MANCHESTER 1824
The University of Manchester

Enhance the Performance of BERT for Authorship Verification by Handling the Data

By Jungwoo Koo

Introduction

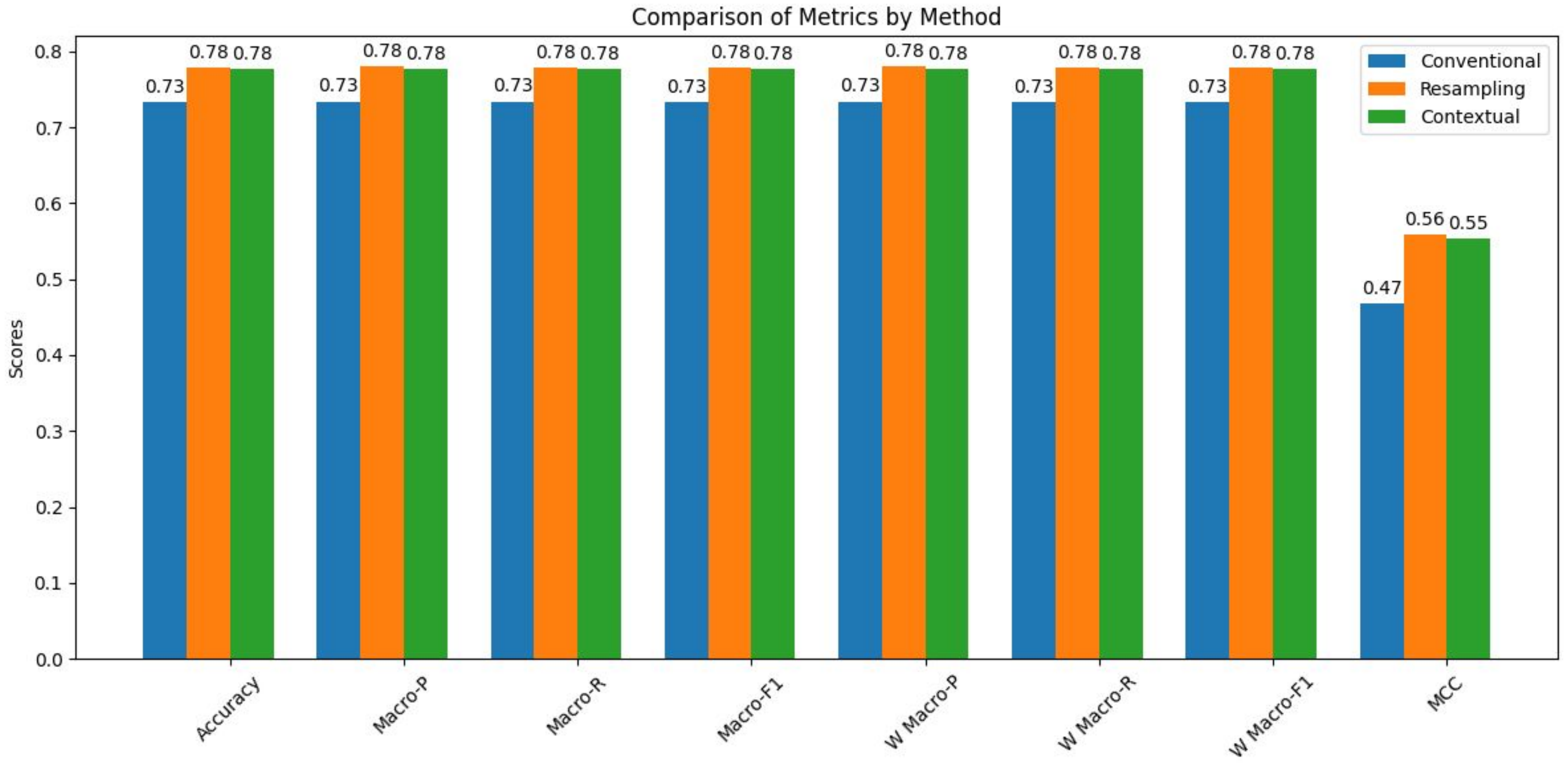
As many of the researchers study the Natural Language Processing or Understanding (NLP or NLU), various novel architectures of the language model are introduced. However not many of the researches focus the data itself. Therefore, in this work, the technologies to handle the data to improve the performance is evaluated. To train the model, the Bidirectional Encoder Representations from Transformers (BERT) is utilized as one of the novel language model. Firstly, the appropriate pre-processing for the BERT model is introduce as the conventional pre-processing in NLP or NLU suggests worse performance. Then, two data augmentations are introduces: Resampling and Contextual Word Augmentor. In the traditional approach to tokenize the text or sentence for the BERT, the maximum number of the tokens are pre-defined and fixed lengths of tokens are extracted from the text. However, this approach has a limitation that not all of the text are utilized as the training data. Therefore, new novel approach, Resampling, is adopted to solve this problem. In addition, Contextual Word Augmentor generates the new texts by inserting or substituting the words that has the similar meaning in the context. Experimental results showed that the techniques introduced in this work can improve the performance of the BERT in the task of pairwise classification, Authorship Verification.

Methodologies

- *Pre-processing*: First pre-processing includes the conventional steps in NLP or NLU including converting to lowercase, expanding contractions (ex. don't -> do not), removing multiple spaces, punctuations, and stopwords, and lemmatization. After this, the pre-processing for BERT is introduced with only three steps: converting to lowercase, expanding contractions, and removing multiple spaces.
- *Data augmentation*
    - Resampling: When the tokens are extracted from the text, if the length of the text is larger than the predefined maximum number of the tokens, then random start index is sampled. From that start index, the maximum number of tokens are extracted.
    - Contextual Word Augmentor: Given the text, the BERT model finds the contextual word embeddings for similar words. Then, this word is either inserted or substituted to the input text to generate the new text. This process is conducted with half of the training data and, at the end, the training dataset becomes 1.5 bigger than the original one.
- *Model*: Pre-trained Bert is fine-tuned during the training.
- *Metrics*: To evaluate the performance, several metricies are computed: accuracy, macro precision, macro recall, macro f1-score, weighted macro precision, weighted macro recall, weighted macro f1-score, and Matthews Correlation Coefficient (MCC)
- *Hyperparameters*
    - Batch size : 32
    - Learning rate: 5e-5
    - Epochs : 3
    - Tokens per text : 256
    - Seed : 42

Results

The bar graph below compares three performances: **Conventional** (traditional preprocessing), **Resampling** (Resampling with modified preprocessing), and **Contextual** (Contextual Word Augmentor with modified preprocessing). The performances are increased from **Conventional** to **Resampling** and **Contextual**, while **Resampling** and **Contextual** show the similar performances. Each figure except MCC increased by 4% after the data handling techniques. In case of MCC, it is increased by 9% in both **Resampling** and **Contextual**. It suggests that the data handling techniques introduced in this work enhance the model performance by providing the better training data. However, the difference between the **Resampling** and **Contextual** is not significant. Therefore, in the final version of the model, only **Resampling** techniques are adopted since the **Contextual** includes more training data and it leads to the increase of training time.



github repo

Conclusions

In this work, novel data handling techniques are introduce to enhance the training of the language model BERT. It is figured out that the BERT model can show better performance with the text including the punctuations, stopwords, and without lemmatization. Since the BERT is complex enough, it can capture more contexts with those information. Additionally, two data augmentations are introduced to generate the different set of training dataset. Resampling randomly samples the tokens from the subset of text in every epochs and Contextual generates new dataset by inserting or substituting the contextually appropriate words. These two data augmentations enhance the training data so that the model can be trained in improved dataset. This work suggests the importance of the data itself in training the language model as well as the model architectures.