

ABSTRACT

SENTIMENT CLASSIFICATION WITH LSTM AND DISTILBERT EMBEDDING (J&T Express Case Study)

By

Gabriel Agape Gananputra

18/430263/PA/18776

Self-training, one of the semi-supervised learning methods, is very useful in situations where there is a shortage of labeled data for a specific task and an abundance of unlabeled data. Sentiment classification can be particularly important; for example, classifying sentiments toward J&T Express on the Twitter platform. This study aims to implement a 3-class sentiment classification system to process sentiment data regarding J&T Express from the Twitter platform using the self-training method with an LSTM model and DistilBERT embeddings. The classification results will be tested with 10-fold validation and analyzed using a confusion matrix through accuracy, *precision*, and *recall* parameters.

During the testing process, using a larger batch size (16) generally provides better performance compared to a smaller batch size (8) in machine learning text classification models, with micro *precision* and micro F1 reaching around 49% at a learning rate of 0.0001. Additionally, preprocessing techniques such as stopword removal and word normalization also affect performance, with stopword removal showing a slight advantage in improving micro F1. Selecting the optimal batch size and applying appropriate preprocessing techniques are keys to enhancing the efficiency and accuracy of the model.

Keywords: *Distilled Bidirectional Encoder Representation from Transformers* (DistilBERT), Long Short-Term Memory (LSTM), Word Embedding

INTISARI

KLASIFIKASI SENTIMEN MENGGUNAKAN LSTM DAN EMBEDDING DISTILBERT (STUDI KASUS J&T Express)

Oleh

Gabriel Agape Gananputra

18/430263/PA/18776

Self-training, salah satu metode semi-supervised *learning*, sangat berguna dalam situasi di mana terdapat kekurangan data terlabel untuk tugas spesifik dan melimpahnya data yang belum terlabel. Klasifikasi sentimen dapat menjadi suatu hal yang penting, Salah satu contohnya yaitu klasifikasi sentimen terhadap J&T Ekspress pada platform tweeter. Penelitian ini bertujuan untuk mengimplementasikan sistem klasifikasi sentimen 3 kelas untuk memproses data sentimen terhadap J&T Express dari platform tweeter dengan metode self-training dengan model *LSTM* dengan bantuan *DistilBERT* embedding. hasil klasifikasi akan diuji dengan 10 fold validation dan dianalisis menggunakan confusion matrix melalui parameter *accuracy*, *precision*, dan *recall*.

Pada proses pengujian penggunaan batch size yang lebih besar (16) umumnya memberikan performa yang lebih baik dibandingkan batch size yang lebih kecil (8) dalam model klasifikasi teks machine learning, dengan presisi mikro dan F1 mikro yang mencapai sekitar 49% pada learning rate 0.0001. Selain itu, teknik pra-pemrosesan seperti penghilangan stopword dan normalisasi kata juga berpengaruh terhadap performa, dengan penghilangan stopword menunjukkan sedikit keunggulan dalam meningkatkan F1 mikro. Pemilihan batch size yang optimal dan penerapan teknik pra-pemrosesan yang tepat adalah kunci untuk meningkatkan efisiensi dan akurasi model.

Kata kunci: *Distilled Bidirectional Encoder Representation from Transformers* (DistilBERT), Long Short-Term Memory (LSTM), Word Embedding