Ekspansi Data Menggunakan Forward-Backward Translation untuk Deteksi Ujaran Kebencian Multi-Label dalam Bahasa Indonesia = Data Expansion Using Forward-Backward Translation for Multi-Label Hate Speech Detection in Bahasa Indonesia

Fairuz Astari Devianty, author

Abstrak

Dengan tumbuh dan berkembangnya platform media sosial, komunikasi bisa menjadi lebih mudah dilakukan. Namun, hal tersebut dapat disalahgunakan, seperti penyebaran hate speech melalui media sosial yang semakin marak terjadi. Meski kebebasan berekspresi adalah hak setiap orang di Indonesia, namun karena dampak negatifnya konten kebencian harus dihilangkan. Salah satu solusinya adalah dengan membangun sebuah model yang dapat mendeteksi hate speech secara otomatis. Untuk membangun model pendeteksian hate speech yang baik, dibutuhkan data beranotasi dengan jumlah yang besar untuk melatih model. Selain itu perlu juga diperhatikan target dan kategori dari hate speech tersebut. Namun, saat ini hanya ada satu multi-label hate speech dataset Bahasa Indonesia yang tersedia dan memiliki kekurangan proposi data dari setiap label yang tidak seimbang. Untuk mengatasi masalah kekurangan data ini, penulis mengusulkan sebuah metode yaitu Forward-Backward Translation untuk menghasilkan data secara otomatis. Metode ini merupakan gabungan dari forward translation dan back-translation. Forward translation dilakukan pada dataset dari high-resource language dan back-translation dilakukan pada dataset dari low-resource language. Dengan digabungkannya kedua proses ini dataset yang dihasilkan akan memiliki jumlah yang besar dan memiliki kualitas terjemahan yang baik. Metode ini digunakan untuk menambahkan data pada deteksi multi-label hate speech Bahasa Indonesia dengan tambahan data dari Bahasa Inggris. Performa pendeteksian multi-label hate speech pada dataset baru hasil penelitian ini berhasil meningkat bila dibandingkan dengan pada dataset hate speech Bahasa Indonesia yang sudah ada. Dataset ini mendapatkan F1-score sebesar 0.76 saat melakukan multi-label classification dan F1-score sebesar 0.78 saat melakukan hierarchical classification.

.....The growth and development of social media platforms make communication easier. However, this can be misused. For example, the spread of hate speech via social media is increasing. Freedom of speech is everyone's right in Indonesia, but malicious content must be eliminated due to its negative impact. One solution is to build a model that can automatically detect hate speech. Building a good hate speech detection model requires a large amount of annotated data to train the model. It is also necessary to pay attention to the target, category, and level of hate speech. However, there is currently only one multi-label hate speech dataset in Bahasa Indonesia available and the proportion of data for each label is unequal. To overcome this data scarcity problem, we propose a forward-backward translation method to generate data automatically. This method combines forward and backward translation. A forward translation is performed for dataset in high-resource languages and a backward translation is performed for dataset in low-resource languages. By combining these two processes, the resulting dataset will have a large amount of data and good translation quality. This method will be used to add data on multi-label hate speech detection in Bahasa Indonesia with additional data from English. As a result of this study, the performance of multi-label hate speech detection in the new dataset improved compared to the existing Bahasa Indonesia hate speech dataset. This dataset

gets an F1-score of 0.76 for multi-label classification and an F1-score of 0.78 for hierarchical classification.