ABSTRAK

IDENTIFIKASI SUSTAINABLE DEVELOPMENT GOALS TERHADAP TWEET BERBAHASA INDONESIA

GATHOT HERMAWAN NUGROHO

19/442474/PA/19223

Sustainable Development Goals merupakan tujuan pembangunan yang diajukan oleh PBB dan terdiri dari 17 tujuan utama. Dalam penerapannya, sering ditemukan bahwa setiap isu dapat memiliki lebih dari satu SDG. Melalui klasifikasi, isu tersebut dapat digolongkan menjadi lebih dari satu SDG menggunakan klasifikasi multi-label. Metode penelitian yang dilakukan sebelumnya belum menggunakan data berbahasa Indonesia. Data yang didapat dari twitter dapat di proses melalui tokenisasi dan TF-IDF dan dibuat model klasifikasi multi-label dengan *classifier* Multinomial Naive Bayes(MNB) dan Complement Naive Bayes(CNB). Melalui skenario eksperimen yang didasarkan pada *classifier* dan penggunaan *preprocessing*, model dapat dievaluasi berdasarkan skenario dan metode evaluasi seperti *hamming loss*, *Subset Accuracy*(EMR), dan *F1-Score*. Hasil yang didapatkan berupa nilai hamming loss sebanyak 0,124 pada MNB dan 0.125 pada CNB dimana CNB memiliki hasil yang baik secara konsisten dibanding MNB. Penggunaan stopword removal terlihat efektif dalam meningkatkan kualitas training data dengan peningkatan nilai EMR dan *F1-Score*. Namun, menggunakan data twitter untuk klasifikasi SDG terbukti sulit dikarenakan penggunaan bahasa yang kurang baku dan isi data yang sedikit setiap tweetnya.

ABSTRACT

IDENTIFICATION OF SUSTAINABLE DEVELOPMENT GOALS BASED ON INDONESIAN TWEETS

GATHOT HERMAWAN NUGROHO

19/442474/PA/19223

Sustainable Development Goals is a goal proposed by the UN to help governments develop their countries. In practice, it's often found to be difficult to classify one issue into each class of SDGs for they have similar properties. Using classification, we could classify each of them into multiple labels. Previous studies proved it can be done in multiple languages, but it hasn't been implemented in the Indonesian language yet. Using twitter data, we could gather much data necessary for the classification. It is then pre-processed through data cleaning and tokenization then vectorized using TF-IDF before going into the Multinomial and Complement Naive Bayes classification model. Through scenarios based on classifiers and preprocessing methods, the model could be evaluated through comparisons between scenarios and using evaluation methods like hamming loss, subset accuracy(EMR), and F1-Score. Based on the results, MNB has lower loss than CNB by having 0,124 hamming loss compared to CNB's 0,125 though CNB results have been consistently better than MNB despite having lower max performance. The usage of stopword removal as a preprocessing method proved to increase the training data quality by increasing EMR and F1-Score. It's also found that using twitter data for SDG classification is not ideal with the abundance of slang usage and lower word count found on every document.