INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

Fine-grained domain classification using Transformers

ABSTRACT

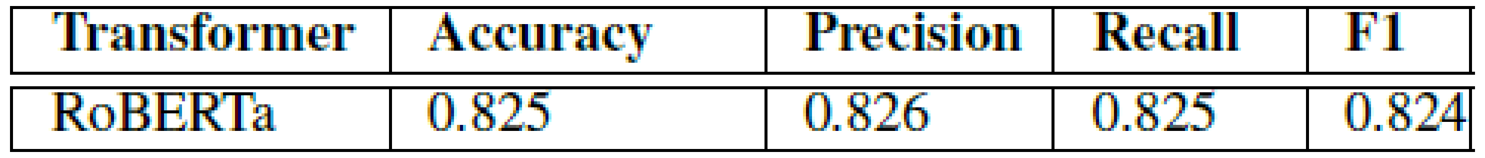
The introduction of transformers in 2017 and successively BERT in 2018 brought about a revolution in the field of natural language processing. Such models are pretrained on vast amounts of data and are easily extensible to be used for transfer learning. Continual work on transformer-based architectures has led to a variety of new models with state-of-the-art results. RoBERTa is one such model, which brings about a series of changes to the BERT architecture and can produce better quality embeddings at an expense of functionality. In this paper, we attempt to solve the well-known text classification task of fine-grained domain classification using BERT and RoBERTa and perform a comparative analysis of the same.
We also attempt to evaluate the impact of data preprocessing specially in the context of fine-grained domain classification. The results obtained outperformed all the other models at the ICON TechDOfication 2020(subtask-2a)Fine-grained domain classification task and ranked first.

OBJECTIVE

A short document consist of English text, participants have to develop an algorithm to identify which sub-domain it belongs to from following (Fine-Grained Domain Classification) :-

- Artificial Intelligence (ai)
- Algorithm (algo)
- Computer Architecture (ca)
- Computer Networks (cn)
- Database Management system (dbms)
- Programming (pro)
- Software Engineering (se)

.



Transformer	Accuracy	Precision	Recall	F1
RoBERTa	0.825	0.826	0.825	0.824

METHOD

A comparative study between RoBERTa and BERT was done using two approaches.

In the first approach, only one-hot-encodings for the raw text was fed as it is to both the transformers.

In the second approach the raw data was preprocessed keeping in mind the nature of finegrained domain classification task. First, tokenization was done on the text using spaCy and the stop words were filtered out. Next, the tokens were passed through a counter and the top 20 tokens from the entire corpus were identified and then removed. As domain classification relies more on the keywords than the sentence structures, the data was cleaned. Lastly, the text was reconstructed from the remaining tokens. This was done to reduce the generalization amongst the sub-domains as the text had a lot of common terms from the higher level computer science domain itself.

Authors: Akshat Gahoi ; Akshat Chhajer

Research Center Name: LTRC