Medication Mention Extraction in Tweets using DistilBERT with Bootstrapping

Peijin Han,[1] Deahan Yu,[2] V.G.Vinod Vydiswaran[3,2]
[1]Department of Computational Medicine and Bioinformatics, Medical School
[2]School of Information, [3]Department of Learning Health Sciences, Medical School
University of Michigan
Ann Arbor, 48109, MI, USA
{pehan, deahanyu, vgvinodv}@umich.edu

***Abstract*—A large volume of layperson-authored messages get posted and consumed on Twitter, which makes it an important source for public health-related studies. The Task 3 of the BioCreative VII shared tasks challenged participants to detect and extract mentions of medications or dietary supplements from health-related tweets. In this article, we describe the runs we submitted for our participation in this task. Our system exploited two BERT embedding models – DistilBERT and BERTweet – for drug name extraction. On the test set, our best run achieved an overlapping precision, recall, and F1 of 77.2%, 78.2%, and 77.7% respectively, and a strict precision, recall, F1 of 69.9%, 69.4%, and 69.6% respectively. Our best run achieved a better overlapping F1 score compared to the mean of all 16 submitting teams.**

***Keywords*—medication extraction, DistilBERT, Bootstrapping**

I. Introduction

Twitter posts are now recognized as an important source of layperson-authored data, that has the potential to provide unique insights into population health (1). A fundamental step towards incorporating Twitter data in pharmacoepidemiological research is to automatically recognize medication mentions in tweets. A common approach is to search for tweets containing lexical matches of drug names from a manually-compiled dictionary. Even allowing for variants and misspellings, this approach has a low recall when applied to a corpus where drug names are rare (2). Numerous shared task challenges, such as the Social Media Mining for Health, have been organized, to encourage the NLP research community to develop state-of-the-art approaches for medication mention extraction.

In recent years, participants have largely adopted transformer-based models for this task, but their use in "real world" applications is still limited (3). This is in part because previous models were trained and tested on balanced training corpora, while the real-life data is highly imbalanced.

In this year's challenge, the dataset consists of all tweets posted by 212 Twitter users during their pregnancy and after. This data represents the natural and highly imbalanced distribution of drug mentions on Twitter, with only approximately 0.2% of the tweets mentioning any medication. Training transformer-based models on extremely imbalanced corpora is difficult and takes a long time if the training corpus is large (3). As our participation in this year's task, we proposed DistilBERT with bootstrapping for drug extraction. DistilBERT is a general-purpose pre-trained version of BERT (4). It is 40% smaller, 60% faster, and retains 97% of the language understanding capabilities (4). We trained three versions of the model by applying bootstrapping methods to create ensemble models, and finalized the predictions using majority voting. We believe that our system provides new methods to improve the extraction of drugs mentioned in posts and enhance the utility of social media for public health research.

II. Methodology

A. Datasets

The training data consists of 89,004 tweets from 212 users with only 234 tweets mentioning at least one drug. This dataset was previously shared with the research community as training data for SMM4H'21 (5) (SMM4H'21 training). The validation data consists of 38,149 tweets by the same 212 users, with 105 tweets mentioning at least one drug (SMM4H'21 validation). The test data consists of 54,482 tweets. To better train the model, we used the additional data from SMM4H'18 shared tasks, which was a balanced set of 9,622 tweets that either mention drugs or contain phrases with ambiguous drug names (SMM4H'18 training). We combined SMM4H'21 training and SMM4H'18 training sets as our final training dataset.

B. Data preprocessing

To clean the data, we first removed all the emojis, URLs, and special characters '@' and '#' that prefix usernames and hashtags. Next, we removed all the punctuation from the tweets, except for '-', '(' and ')'. We tokenized individual tweet using `nltk` package. Word tokens were split by space. A word with '-', '(', or ')' was counted as one token. For example, "Razberi-K" and "(G2T)" were considered as one token and not split into sub-tokens.

In the training data, the output labels "B-drug", "I-drug", and "O" were mapped to each token according to the position of the drug name span, where "B-drug" indicates the first word of the drug name, "I-drug" indicates the following word(s) of the drug name, and "O" indicates words that are not part of a

TABLE I
PERFORMANCE ON THE VALIDATION AND TEST DATASETS

Metrics	Validation Set			Test Set		
	D[a]	D+B5[b]	D+B10[c]	D[a]	D+B5[b]	D+B10[c]
Overlapping P	0.795	0.805	0.737	0.772	0.755	0.772
Overlapping R	0.848	0.867	0.829	0.762	0.796	0.782
Overlapping F1	0.820	**0.835**	0.780	0.767	0.775	**0.777**
Strict P	0.748	0.759	0.692	0.690	0.678	0.699
Strict R	0.790	0.810	0.771	0.667	0.701	0.694
Strict F1	0.769	**0.783**	0.730	0.678	0.689	**0.696**

[a]DistilBERT without bootstrapping.
[b]DistilBERT with bootstrapping 5 times.
[c]DistilBERT with bootstrapping 10 times.

drug name. These labels were all converted to numeric values when they were entered into the DistilBERT model.

C. DistilBERT

DistilBERT is a model pre-trained with knowledge distillation. Knowledge distillation (6, 7) is a compression technique where a compact model – the student – is trained to reproduce the behaviour of a larger model – the teacher – or an ensemble of models. The student is trained with a distillation loss over the soft target probabilities of the teacher: $L_{ce} = \sum_i t_i \times \log(s_i)$ where t_i and s_i are probability estimated by the teacher and the student, respectively. DistilBERT has the same general architecture as BERT, but has fewer layers. Sanh et al. showed that when compared to BERT, DistilBERT is 40% smaller, 60% faster, and retains 97% of the BERT performance (4).

To encode the tokens, we used a pre-trained DistilBERT tokenizer on ready-split tokens rather than the full sentence. We also applied padding and truncation, to normalize the token sequence for each tweet to be the same length as the maximum sequence length in the dataset.

The input of the DistilBERT model are *input_ids*, *labels*, and *attention mask*. The *input_ids* is the pre-trained embedding of the tokens. The *attention mask* masks the padding tokens to prevent them being used in the attention layer. DistilBERT uses WordPiece Tokenization, which can split single words into multiple tokens such that each token is likely to be in the vocabulary. To avoid the mismatch between labels and sub-tokens, we only train on the labels for the first sub-token of a split token. For all tokens we wanted to ignore, such as padding tokens or sub-tokens that were not the first sub-token, the label was set to -100.

D. Bootstrapping

In statistics, *bootstrapped sampling* is a method that involves drawing of sample data repeatedly with replacement from a data source to estimate a population parameter (8). In machine learning, *bootstrapped sampling* is used to generate multiple versions of dataset used to train predictive models. The predictions from these models are then aggregated to get the final predictions. Previous studies on bootstrapped models have shown improved performance compared to individual models (9). Bootstrapping is a widely used bagging method, which avoids over-fitting and improves the stability of machine learning algorithms (9).

The **research question** underlying our participation in the BioCreative VII shared task was whether bootstrapping will help improve DistilBERT models on the medication mention extraction task. To construct the training dataset for bootstrapping, we re-sampled the SMM4H'21 training data either five or ten times with replacement and combined the SMM4H'18 training data to each of them. We applied DistilBERT model to each dataset and generated individual predictions. We used majority voting to determine the final predictions for each tweet. To check consistency of individual models, we checked the individual predictions against the final voted prediction for all tweets, and didn't find any final prediction where all of the predictions were different from the individual models.

The pre-trained DistilBERT was downloaded from HuggingFace (10). The hyper-parameters used for fine-tuning DistilBERT are as follows: batch size was set as 16; warm up steps as 500; learning rate was 5e-5; and the number of training epochs was set to 10.

In addition to DistilBERT, we also tried BERTweet with similar bootstrap strategy for this model was pre-trained on the tweets corpus. However, the performance of these models on the validation set was significantly lower than the DistilBERT models (0.308 overlapping F1, compared to 0.835 overlapping F1 for a similar DistilBERT model). Therefore, we didn't choose BERTweet models among our official submission runs.

III. RESULTS

The performance of the three model variations are summarized in Table I. For all three experiments, the recall measure was consistently higher than precision in both validation and test sets. On the validation set, the bootstrapped model with the five model ensemble achieved the highest overlapping F1 and strict F1 scores, and outperformed the model with no bootstrapping (submission 1). On the test set, we observe that while bootstrapped model still performed better than the individual model, the bootstrapped model with the ten-model ensemble (submission 3) achieved the best overlapping F1-score of 0.777, and the strict F1 score of 0.696. The bootstrapped model with the five-model ensemble (submis-

sion 2) also outperformed the model with no bootstrapping (submission 1).

The DistilBERT model with the ten-model ensemble was judged as the best model among our runs. This model achieved an overlapping F1 score above the mean score of 0.749 among the best performing runs from the 16 participating teams on this task.

IV. Conclusion

This paper describes our participation in the BioCreative VII shared task 3 on medication mention extraction. We experimented with DistilBERT models with bootstrapping for extracting spans of medication or dietary supplements. Our best model on the test set was the bootstrapped DistilBERT model with a ten-model ensemble, that achieved an overlapping F1-score of 77.7%. This exceeded the mean score of 74.9% of the best runs from the 16 participating teams. Through these experiments, we provide additional evidence to indicate that bootstrapped sampling helps further improve DistilBERT models for extracting medication mentions from health-related tweets.

References

1. Edo-Osagie, O., Iglesia, B. D. L., Lake, I., and Edeghere, O. (2020). A scoping review of the use of twitter for public health research. *Computers in Biology and Medicine*, 122(0010-4825):103770.
2. Weissenbacher, D., Sarker, A., Klein, A., O'Connor, K., Magge, A., and Gonzalez-Hernandez, G. (2019). Deep neural networks ensemble for detecting medication mentions in tweets. *J Am Med Inform Assoc*, 26(12):1618–1626.
3. Weissenbacher, D., Rawal, S., MaggeGraciela, A., and Gonzalez-Hernandez (2021). Addressing extreme imbalance for detecting medications mentioned in twitter user timelines. *Artif Intell Med. AIME 2021. Lecture Notes in Computer Science*, 12721:93–102.
4. Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Computing Research Repository*, arXiv:1910.01108 [cs.CL]. version 4.
5. Magge, A., Klein, A., Miranda-Escalada, A., Al-garadi, M. A., Alimova, I., Miftahutdinov, Z., Farre-Maduell, E., Lopez, S. L., Flores, I., O'Connor, K., Weissenbacher, D., Tutubalina, E., Sarker, A., Banda, J. M., Krallinger, M., and Gonzalez-Hernandez, G., editors (2021). *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, Mexico City, Mexico. Association for Computational Linguistics.
6. Bucila, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of KDD*.
7. Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *Computing Research Repository*, arXiv:1503.02531 [stat.ML]. version 1.
8. Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7(1):1–26.
9. Breiman, L. (1994). Bagging predictors. *Mach. Learn*, 24(2):123–140.
10. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2019). Huggingface's transformers: state-of-the-art natural language processing. *Computing Research Repository*, arXiv:1910.03771 [cs.CL]. version 5.