Named Entity Recognition using Pre-trained Language Models

Mohamad Nabeel

December 2023

Abstract

To solve the task of NER, distilBERT-base-cased was chosen. Performance metrics used were Precision, Recall, and F1 for each class, as well as the average of these metrics. The Coherent Kappa score was also chosen, given that the class distribution of the dataset was unbalanced. The average of the metrics for system A consisted of a Precision of 93.5%, Recall of 94.8%, F1 of 94.2%, and Kappa of 95.9%. For system B: Precision 96.0%, Recall 96.8%, F1 96.4%, and Kappa 97.2%.

The average scores show promising results, and the Kappa score indicates great performance despite the unbalanced dataset. However, individual class evaluations show the systems' limitations. System B performed worse on ANIM and DIS. The F1-score for these classes was 73.8% and 75.3%, respectively. The model may find these two classes ambiguous. The rest of the classes in system B averaged approximately 98.7%. System A shares a similar pattern and performs slightly worse than system B, as it manages more labels.

The chosen metrics have limitations in capturing the models' contextual understanding, as some words can be categorized into multiple entities. There is also a limitation in assessing the model's performance based on the length of the given sequences.