UIT-HSE at WNUT-2020 Task 2: Exploiting CT-BERT for Identifying COVID-19 Information on the Twitter Social Network

Khiem Vinh Tran*, Hao Phu Phan**, Kiet Van Nguyen*, Ngan Luu-Thuy Nguyen*

*University of Information Technology VNU-HCM, Vietnam
**National Research University HSE, Russia

Abstract

Recently, COVID-19 has affected a variety of real-life aspects of the world and has led to dreadful consequences. More and more tweets about COVID-19 has been shared publicly on Twitter. However, the plurality of those Tweets are uninformative, which is challenging to build automatic systems to detect the informative ones for useful AI applications. In this paper, we present our results at the W-NUT 2020 Shared Task 2: Identification of Informative COVID-19 English Tweets. In particular, we propose our simple but effective approach using the transformer-based models based on COVID-Twitter-BERT (CT-BERT) with different fine-tuning techniques. As a result, we achieve the F1-Score of 90.94place on the leaderboard of this task which attracted 56 submitted teams in total.

Task Description

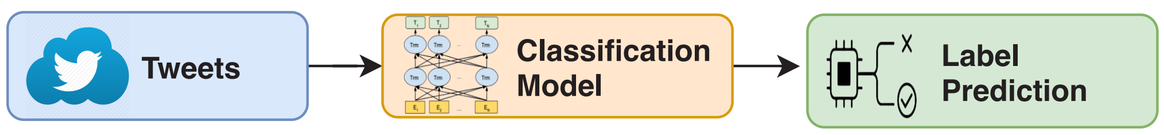

Figure 1: Task Definition

- **Input**: Given English Tweets on the social networking site Twitter.
- **Output**: One of two different labels (INFORMATIVE and UNINFORMATIVE) predicted by classifiers.

Our approach

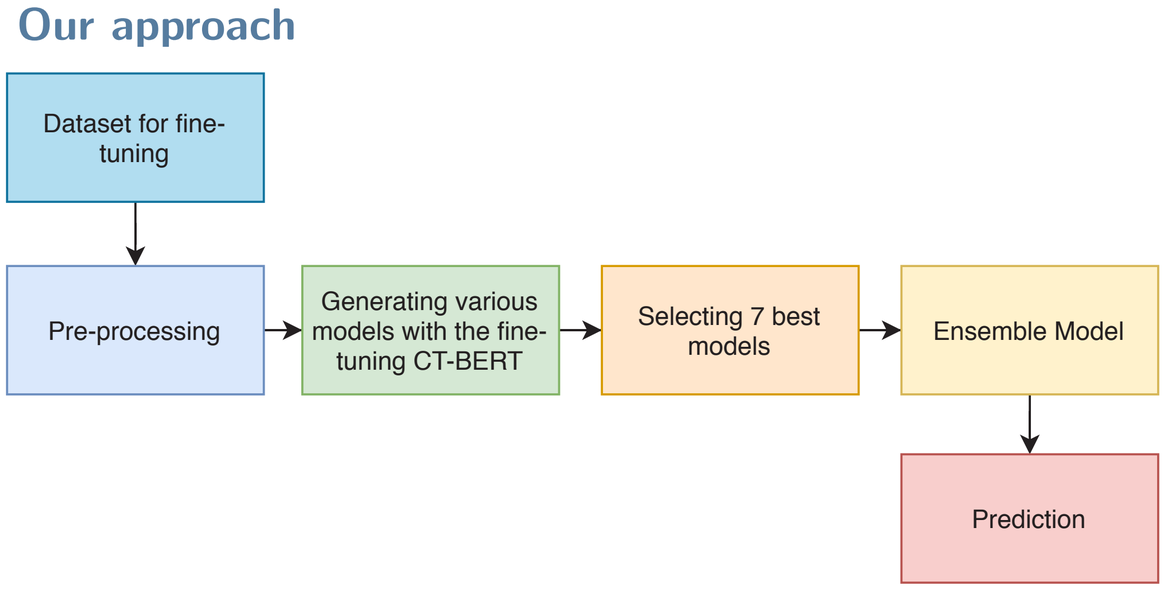

Figure 2: Our Approach

Pre-processing

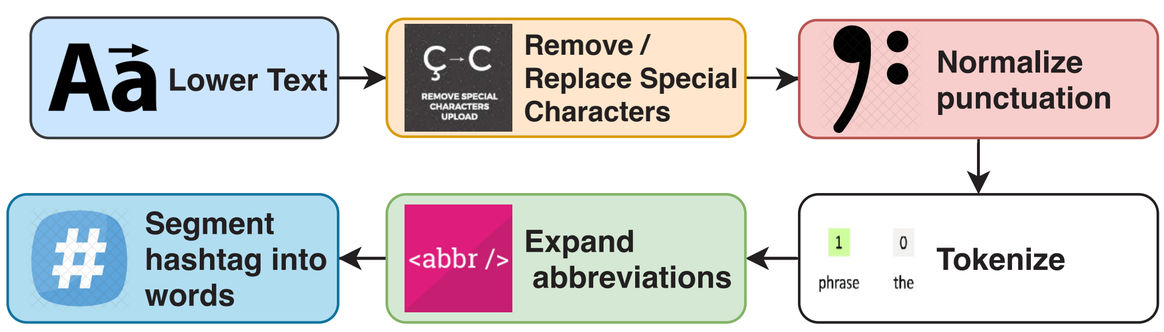

Figure 3: Pre-processing

- **Step 1**: Converting Tweets into lower texts.
- **Step 2**: Removing or replacing special characters (emojis including) with ASCII alternatives.
- **Step 3**: Normalizing punctuation.
- **Step 4**: BERT-tokenizer tokenizes each sentence into a list of tokens.
- **Step 5**: Expanding some common abbreviations.
- **Step 6**: Segmenting each hashtag into words.

Methodology

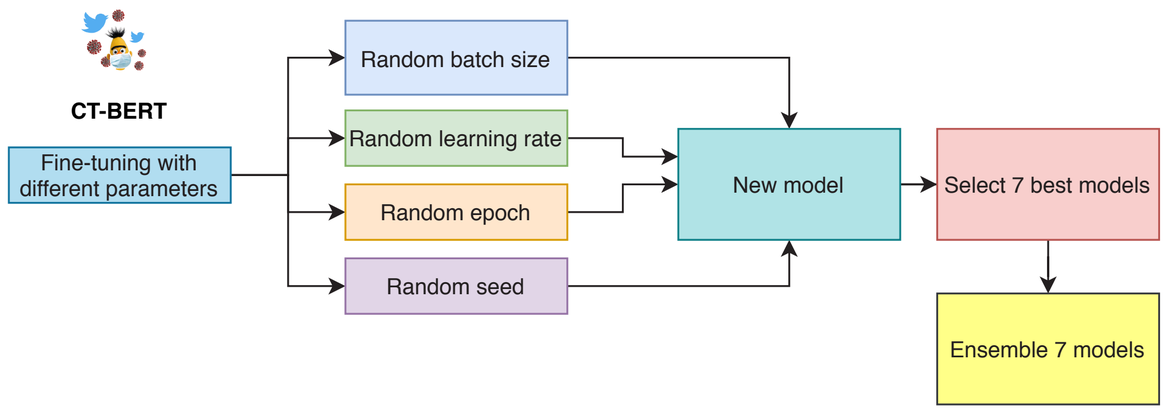

Figure 4: Methodology

Result

Model	P	R	F1	Acc
1	0.9179	0.9237	0.9208	0.9250
2	0.9059	0.9386	0.9220	0.9250
3	0.9202	0.9280	0.9241	0.9280
4	0.9043	0.9407	0.9221	0.9250
5	0.9236	0.9216	0.9226	0.9270
6	0.9076	0.9364	0.9218	0.9250
7	0.9216	0.9216	0.9216	0.9260
Ensemble (SV)	0.9174	0.9407	0.9289	0.9320
Ensemble (HV)	**0.9213**	**0.9428**	**0.9319**	**0.9350**

Table 1: Model performances of our proposed approach on the validation set. HV, SV, P, R, F1, and Acc stand for Hard Voting, Soft Voting, Precision, Recall, F1-score and Accuracy, respectively.

Rank	Team Name	P	R	F1	Acc
1	NutCracker	0.9135	0.9057	0.9096	0.9150
2	NLP_North	0.9029	0.9163	0.9096	0.9140
3	**UIT-HSE**	**0.9046**	**0.9142**	**0.9094**	**0.9140**
4	#GCDH	0.8919	0.9269	0.9091	0.9125
5	Loner	0.8918	0.9258	0.9085	0.9120
48	Baseline	0.7730	0.7288	0.7503	0.7710

Table 2: Performance of our system on the final scoreboard of the W-NUT-2020 Task 2. P, R, F1, and Acc stand for Precision, Recall, F1-score and Accuracy, respectively.

Other information

Figure 5: Full-text article