IDSOU at WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets

Sora Ohashi [†] **Tomoyuki Kajiwara** [‡] **Chenhui Chu** [‡]
Noriko Takemura [‡] **Yuta Nakashima** [‡] **Hajime Nagahara** [‡]
[†] Graduate School of Information Science and Technology, Osaka University
[‡] Institute for Datability Science, Osaka University
ohashi.sora@ist.osaka-u.ac.jp
{kajiwara, chu, takemura, n-yuta, nagahara}@ids.osaka-u.ac.jp

Abstract

We introduce the IDSOU[1] submission for the WNUT-2020 task 2: identification of informative COVID-19 English Tweets. Our system is an ensemble of pre-trained language models such as BERT. We ranked 16th in the F1 score.

1 Introduction

The spread of the COVID-19 is causing fear and panic to people around the world. To monitor the COVID-19 outbreaks in real-time, SNS analysis such as Twitter is attracting much attention. Although there are 4 million COVID-19 English Tweets posted daily on Twitter (Lamsal, 2020), most of them are uninformative. Against this background, WNUT-2020 held a shared task[2] (Nguyen et al., 2020) to automatically identify whether a COVID-19 English Tweet is informative or not.

Our system employs an ensemble approach based on pre-trained language models. Such pre-trained language models (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lan et al., 2020; Conneau et al., 2020; Lewis et al., 2020) have achieved high performance in various text classification tasks (Wang et al., 2019). In addition, we employ domain-specific pre-trained language models (Lee et al., 2019; Alsentzer et al., 2019; Müller et al., 2020) to build models suitable for COVID-19 and Twitter domains. Each model is optimized for three types of loss functions, cross-entropy, negative supervision (Ohashi et al., 2020), and Dice similarity coefficient (Li et al., 2020), which are useful for various text classification tasks. Finally, we ensemble 48 classifiers based on 16 pre-trained language models and 3 loss functions with a random forest classifier (Breiman, 2001).

[1] **I**nstitute for **D**atability **S**cience, **O**saka **U**niversity
[2] http://noisy-text.github.io/2020/

	Train	Dev	Test
Informative	3,303	472	944
Uninformative	3,697	528	1,056
Total	7,000	1,000	2,000

Table 1: Statistics of the dataset.

2 WNUT-2020 Shared Task 2

In the shared task (Nguyen et al., 2020), systems are required to classify whether a COVID-19 English Tweet is informative or not. Such informative Tweets provide information about recovered, suspected, confirmed and death cases as well as location or travel history of the cases. The 10,000 COVID-19 English Tweets[3] shown in Table 1 have been released for the shared task.

The baseline system is based on fastText (Bojanowski et al., 2017). Systems are evaluated by accuracy, precision, recall and F1 score, and are ranked by F1 score, which is the main metric. Note that the latter three metrics are calculated for the informative class only.

3 IDSOU System

We first introduce each base model in Section 3.1 and each loss function in Section 3.2. We then introduce the ensemble model in Section 3.3. Finally, Section 3.4 describes the implementation details.

3.1 Base Models

Recently, the fine-tuning approach for pre-trained language models (Devlin et al., 2019) has achieved the highest performance for many text classification tasks (Wang et al., 2019). We employ the following pre-trained language models of six types of architecture for the shared task.

[3] https://github.com/VinAIResearch/COVID19Tweet

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 428–433

BERT (Devlin et al., 2019) The transformer encoder pre-trained by multitask learning of masked language modeling and next sentence prediction. We employ three types of pre-trained models, BERT-base,[4] BERT-large,[5] and BERT-large-wwm.[6] BERT-base consists of 12 transformer layers, 12 self-attention heads per layer, and a hidden size of 768. BERT-large and BERT-large-wwm consist of 24 transformer layers, 16 self-attention heads per layer, and a hidden size of 1,024.

XLNet (Yang et al., 2019) The transformer encoder pre-trained by permutation language modeling. We employ two types of pre-trained models, XLNet-base[7] and XLNet-large.[8] The parameters of XLNet-base and XLNet-large are the same as BERT-base and BERT-large, respectively.

RoBERTa (Liu et al., 2019) The transformer encoder pre-trained by masked language modeling. RoBERTa has the same architecture as BERT, but pre-trains more steps on larger data with larger batch sizes. We employ two types of pre-trained models, RoBERTa-base[9] and RoBERTa-large.[10]

XLM-RoBERTa (Conneau et al., 2020) The multilingual transformer encoder pre-trained by masked language modeling. We employ a pre-trained model of XLM-RoBERTa-base.[11] XLM-RoBERTa-base consists of 12 transformer layers, 8 self-attention heads per layer, and a hidden size of 3,072.

ALBERT (Lan et al., 2020) The transformer encoder pre-trained by multitask learning of masked language modeling and sentence order prediction. ALBERT has significantly fewer parameters than the traditional BERT architecture due to two parameter reduction techniques, factorized embedding parameterization and cross-layer parameter sharing. We employ two types of pre-trained models, ALBERT-base[12] and ALBERT-large.[13] ALBERT-base and ALBERT-large have the same number of layers, attention heads, and hidden size as BERT-base and BERT-large, respectively, but the embedded size is 128.

BART (Lewis et al., 2020) The denoising autoencoder based on a bidirectional transformer encoder and a left-to-right transformer decoder. We employ two types of pre-trained models, BART-base[14] and BART-large.[15] BART-base consists of 12 transformer layers, 16 self-attention heads per layer, and a hidden size of 768. BART-large consists of 24 transformer layers, 16 self-attention heads per layer, and a hidden size of 1,024.

The language models mentioned above are pre-trained on corpora in the general domain such as the BookCorpus (Zhu et al., 2015) and English Wikipedia. Recent studies (Lee and Hsiang, 2019; Beltagy et al., 2019) have revealed that language models pre-trained on a domain-specific corpus achieve better performance in that domain. We employ the following three types of BERT models pre-trained on large-scale corpora of the medical domain and Twitter domain to build a classifier suitable for COVID-19 English Tweets.

BioBERT (Lee et al., 2019) The BERT encoder pre-trained on corpora in the biomedical domain such as PubMed abstracts (PubMed)[16] and PubMed Central full-text articles (PMC).[17] We employ two types of pre-trained models, BioBERT-base[18] and BioBERT-large.[19]

ClinicalBERT (Alsentzer et al., 2019) The BERT encoder pre-trained on corpora in both biomedical and clinical domains such as PubMed, PMC, and the MIMIC-III v1.4 database (Johnson et al., 2016). We employ a pre-trained ClinicalBERT[20] model with the same architecture as BERT-base.

[4] https://huggingface.co/bert-base-uncased
[5] https://huggingface.co/bert-large-uncased
[6] https://huggingface.co/bert-large-uncased-whole-word-masking
[7] https://huggingface.co/xlnet-base-cased
[8] https://huggingface.co/xlnet-large-cased
[9] https://huggingface.co/roberta-base
[10] https://huggingface.co/roberta-large
[11] https://huggingface.co/xlm-roberta-base
[12] https://huggingface.co/albert-base-v2
[13] https://huggingface.co/albert-large-v2
[14] https://huggingface.co/facebook/bart-base
[15] https://huggingface.co/facebook/bart-large
[16] https://www.ncbi.nlm.nih.gov/pubmed/
[17] https://www.ncbi.nlm.nih.gov/pmc/
[18] https://huggingface.co/dmis-lab/biobert-v1.1
[19] https://huggingface.co/trisongz/biobert_large_cased
[20] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

	XE	NS	DS
BERT-base	0.899	**0.904**	0.890
BERT-large	0.898	**0.899**	0.879
BERT-large-wwm	0.901	**0.910**	0.902
XLNet-base	0.898	**0.909**	0.885
XLNet-large	**0.911**	0.907	0.892
RoBERTa-base	0.906	**0.909**	0.896
RoBERTa-large	**0.916**	0.908	0.887
XLM-RoBERTa	**0.891**	0.879	0.856

	XE	NS	DS
ALBERT-base	**0.907**	0.893	0.883
ALBERT-large	**0.900**	0.897	0.867
BART-base	**0.886**	0.885	0.880
BART-large	**0.908**	0.905	0.855
BioBERT-base	**0.893**	0.889	0.873
BioBERT-large	0.896	**0.897**	0.859
ClinicalBERT	**0.879**	0.877	0.878
COVID-Twitter-BERT	0.922	0.923	**0.926**

Table 2: F1 scores of each loss function on the development set.

COVID-Twitter-BERT (Müller et al., 2020)
The BERT encoder pre-trained on the COVID-19 English Tweets. These are 160M Tweets collected between January 12 and April 16, 2020 containing at least one of the keywords "wuhan", "ncov", "coronavirus", "covid", or "sars-cov-2". We employ a pre-trained COVID-Twitter-BERT[21] model with the same architecture as BERT-large.

3.2 Loss Functions

We train classifiers based on pre-trained language models with the following three loss functions.

XE: Cross Entropy
We employ the following cross-entropy loss commonly used in text classification tasks.

$$L_{\mathrm{XE}} = \frac{1}{N}\sum_{i=1}^{N} \log P_i \tag{1}$$

where $P_i := P(y_i|X_i)$, y_i is the gold label, and X_i is the input text.

NS: Negative Supervision (Ohashi et al., 2020)
This loss function separates the representation of Tweets with different labels.

$$L_{\mathrm{NS}} = L_{\mathrm{XE}} + \frac{1}{NM}\sum_{i=i}^{N}\sum_{n=1}^{M} \cos(\boldsymbol{v}_i, \boldsymbol{v}_n) \tag{2}$$

where $\boldsymbol{v}_i$ is the representation of i-th text and $\boldsymbol{v}_n$ is that of negative examples, *i.e.* text representations that has different labels. We set the number of negative examples $M = 2$.

[21] https://huggingface.co/digitalepidemiologylab/covid-twitter-bert

	Loss	F1
BERT-large	XE	0.898
COVID-Twitter-BERT	XE	0.922
Ensemble: 16 models	XE	0.929
Ensemble: 48 models	XE+NS+DS	**0.933**

Table 3: Performance comparison of single model and ensemble model on the development set.

DS: Dice Similarity Coefficient (Li et al., 2020)
The loss function based on Dice-coefficient. The gap between maximizing F1 score and minimizing DS loss is less than that of minimizing XE loss.

$$L_{\mathrm{DS}} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \frac{2(1-P_i)P_i y_i + \gamma}{(1-P_i)P_i + y_i + \gamma}\right] \tag{3}$$

For smoothing purpose, we simply set $\gamma = 1$ following Li et al. (2020).

3.3 Ensemble Model

We ensemble 48 classifiers (16 pre-trained language models for each 3 loss functions) described above to make prediction stable. The Random Forest Classifier (Breiman, 2001) is trained using k-fold cross-validation on the development with the probabilities of the informative class estimated by each base model as the features.

3.4 Implementation Details

We implemented all models based on the Hugging Face's Transformers (Wolf et al., 2019) with Adam optimizer (Kingma and Ba, 2015). Hyperparameters of each base model were determined from the following combinations based on the F1 score in the development set.

Rank	Team	F1	Precision	Recall	Accuracy
1	NutCracker	0.9096	0.9135	0.9057	0.9150
2	NLP_North	0.9096	0.9029	0.9163	0.9140
3	SupportNUTMachine	0.9094	0.9046	0.9142	0.9140
4	#GCDH	0.9091	0.8919	0.9269	0.9125
5	Loner	0.9085	0.8918	0.9258	0.9120
⋮					
11	Husky	0.8992	0.8959	0.9025	0.9045
12	Hanoi001	0.8991	0.8787	0.9206	0.9025
13	UET	0.8989	0.8891	0.9089	0.9035
14	Emory	0.8974	0.8744	0.9216	0.9005
15	NJU ConvAI	0.8973	0.8751	0.9206	0.9005
16	**IDSOU**	**0.8964**	**0.8988**	**0.8941**	**0.9025**
17	ComplexDataLab	0.8945	0.9195	0.8708	0.9030
18	UPennHLP	0.8941	0.9028	0.8856	0.9010
19	datamafia	0.8940	0.8857	0.9025	0.8990
20	NIT_COVID-19	0.8914	0.8594	0.9258	0.8935
21	NHK_STRL	0.8898	0.8985	0.8814	0.8970
⋮					
48	BASELINE	0.7503	0.7730	0.7288	0.7710
⋮					
55	TMU-COVID19	0.5789	0.5000	0.6875	0.5280

Table 4: Official results in descending order of the F1 score.

- Batch size: [16, 32]
- Learning rate: [1e-5, 3e-5, 5e-5]
- Early stopping: [5]

We implemented the ensemble model based on the scikit-learn (Pedregosa et al., 2011). Hyper-parameters of the random forest classifier were determined through 5-fold cross-validation from the following combinations in the development set.

- `n_estimators`: [50, 100, 150]
- `max_depth`: [2, 4, 8]
- `min_samples_split`: [2, 8, 32]
- `max_samples`: [0.2, 0.5, 0.8, 1.0]

We followed the default data split provided by the task organizers. No external data has been used.

4 Results

Table 2 shows the F1 scores of each base model on the development set. The COVID-Twitter-BERT pre-trained with the in-domain corpus achieved the highest performance as expected. Since non-expert posts make up the majority of SNS, models pre-trained in the biomedical and clinical domains did not outperform that of the general domain.

Regarding the loss function, XE loss showed stable performance. NS loss is effective for 6 out of 16 models and seems to be compatible with BERT. DS loss achieved the best performance in combination with COVID-Twitter-BERT, although overall performance is not high.

Table 3 shows the effect of our ensemble method. These results reveal the effectiveness of the ensemble of both different pre-trained language models and different loss functions.

Table 4 shows the official results. We ranked 16th out of 55 teams in the F1 score.

5 Conclusions

We describe the IDSOU submission for the WNUT-2020 task 2. Our system is an ensemble model based on 16 pre-trained language models and 3 loss functions with a random forest classifier. In the official result, we ranked 16th out of 55 teams.

Acknowledgments

This work was supported by Innovation Platform for Society 5.0 from Japan Ministry of Education, Culture, Sports, Science and Technology.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45:5–32.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3(160035).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Rabindra Lamsal. 2020. Coronavirus (COVID-19) Tweets Dataset.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the Eighth International Conference on Learning Representations*.

Jieh-Sheng Lee and Jieh Hsiang. 2019. PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model. *arXiv:1906.02124*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, pages 1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice Loss for Data-imbalanced NLP Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv:2005.07503*.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Sora Ohashi, Junya Takayama, Tomoyuki Kajiwara, Chenhui Chu, and Yuki Arase. 2020. Text Classification with Negative Supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 351–357.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 7th International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, pages 5753–5763.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv:1506.06724*.