

ABSTRAK

Berita media massa daring merupakan salah satu sumber utama informasi bagi masyarakat, sehingga diperlukan sistem yang mampu mengelompokkan berita secara otomatis untuk mendukung analisis topik yang sedang berkembang. Penelitian ini bertujuan mengembangkan model klasifikasi multi-label berita berbahasa Indonesia menggunakan kombinasi Bidirectional Long Short-Term Memory (Bi-LSTM) dan representasi teks berbasis embedding IndoBERT. Dataset yang digunakan terdiri atas lima kategori, yaitu politik, hukum, pendidikan, olahraga, dan ekonomi. Representasi teks dilakukan menggunakan embedding IndoBERT dengan variasi panjang token max length sebagai skenario pengujian. Hasil embedding kemudian digunakan sebagai masukan bagi model Bi-LSTM untuk melakukan klasifikasi multi-label dengan variasi nilai ambang probabilitas threshold. Kinerja model dievaluasi menggunakan metrik precision, recall, F1-score, dan Hamming Loss pada data judul dan isi berita. Hasil penelitian menunjukkan bahwa model Bi-LSTM dengan embedding IndoBERT mampu memberikan kinerja yang baik dalam klasifikasi multi-label berita daring. Pada data judul berita, model memperoleh F1-score terbaik sebesar 0,7588 dan Hamming Loss terendah sebesar 0,1053. Sementara itu, pada data isi berita, model mencapai F1-score terbaik sebesar 0,7677 dan Hamming Loss terendah sebesar 0,0947. Hasil tersebut menunjukkan bahwa penggunaan isi berita yang memiliki informasi kontekstual lebih panjang dapat meningkatkan kemampuan model dalam mempelajari keterkaitan antar kata dan label. Selain itu, penerapan mekanisme early stopping mampu meningkatkan efisiensi proses pelatihan tanpa menurunkan performa model secara signifikan. Secara keseluruhan, kombinasi Bi-LSTM dan embedding IndoBERT efektif digunakan untuk klasifikasi multi-label berita berbahasa Indonesia.

Kata kunci: Berita Daring, Bi-LSTM, Embedding, IndoBERT, Klasifikasi Multi-Label.

ABSTRACT

Online mass media news has become one of the primary sources of information for society, creating the need for systems capable of automatically categorizing news articles to support the analysis of emerging topics. This study aims to develop a multi-label classification model for Indonesian news articles using a combination of Bidirectional Long Short-Term Memory (Bi-LSTM) and IndoBERT-based text embeddings. The dataset consists of five categories: politics, law, education, sports, and economics. Text representation was performed using IndoBERT embeddings with various token lengths max length as experimental scenarios. The resulting embeddings were then used as input for the Bi-LSTM model to perform multi-label classification with different probability threshold values. The model performance was evaluated using precision, recall, F1-score, and Hamming Loss metrics on both news titles and news content. The results indicate that the Bi-LSTM model with IndoBERT embeddings performs well in multi-label news classification tasks. For news titles, the model achieved the best F1-score of 0.7588 and the lowest Hamming Loss of 0.1053. Meanwhile, for news content, the model obtained the best F1-score of 0.7677 and the lowest Hamming Loss of 0.0947. These findings suggest that longer contextual information contained in news content improves the model's ability to learn relationships between words and labels. Furthermore, the implementation of the early stopping mechanism improved training efficiency without significantly reducing model performance. Overall, the combination of Bi-LSTM and IndoBERT embeddings is effective for multi-label classification of Indonesian news articles.

Keyword: Embedding, Bi-LSTM, IndoBERT, Multi-label Classification, Online News.