

ABSTRAK

CLASSIFYING NATURAL DISASTER TWEETS USING A CONVOLUTIONAL NEURAL NETWORK AND BERT EMBEDDING

By

Lucas Satria Aji Dharma

17/408292/PA/17645

Platform media sosial telah menjadi sarana untuk menemukan sumber informasi yang luas di seluruh internet. Twitter telah menjadi salah satu platform microblogging yang lebih populer di sekitar kita, dan semakin banyak pengguna di platform ini berarti semakin beragam jenis informasi yang dapat ditemukan dalam sehari. Di Twitter pengguna dapat mengekspresikan diri mereka melalui tweet, ini kemudian akan ditampilkan di timeline twitter dan pengguna lain dapat melihat tweet tersebut. Jika sebuah tweet tiba-tiba menjadi viral, Twitter akan menempatkan tweet pengguna ke halaman tren yang memungkinkan lebih banyak pengguna untuk melihat tweet tersebut. Saat terjadi bencana alam seringkali banyak tweet yang menyebutkan tentang bencana sehingga tweet tersebut menjadi trending topik di Twitter. Dari sini, sejumlah besar tweet tentang bencana dapat dikumpulkan sebagai data, tetapi tidak selalu tweet tersebut berisikan informasi tentang bencana. Seringkali beberapa tweet menggunakan kata-kata bencana alam tetapi tidak berbicara tentang bencana itu sendiri, sehingga tidak informatif dan dapat diklasifikasikan sebagai tweet non-bencana.

Makalah penelitian ini bertujuan untuk mengusulkan sebuah sistem untuk mengklasifikasikan tweet bencana dan tweet non-bencana pada saat terjadi bencana. Metode yang diusulkan didasarkan pada Convolutional Neural Network (CNN), menggunakan Bidirectional Encoder Representation from Transformers (BERT) sebagai sebuah Embedding. Sebagai pembandingan, akan digunakan metode embedding lain yang bernama Word2Vec.

Hasil Evaluasi setelah pelatihan dan pengujian CNN dengan embedding BERT memberikan hasil yang paling konsisten mencapai precision 0,963, recall 0,960, dan f1-score 0,961.

ABSTRACT

CLASSIFYING NATURAL DISASTER TWEETS USING A CONVOLUTIONAL NEURAL NETWORK AND BERT EMBEDDING

Social media platforms have become a medium to find a vast source of information throughout the internet. Twitter has become one of the more popular microblogging platforms out there, and the more users there are in these platforms means the more various types of information can be sent out in a day. On Twitter users can write their expression in the form of tweets, this will then create a post on twitter's timeline and other users can see these tweets. If a tweet suddenly gets viral, Twitter will put the user's tweets into the trending page allowing even more users to view the said tweet. During an event of a natural disaster often a lot of the tweets that are being posted, have mention of the disaster making it a trending topic on Twitter. From this, a vast amount of tweets about a disaster can be collected as data, but not always are the tweets containing information about the disaster. Often some tweets use natural disaster words but do not talk about the disaster itself, hence are not informative and can be classified as a non-disaster tweet.

This research paper aims to propose a system to classify the disaster tweets and the non-disaster tweet during a disaster. The proposed method is based on Convolutional Neural Network (CNN), using a Bidirectional Encoder Representation from Transformers (BERT) as an Embedding. As a comparison, it will then be compared with another embedding method named Word2Vec.

The Evaluation result after training and testing of the CNN with BERT embeddings gave the most consistent results attaining a precision of 0.963, a recall of 0.960, and an f1-score of 0.961.

Keywords: Text Classification, Deep Learning, Natural Language Processing (NLP), Convolutional Neural Network (CNN), Bidirectional Encoder Representation from Transformers (BERT), Word Embedding.