Analisis Perbandingan Performansi CNN dan LSTM dalam Mendeteksi Ujaran Kebencian di Twitter

Artisa Bunga Syahputri[1], Yuliant Sibaroni[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]artisabunga@students.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id,

Abstrak

Perkembangan internet pada saat ini menjadi salah satu faktor yang memberi peluang bagi pengguna media sosial untuk meninggalkan komentar dan postingan yang mengandung ujaran kebencian. Mendeteksi ujaran kebencian pada media sosial terutama pada Twitter menjadi topik yang banyak diteliti beberapa waktu belakangan. Penelitian yang telah dilakukan biasanya menerapkan pendekatan machine learning biasa, dan saat ini pendekatan deep learning menjadi populer karena dianggap dapat memberikan hasil yang lebih baik dan lebih efektif, namun masih jarang diterapkan untuk melakukan deteksi ujaran kebencian pada teks yang berbahasa Indonesia. Penelitian ini menunjukan hasil perbandingan performansi dari pendekatan deep learning dengan menggunakan model arsitektur CNN, LSTM, dan kombinasi CNN+LSTM untuk mendeteksi ujaran kebencian pada postingan Twitter yang menggunakan bahasa Indonesia. Dataset yang digunakan terdiri dari data umum yang merupakan keseluruhan dataset dan data dengan topik yang spesifik membahas mengenai pemerintahan yang diambil dari dataset umum. Dari penelitian yang telah dilakukan menampilkan hasil yang lebih baik saat model arsitektur CNN diimplementasikan pada data tweet berbahasa Indonesia dibandingkan hasil yang diperoleh dari model arsitektur LSTM dan kombinasi CNN+LSTM dengan nilai akurasi dan F1-score mencapai 81%. Selain itu penerapan model deep learning dalam mendeteksi ujaran kebencian memberikan performansi yang lebih baik dibandingkan dengan penelitian sebelumnya dengan dataset yang sama namun menerapkan model machine learning dengan ekstraksi fitur. Penelitian ini juga menunjukan bahwa data yang spesifik membahas satu topik khusus dan pemilihan parameter model memberikan pengaruh yang signifikan pada performansi model sehingga performa model menjadi lebih meningkat dibandingkan saat diterapkan pada data dengan topik yang umum dan memiliki kamus kata yang lebih banyak.

Kata kunci : ujaran kebencian, Twitter, deep learning, CNN, LSTM

Abstract

The internet's current development has become one factor that gives social media users opportunities to leave comments and posts containing hate speech. Detecting hate speech on social media, particularly on Twitter, has recently become a widely researched topic. Research that has been conducted usually applies a standard machine learning approach. The deep learning approach has become popular because it provides better and more effective results. However, it's still rare to be applied to detect hate speech in Indonesian language texts. This research shows the results of a performance comparison from the deep learning approach using CNN, LSTM, and CNN+LSTM architecture models for detecting hate speech in tweets using the Indonesian language. The dataset used is divided into a general dataset which is the entire dataset and a specific topic dataset that deals with the topic of government, which was taken from the general dataset. The research shows better results when the CNN architecture model is implemented on Indonesian language tweet data compared to the results obtained from the LSTM architecture model and the combination of CNN+LSTM with accuracy and F1 score reaching 81\%. Furthermore, the implementation of deep learning models in detecting hate speech performs better than previous research using the same dataset but applying machine learning models with feature extraction. This research also shows that specific data discussing a particular topic and selection of model parameters significantly impact the model's performance. Thus the version of the model becomes better when applied to data with a general topic and a more extensive vocabulary.

Keywords: hate speech, Twitter, deep learning, CNN, LSTM.

1. Pendahuluan

Latar Belakang

Sosial media menjadi salah satu media informasi dan komunikasi yang digunakan hampir seluruh masyarakat di dunia untuk dapat saling berkomunikasi dan memberikan kemudahan dalam mengunggah berbagai postingan berupa teks, gambar, video, dan audio. Hal ini menjadi salah satu peluang bagi beberapa orang untuk dengan bebas memberikan ujaran kebencian dalam bentuk postingan maupun komentar pada orang lain [1]. Twitter menjadi salah satu media sosial yang banyak digunakan saat ini. Twitter membatasi penggunaan karakter dalam sebuah

postingan serta memungkinkan pengguna untuk memposting gambar, video, dan meninggalkan komentar pada postingan dari pengguna lain yang diikuti[2]. Beberapa Perusahaan besar, salah satunya Twitter, telah berusaha untuk mengurangi dan menghapus konten-konten yang mengandung ujaran kebencian[3] dengan menerapkan beberapa pendekatan seperti machine learning.

Beberapa penelitian telah dilakukan untuk melakukan deteksi terhadap postingan yang berisi ujaran kebencian pada media sosial terutama Twitter dengan menerapkan pendekatan machine learning dan telah dibuktikan efektif dalam mendeteksi teks berisi ujaran kebencian [3], [4]. Namun saat ini penggunaan metode deep learning menjadi populer karena mampu memberikan performa yang lebih baik dibandingkan metode machine learning traditional dan lebih baik dalam *topic classification, sentiment analysis, question answering*, dan *language translation* [5]. Algoritma deep learning memungkinkan model komputasi dengan lapisan-lapisan pemprosesan untuk dapat mempelajari representasi data dengan beberapa tingkat abstraksi [5] dan mampu mempelajari fitur-fitur penting dari data secara otomatis[6].

Beberapa metode deep learning diantaranya adalah Convolutional Neural Network (CNN) yang merupakan salah satu algoritma deep learning yang mampu memberikan hasil yang memuaskan saat diaplikasikan pada beberapa bidang salah satunya pada Natural Language Processing (NLP)[7], dan Long-Short Term Memory (LSTM) yang terkenal karena merupakan modifikasi dari RNN yang mampu melakukan pemrosesan terhadap data sekuensial yang cukup panjang[2].

Pada penelitian yang dilakukan pada tahun 2020 [2], dengan menerapkan metode CNN dalam mendeteksi ujaran kebencian dan menunjukan hasil yang bagus saat dibandingkan dengan baseline model yang digunakan yaitu model SVM, namun penelitian ini menggunakan dataset berbahasa inggris yang memiliki struktur bahasa yang cukup kompleks dibandingkan bahasa lain. Penerapan metode deep learning saat ini kebanyakan dilakukan untuk data set berbahasa inggris, sedangkan data set yang menggunakan bahasa lain masih jarang digunakan[8]termasuk bahasa Indonesia. Penelitian [9] melakukan deteksi untuk ujaran kebencian di Facebook menggunakan SVM dan LSTM dimana nilai F1 score dari model LSTM mencapai 72% dan lebih baik dari baseline model yang sama yaitu model SVM. Dari dua penelitian sebelumnya ditampilkan bahwa model CNN dan LSTM memberikan performa yang lebih baik dibandingkan baseline model yang sama yaitu model machine learning SVM.

Pada penelitian ini akan dilakukan analisis terhadap perbandingan performansi antara dua model deep learning yaitu CNN, LSTM, dan membandingkan performansi dari kombinasi CNN+LSTM dalam mendeteksi ujaran kebencian pada media sosial Twitter namun terbatas pada data berbahasa Indonesia yang diambil dari postingan dan komentar pada media sosial Twitter. Hasil riset ini juga akan dibandingkan dengan penelitian [10] yang menggunakan dataset yang sama.

Topik dan Batasannya

Permasalahan yang diangkat dalam penelitian ini adalah mengenai analisis perbandingan performansi antara model deep learning CNN, LSTM, dan kombinasi antara CNN dan LSTM dalam mendeteksi ujaran kebencian berbahasa Indonesia pada media sosial Twitter. Dari permasalahan tersebut maka hal yang akan dijabarkan dalam penelitian ini adalah pertama, bagaimana perbandingan performansi dari model CNN dan LSTM dalam kasus mendeteksi ujaran kebencian, dan kedua, bagaimana performansi dari model CNN dan LSTM ketika dikombinasikan dalam mendeteksi ujaran kebencian.

Penelitian ini terbatas pada data set berupa teks berbahasa Indonesia yang diambil dari postingan dan komentar pada media sosial Twitter dan performansi diukur berdasarkan nilai perbandingan akurasi, *F1 score*, *Precision*, dan *Recall* dari setiap model.

Tujuan

Tujuan dari penelitian ini adalah pertama, menunjukan perbandingan performansi dari model CNN dan LSTM dalam kasus mendeteksi ujaran kebencian berbahasa Indonesia pada media sosial Twitter, dan kedua, menunjukan performansi antara model CNN dan LSTM saat di kombinasikan dalam mendeteksi ujaran kebencian di media sosial Twitter yang berbahasa Indonesia.

Organisasi Tulisan

Bagian selanjutnya adalah studi terkait yang memaparkan beberapa penelitian yang pernah dilakukan dan selaras dengan penelitian yang dikerjakan saat ini; bagian ketiga berisi sistem yang dibangun dan digunakan dalam penelitian; bagian keempat berisi evaluasi dan diskusi dari hasil penelitian yang diperoleh; bagian kelima berisi kesimpulan dari penelitian yang dikerjakan.

2. Studi Terkait

Menurut Mondal[1], Ujaran kebencian didefinisikan sebagai komentar ataupun postingan yang bersifat ofensif, baik seluruh maupun sebagian isi kontennya, dan berisi makna negatif dari penulis komentar terhadap aspek