Automatic classification of tweets mentioning a medication using pre-trained sentence encoders

Laiba Mehnaz, MIDAS-IIITD, Delhi Technological University, India

Task

Objective: Automatic classification of tweets that mention medications. [1]

Dataset: Tweets posted by 112 pregnant women, with only 0.2 % of the tweets mentioning a medication, making the dataset highly imbalanced.

Approach

Aim: To analyze the performance of pretrained encoders on a dataset consisting of tweets, in a biomedical domain, and having a high class imbalance.

Pretrained encoders used: BERT, BioBERT, ClinicalBERT, SciBERT, RoBERTa, BioMed RoBERTA, ELECTRA, and ERNIE 2.0.

Experimental Results

Model	F1 Score	Precision	Recall
BERT base	0.78	0.83	0.74
BioBERT base	0.80	0.84	0.77
ClinicalBERT base	0.81	0.82	0.80
SciBERT base	0.83	0.90	0.77
RoBERTa base	0.81	0.82	0.80
BioMed RoBERTa base	**0.85**	0.90	0.80
ELECTRA base	0.79	0.81	0.77
ERNIE 2.0	0.83	0.92	0.74

Table 1: Perfomance of pretrained encoders on the validation dataset. BioMed-RoBERTa perfoms the best, followed by SciBERT and ERNIE 2.0.

Model	F1 Score	Precision	Recall
BioMED-RoBERTa base	**0.76**	**0.77**	**0.74**
Mean scores for Task 1	0.66	0.7	0.7

Table 2: Final submission(BioMed RoBERTa) test scores on the test set.

The original training dataset consisted of 146 tweets that mention a drug, and 55,273 tweets do not mention any. Due to the class imbalance in the dataset, we oversampled the positive tweets(i.e., mention a drug) by simply copying them. After oversampling, our training dataset consisted of 110,546 tweets in total.

Conclusions

1. BioMed RoBERTa is the best performing model, which is also the most computationally expensive model in the list.

2. All the BERT variants that were pretrained on medical corpus, show a consistent increase in performance compared to BERT, proving the importance of domain-specific pretraining.

3. ERNIE 2.0 matches the performance of SciBERT without any domain-specific prertraining, leading to a very interesting question of the possibility of universal pretrained encoders, that can learn domain agnostic linguistic features.

References

[1] Ari Z. Klein et. al. "Overview of the fifth Social Media Mining for Health Applications (SMM4H) Shared Tasks at COLING 2020". In: *Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task. 2020*