NCU-IISR/AS-GIS: Detecting Medication Names in Imbalanced Twitter Data with Pretrained Extractive QA Model and Data-Centric Approach

Yu Zhang[1], Jong Kang Lee[1], Jen-Chieh Han[1] and Richard Tzong-Han Tsai [123]*
[1]Department of Computer Science and Information Engineering, National Central University, Taiwan
[2]IoX Center, National Taiwan University, Taiwan
[3]Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan
*Corresponding author: thtsai@g.ncu.edu.tw

Abstract——**In this paper, we introduce our system for the BioCreative VII Track 3 - Automatic extraction of medication names in tweets. Automatically extracting medication names from imbalanced data is challenging for deep learning models. Also, the length of the tweets is very short, which makes it hard to recognize medication names from the limited context. Here, our system combines classification and extractive question answering to solve the above problem. Moreover, domain-specific and task-specific pre-trained language models, as well as data-centric approaches are used to enhance our system. By combining the dictionary filtering and ensemble method, our system achieved 0.804 Strict F1 score far above the average performance 0.696 of 16 participating teams. Without using the dictionary and ensemble method, the single model we submitted achieved 0.797 Overlapping F1 which outperforms the result 0.773 of baseline system.**

Keywords—social media; medication detection; imbalanced data; text classification; data-centric; extractive question answering

I. Introduction

In recent years, twitter has become an important and popular resource in health informatics for disease surveillance, virus spread monitoring, and medication detection. In the BioCreative VII task 3, the participants have to develop text-mining systems to extract the spans of the drug and dietary supplement from the given tweet. The task expects participants to propose effective medication recognition methods to go beyond dictionary matching and thus facilitate the use of social media materials in public health research.

There are two main challenges: First, the distribution of medication names in the tweets is very sparse. The training data has about 89,200 twitters, but only 218 tweets contain at least one drug name. Second, the limited length of tweets makes it difficult to disambiguate the word sense from context. For instance, many drug names, such as Pain Killer and BOTOX, look like cultural product names. Moreover, Twitter limited the maximum length of each tweet to 140 characters until November 2017. The task dataset, however, are mostly posted before November 2017, thus direct applying named entity recognition methods from general domains to twitter did not yield good results.

To tackle these challenges, we followed up the strategy of the previous research (1), and we developed a two-stage system. First, all tweets were classified for the presence or absence of medication or dietary supplement names. Then, we converted the screened tweets that are likely to contain drug names into an extractive question-and-answer data format similar to SQuAD dataset (2) and used a pre-trained model for few-shot question answering called Splinter (3) to extract drug names

Although the same two-stage approach was used, we made some changes: First, we transformed the NER problem into a QA problem, which yields better results; Second, we applied multiple pre-trained language models for experiments; Third, data-centric approach was used to augment the training data.

This paper consists of the following contents. In Section II, we describe our system for BioCreative VII task 3. The results of internal experiments and test set submissions are detailed in Section III and Section IV. In Section V, we discuss the advantages, limitations, and future research directions of this approach.

II. System Description

In this section, we introduce our system, which consists of three parts. The first part is our medication text classification component. We introduce the used pre-trained models for classification and data preprocessing. The second part is an extractive QA component and postprocessing method. Thirdly, we describe the training data and data-centric approach, which is used in the above two components. These data augment methods use several external datasets which were published in previous studies. Figure 1 illustrate the architecture of our system.

A. Tweet Classification

Although the task of medication name extraction can be performed directly using a sequential labeling approach, this

TABLE I. Pre-trained Models used for Classification

Model	Pretraining Corpus	Text Size
BERTweet	Tweet	16B words/ 80GB
DeBERTa	Wiki + Book + Web	78 GB
BioBERT	Wiki + Book + PubMed	7.8B words
BioELECTRA	PubMed + PMC	13.8B words /84GB

This study is supported by the Ministry of Science and Technology, Taiwan [No.: MOST 109-2221- E-008-062-MY3].

approach is not as effective as the two-stage approach of classification followed by extraction. Due to the severely unbalanced classes of the dataset , the previous research (1) shows that adding a classification step before medication name extraction to classify whether a tweet contains medication names yields better results.

Among the different kinds of deep learning classification methods, classification methods based on pre-trained language models, like BERT, are shown to have the high performances on tweets (4). The effectiveness of different pre-trained language models is influenced by the pre-training text and the model architecture. If a pre-trained corpus is more similar to the text of its downstream task, it will have better result. Therefore, we surveyed the related literature and experimental reports to select the following pre-trained models: BERTweet (5), DeBERTa (6), BioBERT (7) and BioELECTRA (8). Table I shows the pretraining resources and corpus size of these models.

Also, sequential fine-tuning on similar tasks has been shown to improve the effectiveness of the models (9, 10). Natural language inference datasets, like MultiNLI (11), have been found to be effective in improving the final results of classification tasks. So, we fine-tuned above models on MNLI before fine-tuned them on the tweet classification task.

For model training and inference, Google BERT's architecture (12) is employed. We concatenated the prompting sentence "This tweet mention a drug, medication or dietary supplement in it" and each tweet as shown in Figure 1. [CLS] token embedding is used as the features. We then appended a softmax linear layer to output the logits of positive and negative.

Some preprocessing methods were applied to tweets including converting emojis to text, replacing user-names and URLs with @USER and HTTPURL. These methods are consistent with the preprocessing used by BERTweet (5).

B. *Extractive Question Answering*

Extracting spans containing medication names from text is usually formulated as a named entity recognition (NER) task. However, the previous study demonstrated that better performance was achieved by using a reading comprehension approach (13), i.e., an extractive question answering approach. This approach can encode entity context or keywords into the query/question, therefore the attention mechanism can utilize the information for the entity extraction.

For transforming entity recognition problem to QA dataset, we designed a targeted query "Extract the spans that mention a drug, medication or dietary supplement in the tweet" for each tweet that could potentially contain a drug name.

We use Splinter as the model to extract the spans of drug names. Splinter's architecture is the same as BERT which uses multiple layers of transformer encoder component, but it uses a pre-training method, called recurring span selection, specifically designed for the Extractive QA task (3). Splinter is shown to perform best on almost all SQuAD-like QA datasets of different fields, especially when the sizes of training set are small. This demonstrated that the Splinter model can fully exploit prior knowledge of downstream tasks.

Following the instructions in the Splinter paper, we appended a [Question] token to the end of the input sequence in fine-tuning. In the output layer of the model, Splinter computed a start vector and an end vector of the [Question] token through using the parameter matrices **S** and **E**. Each token's start and end position probabilities are calculated by the inner product of the

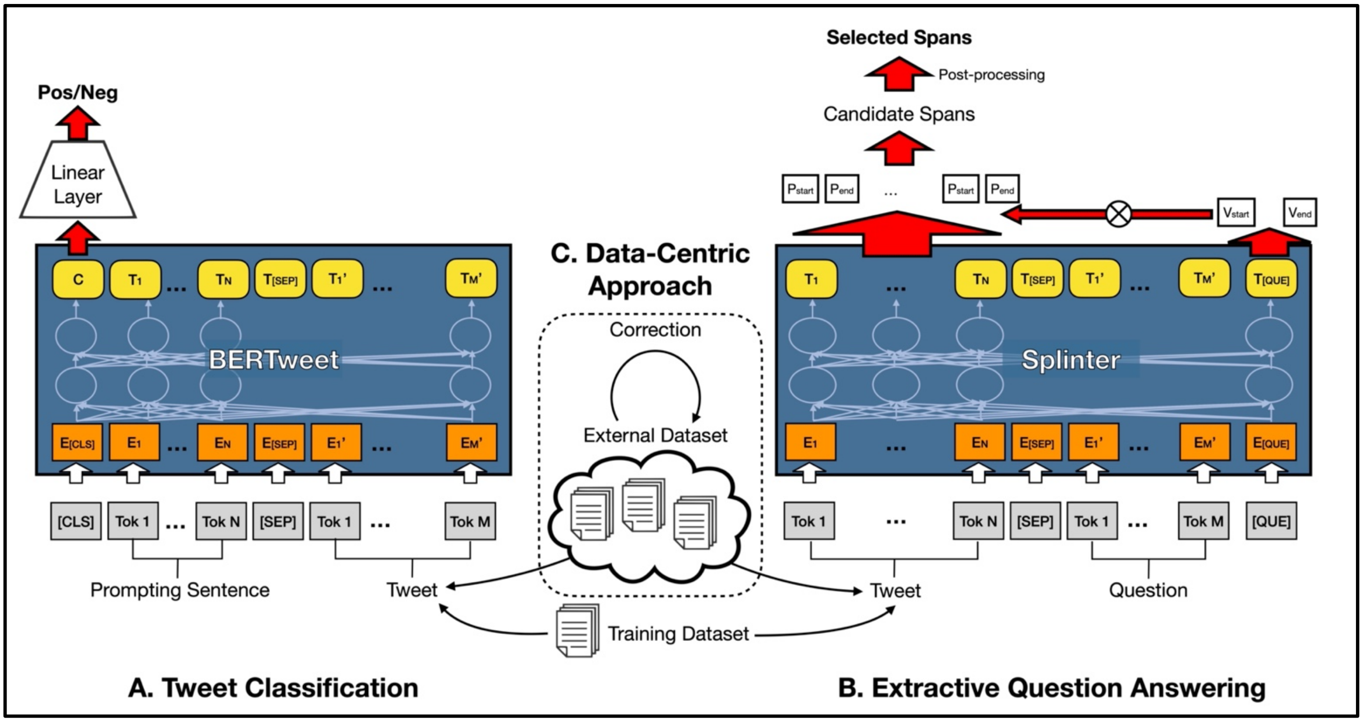

Fig. 1. Overview of our two-stage system combining text classification, extractive question answering, and data-centric approach.

start/end vector with the token's representation $\boldsymbol{x_i}$. The equations below are from Splinter paper.

$$P(s = i \mid T, q) = \frac{\exp\left(\boldsymbol{x}_{\boldsymbol{i}}^{\top} \boldsymbol{S}\boldsymbol{x}_{\boldsymbol{q}}\right)}{\sum_j \exp\left(\boldsymbol{x}_{\boldsymbol{j}}^{\top} \boldsymbol{S}\boldsymbol{x}_{\boldsymbol{q}}\right)}$$

$$P(e = i \mid T, q) = \frac{\exp\left(\mathbf{x}_{\mathbf{i}}^{\top} \mathbf{E}\mathbf{x}_{\mathbf{q}}\right)}{\sum_j \exp\left(\mathbf{x}_{\mathbf{j}}^{\top} \mathbf{E}\mathbf{x}_{\mathbf{q}}\right)}$$

We used Viterbi searching as the post-processing method to extract potential multiple drug names in a tweet. Each candidate span was filtered by a threshold value of probability. We tried multiple threshold values in the range of 0.1-0.15 to obtain the best results on the development dataset, and this threshold was also applied to the test set. Finally, for the extracted answer spans, we used regular expression to match all the spans that match in the corresponding tweet as our submissions.

C. Data-centric Approach

The data-centric approach was proposed by Andrew Ng in a workshop (14), and emphasizes the importance of training data, which forms a dichotomy terminology with the model-centric approach emphasized by current researchers. We considered that data-centric approaches can be divided into two parts: data augmentation, which focuses on collecting more relevant or similar data for training, and data quality enhancement, which focuses on orientations such as label consistency or data filtering. Data-centric approach could achieve greater performance improvements than a model-centric approach for datasets with scarce resources or imbalanced labels, which is exactly the problem with the tweet medication name extraction dataset.

External Datasets for Data Augmentation: Many researchers have previously conducted studies on the topic of tweets and drug usage, and some of them have published corresponding datasets. Hence, in addition to the dataset provided by the task organizer, we collected and processed the following additional datasets to fine-tune the classification and Extractive QA tasks:

- #1 TwiMed Dataset (15)

  TwiMed consists of 1000 tweets and 1000 PubMed sentences. Data are selected and annotated by the same pharmacists using the same guidelines at the entity level. We only used tweet corpus for classification and Extractive QA. A total of 508 annotated tweets were collected via the Twitter API. Other tweets are either deleted or not available for viewing.

- #2 SMM4H 2018 - Task 2 Medication Intake Classification Dataset (16)

  This dataset contains tweets that have been manually categorized into three classes: definite intake, possible intake, and no intake. The dataset authors provided 17,773 Tweet IDs. We excluded the "no intake" category and finally obtained 5,453 data for the training of the classification model.

- #3 SMM4H 2017 - Task 1 ADR Classification Dataset (17)

  The dataset included 25,678 tweets annotated to indicate the presence or absence of ADRs (Adverse Drug Reaction). We ended up with 8,554 tweets containing drug names for classification.

- #4 Large-scale Drug Usage-related Twitter Dataset (18, 19)

  This dataset uses Twitter data stored in the Internet Archive. They used a dictionary rule-based approach to filter the data that may contain drug names. We got a total of 2,162,822 tweets from this dataset. Because the rule matching approach is not as accurate as expert annotation, we did not use this dataset directly but used it as a supplemental source after filtering it with a BERTweet classification model, which performed best on dev dataset, and medication name lexicon made from training data.

Data Quality Enhancement: For the data augmentation of the Extractive QA task, we want the annotation of the external dataset to be consistent with the annotation of the training dataset. However, we found problems with the two external datasets used for drug name extraction, SMM4H 2018 Task 1 dataset and TwiMed dataset. A total of 343 extracted drug names for SMM4H 2018 Task 1 could not be matched to the corresponding tweet. For example, there is a tweet "Click here for $1.50 coupon #TeethingDoesntHaveToBite with **Infants' Advil** #FreeSample …" in the dataset, but the given extracted span is **"infant's advil"**. We used both manual examination and the Levenshtein distance algorithm to correct these problems.

In addition, we also found that the extracted drug Span was longer for the SMM4H 2018 Task 1 dataset compared to the training dataset, with an average of 11.14 for the former and 9.15 for the latter. Thus, we filtered the SMM4H 2018 Task 1 dataset by extracting data with drug span lengths longer than 27 for manual evaluation, which is the maximum length of drug span for the training dataset. A total of 189 instances were reviewed, 39 of which were removed and the rest were corrected. An example of a correction is to change **"cortisone 10 maximum strength"** to **"cortisone"**. A similar approach has been applied to the TwiMed dataset to improve the quality of the data.

III. Experiment Results

We used an RTX 3090 GPU for fine-tuning the model. For the classification model, we tried the following range of fine-tuning parameters: learning rate [4e-6, 8e-6, 1e-5, 3e-5], batch

TABLE II. INTERNAL EXPERIMENT RESULTS FOR CLASSIFICATION

Model	External Data	Best F1 on Dev
BioBERT-base	/	78.39
BioELECTRA	#0, #1, #2	80.67
DeBERTa-large	#0, #1, #2	88.37
BERTweet-base	/	84.87
BERTweet-base	#0	86.96
BERTweet-base	#0, #1, #2	87.16
BERTweet-large	/	89.72
BERTweet-large	#4 filtered	90.65
BERTweet-large	#0, #1, #2	**92.52**
BERTweet-large	#0, #1, #2, #3	91.58

TABLE III. EXPERIMENT RESULTS FOR SPAN EXTRACTION

Model	Recall	Precision	Strict F1
BERTweet-base for sequence labeling	64.8	80.0	71.6
BERTweet-base + Splinter	81.9	84.3	83.1
BERTweet-large + Splinter	90.5	89.6	90.0

size [16, 18, 20 ,32, 64]. Part of the internal experiment results for classification is shown in Table II.

The limited data from Table II shows the classification performance of each pretrained models and boosting effect of adding external datasets. #0 external data is SMM4H 2018 Task 1 dataset provided by the task organizer. Among different classification models, BERTweet-large was found to achieve the highest performance. This is expected, as BERTweet is the only model that has been pre-trained specifically on the Twitter corpus, and the large version has more parameters and will outperform the base version. The effect of sequential fine-tuning on the MultiNLI dataset, on the other hand, is inconsistent across models. Due to page limitation, the experiment results of sequential fine-tuning and the effect of adding prompting sentence are not shown here. We believe that the lack of positive cases in the development dataset may make comparisons between high-performance models difficult.

The internal experiment results of the two-stage system combined classification and extractive QA model are shown in Table III. We only send the data labeled as positive by the classification model, i.e., tweets that may contain drugs, to the Splinter model for drug name extraction. These experimental results demonstrate the higher performance by using the Splinter. We did not try to use Splinter alone to extract drug names directly without the classification because the model was not designed to predict the presence or absence of spans.

IV. Evaluation

In the evaluation phase, we selected seven BERTweet-large classification models with F1 scores higher than 90 on the development dataset and tried to improve the performance using a voting-based ensemble method. For the extractive QA component, we only use the Splinter model which has the highest performance. The configurations of collected classification models are shown in Table IV.

The configurations of the submitted three runs are as follows:

TABLE IV. CLASSIFICATION MODELS USED FOR EVALUATION

Model	External Data	Sequential Fine-tuning	F1 on Dev
BERTweet-large	#0	Y	91.08
BERTweet-large	MultiNLI	Y	91.07
BERTweet-large	#0, #1, #2	N	92.52
BERTweet-large	#0, #1, #2, Predicted Dev	N	92.45
BERTweet-large	#0, #1, #2, #3	N	91.58
BERTweet-large	#4 filtered	N	90.65
BERTweet-large	#0	N	91.71

TABLE V. EVALUATION RESULTS

Run	Strict R	Strict P	Strict F1
1	74.1	73.2	73.6
2	**83.0**	67.4	74.4
3	81.0	**79.9**	**80.4**
Mean	65.8	75.4	69.6
Run	**Overlapping R**	**Overlapping P**	**Overlapping F1**
1	80.3	79.2	79.7
2	**87.8**	71.3	78.7
3	84.4	**83.2**	**83.8**
Mean	70.9	81.1	74.9

- Run 1: It only used the BERTweet-large model which achieved the 0.9252 F1 score on the development dataset for classification and the Splinter model for medication name span extraction.
- Run 2: It used all 7 classification models and filtered the tweets classified as positive by at least one model for span extraction. In addition, we used the medication name dictionary (20) from the baseline method provided by the task organizer to exclude data for which none of the tokens of selected spans belong to the dictionary. The lexicons of medication names in the dictionary were tokenized by spaces.
- Run 3: It used similar method as Run 2, the only difference is that we use majority voting for ensemble here.

Table V shows the evaluation results of our submitted data and the mean scores of all 16 teams that participated in the task.

V. Discussion

As the evaluation results shown, almost all the F1 scores of our three submissions exceeded the mean score, which indicates that our model has generally good performance. It is worth noting that the average recall scores of all teams participating in the task are low compared to the average precision scores, both in terms of strict recall and overlapping recall, while our submissions are relatively balanced, performing well above the average in terms of recall. In addition, our run 1 submission without dictionary filtering and ensemble achieved decent results outperforming the highest overlapping F1 score 0.773 in baseline system (1), which shows that deep learning techniques have great potential for exploitation.

However, we also noticed a large discrepancy between the performance of the system on the development dataset and the test dataset. Given that the ensemble method used in run 3 is effective in improving F1 scores, we speculate that the classification models over-fitted the development dataset during the fine-tuning process.

In future work, we would like to find ways to overcome the over-fitting problem. We hope to be able to achieve good results with just a single model without relying on ensemble methods. Automated methods for the data-centric approach are also worthy of further research.

References

1. Weissenbacher, D., S. Rawal, A. Magge, and G. Gonzalez-Hernandez. (2021) *Addressing Extreme Imbalance for Detecting Medications Mentioned in Twitter User Timelines*. in *International Conference on Artificial Intelligence in Medicine*. Springer.

2. Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. (2016) *Squad: 100,000+ questions for machine comprehension of text.* arXiv preprint arXiv:1606.05250.

3. Ram, O., Y. Kirstain, J. Berant, A. Globerson, and O. Levy. (2021) *Few-shot question answering by pretraining span selection.* arXiv preprint arXiv:2101.00438.

4. Klein, A., et al. (2020) *Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020.* in *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*.

5. Nguyen, D.Q., T. Vu, and A.T. Nguyen. (2020) *BERTweet: A pre-trained language model for English Tweets.* arXiv preprint arXiv:2005.10200.

6. He, P., X. Liu, J. Gao, and W. Chen. (2020) *Deberta: Decoding-enhanced bert with disentangled attention.* arXiv preprint arXiv:2006.03654.

7. Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, and J. Kang. (2020) *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics. **36**(4): p. 1234-1240.

8. raj Kanakarajan, K., B. Kundumani, and M. Sankarasubbu. (2021) *BioELECTRA: Pretrained Biomedical text Encoder using Discriminators*. in *Proceedings of the 20th Workshop on Biomedical Language Processing*.

9. Vu, T., T. Wang, T. Munkhdalai, A. Sordoni, A. Trischler, A. Mattarella-Micke, S. Maji, and M. Iyyer. (2020) *Exploring and predicting transferability across NLP tasks.* arXiv preprint arXiv:2005.00770.

10. Zhang, Y., J.-C. Han, and R.T.-H. Tsai. (2021) *NCU-IISR/AS-GIS: Results of Various Pre-trained Biomedical Language Models and Linear Regression Model in BioASQ Task 9b Phase B*. in *CEUR Workshop Proceedings*.

11. Williams, A., N. Nangia, and S.R. Bowman. (2017) *A broad-coverage challenge corpus for sentence understanding through inference.* arXiv preprint arXiv:1704.05426.

12. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2018) *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805.

13. Li, X., J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li. (2019) *A unified MRC framework for named entity recognition.* arXiv preprint arXiv:1910.11476.

14. Ng, A.Y., *A chat with andrew on mlops: From model-centric to data-centric ai*. 2021.

15. Alvaro, N., Y. Miyao, and N. Collier. (2017) *TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations.* JMIR public health and surveillance. **3**(2): p. e6396.

16. Weissenbacher, D., A. Sarker, M. Paul, and G. Gonzalez. (2018) *Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018*. in *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*.

17. Sarker, A. and G. Gonzalez-Hernandez. (2017) *Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017.* Training. **1**(10,822): p. 1239.

18. Tekumalla, R. and J.M. Banda. (2020) *A large-scale Twitter dataset for drug safety applications mined from publicly existing resources.* arXiv preprint arXiv:2003.13900.

19. Tekumalla, R., J.R. Asl, and J.M. Banda. (2020) *Mining Archive. org's twitter stream grab for pharmacovigilance research gold*. in *Proceedings of the International AAAI Conference on Web and Social Media*.

20. Weissenbacher, D., A. Sarker, A. Klein, K. O'Connor, A. Magge, and G. Gonzalez-Hernandez. (2019) *Deep neural networks ensemble for detecting medication mentions in tweets.* Journal of the American Medical Informatics Association. **26**(12): p. 1618-1626.