Handling Variance of Pretrained Language Models in Grading Evidence in the Medical Literature

Fajri Koto* and Biaoyan Fang*
The University of Melbourne

The 19th Annual Workshop of the Australasian Language Technology Association, 8-10 December 2021

I. Introduction

Evidence-Based Medicine (EBM):

clinical expertise

medical literatures

Best treatments

Strength of Recommendation Taxonomy (SORT) (Ebell et al., 2004)

A : consistent and good quality patient-oriented evidence

B : inconsistent

C : other evidence, such as consensus guidelines, usual practice and opinion

Obtaining these grades on a wide-scale is expensive and requires in-depth medical expertise. **We focus on classification model of SORT framework**

II. Dataset

ALTA 2021 Shared-Task

```
00667 A 10796398 11508437
00668 A 9036306
00669 C 7391096 11204962 7790481 6863528
00670 B 9569395 12069675
00671 B 11083602 10875559 15283004
```

1 row is 1 data / evidence

ID **Class** **Literatures (PubMed) ID**

	Train	Dev	Test
Evidences	677	178	183
in A	212	48	-
in B	311	80	-
in C	154	50	-
Ave. resources per evidence	2.4	2.5	2.3
Ave. words per abstract	269.9	262.6	274.1
Ave. words per evidence	655.9	653.7	643.9

ALTA 2021 Shared-Task

- ~45% in train and development set are class B
- No significant differences in terms of number of resources and words between each dataset

III. Methods

Domain-specific pretrained models:

1. Biomed BERT
2. Biomed RoBERTa
3. Biomed RoBERTa (TAPT) → further pretrained with train set for 400 epochs

Optimized for medical literatures

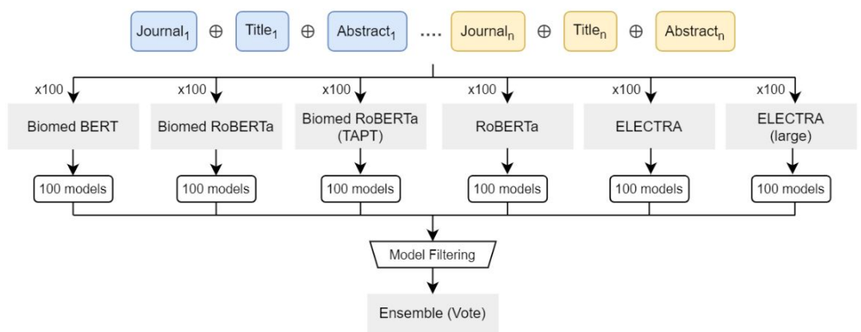
VS

Domain-generic pretrained models:

1. RoBERTa
2. ELECTRA
3. ELECTRA (large)

Contains rich latent discourse rep. (Koto et al., 2021)

Improved with filtered ensemble methods:



IV. Results

Results on **development** set over 100 random seeds:

Model	Accuracy			
	Mean	Max	Min	Std
Biomed BERT	58.7	66.9	52.8	2.9
Biomed RoBERTa	**59.5**	67.4	55.1	2.5
Biomed RoBERTa (TAPT)	58.3	65.7	52.8	2.6
RoBERTa	59.1	64.6	53.9	2.2
ELECTRA	59.2	65.7	44.9	3.6
ELECTRA (large)	53.3	64.6	44.9	6.7

High variance

Results on **development** set, filtered ensemble (majority voting)

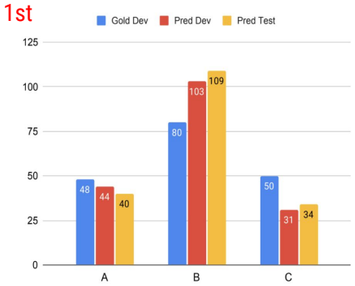
Model	Filtered models (n)	Acc.
Baseline		
Naive Bayes (unigram+bigram)	–	46.1
Logistic Regression (unigram+bigram)	–	51.1
Ensemble method		
All 500 "base" models	8	69.7
Biomed BERT	11	68.5
Biomed RoBERTa	7	67.4
Biomed RoBERTa (TAPT)	11	66.3
RoBERTa	3	67.9
ELECTRA	6	**70.2**
ELECTRA (large)	18	67.4

Domain-generic models are better

Results on **test** set using filtered ensemble:

Model	Accuracy	
	Dev	Test
All 500 "base" models	69.7	49.7
ELECTRA	70.2	50.2 → rank 2nd
ELECTRA (large)	67.4	**53.6** → rank 1st

Label distribution (Gold, Dev, and Test) using ELECTRA (large)



V. Conclusions

1. We show that morden pretrained language models suffer from high-variance issues on evidence grading task in medical literature.
2. We propose an ensemble method to handle the high-variance issues. With model filtering, we achieve competitive result