ABSTRAK

PENDEKATAN BARU EKSTRAKSI INFORMASI *BIOMEDICAL* BIG DATA *REPORT* DENGAN BIOWORDVEC MENGGUNAKAN MODEL *HYBRID LONG SHORT-TERM MEMORY – CONVOLUTIONAL NEURAL NETWORK* (LSTM – CNN)

Oleh

DIAN KURNIASARI

Peningkatan angka kematian akibat leukemia telah mendorong pesatnya pertumbuhan publikasi mengenai penyakit ini. Lonjakan publikasi tersebut berdampak signifikan pada peningkatan literatur biomedis, yang membuat ekstraksi informasi relevan tentang leukemia secara manual semakin menantang. Hal ini dikarenakan penelitian yang ada sebelumnya umumnya hanya memperhitungkan komponen leksikal dan sintaksis teks tanpa mempertimbangkan makna semantiknya.

Tujuan dilakukannya penelitian ini antara lain adalah untuk menemukan model *Deep Learning* (DL) terbaik dalam melakukan ekstraksi informasi yang relevan secara semantik pada sejumlah besar data biomedis yang disebut *biomedical* big data *report*. *Semantic Text Similarity* (STS) adalah salah satu bidang penelitian penting dalam aplikasi saat ini yang terkait dengan analisis semantik teks. Metode tersebut memungkinkan ekstraksi informasi dari suatu teks menjadi lebih bermakna karena melibatkan penerapan representasi distribusi kata-kata atau sumber eksternal pengetahuan semantik terstruktur seperti *word embedding*. Namun perlu diperhatikan bahwa *word embedding* yang digunakan harus sesuai dengan domain penelitian.

Penelitian ini mengusulkan penerapan arsitektur *Siamese Manhattan* pada model DL, yaitu model CNN, LSTM, *hybrid* CNN-LSTM, dan *hybrid* LSTM-CNN, untuk melakukan analisis semantik teks biomedis. Teks biomedis yang memiliki makna semantik atau berada pada konteks yang sama direpresentasikan ke dalam bentuk vektor berdasarkan *word embedding* khusus domain biomedis, yaitu BioWordVec. Lebih lanjut, model tersebut dibangun dan dibandingkan berdasarkan jumlah lapisan tersembunyi dan metode pelabelan yang digunakan. Jumlah lapisan tersembunyi yang digunakan adalah dua dan tiga, sedangkan metode pelabelan yang digunakan adalah metode *Cosine Similarity* (CS) dan metode *Word Mover's Distance* (WMD). Hasil analisis semantik menunjukkan bahwa setiap kalimat memiliki makna semantik yang identik dengan tingkat *similarity* 1.

Hasil tersebut selanjutnya menjadi landasan untuk dilakukan klasifikasi teks sebagai bentuk aplikasi langsung dari STS. Model klasifikasi teks dibangun dan

dibandingkan berdasarkan dua skema pembagian data, yaitu *train-test split* dan *k-fold Cross Validation*. Masalah ketidakseimbangan kelas yang muncul selama proses klasifikasi kemudian diatasi melalui prosedur *resampling* menggunakan kombinasi metode *Random Undersampling* dan *Random Oversampling*. Sama seperti tahap sebelumnya yaitu STS, tahap klasifikasi teks juga menerapkan BioWordVec sebagai metode representasi kata.

Secara keseluruhan, model *hybrid* LSTM – CNN yang diusulkan untuk ekstraksi informasi memiliki performa yang lebih baik dibandingkan model CNN, LSTM, dan *hybrid* CNN – LSTM dengan nilai akurasi mencapai 100% pada tugas STS dan mencapai 99% untuk tugas klasifikasi teks. Dengan demikian, dapat disimpulkan bahwa model DL terbaik untuk melakukan ekstraksi informasi pada penelitian ini adalah model *hybrid* LSTM – CNN dengan implementasi *word embedding* khusus domain biomedis, yaitu BioWordVec.

Kata Kunci: Klasifikasi Teks, *Semantic Text Similarity*, BioWordVec, *Hybrid* LSTM – CNN.

ABSTRACT

NEW APPROACH TO BIOMEDICAL BIG DATA REPORT INFORMATION EXTRACTION WITH BIOWORDVEC USING HYBRID LONG SHORT-TERM MEMORY – CONVOLUTIONAL NEURAL NETWORK (LSTM – CNN) MODEL

By

DIAN KURNIASARI

The rise in death rates associated with leukemia has fueled the rapid expansion of publications focused on this disease. The rise in the number of publications has substantially affected the growth of biomedical literature, making it more challenging to manually extract pertinent information concerning leukemia. That is because prior studies typically focused solely on the lexical and syntactic aspects of the text, neglecting its semantic significance.

This study aims to identify the most effective Deep Learning (DL) model for extracting semantically significant information from a substantial volume of biomedical data called the Biomedical Big Data Report. Semantic Text Similarity (STS) is a crucial field of research in contemporary applications that deal with the semantic analysis of texts. This approach enhances extracting information from a text by utilizing a distributed model of words or an external source of organized semantic knowledge, such as word embedding. Nevertheless, it is essential to acknowledge that word embedding must suit the specific research field.

This study suggests implementing the Siamese Manhattan architecture in Deep Learning (DL) models, namely Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, hybrid CNN-LSTM models, and hybrid LSTM-CNN models, to do semantic analysis on biomedical text. Biomedical language with semantic significance or in the same context is transformed into vector representation using word embedding techniques specifically designed for the biomedical field, known as BioWordVec. In addition, the models are constructed and evaluated based on the number of hidden layers and labelling techniques employed. Two to three hidden layers are utilised, along with the Cosine Similarity (CS) and Word Mover's Distance (WMD) tagging methods. The findings of the semantic analysis indicate that every sentence has the same semantic meaning, with a similarity level of 1.

These results are the foundation for text classification, which directly implements STS. Text classification models are constructed and evaluated using two data partitioning methods: train-test split and k-fold Cross Validation. The class imbalance issue during the classification process is addressed using a resampling technique that combines Random Undersampling and Random Oversampling

approaches. Like the previous step, STS, the text categorization stage utilizes BioWordVec to represent words.

The hybrid LSTM – CNN model outperforms the CNN, LSTM, and hybrid CNN – LSTM models in information extraction, achieving accuracy rates of 100% on the STS task and 99% on the text classification task. Thus, it can be concluded that the best DL model for extracting information in this research is a hybrid LSTM – CNN model with the implementation of word embedding specifically for the biomedical domain, namely BioWordVec.

Keywords: Text Classification, Semantic Text Similarity, BioWordVec, Hybrid LSTM – CNN.