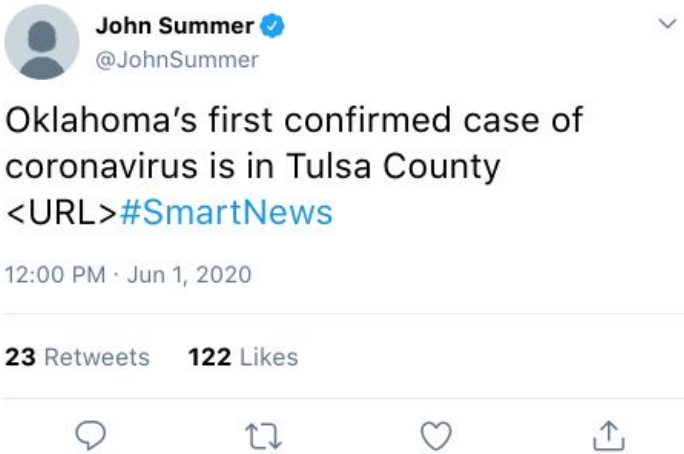
NLP North at WNUT-2020 Task 2: Pre-training versus Ensembling for Detection of Informative COVID-19 English Tweets

Barbara Plank **Anders Giovanni Møller** **Rob van der Goot**

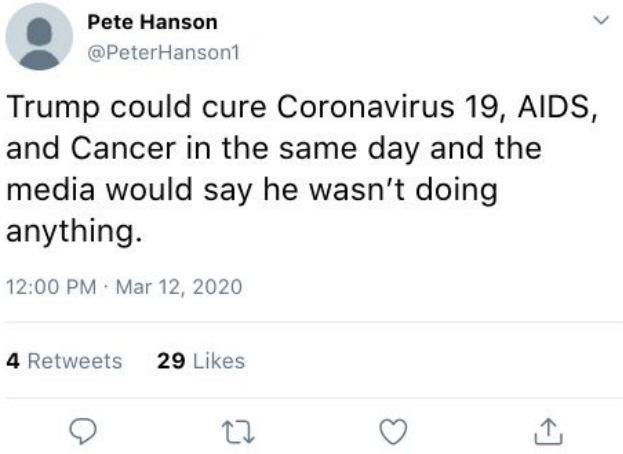
IT UNIVERSITY OF COPENHAGEN

Task: Identification of informative COVID-19 English Tweets*

Informative

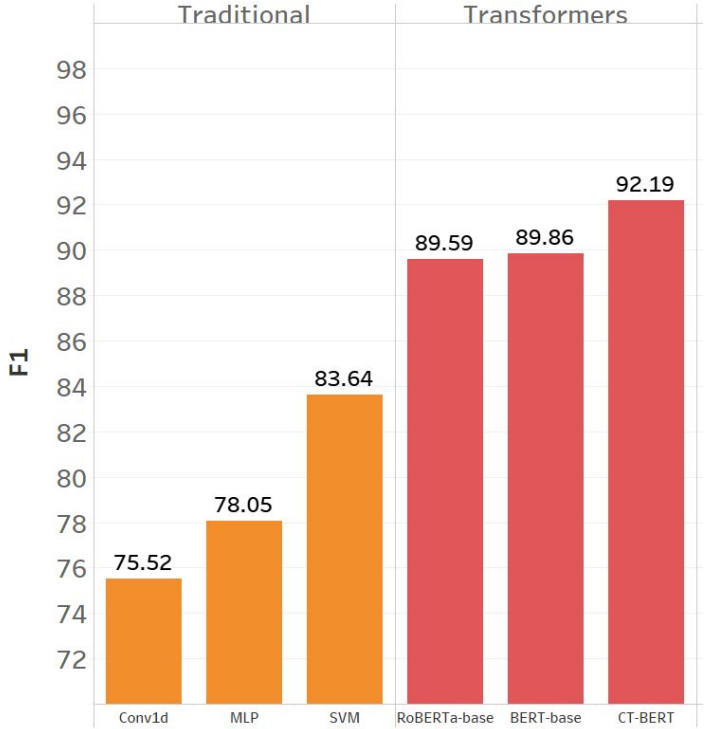


Uninformative



*Nguyen et al., 2020

Transformer-based models outperform traditional methods

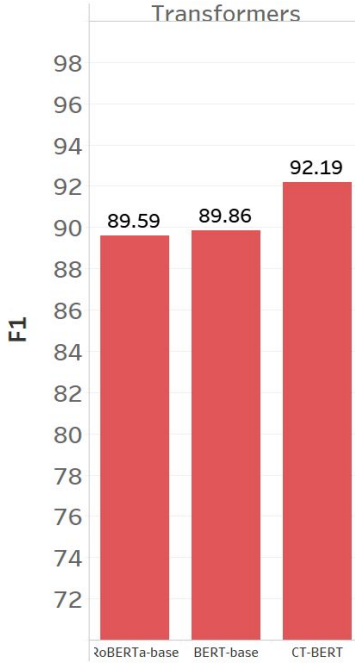


Domain-specific pre-training is important in this classification task

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



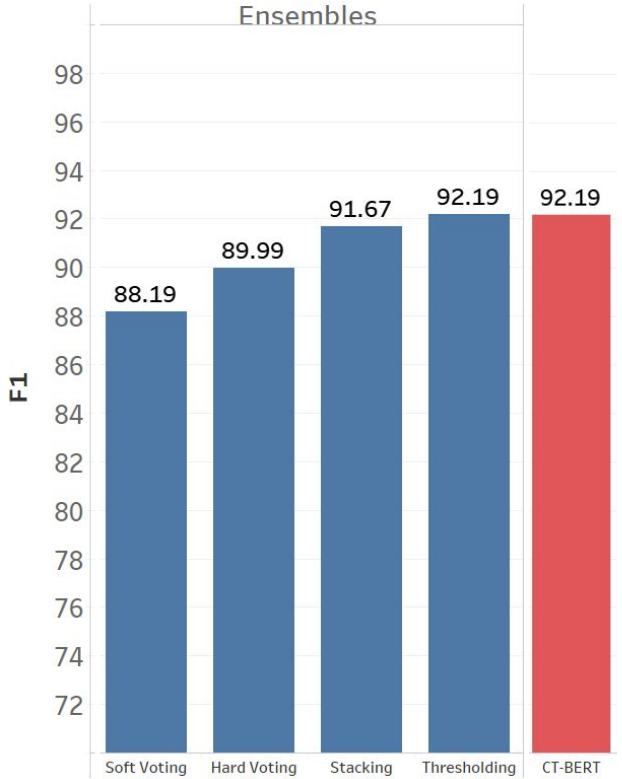
Our ensemble methods did not beat stand-alone CT-BERT

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