Database, 2022, 1–9
DOI: https://doi.org/10.1093/database/baac067
Original article

Task reformulation and data-centric approach for Twitter medication name extraction

Yu Zhang[1], Jong Kang Lee[1], Jen-Chieh Han[1] and Richard Tzong-Han Tsai[1,2,3,*]

[1]Department of Computer Science and Information Engineering, National Central University, No. 300, Zhongda Rd., Zhongli Dist., Taoyuan City 32001, Taiwan
[2]IoX Center, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan
[3]Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan

[*]Corresponding author: Tel: +886-3-4227151 ext. 35203; Fax: +886-3-4222681; Email: thtsai@g.ncu.edu.tw

Citation details: Zhang, Y., Lee, J.K., Han, J. *et al.* Task reformulation and data-centric approach for Twitter medication name extraction. *Database* (2022) Vol. 2022: article ID baac067; DOI: https://doi.org/10.1093/database/baac067

Abstract

Automatically extracting medication names from tweets is challenging in the real world. There are many tweets; however, only a small proportion mentions medications. Thus, datasets are usually highly imbalanced. Moreover, the length of tweets is very short, which makes it hard to recognize medication names from the limited context. This paper proposes a data-centric approach for extracting medications in the BioCreative VII Track 3 (Automatic Extraction of Medication Names in Tweets). Our approach formulates the sequence labeling problem as text entailment and question–answer tasks. As a result, without using the dictionary and ensemble method, our single model achieved a Strict F1 of 0.77 (the official baseline system is 0.758, and the average performance of participants is 0.696). Moreover, combining the dictionary filtering and ensemble method achieved a Strict F1 of 0.804 and had the highest performance for all participants. Furthermore, domain-specific and task-specific pretrained language models, as well as data-centric approaches, are proposed for further improvements.

Database URL: https://competitions.codalab.org/competitions/23925 and https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/

Introduction

Twitter has become an important and popular resource in health informatics for disease surveillance, monitoring the spread of viruses, and detection of medication usage in recent years. BioCreative VII Task 3 Automatic Extraction of Medication Names in Tweets expects participants to propose effective medication recognition methods to go beyond lexical matching and thus facilitate using social media data in public health research. The task provides a Twitter corpus of all timeline tweets posted by 212 users who have self-reported their pregnancies, with the spans of drugs and dietary supplements (for simplicity, they will be referred to as drug names in the following) in the text annotated by experts. The corpus represents a natural and highly uneven distribution of drug mentions on Twitter, with 181 607 tweets not mentioning a drug (negative tweets) and only 442 tweets mentioning at least one drug (positive tweets, approximately 0.24%). The participants have to develop text-mining systems to extract drug names from the given tweets.

However, extracting the names of medications mentioned in tweets is a challenging task. First, the distribution of medication names in tweets is very sparse. The training data consisted of 89 200 tweets, but only 218 tweets contained at least one drug name. Second, the limited length of tweets makes it hard to disambiguate the word sense based on the context. For instance, many drugs, such as Pain Killer and BOTOX, may also be used to refer to cultural products like bands or songs. Second, Twitter limited the maximum length of each tweet to 140 characters until November 2017. Most tweets are only one sentence or even just a few words. This makes it difficult to discern the semantics of words from their contexts. Third, many tweets are not written with proper grammar, and many emoticons, special symbols and slang words are included. Therefore, direct applying named entity recognition (NER) methods from general domains to Twitter has not yielded good results (1).

Since tweet data have the above characteristics, it is not easy to train a sequence labeling model directly on the dataset. The previous research (2) breaks the task down into two subtasks, proposed a two-stage approach and demonstrated that it is effective for tweet NER in general domains. Thus, we developed a two-stage system for medication extraction, as illustrated in Figure 1 (Procedure I). First, all tweets were classified for the presence or absence of drug names. Then, we converted the screened tweets that are predicated to contain drug names into an extractive question-and-answer (QA)

Received 26 February 2022; Revised 26 July 2022; Accepted 20 August 2022

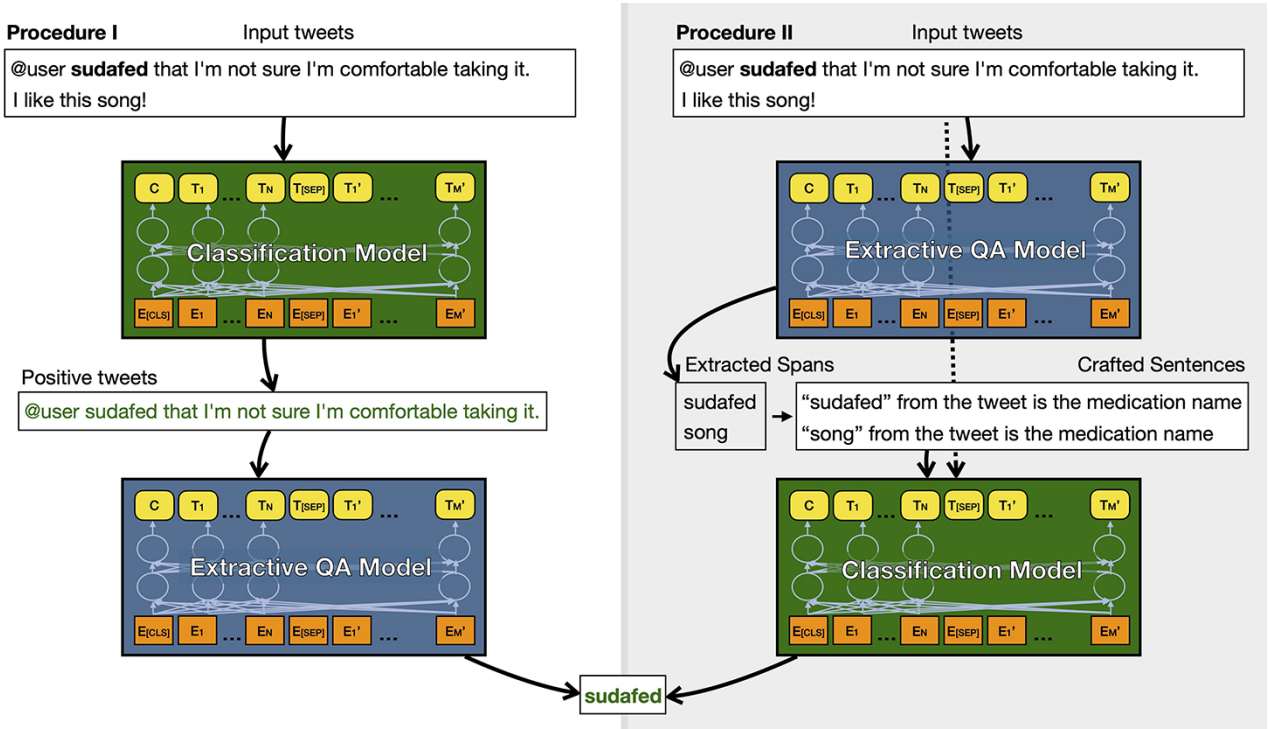

Figure 1. Overview of our two procedures for extracting medication names from tweets.

data format similar to the SQuAD dataset (3) and used a pre-trained language model (PLM) for few-shot question answering (learning from just a few questions) called Splinter (4) to extract drug names.

In addition, we tried a different order to concatenate our two-stage, as shown in Figure 1 (Procedure II). First, Splinter extracted all possible drug names from input tweets. Then, we used the extracted drug names to create short sentences, which were fed into a PLM along with the original tweets to determine whether the extracted names are correct. Both procedures achieved better results compared to direct sequence tagging. The latter approach can achieve relatively good performance using only two Bidirectional Encoder Representations from Transformers (BERT)-base-sized PLMs. This makes it possible to apply this system to the monitoring of large-scale social media content.

We have also adopted data-centric approaches to improve the model performances. These include adding external datasets to increase the proportion of positive cases and the use of 'confident learning' and span annotation correction to improve the quality of external datasets. We used the Cleanlab tool (https://github.com/cleanlab/cleanlab), which implements state-of-the-art confident learning algorithms, to remove noise data in our dataset (5). Confident learning uses a portion of the training data to train a model, which will be used to estimate the errors in the rest data. Then, it calculates an error matrix and uses the matrix to filter the data that are most likely to be noisy.

The highlights of this paper are as follows:

- We introduced our systems for BioCreative VII Task 3, and it achieved the best performance among all participants.
- We proposed task reformulation and data-centric approaches for improving the performance of PLMs.
- Specifically, our results suggest that using a two-stage approach (text entailment and extractive QA) has better performance than using sequence labeling directly in medication extraction from tweets.
- Moreover, we evaluated the methods for improving data quality by Cleanlab (5) and showed that it could further improve our best results in the BioCreative VII Task 3.

Related work

Identifying drug names in the text is a typical NER task. In this section, we first review recent NER approaches applied to Twitter text. Then, we discuss two recent approaches for enhancing the effectiveness of the PLMs.

Twitter NER datasets

NER identifies the spans of real-world entities in text and is one of the fundamental NLP tasks. In recent years, Twitter has shown its global social media status and the relative openness of data and has gained the attention of many researchers. Therefore, many datasets for NER in Twitter have been proposed in recent years (6–9). These Twitter NER datasets typically contain thousands of tweets. In addition to the named entities such as people names, places and organizations often annotated in common NER datasets, Twitter NER datasets tend to tag miscellaneous entertainment or cultural product entities, such as TV shows, electronic products, movies, music artists and so on.

Meanwhile, researchers in health-related domains have noted the diversity of user-generated content on Twitter. In the context of this trend, the Social Media Mining for Health Research and Applications (#SMM4H) shared tasks that were started in 2016 (10) and continue to be held annually. They

organized some Twitter NER tasks, like extracting adverse drug effect mentions, identifying professions and occupations, etc. Below we highlight some SMM4H participants' approaches for Twitter NER tasks.

Twitter NER approaches

Like the CRFClassifier (11), traditional NER tools usually suffer from the diverse and noisy style of tweets. Some researchers have combined external databases or mixed multiple machine learning methods to improve the results. Ritter *et al.* used LabeledLDA to leverage the Freebase dictionary as a source of distant supervision (6). Van Erp *et al.* combined a web extractor, Ritter *et al.*'s system and Stanford NER with generated features such as Part-of-speech (POS) to perform classification using the Sequential minimal optimization (SMO) machine learning algorithm (12).

After 2015, most researchers started using neural networks for Twitter NER tasks. The CambridgeLTL used bidirectional Long short-term memory (LSTM) and won first place in the 2016 WNUT (9). After 2018, the recurrent neural network-based models were gradually replaced by PLMs such as BERT (13–15). BERT is pretrained on large datasets such as Wikipedia and Google Books datasets. Its Transformer provides a more robust architecture against different NLP tasks. In SMM4H 2021, the BERTweet and RoBERTa models based on BERT were widely used by the participants (16). This paradigm shift is in line with the overall NLP research community trends.

Task reformulation for PLMs

PLMs have demonstrated their robustness against traditional statistical and machine learning approaches in NER tasks. However, the PLMs are also not easily modified because the cost of pretraining is too high for many researchers. Therefore, it is common to fine-tune the parameters of the PLM and then modify the last output layer. Task reformulation is one of the recent directions to solve this problem. For example, template-based learning is an approach that transforms a text classification task into a cloze task, which is the common language models' pretraining task (17). This approach makes better use of the knowledge captured in the pretraining tasks, thus reducing the number of task-specific training instances required to achieve performance similar to that of previous approaches.

In addition to converting text classification to cloze tasks, a recently proposed task reformulation approach in the NER task is to convert sequence tagging tasks to extractive QA tasks. Li *et al.* designed natural language questions for named entities and used the BERT model to find the answer spans in the text (18). This approach has achieved state-of-the-art results on most NER datasets, such as CoNLL 2003 and OntoNotes 5.0. Although task reformulation has achieved good results in few-shot learning and zero-shot learning, its performance in imbalanced class datasets lacks validation. Our study can be a contribution in this regard.

Data-centric approach

Andrew Ng (19) proposed a data-centric approach and emphasized the importance of training data, which forms a dichotomous terminology with the model-centric approach emphasized by most researchers. In NLP, many data augmentation methods have been proposed, including back-translation, reversing the order of words, etc., but there has been relatively little research on data quality enhancement methods. However, in the field of computer vision, the approach called 'confident learning' has been successfully applied to improve the quality of datasets and find hundreds of labeling errors in the ImageNet dataset (5). A 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. We view data-centric approaches as divisible 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. We also tried to employ confident learning in this study to remove data that may be problematic.

Materials and methods

In this section, we introduce our system, which consists of four parts. The first part is our medication text classification component. We introduce the pretrained models used for classification, data preprocessing and task reformulation. The second part is about the extractive QA component and post-processing method. Third, we describe the training data and data-centric approach, which is used in the above two components. Finally, we will introduce our post-processing methods, including the voting-based ensemble method, filtering of extracted spans, etc. Figure 2 illustrates the main architecture of our system.

Tweet classification

Previous research (2) shows that better results are obtained by adding a classification step before medication name extraction to classify whether a tweet contains medication names. In this paper, in addition to classifying the tweet in the first step, we also reversed the order and checked whether the extraction was correct by classifying it with PLM after the drug name was extracted.

Among the different kinds of deep learning classification methods, classification methods based on PLMs, like BERT, are shown to have high performance on tweets (20). The effectiveness of different PLMs is influenced by the pretraining text and the model architecture. If the pretraining corpus is more similar to the text of the downstream task, it will have better results. Therefore, we surveyed the related literature and experimental reports to select the following pretrained models: BERTweet (21), DeBERTa (22), BioBERT (23) and BioELECTRA (24). Table 1 shows the pretraining resources and corpus size of these models.

For model training and inference, Google BERT's architecture (25) is employed. [CLS] token embedding is used as the features. We then appended a SoftMax linear layer to output the logits of positive and negative. We use an additional prompting sentence for each sample, which is separated from the tweet by a special separator token. This turns the original single-sentence classification task into a textual entailment task. For Procedure I, We crafted the prompting sentence 'This tweet mention a drug, medication or dietary supplement in it' based on the task description. For Procedure II, we add the

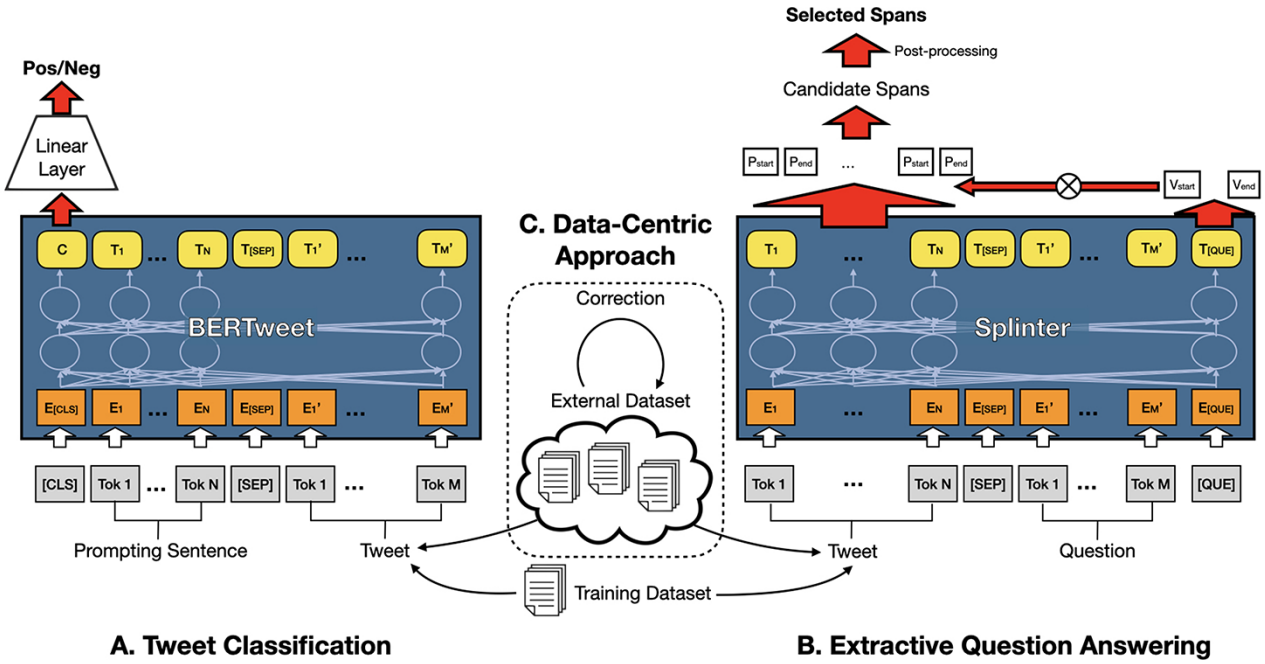

Figure 2. Overview of our two-stage system combining text classification, extractive question answering and data-centric approach.

Table 1. PLMs used for classification

Model	Pretraining corpus	Text size
BERTweet	Tweet	16 B words/80 GB
DeBERTa	Wiki + Book + Web	78 GB
BioBERT	Wiki + Book + PubMed	7.8 B words
BioELECTRA	PubMed + PubMed Central	13.8 B words/84 GB

Table 2. Task datasets overview

Data type	Tweets number	Tweets with at least one drug	Tweets with more than one drugs
Training	88 988	218 (0.25%)	16 (0.02%)
Validation	38 137	93 (0.24%)	12 (0.03%)
Test	54 482		
#0 SMM4H'18 Task 1	9622	4975 (51.70%)	790 (8.21%)

potential drug names extracted by the Splinter model to the prompting sentence. The sentence template is ' from the tweet is the name of a drug, medication or dietary supplement'. If a tweet has more than one potential drug name, we create a prompting sentence for each potential drug. The key idea of the task reformulation is to convert the class labels into natural language sentences that can be used to describe the labels, and PLMs need to determine whether the example entails a label description (26).

Extractive question answering

Extracting spans containing medication names from text is usually formulated as a sequencing tagging task. However, we transformed it into an extractive QA task by adding questions. This approach can encode entity context or key words into the question; therefore, the transformer's self-attention mechanism can utilize the information for the entity extraction.

For transforming entity recognition problem to a QA dataset, we designed a targeted query 'The medications, drugs and dietary supplements in the text are: [Question]' for each tweet that could potentially contain a drug name. A Splinter model was used to extract the spans of drug names. Splinter's architecture is the same as BERT, which uses multiple layers of transformer encoder components, but it uses a pretraining method, called recurring span selection, specifically designed for the extractive QA task (4). Splinter has been shown to perform best on almost all SQuAD-like QA datasets of different fields, especially when the sizes of training sets are small. This demonstrates 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 start/end vector with the token's representation x_i. The equations below are from the Splinter paper.

$$P(s=i|T,q)=\frac{\exp\left(x_i^{\top}Sx_q\right)}{\sum_j \exp\left(x_j^{\top}Sx_q\right)}$$

$$P(e=i|T,q)=\frac{\exp\left(x_i^{\top}Ex_q\right)}{\sum_j \exp\left(x_j^{\top}Ex_q\right)}$$

We used Viterbi searching as the post-processing method to extract potential multiple drug names from a tweet. Each candidate span is filtered with a threshold of probability 0.1, and overlapping spans are removed. Finally, we used regular

Table 3. External datasets

No.	Dataset	Tweets number	Data annotation	Annotation method
#1	TwiMed (27)	508	Drug span	Expert
#2	SMM4H'18 Task 2 (14)	5453	Drug intake classification	Expert
#3	SMM4H'17 Task 1 (13)	8554	Adverse drug reaction classification	Expert
#4	Large-scale drug usage-related Twitter dataset (28, 29)	2 162 822	Contain drug names	Lexicon

expression to find all the spans in the corresponding tweet which match our extracted spans.

Datasets and data-centric approach

Datasets: As shown in Table 2, there are four different types of datasets: training data, validation data, test data and additional data from #SMM4H'18 shared tasks. Except for the additional dataset, all tweets in the corpus are from 212 pregnant Twitter users. Here, we number SMM4H'18 Task 1 as #0 to facilitate subsequent discussions.

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 datasets provided by the task organizer, we collected and processed the following additional datasets (Table 3) to fine-tune the classification and extractive QA tasks. All tweets were collected via the Twitter API.

We used the first three datasets directly as training data for model training. But the #4 dataset was not used directly since it uses a dictionary rule-based approach to filter the data that may contain drug names, and this approach is not as accurate as an expert annotation. For this dataset, we took a special filtering approach. First, we used the trained BERTweet classification model to filter out tweets that may contain drugs. Then, we used a medication name list to filter these tweets. The medication name list is derived from the medication names of the training data, and we only use those drug names that appear fewer than five times in the training data. We believe this allows for a more balanced frequency of drugs in the dataset and avoids adding too many new additional tweets, which would disrupt the original data distribution. Eventually, 187 tweets containing drug names were added to the training dataset.

Span Annotation Correction for External Datasets: 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, the 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 than for 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.

Automatic Data Improvement using Cleanlab: Cleanlab is a tool that applies confident learning (6) to machine learning datasets. CL is an approach that focuses on label quality by characterizing and identifying label errors in a dataset. It is based on the principles of pruning noisy data, estimating noise with probabilistic threshold counts and ranking examples. The input to CL is the out-of-sample predicted probabilities and the vector of noisy labels. Here, we assumed that the task training data are the golden standard and do not contain labeling errors. Therefore, we only used the training dataset to train the classification model and used it to make predictions on the external dataset. The predicted probabilities were fed into Cleanlab to filter the data with labeling problems. The filtered tweets were not used as training data.

Post-processing

During the challenge, we selected seven BERTweet-large classification models with F1 scores higher than 90 on the validation dataset and tried to improve the performance using a majority-voting-based ensemble method. For the extractive QA component, we only use the Splinter model, which has the highest performance. In addition, we used the medication name dictionary (2) 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 lexicon of medication names in the dictionary was tokenized by spaces. The cases described here are for Procedure I.

After the challenge, we tried to improve the performance of the model using Procedure II. Since this system is targeted for large-scale social media content monitoring, inference speed is important. The ensemble approach based on multiple models cannot satisfy this need. Therefore, we use only one Splinter model and one BERTweet-based model with the best performance on validation data to compose the Procedure II system.

Results and discussion

The experiments were conducted on the BioCreative VII Track 3 task dataset. At first, we evaluated the performance of the whole system on the test dataset, and the official F1 score script is used as the evaluation metric. Among them, the most important evaluation metrics are overlapping F1 and Strict F1. According to the task organizers' definition (1), in strict evaluation, a system was rewarded only when it predicted the

Table 4. Whole system and procedure evaluation on test data

Procedure	System	Strict			Overlapping		
		R	P	F1	R	P	F1
I	BERTweet-base + Splinter	74.1	69.9	71.9	79.6	75	77.2
	BERTweet-large + Splinter	74.1	73.2	73.6	80.3	79.2	79.7
	7 BERTweet-large ensemble + Splinter + Lexicon	81	79.9	80.4	84.4	83.2	83.8
II	Splinter + BERTweet-base	79.6	74.5	77	83	77.7	80.3
	Splinter + BERTweet-large	77.6	78.1	77.8	81	81.5	81.2
	Splinter + 7 BERTweet-large ensemble + Lexicon	80.3	78.7	79.5	83.7	82	82.8
	Baseline (2)	66	89	75.8	67.3	90.8	77.3
	Mean (1)	65.8	75.4	69.6	70.9	81.1	74.9

exact start and end positions of the annotated drug name. In the overlapping evaluation, this restriction was relaxed and a system was rewarded when the span of its prediction overlapped with the span of the annotated drug name. Since the test dataset is not publicly available, we use the validation dataset to evaluate the key parts of the system in Analysis of classification and extractive QA model Section, Effects of task reformulation, Effects of Cleanlab Section. Next, we conducted extensive experiments on evaluating the effects of task reformulation and confident learning by using the base version of BERTweet in sections C and D. Lastly, we conducted an error analysis on our best configuration and illustrated some representative types of our error cases.

Whole system and procedure evaluation on test data

In Table 4, we compare the performance of our whole systems on test data. The details of the relevant configuration and components used will be described in later sections. The baseline system is a combination of the lexicon and the two-stage BERT model. Mean refers to the average score of the 16 participating teams during the challenge.

As the evaluation results show, our system outperforms the baseline system in almost all F1 metrics and is well above average. And, our system, for both procedures, has more balanced performance on test data, with precision and recall close to each other. Both the baseline system and mean performances of all participants tend to have higher precision and lower recall scores. The higher recall rate means that our system is less likely to miss mentions of drug names, which is an advantage we believe is more practical in real-world applications. In application scenarios, deep learning models are mostly used to assist human decision-makers. Reviewing and removing false-positive cases is a relatively simple task for humans, especially if the proportion of positive cases is extremely low. Using the proportion of classes in the task dataset as an example, the human decision-maker only needs to review about 20 out of 10 000 tweets that are predicted by the model to contain drug names. However, if the model misses a large number of tweets containing drug names, this would be difficult and time-consuming for the human decision-makers to remedy.

The performance of the Procedure II system is particularly noteworthy, as it exceeds that of the corresponding systems with Procedure I with the same number of parameters using only one BERTweet model. Just changing the flow can improve the performance of the system, which we believe is attributable to the fact that the prompting sentences generated under Procedure II contain specific spans that are more in line with the context and therefore of higher quality. The performance of the multi-model ensemble system of Procedure II is slightly lower than that of Procedure I; however, the scores are very close. We considered that the PLMs might reach their limits on this task dataset, and we will discuss more in the error analysis section. On the other hand, the system of Procedure II is more suitable for real-world social media analysis applications since it can achieve better results with the base model. The amount of data in this scenario are very large and are constantly increasing, the number of model parameters, and the speed of inference must be taken into account. In the case of limited resources, it is difficult to ensemble multiple models to meet the inference requirements.

Table 5. Performance of PLMs in classification task

PLM	F1 on development data
BERTweet-base	86.96
DeBERTa-base	82.96
BioBERT-base	82.3
BioELECTRA-base	76.42

Detailed analysis of classification and extractive QA model on validation dataset

The limited data from Tables 5 and Table 6 show the classification performance of each PLM and boosting effect of adding external datasets. We used an RTX 3090 GPU for fine-tuning the model and tried the following range of fine-tuning parameters: learning rate [4e – 6, 8e – 6, 1e – 5 and 3e – 5], and batch size [16, 18, 20, 32 and 64]. The evaluation score is the highest F1 score on the validation dataset. In Table 5, all models were trained with the training data and #0 external data, since #0 is provided by the task organizer. Among different PLMs, BERTweet-large was found to achieve the highest performance. This is expected, as BERTweet is the only model that has been pretrained specifically on the Twitter corpus, and the large version has more parameters and will outperform the base version. The poor performance of BioBERT and BioELECTRA is probably due to the fact that they were mostly pretrained on PubMed texts. The language style of these biomedical research papers is formal and the discussion topics are academically related, which is different from the social media texts of Twitter. We believe that the lack of positive cases in the validation dataset may make comparisons between high-performance models difficult.

Next, we will discuss the main challenge of the task dataset, namely class imbalance. There are only about 0.25%

Table 6. Effect of adding external datasets in classification task

Method	BERTweet model size	External datasets					F1
		#0	#1	#2	#3	#4	
2	Base						84.87
3	Base	✓					86.96
5	Base	✓	✓	✓			87.16
7	Large	✓	✓	✓	✓		91.58
8	Large	✓	✓	✓	✓	✓	91.07
9	Large	✓	✓	✓			92.52

of cases in the original training dataset that are positive. Before the experiment, we thought that if we did not increase the proportion of positive cases, the model would perform very poorly. However, the experiments show that BERTweet-base maintains a good performance of 84.87 F1 score even without adding external datasets or balancing positive/negative cases. We were surprised by this result. However, there are similar experimental results from other teams participating in this task. Qing Han *et al.* used PubMedBERT to sequence-tag the original task dataset directly and achieved an F1 score of 70.9(30). It is reasonable that we achieved a better performance using BERTweet pretrained on the Twitter dataset and performing only the classification task. This means that class imbalance may not be a major issue for the BERT model, especially when the task organizer has provided an additional dataset (#0 SMM4H'18 Task 1). This external dataset increases the percentage of positive cases from 0.25% (218 tweets) to 5.26% (5193 tweets). The data augmentation approach we have adopted further alleviates this problem. The benefit of this approach is that the positive examples in the additional dataset provide more diversity in language text content and some new drug names, whereas common solutions to class imbalance such as under sampling detract from this. This is perhaps even more important for Twitter text, which is highly variable and lacks uniform writing rules. In addition, our experiments show that using a larger model with more parameters also improves performance on the imbalanced class dataset (BERTweet-large improves the F1 score by 4.85% compared to BERTweet-base).

It is also worth noting that the effect of the model is rather weakened by the addition of the #3 and #4 datasets. We believe this could be due to two possible reasons. One possibility is that there are too many positive data, and the class distribution does not match the distribution in the original task dataset. After adding the #3 and #4 datasets, there will be more than 19 000 positive tweets in the training data, with a proportion of 17.4%. The other possibility is that the content of the tweets in the #3 dataset is not only references to medication or taking medication but rather discussions of side effects related to medication. Although it still contains drug names, the distribution of text semantics is different from the original task dataset.

The experiment results of the two-stage system combined classification and extractive QA model are shown in Table 7. Sequence tagging means that we use the traditional BIO NER annotation method for model training and validation. For the experiments combining the classification model with the Splinter model, 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 high performance of using Splinter. And, because it scored close to exactly right, we did not try more comparisons of other PLMs.

Table 7. Performance of two-stage system on validation dataset

Model	Recall	Precision	Strict F1
BERTweet-base for sequence tagging	64.8	80	71.6
BERTweet-base + Splinter	81.9	84.3	83.1
BERTweet-large + Splinter	90.5	89.6	90

Table 8. Results of task reformulation on tweet classification

Model	Task type	F1 on development data
BERTweet-base	Textual entailment	84.33
BERTweet-base	Single-sentence classification	83.06

Effects of task reformulation on tweet classification

Table 8 shows the improvement obtained after we transformed the tweet single-sentence classification into a textual entailment task. The results are averaged over six runs with the same six distinct random seeds. The training data for this experiment are the task training dataset plus the external datasets #0, #1 and #2. This combination of training datasets achieves the best results in our experiments. In the experiments, the batch size and learning rate were fixed at 32 and 1e – 5, respectively. The comparison experiments here are limited to the classification task in Procedure I because, in Procedure II, we need the potential drug spans extracted by Splinter to make a prompting sentence for each span, and this would result in inconsistent sample sizes.

From the results, we can see that improvement of the model effect can be obtained by simply adding a prompting sentence to the input sentences. This is a low cost and convenient way to incorporate external knowledge into the model compared to modifying the model structure.

Effects of Cleanlab

Table 9 reflects the improvements in model performance after filtering the training data using Cleanlab. In Procedure II, we used the trained Splinter model to predict the potential spans and used the prediction results to generate classification training data. Among them, if the prediction result is empty or the predicted span is too long (more than 25 characters), we will discard the example. Other processing methods for Cleanlab

Table 9. Results of using Cleanlab on tweet classification

Procedure	Model	Training data	Cleanlab filter	Error labels	F1 on development data
I	BERTweet-base	Train + #0 #1 #2	N		84.33
	BERTweet-base	Train + #0 #1 #2	Y	1178	88.19
II	BERTweet-base	Train + #0	N		86.61
	BERTweet-base	Train + #0	Y	474	87.8

Table 10. Representative examples of different error types

Category	No.	Text	Predicted span	Golden span
FP	1	You want **acrygel**? Hm k 6$ extra @User	Acrygel	
	2	You had my at **biotin** infusion…the pretty gold packaging was just a bonus ◆◆◆ suavebeauty… [URL]	Biotin	
	3	My organic red raspberry leaf tea gets delivered today (AKA **steroids** for my lady parts) time to get this uterus ready to expell human life	Steroids	
	4	This **flu shot** has my arm feeling like someone took a bat to it☺☺	Flu shot	
FN	5	@User make snacking delicious! They are packed with **vitamins**!!!! Hey $1 off here! [URL]… [URL]		Vitamins
	6	@User I am. We are going through 14pts a week & I'm pretty much the only person who has any. And orange **Rennies**. It's just blurgh.		Rennies
	7	@User @User Got my **Tdap** at 37 weeks… Hope it wasn't too late. Article states 27–36 weeks. 😀		Tdap
	8	No first thing in the morning and still have this migraine!!! It's been now 4 days and nothing is working! Including a **blood patch**		Blood patch
SE	9	Ugh. My next **syringe** change is going to be around 4 am. Whyyy. #LifeWithA**Zofran**Pump	Syringe	Zofran
	10	□I hate the sleepy side effect when your eggo is preggo or maybe it these **vitamins**□**never** wear panties ever but I… [URL]	Vitamins□never	Vitamins

were the same as Procedure I. 'Error labels' means the number of instances that Cleanlab identifies as having possible label errors, not that their labels are necessarily wrong.

The experimental results demonstrate the effectiveness of Cleanlab as well as confidence learning. This approach is faster than manually reading and reviewing label issues one by one. We believe that it is well worthwhile to adopt confident learning to enhance effectiveness in various studies and practices in the field of NLP.

Error analysis

Table 10 shows some representative examples of model prediction errors on validation data, where the user names and web links are masked. We divided these errors into three categories for analysis. The first two categories are related to classification model errors, which are false-positive classification (FP) and false-negative classification (FN), and the last category is related to the span extraction model, which we name span extraction error (SE).

We found that textual context is an important influencing factor in classification errors. Both the lack of context and complex rhetoric may lead to model prediction errors. In the No. 1 FP example, 'acrygel' is a nail polish, but this is difficult to be identified from the very short text. The No. 7 FN example is also related to the lack of context. The 'Tdap' is a vaccine and refers to the three diseases Tetanus, Diphtheria and Pertussis. In the absence of other medically relevant cues in the context, the category of Tdap will be difficult to be determined. In addition, we believe that the error in the No. 3 example is rhetorically related. The No. 3 example uses allusion, and here, the 'steroids' is not therapeutic drugs but red raspberry leaf tea. In the No. 6 example, 'Rennie' is a stomach medicine, but the text uses 'Rennies', an uncommon 's' that is, in fact, a spelling error that causes the model to omit it. The few remaining errors are related to the ambiguity of the classification of drugs themselves, which involves whether some drugs or dietary supplements should be labeled as drugs when they are added to another item (No. 2 and No. 5). Also, there are some errors related to the drug definitions. For example, should specific treatments be identified as drug names? In the validation data, 'flu shot' is not annotated, but 'blood patch' is considered a drug name. We found that the annotation of this kind of case is also complex and inconsistent in the training data. It often requires additional knowledge.

On the other hand, SEs are relatively straightforward. They are caused by the drug span being concatenated to other unrelated content in the original text, either in the tag text or by a special symbol. This problem can be partially solved by collecting a list of possible special symbols and preprocessing the text using uppercase breaks.

According to our error analysis, most of the prediction error cases (∼66.67%) of our system are related to the diversity of the tweet text itself. Prediction errors caused by lack of context, metaphorical rhetoric and spelling errors need more diverse domain texts as training data to enhance the system's performance. At the same time, we consider that relying only on PLMs is not enough to solve these problems. In many cases, authors' language preferences and community relations may influence the use of language. It may be possible to collect more data related to the author and social media interactions and combine the social network and graph neural network models to further improve the system performance. Such data may include the author's history of tweets, replies to the tweet, retweeters, etc.

Conclusion

In this paper, we develop a system that can extract medication names from noisy and class-imbalanced tweets. This system contains two modules, classification and span extraction. In each module, PLMs act as the main model structure. During the training process, we enhance the performance of the model through a data-centric approach and task reformulation. Experimental results on the BioCreative VII Task 3 dataset demonstrate that the proposed approach outperforms the existing state-of-the-art systems. This paper also shows that adapting the system procedure to produce more contextualized prompting sentences can effectively improve the system performance.

As future work, we hope to address this problem at the level of large-scale social media content. In this case, it is a very challenging problem to process the large amount of social media content more efficiently. We need to face issues such as model inference speed, preprocessing and filtering. In addition, we are interested in developing this system to a more fine-grained level, such as distinguishing different kinds of drugs, health supplements, vaccines, etc.

Funding

Ministry of Science and Technology, Taiwan (No. MOST 109-2221-E-008-062-MY3).

Conflict of interest

None declared.

References

1. Weissenbacher,D., O'Connor,K., Rawal,S. *et al.* (2021) BioCreative VII–Task 3: automatic extraction of medication names in tweets. In: *BioCreative VII Workshop*.
2. Weissenbacher,D., Rawal,S., Magge,A. *et al.* (2021) Addressing extreme imbalance for detecting medications mentioned in twitter user timelines. In: *International Conference on Artificial Intelligence in Medicine*. Springer.
3. Rajpurkar,P., Zhang,J., Lopyrev,K. *et al.* (2016) Squad: 100,000+ questions for machine comprehension of text. *arXiv Preprint arXiv:1606.05250*. 10.48550/arXiv.1606.05250.
4. Ram,O., Kirstain,Y., Berant,J. *et al.* (2021) Few-shot question answering by pretraining span selection. *arXiv Preprint arXiv:2101.00438*. 10.48550/arXiv.2101.00438.
5. Northcutt,C., Jiang,L. and Chuang,I. (2021) Confident learning: estimating uncertainty in dataset labels. *J. Artif. Intell. Res.*, **70**, 1373–1411. 10.1613/jair.1.12125.
6. Ritter,A., Clark,S. and Etzioni,O. (2011) Named entity recognition in tweets: an experimental study. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.
7. Finin,T., Murnane,W., Karandikar,A. *et al.* (2010) Annotating named entities in Twitter data with crowdsourcing. In: *Proceedings of the NAACL Workshop on Creating Speech and Text Language Data with Amazon's Mechanical Turk*. Los Angeles, US.
8. Cano Basave,A.E., Varga,A., Rowe,M. *et al.* (2013) Making sense of microposts (# msm2013) concept extraction challenge.
9. Strauss,B., Toma,B., Ritter,A. *et al.* (2016) Results of the wnut16 named entity recognition shared task. In: *Proceedings of the 2nd Workshop on Noisy User-Generated Text (WNUT)*. Osaka, Japan.
10. Sarker,A., Nikfarjam,A. and Gonzalez,G. (2016) Social media mining shared task workshop. In: *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific. Hawaii, US.
11. Derczynski,L., Maynard,D., Rizzo,G. *et al.* (2015) Analysis of named entity recognition and linking for tweets. *Inf. Process Manag.*, **51**, 32–49. 10.1016/j.ipm.2014.10.006.
12. Van Erp,M., Rizzo,G. and Troncy,R. (2013) Learning with the web: spotting named entities on the intersection of NERD and machine learning. In # MSM. Citeseer.
13. Sarker,A. and Gonzalez-Hernandez,G. (2017) Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017. *Training*, **1**, 1239. http://ceur-ws.org/Vol-1996/paper8.pdf.
14. Weissenbacher,D., Sarker,A., Paul,M. *et al.* (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*. Brussels, Belgium.
15. Weissenbacher,D., Sarker,A., Magge,A. *et al.* (2019) Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In: *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*. Florence, Italy.
16. Magge,A., Klein,A., Miranda-Escalada,A. *et al.* (2021) Overview of the Sixth Social Media Mining for Health Applications (# SMM4H) shared tasks at NAACL 2021. In: *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*. Mexico City, Mexico.
17. Scao,T.L. and Rush,A.M. (2021) How many data points is a prompt worth? *arXiv Preprint arXiv:2103.08493*. 10.48550/arXiv.2103.08493.
18. Li,X., Feng,J., Meng,Y. *et al.* (2019) A unified MRC framework for named entity recognition. *arXiv Preprint arXiv:1910.11476*. 10.48550/arXiv.1910.11476.
19. Ng,A.Y. (2021) A chat with Andrew on MLOps: from model-centric to data-centric ai.
20. Klein,A., Alimova,I., Flores,I. *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*.
21. Nguyen,D.Q., Vu,T. and Nguyen,A.T. (2020) BERTweet: a pre-trained language model for English tweets. *arXiv Preprint arXiv:2005.10200*. 10.48550/arXiv.2005.10200.
22. He,P., Liu,X., Gao,J. *et al.* (2020) DeBERTa: decoding-enhanced BERT with disentangled attention. *arXiv Preprint arXiv:2006.03654*. 10.48550/arXiv.2006.03654.
23. Lee,J., Yoon,W., Kim,S. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240. 10.1093/bioinformatics/btz682.
24. Raj Kanakarajan,K., Kundumani,B. and Sankarasubbu,M. (2021) BioELECTRA: pretrained biomedical text encoder using discriminators. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*.
25. Devlin,J., Chang,M.-W., Lee,K. *et al.* (2018) BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*. 10.48550/arXiv.1810.04805.
26. Wang,S., Fang,H., Khabsa,M. *et al.* (2021) Entailment as few-shot learner. *arXiv Preprint arXiv:2104.14690*. 10.48550/arXiv.2104.14690.
27. Alvaro,N., Miyao,Y. and Collier,N. (2017) TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill.*, **3**, e6396. 10.2196/publichealth.6396.
28. Tekumalla,R. and Banda,J.M. (2020) A large-scale Twitter dataset for drug safety applications mined from publicly existing resources. *arXiv Preprint arXiv:2003.13900*. 10.48550/arXiv.2003.13900.
29. Tekumalla,R., Asl,J.R. and Banda,J.M. (2020) Mining archive.org's Twitter stream grab for pharmacovigilance research gold. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Atlanta, U.S.