Using BERT embeddings with CNN & RNN architectures for authorship verification

Authors:
Elliot Brooks & Liam Patterson

MANCHESTER 1824
The University of Manchester

Abstract

This project explores the application of *BERT* embeddings for authorship verification, aiming to discern the author of a given text. Leveraging the rich contextual information captured by *BERT*, the study develops and evaluates two models: a *RNN Bi-LSTM* and a *CNN*. These models exhibit notable improvements over the baseline, achieving accuracies of 54% and 60%, respectively. The findings underscore the effectiveness of *BERT* embeddings in enhancing authorship verification tasks, offering promising avenues for further research in natural language processing.

Introduction

Authorship verification seeks to determine the true authorship of a given text, particularly in cases of disputed or anonymous documents.

With the proliferation of digital communication and the ease of content creation, the need for reliable authorship attribution methods has grown significantly. This task holds relevance across various domains, including plagiarism detection, fraud detection, and historical document analysis. In this study, we explore the application of *BERT* embeddings for authorship verification, aiming to enhance accuracy and robustness in determining text authorship.

The 2 methods presented in this report are *CNN* and *RNN Bi-LSTM*. These models are inspired by 2 papers we found during our research, which detail how *RNN* and *CNN* models can have their performance improved for sentiment analysis by using *BERT* embeddings. [1][2]

Methodology

CNN Model

- Dataset of 30000 examples of labelled text-pairs are loaded and pre-processed, case folding and structuring pairs to be separated with *BERT* 〈SEP〉 tokens.
- Processed data is tokenized with a *distilBERT-based-case* tokenizer into *BERT* embeddings with a max length of 512 tokens.
- An 11-layer *CNN* model is defined and trained in batches of 32 over 100 epochs.

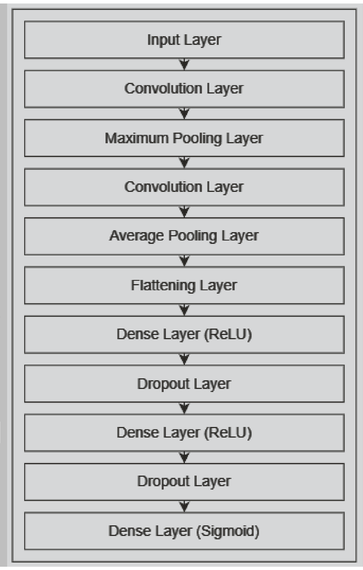

Figure 1:
Representation of CNN model layers

Methodology

RNN Bi-LSTM Model

- Dataset of 30000 examples of labelled text-pairs are loaded and pre-processed, case folding and structuring pairs to be separated with *BERT* 〈SEP〉 tokens.
- Processed data is tokenized with a *distilBERT-based-case* tokenizer into *BERT* embeddings with a max length of 512 tokens.
- A 9-layer *RNN Bi-LSTM* model is defined and trained in batches of 512 over 50 epochs.

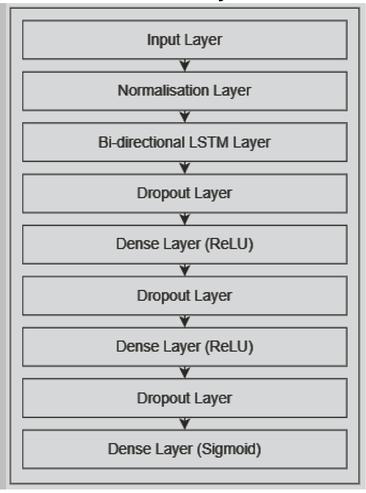

Figure 2:
Representation of RNN Bi-LSTM model layers

Results

- Results for the *CNN* show an 7% improvement in accuracy when compared to a baseline *SVM* and are further detailed reports are in figure 3.
- Results for the *RNN Bi-LSTM* show a 4% improvement in accuracy compared to the baseline *Bi-LSTM* and further detailed reports are in figure 3.

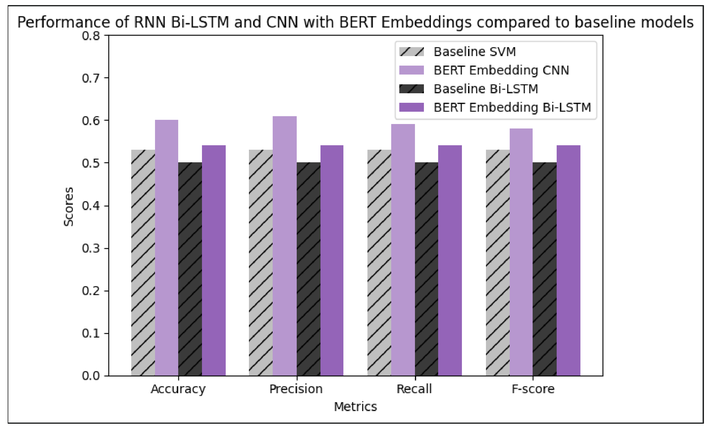

Figure 3:
Graph showing the performance of the *BERT* embedding models against their comparative baselines

Conclusion

In conclusion, our study demonstrates the efficacy of utilizing *BERT* embeddings for enhancing the performance of *CNN* and *Bi-LSTM* models in authorship verification tasks. By leveraging the contextualized representations provided by *BERT*, we achieved notable improvements in accuracy while requiring fewer computational resources compared to a full *BERT* pipeline. This highlights the potential of leveraging pre-trained language models for augmenting the capabilities of traditional deep learning architectures in *NLP* tasks, paving the way for more efficient and accurate authorship attribution methodologies.

References

1. Mutinda, J.; Mwangi, W.; Okeyo, G. Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network. Appl. Sci. 2023, 13, 1445. https://doi.org/10.3390/app13031445

2. Dessì, D.; Recupero, D.R.; Sack, H. An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments. Electronics 2021, 10, 779. https://doi.org/10.3390/electronics10070779