INTISARI

AUTHORSHIP ATTRIBUTION UNTUK TEKS BERBAHASA INDONESIA DENGAN METODE MULTI-TASK LEARNING BERBASIS LSTM

Oleh

Ricky Setiawan
17/412652/PA/17971

Authorship attribution merupakan permasalahan klasifikasi yang bertujuan untuk menentukan penulis dari suatu teks berdasarkan kumpulan data yang terdiri dari penulis dan tulisannya. Penelitian terkait dengan topik ini masih terus dipelajari karena terkait dengan aplikasi forensik yang penting seperti identifikasi penulis dari pesan anonim, dll. Salah satu metode yang sering digunakan dalam permasalahan ini yaitu LSTM. LSTM mengekstraksi fitur pada data sekuensial untuk mendapatkan konteks informasi secara efektif sehingga dapat mengenali pola tulisan seseorang dengan baik. Performa dari LSTM dapat ditingkatkan dengan menggunakan model *Multi Task Learning* (MTL). MTL memanfaatkan korelasi diantara keterkaitan *task* dengan belajar *task* secara paralel sehingga membuat model dapat menggeneralisasi lebih baik pada *task* utama yang berdampak pada peningkatan akurasi klasifikasi.

Pada penelitian ini, model *multi-task learning* (MTL) berbasis LSTM diusulkan untuk memecahkan permasalahan *authorship attribution* sebagai *task* utama dan identifikasi gender sebagai *task* pembantu. Pada setiap jenis kelamin mempunyai karakteristik penulisan yang berbeda, berdasarkan hal tersebut diketahui bahwa terdapat potensi penggunaan MTL membantu meningkatkan performa model. Penelitian ini diimplementasikan pada dataset berbahasa Indonesia, menggunakan GloVe *embedding* dengan dimensi 300, 1 LSTM layer dengan *hidden layer* sebanyak 256 *node*, 4 *dense layer* yang secara berurutan mempunyai 256, 128, 64, dan 32 *node*. Hasil performa dari Model MTL kemudian dibandingkan dengan arsitektur *single task*.

Percobaan diterapkan pada dataset Twitter dan situs berita daring. Pada setiap dataset banyak data untuk masing-masing kelas terdistribusi secara merata. Dataset dibagi menjadi 72% *train*, 8% val, dan 20% *test*. Dari hasil pengujian akurasi terhadap data uji didapatkan peningkatan akurasi sebesar rata-rata 0.94% untuk seluruh dataset. Meskipun hasil peningkatan akurasi yang didapatkan tidak terlalu signifikan, namun dari penelitian ini dapat dibuktikan bahwa penggunaan *authorship attribution* dengan identifikasi gender dalam arsitektur MTL dapat meningkatkan performa.

Kata kunci : *author attribution, multi task learning, LSTM, word embedding*

ABSTRACT

AUTHORSHIP ATTRIBUTION FOR INDONESIAN TEXT USING MULTI-TASK LEARNING BASED ON LSTM

By

Ricky Setiawan
17/412652/PA/17971

Authorship attribution is a classification problem to identify the author from a text based on a dataset consisting of the author and his writing. Research related on this topic is still being studied because it is related to important forensic applications such as to identify the author of anonymous messages, etc. One method that is often used in this problem is LSTM. LSTM can extracts features on sequential data to get the context of information effectively so that it can recognize a person's writing pattern well. The performance of the LSTM can be improved by using the Multi Task Learning (MTL) model. MTL utilizes the correlation between related tasks and learning tasks in parallel so as to make the model generalize better on the main task which results in an increase in classification accuracy.

In this study, an LSTM-based multi-task learning (MTL) model is proposed to solve the problem of authorship attribution as the main task and gender identification as an auxiliary task. Each gender has different writing characteristics, based on this it is known that there is a potential for using MTL to help improve model performance. This research is implemented in Indonesian language dataset, using GloVe embedding with dimension of 300, 1 LSTM layer with 256 nodes, and 4 dense layers which have 256, 128, 64, and 32 nodes respectively. The performance results of the MTL model are then compared with the single task architecture.

The experiments have been applied to the twitter and online news dataset. For each dataset all of the data for each class is evenly distributed. The dataset is divided into 72% train, 8% val, and 20% test set. From the results of model testing using test data, it was found that an increase in accuracy of an average of 0.94% for all datasets. Although the results of the increase in accuracy obtained are not very significant, from this research it can be proven that the use of authorship attribution with gender identification in MTL architecture can improve the model performance.

Keywords : *author attribution, multi task learning, LSTM, word embedding*