Introduction

Research reporting quality is highly dependent on an author’s ability to concisely and explicitly outline how they execute experimental procedures. To justify the acceptance of articles, authors need to submit a CONSORT guideline checklist, indicating what page number each element can be found. For this project, we proposed to create a tool that can automatically identify relevant elements of the CONSORT guidelines from the methods section of a submitted article.

Exploratory Data Analysis

Before deciding specific method, we explored and analyzed the dataset. For those unlabeled sentences, we labeled them as 0 for better analysis. We visualized label’s distributions and calculated average word counts for each label. Since some sentences may have multiple labels, we also analyzed correlation between labels. In addition, we counted word frequencies and extracted key words for each label to better understand our labels. Because the label 0 takes quite a large fraction in our dataset, we also performed analysis among unlabeled sentences. Here are two data visualizations we included in our report.

Method, Result and Display

Our main method is based on fine-tuning SCIBERT, a BERT model pre-trained on scientific text. First, we tokenized and lemmatized the corpus. To enforce consistency, we truncated and padded token 0 to the tensors. We inserted the token 101 ([CLS]) indicating the beginning of a sentence and the token 102 ([SEP]) indicating the end of a sentence. In addition, we build the attention masks to distinguish the actual input tokens and the padding ones. We set 11 layers of Transformer encoders stacked one over the other and appended a dense linear layer and SOFTMAX function after that. For unbalanced classes, we use systematic sampling based on MOD to oversample them. To simplify the problem for better predictions, we used One-vs-Rest method to split the dataset into two categories. After selecting a specific label, all other labels will be combined as a label ‘Other’, and a Binary Classifier is built on this processed dataset. Since we have total 18 labels and one of the label stands for unlabeled sentence, we built 17 binary classifiers in total. Among all the binary classifiers, our best average validation accuracy reached 99.6% and the worst reached 88.6%. We also built an HTML webpage to display our result dynamically, which is rendered by FLASK and JINJA2.

Acknowledgments

We would like to express our deepest gratitude to our mentor Joshua Fisher and supervisor Adam Kelleher who made this work possible. Their guidance carried us through all the stages of our work.

References