BioCreative VII – Task 3: Automatic Extraction of Medication Names in Tweets

Davy Weissenbacher[1], Karen O'Connor[1], Siddharth Rawal[1], Graciela Gonzalez-Hernandez[1]
1. DBEI, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

***Abstract*—We present the BioCreative VII Task 3 which focuses on drug names extraction from tweets. Recognized to provide unique insights into population health, detecting health related tweets is notoriously challenging for natural language processing tools. Tweets are written about any and all topics, most of them not related to health. Additionally, they are written with little regard for proper grammar, are inherently colloquial, and are almost never proof-read. Given a tweet, task 3 consists of detecting if the tweet has a mention of a drug name and, if so, extracting the span of the drug mention. We made available 182,049 tweets publicly posted by 212 Twitter users with all drugs mentions manually annotated. This corpus exhibits the natural and strongly imbalanced distribution of positive tweets, with only 442 tweets (0.2%) mentioning a drug. This task was an opportunity for participants to evaluate methods robust to class-imbalance beyond the simple lexical match. A total of 65 teams registered, and 16 teams submitted a system run. We summarize the corpus and the tools created for the challenge, which is freely available at *https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/*. We analyze the methods and the results of the competing systems with a focus on learning from class-imbalanced data.**

Keywords—social media; pharmacovigilance; named entity recognition; drug name extraction; class-imbalance.

I. MOTIVATION

Twitter posts are now recognized as an important source of patient-generated data, providing unique insights into population health. A fundamental step towards incorporating Twitter data in pharmacoepidemiological research is to automatically recognize drug mentions in tweets. A common approach is to search for tweets containing lexical matches of drug names occurring in a manually compiled dictionary. This approach has several limitations, even when allowing for variants and misspellings. In our prior study (1), when using the lexical match approach on a corpus where names of drugs are rare, we retrieved only 71% of the tweets that we manually identified as mentioning a drug, and more than 45% of the tweets retrieved were false positives. For example, when tweets mention the word *propel* it denotes predominantly the verb and not the brand name of a corticosteroid. In addition, descriptive text and medication class mentions (such as 'my blood pressure med' or 'my anti-seizure pill'), as well as compounds and 'street' names for medications ('the blue pill') present additional challenges. This competition was an opportunity to go beyond the lexical match approach, providing new methods to improve the extraction of drugs mentioned in posts and enhancing the utility of social media for public health research.

Existing works tackling the problem of detecting drug names on Twitter mainly focused on collecting large corpora suitable to train machine learning systems. However, their method to collect the tweets often biased their collection. In (2,3,4), the authors collected all tweets mentioning a drug from a predefined list of drugs. To reduce the noise in their collection, (5) removed all tweets mentioning common phrases ambiguous with drug names, and (6) imposed that a drug name co-occurs with the name of a disease. These methods missed all tweets mentioning drugs not occurring in their initial lists and discarded ambiguous tweets that are valuable for training machine learning algorithms, since these tweets are negative examples easily mislabeled by automatic systems that show too much confidence in their features representing the drug names.

To reduce bias, we collected our corpora for Task 3 from a separate Twitter corpus (9) that imposes a health-related criterion on the selected users, those self-reporting a pregnancy. The corpus was collected by first identifying users self-reporting a pregnancy, then, for those classified as true pregnancy announcements, collect all their publicly available tweets (their timelines) using the Twitter API. For a study using that corpus (9) we manually annotated all mentions of medications in the timelines. This method ensures that our corpus is representative of the way Twitter users speak about drugs on the platform and exhibits a natural distribution of tweets mentioning drugs from user-centered perspective. A limitation of this method is that this distribution is extremely imbalanced, with only 0.2% of the tweets collected mentioning a drug. Such class imbalance is known to degrade the performance of machine learning systems when modifications are not made to the training process to account for the imbalance, (7,8). Consequently, the class imbalance of our corpus was the main concern of the 16 participants of our competition who proposed concrete solutions to train their systems on our challenging dataset, thus developing systems capable of closely modeling the detection of drugs in tweets as one would do in practice.

II. TASK DESCRIPTION AND CORPORA

Task 3 is a named entity recognition task that involves detecting tweets mentioning drug names (prescriptions and over the counter), or dietary supplements, and extracting the spans of text denoting the drug names. The dataset consists of

212 Twitter users' timelines, collected during our past project described in (9). Using the official Twitter API, we detected 44,825 users publicly announcing a pregnancy via a tweet. We collected all publicly available tweets for those users, both before and after the announcement. We then continued to collect the tweets posted by the users during and after their pregnancies. We manually identified the pregnancy timeframe in 212 users' timelines collected and annotated the spans of drugs mentioned in all tweets posted during that timeframe and one month before and one month after the pregnancies. Our senior annotator (KO) and a staff annotator double annotated 12 timelines and computed an IAA of 0.88 Cohen's Kappa. Our corpus represents the natural and highly imbalanced distribution of drug mentions on Twitter, with 181,607 tweets not mentioning a drug (the negative tweets), and only 442 tweets mentioning at least one drug (the positive tweets), that is, approximately 0.2% of the tweets.

Table 1 shows some examples of tweets annotated in the tabular separated value format. Each tweet is represented by its unique tweet ID, its text, and if a tweet mentions a drug the indexes of the characters at the starting and ending positions of the mention followed by the span itself and the drug name normalized by the annotator. These values are left empty for the tweets not mentioning drugs. In cases where the tweet mentions multiple drug names, such as tweet 42444 in our list of examples, the tweet is duplicated with each occurrence of the tweet indicating the span of one drug name.

TABLE I. EXAMPLES OF TWEETS ANNOTATED WITH DRUG MENTIONS

Tweet ID	text	Begin	End	Span	Drug normalized
39778	Only 3 Arnica Balms left…	8	19	Arnica Balms	arnica balm
40428	@user sudafed that I'm not sure I'm comfortable taking it	7	13	sudafed	sudafed
34396	I like this song!	-	-	-	-
42444	@user no my body hurts, they prescribed me hydros and moltrin	44	49	Hydros	hydrocodone
42444	@user no my body hurts, they prescribed me hydros and moltrin	55	61	Moltrin	motrin

For Task 3, we split our corpus in three sets, a training set (218 positive and 88,770 negative tweets), a validation set (93 positive and 38,044 negative tweets), and test set (131 positive and 54351 negative tweets). We split our corpus by randomly selecting the tweets from all timelines; therefore, the training, validation and test set are likely to have tweets from all timelines.

In addition to the training and validation set, we provided the participants with an extra set of 9,622 tweets annotated with drug names, 4975 positive and 4647 negative tweets. This set, hereafter the SMM4H'18 set, is smaller than the training set but it is more balanced. It was provided to the participants to help them train their machine learning systems with supervision. This dataset was released in 2018 during the #SMM4H shared tasks (10). To collect these tweets, we developed four weak classifiers and used their predictions to select tweets likely to mention a drug in our initial collection of 44,825 users' timelines. We randomly selected 9,622 tweets for manual annotation. The IAA was high with a score of 0.892 Cohen's Kappa, see (1) for details.

We released the training, validation, and extra sets during March 2021. On September 15th, 2021, we released the test set to the participants who had four days to automatically predict the spans of drugs in the test set. Their predictions were submitted to our competition hosted in Codalab (*https://competitions.codalab.org/competitions/23925*) where our evaluation script evaluated their systems. Each team of participants were allowed three runs on the test sets.

III. EVALUATION

A. Metrics

We evaluated the competing systems with the strict and overlapping Precision, Recall, and F1-score for the positive class, which in our task is the annotated spans of drug names. In the strict evaluation, we rewarded a system only if it predicted the exact beginning and end positions of the spans of the drug names annotated. In the overlapping evaluation, we relaxed this constraint and rewarded the system when it predicted a span that overlapped with a span of a drug name annotated. Assuming the system predicted *Arnica* to be a drug name in the tweet 39778 in our list of examples (Table 1), when strictly evaluated, the system is penalized with a False Positive since *Arnica* is not a drug name and a False Negative since it missed *Arnica Balms*. Whereas with the overlapping evaluation the system is rewarded with a True Positive prediction since *Arnica* is a substring of *Arnica Balms*. We ranked the competing systems according to their strict F1-scores since these scores reflect the quality of the outputs of the systems for downstream applications more accurately than the overlapping F1-scores. In cases where two systems achieved equal strict F1-scores, we used their overlapping F1-scores to decide the rank.

B. Baseline System

We released the code, the documentation, and trained models of a baseline labeler to help participants start their development. We chose a standard architecture for our extractor, a BERT embedding layer followed by a bidirectional LSTM layer predicting for each token of a tweet if the token was inside or outside a drug name. We accounted for the class-imbalance of our corpus when training our extractor. We combined three common training heuristics: undersampling, fine-tuning, and filtering. We first pre-trained our extractor on the SMM4H'18 corpus to provide our system with examples of the linguistic patterns for mentioning drug names or phrases ambiguous with drug names. We then fine-tuned our model only on the tweets of the training set filtered in by a classifier. Separating the training of the classifier and

the extractor seems to facilitate the optimizations of their loss functions, one focusing on the semantics of health-related tweets and one focusing on extracting the spans of the drugs. A detailed description and evaluation of our classifier and our extractor can be found in (11).

IV. Systems

A. Results

Sixty-five teams registered to participate in the shared task and sixteen submitted at least one prediction file. We kept the best predictions for each submitting team. Table II presents the performances of each team and summarizes the architectures of the systems, the type of embeddings when available, as well as the strategies applied to train the systems despite the class-imbalance. Four systems achieved better performances than the baseline system, showing technical improvement over the past year.

TABLE II. Task 3 system summaries and Strict/Overlapping F1-scores (F1), precision (P), recall (R)

	Strict			Overlapping			
Team	**F1**	**P**	**R**	**F1**	**P**	**R**	**System Summary**
1	**.804**	.799	.810	**.838**	.832	.844	Classifier + Question Answering; Classifier: ensemble BERTweet-large, data augmentation with MultNLI, TwiMed, SMM4H corpora, 2 million silver-standard tweets
2	**.804**	.799	.810	.824	.819	.830	Ensemble of BERT based muli-task classifiers/extractors; data augmentation & generation with SMM4H'18 and 78,000 silver-standard tweets
3	.764	.805	.728	.793	.835	.755	Ensemble of Megatron-BERT-345M extractors trained with out-of-fold
4	.762	.714	**.816**	.794	.744	**.850**	PubMedBERT-based extractor; data augmentation & generation with SMM4H'18 and 18,800 silver-standard tweets
Baseline	.758	.890	.660	.773	.908	.673	
5	.738	.850	.653	.762	.876	.673	BERT-base + fasttext embeddings + biLSTM + CRF extractor; data augmentation with SMM4H'18
6	.725	.752	.701	.804	.827	.782	Ensemble of BERTweet and Twitter-RoBERTa extractors trained with out-of-fold; oversampling and data generation with SMM4H'18
7	.725	.786	.673	.777	.841	.721	BioRedditBERT extractor with post-filtering using a lexicon; undersampling and data augmentation with SMM4H'18
8	.705	.748	.667	.755	.802	.714	Extracter based on manually curated lexicons
9	.689	.678	.701	.775	.755	.796	DistilBERT extractor trained with bootstrapping; oversampling with SMM4H'18
10	.687	.771	.619	.737	.831	.662	Classifier + Lexicon; Classifier: BERT-large; data augmentation with 200,000 silver-standard tweets from SMM4H'17
11	.683	.629	.748	.739	.680	.810	Collaborative recurrent modules extractor, modules encode various features such as word clinicalBERT embedding, lexicon, POS and morphology
12	.681	.738	.633	.747	.810	.694	Twitter-RoBERTa + FCNs + CRF extractor with a weighted loss function
13	.631	**.910**	.483	.640	**.923**	.490	BERTweet-based extractor; data augmentation with 160,000 positive tweets from SMM4H'18, TwiMed, CADEC, and silver-standard tweets
14	.606	.731	.517	.704	.840	.605	BERT-based extractor; data augmentation with SMM4H'18 and 10,500 silver-standard tweets
15	.585	.727	.490	.659	.812	.554	Classifier + extractor; Classifier: ensemble of BERT-based; Extractor: BERT-based; data augmentation with SMM4H'18 and 326,000 silver-standard tweets from past projects
16	.548	.634	.483	.638	.741	.561	Not Available

B. Analysis

In all the systems but one, the transformer-based networks dominate this competition, although it remains unclear from the results which type of corpora is the best for pretraining the embeddings. The ten best systems chose input embeddings trained on corpora of various genres and domains. Some systems were trained on general domain corpora (ex. Wikipedia and books), others on PubMed abstracts and PMC full-text articles, or on large number of tweets.

The most efficient architecture seems to rely on a filter to remove tweets unlikely to mention drug names and only perform the extraction on the tweets filtered in. Whereas the first ranked system follows the strategy of the baseline system by training a dedicated classifier and applying it upstream from the extractor, the second ranked system proposed a multi-task where the classification and the extraction were performed by the same neural network.

The main challenge of Task 3 was to train machine learning systems on the Twitter timelines which exhibit the natural distribution of tweets mentioning drug names. In past shared

tasks for the classification of tweets mentioning drugs from tweets which do not (not the extraction of their spans), we observed of drop of 6.4 points F1-score between the 0.918 F1-score of best classifier of the SMM4H shared task in 2018 working on a balanced corpus (12) and the 0.854 F1-score of the best classifier of the SMM4H shared task in 2020 working on an imbalanced corpus (13), despite the strategy proposed to address the high degree of class-imbalance (a combination of keyword based pre-filter and an ensemble of classifiers trained with out-of-fold). To address the class-imbalance of our corpus in this task most participants opted for data level preprocessing methods and/or ensemble learning (7); only one system experimented a cost-sensitive learning method with a weighted loss function (24).

Data level preprocessing methods modify the distribution of the examples in the training to improve the training process. This can be done either by removing negative tweets, extending the initial training set with additional positive tweets, or by choosing a hybrid approach. Six systems chose lexicon-based filters or dedicated classifiers to remove negative tweets for this task, as it was relatively easy to detect tweets related to non-medical topics. Given the few positive examples in the training set, the most common approach was to add positive tweets, thus providing examples of the linguistic patterns where drugs are mentioned.

Oversampling, which consists of duplicating positive tweets of the initial training set, was rarely used with only 2 systems opting for this method. Data augmentation was the most popular method with 11 systems out of 16 using it. Besides adding the examples of the SMM4H'18 set we provided, participants looked for existing corpora where drug names were annotated or easy to retrieve automatically. For example, the participants added examples from corpora annotated with adverse drug events or self-report of drug intakes released during past events of the SMM4H shared tasks series. They also proposed various heuristics to create a silver standard corpus. Their two main approaches were to collect a large number of tweets and apply either a lexicon or an extractor trained on a small training corpus to extract the drug names. These additional tweets contained false positive annotations; nonetheless, they were beneficial when the participants added them to the initial training set to (re-)train their extractors.

An alternative to data augmentation was to generate artificial tweets by modifying existing positive tweets. This method was chosen by 3 teams, two of them ranked in the top four positions. The most intuitive way to generate new tweets was to substitute the drug names mentioned in existing tweets with other drug names. The new drug name could be selected from the same drug class or not. Other ways were to concatenate two tweets into one or distorting a tweet by removing randomly words or characters. External tools were also used to paraphrase or to translate the tweets first in German and then use the tweet after translating it back in English.

Table III. TEAM NUMBERS AND SYSTEM DESCRIPTION PAPERS

Team	System description paper
1	Zhang Y. et al. (14)
2	Xu D. et al. (15)
3	Anderson C. et al. (16)
4	Han Q. et al. (17)
5	NA
6	Kulev I. et al. (18)
7	Roller R. et al. (19)
8	Piccolo S. (20)
9	Han P. et al. (21)
10	Tekumalla R. & Banda J. (22)
11	Bagherzadeh P. & Bergler S. (23)
12	Silva J. et al. (24)
13	Zavala R. et al. (25)
14	Lee Y-Q. et al. (26)
15	Hernandez L. et al. (27)
16	NA

V. CONCLUSION

In this paper we presented an overview of the results of the Task 3 of BioCreative VII which focuses on the extraction of drug names in the timelines of 212 Twitter users. Given a tweet posted by a user, the task consists of identifying the spans of text of all drug names mentioned in the tweet. Beside the colloquial style of tweets, our corpus presents an additional challenge to natural language processing systems since it exhibits the natural distribution of tweets with a very low percentage of tweets mentioning drugs. Among the 16 systems proposed for the task, the most popular approaches to improve learning on our imbalanced corpus were assembling different extractors and preprocessing the data to modify the distribution of the training examples. One key to success for the top ranked systems was to filter out tweets unlikely to contain drug names with a dedicated classifier and identify the spans of drugs on the remain tweets with an extractor trained on a dataset extended with both, real and generated, tweets mentioning drugs.

The advance in natural language processing models, thanks to transformers and the clever use of heuristics to rebalance the distribution of the training data improved the performances of extractors when applied on a corpus of tweets with a high class-imbalance. With 0.804 strict F1-score, the performances of the best systems of our challenge are getting very close to the performances achieved by recent named entity recognizers when extracting on Twitter common named entities such as persons, locations, or organizations (28).

REFERENCES

1. 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. Journal of the American Medical Informatics Association, **26**(12), 1618-1626.

2. Carbonell, P., Mayer, M.A. and Bravo, A. (2015) Exploring brand-name drug mentions on twitter for pharmacovigilance. Studies in Health Technology and Informatics **210**, 55-59.

3. Sarker, A. and Gonzalez-Hernandez, G. (2017) A corpus for mining drug-related knowledge from twitter chatter: language models and their utilities. Data Brief **10**, 122-131.

4. Alvaro, N., Miyao, Y. and Collier, N. (2017) TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. *JMIR public health and surveillance*, ***3***(2), e24.

5. Batbaatar, E. and Ryu, K.H. (2019) Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. International Journal of Environmental Research and Public Health **16**(16:3628).

6. Jimeno-Yepes, A., MacKinlay, A., Han, B. and Chen, Q. (2019) Identifying diseases, drugs, and symptoms in twitter. Studies in Health Technology and Informatics **216**, 643-647.

7. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. and Herrera, F. (2018) Learning from Imbalanced Data Sets. Springer.

8. Limsopatham, N. and Collier, N. (2016) Bidirectional LSTM for named entity recognition in twitter messages. Proceedings of the 2nd Workshop on Noisy User-generated Text. **2016**:145-152.

9. Golder, S., Chiuve, S., Weissenbacher, D., Klein, A., O'Connor, K., Bland, M., Malin, M., Bhattacharya, M., Scarazzini, L.J. and Gonzalez-Hernandez, G. (2019) Pharmacoepidemiologic Evaluation of Birth Defects from Health-Related Postings in Social Media During Pregnancy. Drug Safety **42,** 389-400.

10. Weissenbacher, D., Sarker, A., Paul, M.J. and Gonzalez-Hernandez, G. (2018) Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task. **2018**:13-16.

11. Weissenbacher, D., Rawal, S., Magge, A. and Gonzalez-Hernandez, G. (2021) Addressing Extreme Imbalance for Detecting Medications Mentioned in Twitter User Timelines. In: Tucker A., Henriques Abreu P., Cardoso J., Pereira Rodrigues P., Riaño D. (eds) Artificial Intelligence in Medicine. AIME 2021. Lecture Notes in Computer Science, vol 12721. Springer, Cham.

12. Wu, C., Wu, F., Liu, J., Wu, S., Huang, Y. and Xie, X. (2018) Detecting Tweets Mentioning Drug Name and Adverse Drug Reaction with Hierarchical Tweet Representation and Multi-Head Self-Attention. Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. **2018**:34-37.

13. Dang, H. N., Lee, K., Henry, S. and Uzuner, O. (2020) Ensemble BERT for classifying medication-mentioning tweets. In Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. **2020**:37-41.

14. Zhang, Y., Lee, J.K., Han, J-C. and Tsai, R.T-H. (2021) NCU-IISR/AS-GIS: Detecting Medication Names in Imbalanced Twitter Data with Pretrained Extractive QA Model and Data-Centric Approach. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

15. Xu, D., Chen, S. and Miller, T. (2021) BCH-NLP at BioCreative VII Track 3 – medications detection in tweets using transformer networks and multi-task learning. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

16. Anderson, C., Liu, B., Abidin, A., Shin, H-C., Adams, V. (2021) Automatic Extraction of Medication Names in Tweets as Named Entity Recognition. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

17. Han, Q., Tian, S. and Zhang, J. (2021) A PubMedBERT-based Classifier with Data Augmentation Strategy for Detecting Medication Mentions in Tweets. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

18. Kulev, I., Köprü, B., Rodriguez-Esteban, R., Saldana, D., Huang, Y., La Torraca, A. and Ozkirimli, E. (2021) Extraction of Medication Names from Twitter Using Augmentation and an Ensemble of Language Models. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

19. Roller, R., Ayach, A. and Raithel, L. (2021) Boosting Transformers using Background Knowledge, or how to detect Drug Mentions in Social Media using Limited Data. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

20. Piccolo, S.R. (2021) A lexicon-based approach to predicting pregnancy-related medication mentions by Twitter users. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

21. Han, P., Yu, D. and Vydiswaran, V.G.V. (2021) Medication Mention Extraction in Tweets using DistilBERT with Bootstrapping. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

22. Tekumalla, R. and Banda, J.M. (2021) An Enhanced Approach to Identify and Extract Medication Mentions in Tweets via Weak Supervision. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

23. Bagherzadeh, P. and Bergler, S. (2021) Extraction of Medication Names from Tweets CLaC at BioCreative VII Track 3. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

24. Silva, J.F., Almeida, T., Antunes, R., Almeida, J.R. and Matos, S. (2021) Drug Mention Recognition in Twitter Posts Using a Deep Learning Approach. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

25. Zavala, R.R., Martinez, P. and Martinez, J.L. (2021) Creating Domain Specific Embeddings toWork with Imbalanced Datasets in Automatic Extraction of Medication Names in Tweets. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

26. Lee, Y-Q., Wang, C-K., Lee, C-H., Tseng, V.S. and Dai, H-J. (2021) Data Augmentation for BERT in the Medication Extraction Task of BioCreative VII. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

27. Hernandez, L.A.R., Srinivasa, R.C. and Banda J.M. (2021) An ensemble approach for classification and extraction of drug mentions in Tweets. Proceedings of the BioCreative VII Challenge Evaluation Workshop.

28. Suman, C., Reddy, S.M., Saha, S. and Bhattacharyya, P. (2021) Why pay more? A simple and efficient named entity recognition system for tweets. Expert Systems with Applications. **167**(2021) 114101.