pengujian yang dilakukan pada tugas akhir ini.

2. Kajian Pustaka

2.1 Penelitian Terkait

Beberapa penelitian untuk mengklasifikasikan toxic comment sudah dilakukan, baik dalam Bahasa Indonesia maupun Bahasa Asing. Terdapat penelitian klasifikasi toxic comment dengan menggunakan pendekatan Multi-class Multi-label Word Embedding sebagai feature extraction. Mengguanakan dua dataset yang sama dengan dua metode klasifikasi yang berbeda yaitu, Logistic Regression dan KKross dengan k =10 sebagai classifier. Hasil yang didapatkan dari penelitian ini mendapatkan hasil akurasi dari Logistic Regression dan K-Kross mengalami peningkatan setelah menggunakan Word Embedding sebagai feature extraction untuk klasifikasi teks [7]. Terdapat penelitian lainnya yang menggunakan metode deeplearning dan data augmentation. Metode – metode yang digunakan sebagai classifier adalah Convusional Neural Network, CNN Ensemble, Bidiretional LSTM dan Bidirectional GRU. Hasil yang ditunjukan dari penelitian tersebut adalah bahwan CNN Ensemble memiliki nilai f-score yang lebih tinggi dibandingkan dengan metode – metode lainnya [8]. Terdapat juga penelitian yang menggunakan Logixtic Regression dan Neural Network Model untuk melakukan klasifikasi teks. Untuk Neural Network Model metode yang digunakan sebagai classifier adalah Convusional Neural Network, LTSM dan gabungan antara Convusional Neural Network Model (Conv) dengan LTSM. Hasil yang ditunjukan dari penelitian tersebut adalah kombinasi metode Conv dengan LTSM memiliki nilai akurasi paling tinggi diantara metode yang lain [9]. Terdapat juga penelitian lainnya tentang klasifikasi toxic comment dengan menggunakan pendekatan algoritma machine learning. Pada penelitian ini terdapat enam metode machine learning yang digunakan sebagai classifier. Hasil yang didapatkan dari penelitian ini menunjukkan bahwa metode Logistic regression memiliki performansi yang lebih baik dari kelima metode lainnya dengan nilai akurasi maksimum yaitu, 89.64% dengan Hamming Loss sebesar 2.43% [10].

2.2 Klasifikasi Multilabel

Klasifikasi Multilabel adalah suatu metode yang digunkan untuk melakukan klasifikasi data yang digunakan dalam menentukan label kelas untuk data yang memiliki lebih dari satu label kelas. Kategori untuk metode klasifiaksi multilabel dapat dibagi menjadi yaitu, problem transformation methods dan algorithm adaptation method [11]. Problem transformation methods adalah metode yang mentransformasi klasifikasi multilabel menjadi satu atau lebih atau regresi label tunggal. Sedangkan algorithm adaptation method adalah metode yang memperluas algoritma pembelajaran secara khusus untuk menangani data multilabel secara langsung [11].