INTISARI

AUTHORSHIP IDENTIFICATION BERITA DENGAN *RECURRENT NEURAL NETWORKS*

Oleh
Pandya Athallah Erlambang
19/444314/PA/19376

Authorship identification telah berkembang dari fokus pada teks tradisional seperti sastra dan dokumen sejarah hingga mengatasi kompleksitas konten digital, termasuk berita online dan media sosial. Perluasan ini didorong oleh kebutuhan untuk memverifikasi sumber berita dan memerangi misinformasi di lanskap digital saat ini. Artikel-artikel berita, yang dicirikan oleh nada formal dan struktur standar, menghadirkan tantangan unik karena perbedaan gaya yang halus dan potensi melemahnya suara penulis individu melalui proses penyuntingan. *Recurrent Neural Network* (*RNN*) unggul dalam menangkap gaya penulisan dan hubungan kontekstual dengan belajar dari rangkaian teks, menjadikannya cocok untuk identifikasi kepenulisan.

Dalam penelitian ini, model *deep learning–Long Short-Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU)–yang divektorkan dengan GloVe *word embeddings* dilatih pada *dataset* '*All the news*'. Model *machine learning–Support Vector Machine* (SVM) dan *Logistic Regression*–yang divektorkan dengan TF-IDF berfungsi sebagai *baseline*. Keempat model tersebut dilatih pada level kalimat dan artikel. Performa model kemudian dievaluasi dan dibandingkan.

Hasilnya menunjukkan bahwa pada pelatihan tingkat artikel, rata-rata akurasi makro model *LSTM*, *GRU*, *SVM*, dan *Logistic Regression* masing-masing adalah 82%, 85%, 91%, dan 89%. Pada pelatihan tingkat kalimat, akurasi rata-rata makro model *LSTM*, *GRU*, *SVM*, dan *Logistic Regression* masing-masing adalah 46%, 47%, 47%, dan 48%. Hal ini menunjukkan bahwa model *machine learning* yang dilatih di tingkat artikel memiliki performa terbaik, diikuti oleh model *deep learning* juga dilatih di tingkat artikel. Sementara itu, semua model yang dilatih pada tingkat kalimat memiliki performa yang tidak baik.

Kata kunci: *news, authorship identification, deep learning, machine learning, GloVe, TF-IDF*

ABSTRACT

NEWS AUTHORSHIP IDENTIFICATION USING RECURRENT NEURAL NETWORKS

By
Pandya Athallah Erlambang
19/444314/PA/19376

Authorship identification has evolved from focusing on traditional texts like literature and historical documents to tackling the complexities of digital content, including online news and social media. This expansion is driven by the need to verify news sources and combat misinformation in today's digital landscape. News articles, characterized by their formal tone and standardized structure, present unique challenges due to the subtlety of stylistic differences and the potential dilution of individual authorial voices through editing processes. Recurrent Neural Networks (RNNs) excel in capturing writing styles and contextual relationships by learning from sequences of text, making them well-suited for authorship identification.

In this research, deep learning models–Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)–vectorized with GloVe word embeddings are trained on the 'All the news' dataset. Machine learning models–Support Vector Machine (SVM) and Logistic Regression–vectorized with TF-IDF serve as the baseline. The four models are trained at sentence and article levels. The model performance is then evaluated and compared.

The result shows that at article-level training, the macro averaged accuracy of the LSTM, GRU, SVM, and Logistic Regression models are 82%, 85%, 91%, and 89% respectively. At sentence-level training, the macro averaged accuracy of the LSTM, GRU, SVM, and Logistic Regression models are 46%, 47%, 47%, and 48% respectively. Thus indicating that the machine learning models vectorized with TF-IDF trained at article level performed best, followed by the deep learning models vectorized with GloVe also trained at article level. Meanwhile all the models trained at sentence level severely underperformed.

Keywords: *news, authorship identification, deep learning, machine learning, GloVe, TF-IDF*