ABSTRACT

DEEP SUPPORT VECTOR DATA DESCRIPTION (DEEP SVDD) BASED ON BIOWORDVEC FOR LEUKEMIA TEXT PAIR ANOMALY DETECTION IN THE GENIA BIOMEDICAL EVENT CORPUS

By

Khusni Sinta Rodiyah

Anomaly detection in biomedical text pairs is a critical challenge in natural language processing, particularly in data with complex semantic relationships such as the GENIA Biomedical Event dataset. This study aims to develop an anomaly detection model based on Deep Support Vector Data Description (Deep SVDD) for biomedical text pairs: Sentence–TriggerWord, Sentence–EventType, and TriggerWord–EventType. In the initial stage, pseudo-labeling was performed using a modified Word Mover's Distance (WMD) approximation approach with Part-of-Speech (POS) weighting to classify normal and anomalous texts based on semantic similarity. Feature representations were constructed using BioWordVec embeddings, which were then processed using LSTM and BiLSTM encoders to generate latent representations. The results showed that the Deep SVDD model with an LSTM encoder provided the best performance, with an ROC AUC of 0.99 on the Sentence–TriggerWord dataset and 0.98 on the other two datasets, while BiLSTM performed lower. Furthermore, a threshold sensitivity analysis revealed a trade-off between sensitivity and specificity. Increasing the threshold increased specificity but decreased sensitivity. Therefore, the P95 threshold was chosen because it yielded the highest sensitivity and was therefore more effective in minimizing false negatives.

Keywords: Anomaly detection, Deep SVDD, GENIA Biomedical Event, BioWordVec, LSTM, BiLSTM, ROC AUC, Threshold Sensitivity.

ABSTRAK

DEEP SUPPORT VECTOR DATA DESCRIPTION (*DEEP* SVDD) BERBASIS BIOWORDVEC UNTUK DETEKSI ANOMALI PASANGAN TEKS LEUKEMIA PADA GENIA *BIOMEDICAL EVENT CORPUS*

Oleh

Khusni Sinta Rodiyah

Deteksi anomali pada pasangan teks biomedis merupakan tantangan penting dalam *natural language processing*, khususnya pada data dengan hubungan semantik yang kompleks seperti dataset GENIA *Biomedical Event*. Penelitian ini bertujuan mengembangkan model deteksi anomali berbasis *Deep Support Vector Data Description* (*Deep* SVDD) pada pasangan teks biomedis, yaitu *Sentence–TriggerWord*, *Sentence–EventType*, dan *TriggerWord–EventType*. Pada tahap awal, dilakukan *pseudo labeling* menggunakan pendekatan aproksimasi *Word Mover's Distance* (*WMD*) yang dimodifikasi dengan pembobotan *Part-of-Speech* (*POS*) untuk membentuk kelas normal dan anomali berdasarkan tingkat kemiripan semantik. Representasi fitur dibangun menggunakan *embedding* BioWordVec yang kemudian diproses menggunakan *encoder* LSTM dan BiLSTM untuk menghasilkan representasi laten. Hasil penelitian menunjukkan bahwa model *Deep* SVDD dengan *encoder* LSTM memberikan performa terbaik dengan nilai ROC AUC sebesar 0,99 pada dataset *Sentence–TriggerWord* serta 0,98 pada dua dataset lainnya, sedangkan BiLSTM menunjukkan performa yang lebih rendah. Selain itu, analisis sensitivitas *threshold* menunjukkan adanya *trade-off* antara sensitivitas dan spesifisitas. Peningkatan *threshold* meningkatkan spesifisitas namun menurunkan sensitivitas. Oleh karena itu, *threshold* P95 dipilih karena menghasilkan sensitivitas tertinggi sehingga lebih efektif dalam meminimalkan *false negative*.

Kata-kata kunci: Deteksi anomali, *Deep* SVDD, GENIA *Biomedical Event*, BioWordVec, LSTM, BiLSTM, ROC AUC, *Treshold Sensitivity*.