

NLP North at WNUT-2020 Task 2: Pre-training versus Ensembling for Detection of Informative COVID-19 English Tweets



Barbara Plank

Anders Giovanni Møller

Rob van der Goot

Task: Identification of informative COVID-19 English Tweets*

Informative

 **John Summer** ✓
@JohnSummer

Oklahoma's first confirmed case of coronavirus is in Tulsa County
<URL>[#SmartNews](#)

12:00 PM · Jun 1, 2020

23 Retweets **122** Likes

Uninformative

 **Pete Hanson**
@PeterHanson1

Trump could cure Coronavirus 19, AIDS, and Cancer in the same day and the media would say he wasn't doing anything.

12:00 PM · Mar 12, 2020

4 Retweets **29** Likes

Transformer-based models outperform traditional methods

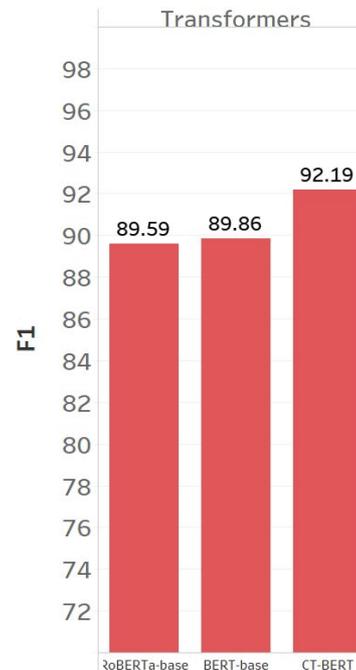


Domain-specific pre-training is important in this classification task

BERT: BookCorpus dataset

RoBERTa: Same as BERT but extended with CC-News, OpenWebText and Stories

CT-BERT: BERT-Large but extended with corona-related tweets



Our ensemble methods did not beat stand-alone CT-BERT



Soft voting: CT-BERT, RoBERTa, BERT



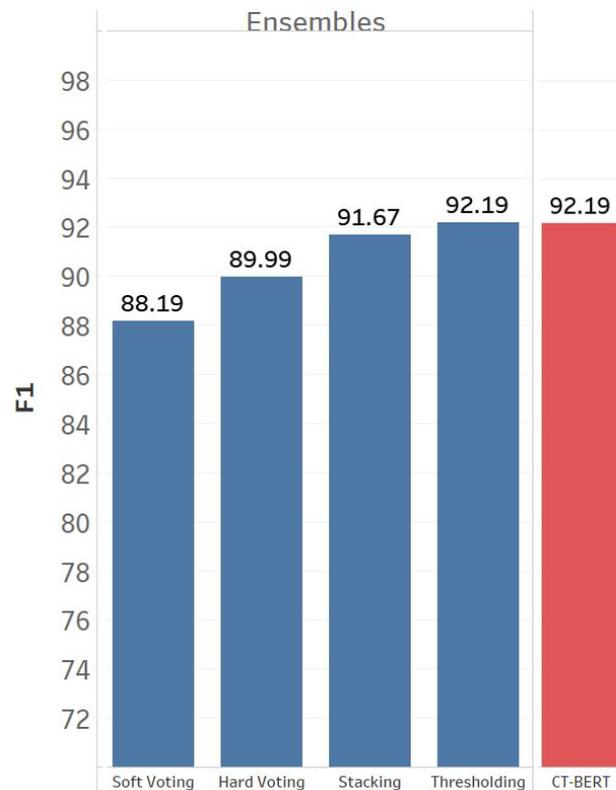
Hard voting: CT-BERT, RoBERTa, BERT



Stacking: Random Forest Classifier on all individual model predictions



Thresholding: CT-BERT, SVM



Summary



Transformer based models, especially CT-BERT, outperform traditional models



Domain-specific pre-training is very important in the classification task



Stacking was the most competitive ensembling method, though still underperforming compared to CT-BERT