Elsevier CONSORT Guideline

Data Science Institute
COLUMBIA UNIVERSITY

Author: Candong Chen, Zhengyi Fang, Tengteng Tao,
Author: Jinyu Wang, Wenbo Zhao
Industry Mentor: Joshua Fisher
Faculty Mentor: Adam S. Kelleher

Data Science Capstone Project with Elsevier

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.

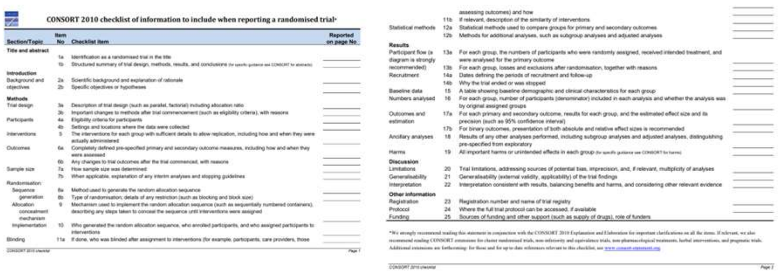
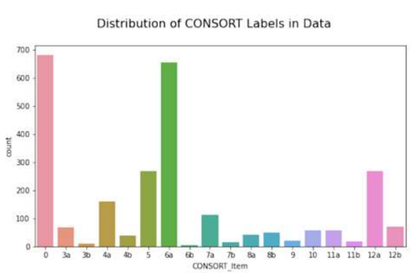


Figure 1. CONSORT 2010 checklist of information to include when reporting a randomized trial

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.



0	3a	3b	4a	4b	5	6a	6b	7a
clinical	double-blind	prescription	year	hospital	dose	measure	secondary	power
	design	oxandrolone	month	centre	receive	blood	death	sample
	control	antibiotic	disease	conduct	daily	assess	cause	difference
	placebo-controlled	clear	within	center	medication	primary	plan	detect
		therefore	exclude	recruit		define	proportion	rate
		compare		site			main	would

7b	8a	8b	9	10	11a	11b	12a	12b
interim	generate	stratify	pack	allocate	blind	capsule	regression	subgroup
committee	list		drug		mask	identical	value	effect
event	computer-generated		pharmacy		staff	appearance		
recommend					investigator	tablet		

Figure 2. Distribution of Labels & Extracted key words for each label

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.

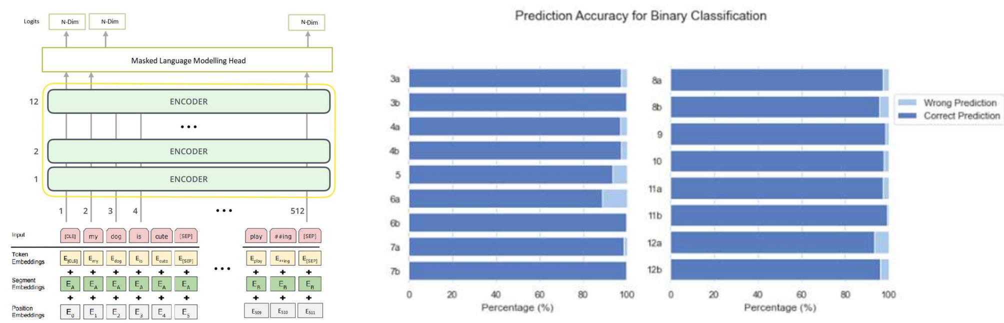


Figure 3. BERT structure & Binary Classifiers' Accuracy

Ensemble

We are still exploring a proper method to ensemble all the binary classifiers. Min-Ling Zhang and Zhi-Hua Zhou provided a series of alternative methods on 2014 IEEE Transactions on Knowledge and Data Engineering, such as Classifier Chains, Calibrated Label Ranking, and Multi-Label k-Nearest Neighbor.

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

[1] Devlin, Jacob , et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." (2018).

[2] Zhang, M. , and Z. Zhou . "A Review on Multi-Label Learning Algorithms." IEEE Transactions on Knowledge & Data Engineering 26.8(2014):1819-1837.