Emory at WNUT-2020 Task 2: Combining Pretrained Deep Learning Models and Feature Enrichment for Informative Tweet Identification

Yuting Guo
Computer Science
Emory University
Atlanta GA 30322, USA
yuting.guo@emory.edu

Mohammed Al-Garadi
Biomedical Informatics
Emory University
Atlanta GA 30322, USA
maalgar@emory.edu

Abeed Sarker
Biomedical Informatics
Emory University
Atlanta GA 30322, USA
abeed@dbmi.emory.edu

Abstract

This paper describes the system developed by the Emory team for the WNUT-2020 Task 2: "Identification of Informative COVID-19 English Tweet". Our system explores three recent Transformer-based deep learning models pretrained on large-scale data to encode documents. Moreover, we developed two feature enrichment methods to enhance document embeddings by integrating emoji embeddings and syntactic features into deep learning models. Our system achieved F_1-score of 0.897 and accuracy of 90.1% on the test set, and ranked in the top-third of all 55 teams.

1 Introduction

Until August 31, 2020, the COVID-19 outbreak has caused nearly 25 million confirmed cases worldwide, including about 800K deaths (WHO, 2020). Recently, much attention has been paid to building monitoring systems to track the development of COVID-19 by aggregating related data from different sources (*e.g.*, the Johns Hopkins Coronavirus Dashboard[1] and the WHO Coronavirus Disease Dashboard[2]). One potentially important source of information is social media, such as Twitter and Reddit, which provides real-time updates of the COVID-19 outbreak. To deal with the massive amount of social media data, several systems have been developed to detect and extract COVID-19 related information from Twitter (Chen et al., 2020; Banda et al., 2020; Sarker et al., 2020). However, only a minority of the data collections contain relevant information (*e.g.*, the recovered, suspected, confirmed and death cases as well as location or travel history of the cases) that are useful for monitoring systems, and it is costly to manually identify these informative data. To help address the problem, WNUT-2020 Task 2 (Nguyen et al., 2020) focuses on attracting research efforts to create text classification systems that can identify whether a COVID-19 English Tweet is informative or not. This could lead to automated information extraction systems for COVID-19, as well as benefit the development of relevant monitoring systems.

This paper describes the system developed by the Emory team for this task. Our solution explores three recent transformer-based models, which are pretrained on large-scale data and achieve great success on different NLP tasks. The pretrained models convert the input document into a embedding matrix, and the first token (*i.e.*, `[CLS]`) embedding is regarded as the document embedding. We also propose two methods to integrate pretrained deep learning models and empirical features by enriching document embeddings with emoji embeddings and syntactic features. The document embedding is fed into a normalization layer and an output layer, which are fine-tuned with the encoder during the training phase. The output is a probability value from 0 to 1, and the class with the highest probability is chosen. The highest F_1-score of our system is 0.897, ranking 14 of all 55 teams in the leaderboard[3].

2 Related Work

Recently, deep learning models pretrained on large-scale data, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and SpanBERT (Joshi et al., 2020), have improved the performance of many downstream NLP tasks such as named entity recognition, semantic role labeling, emotion detection, and question answering. Although these models are trained on open domain data such as news

[1] https://coronavirus.jhu.edu/map.html
[2] https://covid19.who.int/
[3] https://competitions.codalab.org/competitions/25845#results

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 388–393

articles and English Wikipedia, several approaches have been investigated to apply pretrained deep learning models to medical domain tasks. Matero et al. (2019) proposed a dual-context neural model that combines lexical features and personality factors with BERT embeddings for suicide risk assessment. Lee et al. (2019) combined BERT embeddings and BiLSTM (Bidirectional Long Short-Term Memory) to improve the performance of medical text inferences. Roitero et al. (2020) applied machine learning models and BERT to identify medical domain tweets and reported that BERT significantly outperforms some traditional machine learning models such as logistic regression and support vector machines for their task. Encouraged by this considerable progress, our work uses recent pretrained deep learning models as document encoders and fine-tunes the classification model on the dataset provided by the task organizers. Moreover, we develop two feature enrichment methods to enhance the document embeddings generated by pretrained deep learning models.

3 System Description

Figure 1 shows the overall framework of our model for the classification task. In this framework, we use three Transformer-based models described in Section 3.1 as encoders to generate the document embedding $\mathbf{e_d}$, which is the first token (i.e., `[CLS]`) embedding in our model. The document embedding is normalized by layer normalization and is fed into an output layer with Softmax activation. Also, we implement two feature enrichment methods to enhance document embeddings by integrating emoji embeddings and syntactic features (described in Section 3.2).

3.1 Document Encoder

RoBERTa: Devlin et al. (2019) developed a Transformer-based model named `BERT` that has achieved great success in several NLP tasks. Recently, Liu et al. (2019) has released a new pretrained model named `RoBERTa` that is trained with the same model architecture of `BERT` but on different datasets and pretraining tasks. We use `RoBERTa-large` as the encoder which outperforms `BERT-large` on different NLP tasks.

XLNet: Yang et al. (2019) implemented `XLNet`, a generalized autoregressive method that incorporates the autoregressive model into pretraining and optimizes the language model objective using a permutation method that can overcome the limitation of the masked language model in BERT. We use `XLNet-base` as the encoder which has been reported to outperform BERT and RoBERTa on some NLP tasks.

ALBERT: Lan et al. (2020) proposed a variant of BERT named `ALBERT` that applies parameter reduction techniques to reduce memory consumption and to accelerate the training process. ALBERT improves the training efficiency of BERT by factorizing embedding parameters and sharing cross-layer parameters. We use `ALBERT-xxlarge` as the encoder, which has also achieved state-of-the-art results on several NLP tasks.

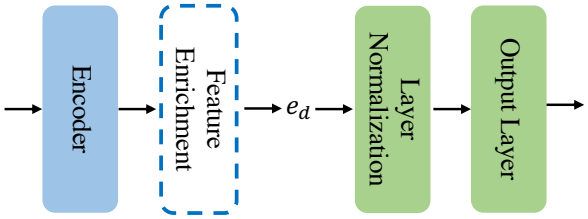

Figure 1: The overall framework of our model.

3.2 Feature Enrichment

Emoji Embedding: Emojis can succinctly represent emotional expressions and have become popular in social media. Recently, several studies have revealed that incorporating emoji information into deep learning models can benefit the performance of tweet classification tasks (Singh et al., 2019; Rangwani et al., 2018). Inspired by that, we employ a pretrained emoji encoder named `emoji2vec`[4] (Eisner et al., 2016) to convert emojis into emoji embeddings. The emoji embeddings are fed into a fully connection layer with $Tanh$ activation, and the output is concatenated to the `[CLS]` token embedding as the document embedding. If multiple emojis appear in one tweet, the emoji embeddings will be concatenated into one fixed-length vector.

Syntactic Feature: Syntactic features have been previously used on many NLP tasks (van der Lee and van den Bosch, 2017; Kurtz et al., 2019; Jabreel and Moreno, 2017), and syntactic dependencies often entail key information relevant to sentence topics. In order to utilize syntactic features, we collected a set of COVID-19 related keywords and extracted their governors (also known as heads) that

[4] https://github.com/uclnlp/emoji2vec

hold grammatical relations with the COVID-19 related keywords using the Stanford Parser (Chen and Manning, 2014), which can indicate what aspects of the keywords are discussed in tweets. The governor embedding and the [CLS] token embedding are then fed into a self-attention layer (Vaswani et al., 2017), and at the output, the new [CLS] token embedding is regarded as the document embedding.

3.3 Ensemble Model

To improve the robustness of our final system, we apply ensemble techniques to combine the results of different models. For each individual model, we take the class with highest output probability as the inference result during the testing phase. The inference results of all models are combined using a majority vote strategy, which is to return the class predicted by most models.

4 Experiments

4.1 Data Preprocessing

The dataset contains 10K COVID-19 English tweets of which 4719 are labeled as informative and 5218 are labeled as uninformative. To clean tweets data, we used the open source tool preprocess-twitter[5] including steps of lowercasing and normalizing numbers, hashtags, capital words and repeated letters. The dataset had been split into training, validation, and test sets with pre-specified sizes. Because the test set is not available when developing our system, we re-split the training set into a new training set and a new validation set with a 90/10 rate, and used the released validation set as the test set. The statistics for the data split are shown in Table 1.

	TRN	DEV	TST	ALL
INFORMATIVE	2928	345	472	3745
UNINFORMATIVE	3310	353	528	4191

Table 1: The statistics of the data split. TRN: the new training set; DEV: the new validation set; TST: The released validation set.

4.2 Training

Our system is implemented in PyTorch and Python 3. We develop five classification models of which three models use different encoders to generate document embeddings, and two models apply emoji embeddings and syntactic features to enrich document embeddings encoded by RoBERTa [6]. Among these models, the batch size of ALBERT is 16, and that of other models is 32. For the model applying emoji embeddings, the dimension of emoji embeddings is 300, and the max number of emojis in one tweet is 3. For the model applying syntactic features, only the governor of the first keyword is considered in order to control model complexity. Other hyperparameters are the same for all models listed in Table 2. Each model is trained separately on the training set and evaluated on the validation set during the training phase. Each experiment runs 3 times with different random initialization, and the model that achieves the highest accuracy on the validation set is selected for the testing phase.

Hyperparameter	Value
Learning rate	4e-5
Adam epsilon	1e-8
Max sequence length	128
Warmup ratio	0.06
Number of epochs	3

Table 2: Hyperparameter configurations of all models.

5 Results and Analysis

5.1 Task Results

We use F_1-score as the primary metric and also report accuracy of each model. Table 3 describes the results of each individual model and the ensemble model on the released validation set.

Model	Precision	Recall	F1	Acc
RoBERTa	86.7	93.3	89.9	90.1
RoBERTa+Emo	86.7	93.6	90.1	90.2
RoBERTa+Syn	87.2	91.7	89.4	89.7
XLNet	85.8	92.4	89.	89.2
ALBERT	88.8	90.3	89.6	90.1
Ensemble	**88.8**	**94.1**	**91.4**	**91.6**

Table 3: The precision, recall, F_1-score and accuracy of each model on the released validation set. RoBERTa+Emo: the model using RoBERTa as the encoder and applying emoji embeddings. RoBERTa+Syn: the model using RoBERTa as the encoder and applying syntactic features.

As we can observe, the F_1-score and accuracy of the ensemble model is marginally higher than any of the individual models. It indicates that the

[5] https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

[6] Due to the limited time and computational resources, we only experimented on adopting emoji embeddings and syntactic features for RoBERTa, and not for XLNet and ALBERT.

Albert Uderzo died in his sleep at his home in **Neuilly**, after a heart attack that was not linked to the coronavirus, his son-in-law **Bernard de Choisy** told the AFP news agency' #asterix
The Indian Express: **New York Zoo** tiger tests positive for coronavirus: Are cats at particular risk?. HTTPURL A group of humans knows everything about this virus. Others don't know anything about it.
Facsism kills in many subtle ways. **Trump** got away with **3,000** deaths in Puerto Rico. He's going to try it again with coronavirus. Time for all members of the media to ovary up and start reporting on the dangerous lies of this regime.

Table 4: Samples of false positives from the released validation set.

document embeddings generated by the individual models can encode information from different perspectives and complement each other.

5.2 Error Analysis

In order to investigate how our system can be improved, we conducted an error analysis on the released validation set. We found that 56 out of 84 error instances were false positives, (*i.e.*, uninformative tweets misclassified as informative), and the remaining 28 instances are false negatives (*i.e.*, informative tweets misclassified as uninformative). We observe that most of the false positives include numbers, locations, and personal names, which can be indicators of informative tweets such as the samples presented in Table 4. These indicators can be the noise that may confuse the model and lead to misclassification of negative instances. It suggests that we still need to improve the ability of our system to understand the context of indicative components in tweets.

5.3 Ablation Study

Given multiple models developed here, we conduct ablation study on each model to see how each individual model can affect the ensemble model. The performance of ensemble models that respectively remove one of the individual models from five models are shown in Table 5. As we can see, removing any of the individual models can increase the precision and decrease the recall of the ensemble model. It indicates that ensemble modeling can better identify negative instances than positive instances. Another notable result is that the recall drops more than other models when removing `RoBERTa`. Combined with the classification results in Table 3, `RoBERTa` might contribute most to the high recall of the ensemble model. It is interesting to note that `RoBERTa+Emo` is the best in the individual models but the least contributed to the ensemble model. The possible reason can be that only 9% of the training data contains emojis so that the emoji features are insufficiently learned during training, which may cause the document embedding of `RoBERTa+Emo` not much different from that of `RoBERTa`.

Model	Precision	Recall	F1	Acc
Ensemble	88.8	**94.1**	**91.4**	**91.6**
w/o ALBERT	89.4	92.6	90.9	91.3
w/o RoBERTa	89.3	91.7	90.5	90.9
w/o RoBERTa+Emo	89.5	92.6	91.0	91.4
w/o RoBERTa+Syn	**89.7**	92.2	90.9	91.3
w/o XLNet	89.5	92.4	90.9	91.3

Table 5: The ablation study of removing each individual model from the ensemble model.

6 Conclusion

This paper describes the system developed by the Emory team for the WNUT-2020 Task 2: "Identification of Informative COVID-19 English Tweets", including system design, implementation, evaluation, and analysis. We explored three pretrained deep learning models to encode documents, and developed two feature enrichment methods to enhance document embeddings by integrating emoji embeddings and syntactic features. For future work, we will continue to develop our system by exploiting other feature enrichment methods, utilizing external knowledge bases, and investigating other pretrained deep learning models.

References

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our preprint manuscript and to not have to update it every week.

Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill*, 6(2):e19273.

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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning Emoji Representations from their Description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Mohammed Jabreel and Antonio Moreno. 2017. SiTAKA at SemEval-2017 Task 4: Sentiment Analysis in Twitter Based on a Rich Set of Features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 694–699, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics 2020*.

Robin Kurtz, Daniel Roxbo, and Marco Kuhlmann. 2019. Improving Semantic Dependency Parsing with Syntactic Features. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 12–21, Turku, Finland. Linköping University Electronic Press.

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 *International Conference on Learning Representations*.

Chris van der Lee and Antal van den Bosch. 2017. Exploring Lexical and Syntactic Features for Language Variety Identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 190–199, Valencia, Spain. Association for Computational Linguistics.

Lung-Hao Lee, Yi Lu, Po-Han Chen, Po-Lei Lee, and Kuo-Kai Shyu. 2019. NCUEE at MEDIQA 2019: Medical Text Inference Using Ensemble BERT-BiLSTM-Attention Model. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 528–532, Florence, Italy. Association for Computational Linguistics.

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 preprint arXiv:1907.11692*.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Harsh Rangwani, Devang Kulshreshtha, and Anil Kumar Singh. 2018. NLPRL-IITBHU at SemEval-2018 Task 3: Combining Linguistic Features and Emoji pre-trained CNN for Irony Detection in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 638–642, New Orleans, Louisiana. Association for Computational Linguistics.

Kevin Roitero, VDMSM Cristian Bozzato, and G Serra. 2020. Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT. In *Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020)*.

Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. 2020. Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource. *medRxiv*.

Abhishek Singh, Eduardo Blanco, and Wei Jin. 2019. Incorporating Emoji Descriptions Improves Tweet Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2096–2101, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

WHO. 2020. Coronavirus disease (COVID-19) Weekly Epidemiological Update. https://www.who.int/docs/default-source/coronaviruse/situation-reports/

20200831-weekly-epi-update-3.pdf?sfvrsn=d7032a2a_4, Last accessed on 2020-09-04.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.