Boosting Transformers using Background Knowledge, or how to detect Drug Mentions in Social Media using Limited Data

Roland Roller[1], Ammer Ayach [1], Lisa Raithel[1,2,3]
[1]Speech and Language Technology Lab, DFKI, Berlin, Germany
[2]Technische Universität Berlin, Berlin, Germany
[3]Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, Orsay, France

***Abstract*—To process natural language and to extract information from text, transformers are currently the model of choice for many different tasks. Conversely, if the number of training examples is very limited, fine-tuning might not achieve the expected results, similarly as for other machine learning methods. In the past, a large range of different techniques have been presented to overcome this challenge, such as data augmentation or using distantly labelled data. In this work, we present our contribution to the drug mention detection of the BioCreative VII Challenge (Track 3), which includes a large number of negative, but only a small proportion of positive documents. In course of this, we explore different techniques to boost performance of a pre-trained transformer model. The combination of our transformer model and usage of background knowledge achieved the best results for our use case.**

Keywords—Drug mention detection; Limited data; Hybrid model; Social Media; User Generated Data

I. Introduction

On social media, people share a lot of personal information, and some of that information is directly or indirectly related to their well-being, lifestyle and health situation, such as alcohol consumption, medication intake and possible adverse drug reactions [1]–[4]. The analysis of this information provides unique insights about population health. However, the automatic analysis of social media text by applying e.g. named entity recognition, might not be straightforward, as relevant information can be described in a more casual or descriptive way [5], including typos and uncommon abbreviations. This work targets the extraction of drug and medication mentions of pregnant women in Twitter messages.

To extract the information of interest, transformer models [6] in many different variations became very popular across the community. Although large transformer models are pretrained, it is still necessary to re-train or fine-tune the model for your target task. Similarly as for any other machine learning approach, using more training data is beneficial, particularly if the task is more complex. Conversely, the generation of labelled examples can be time consuming and therefore expensive. In the past, different techniques have been introduced to overcome the problem of limited data. Techniques include for instance data augmentation [7], back-translation [8], or including additional data using distant supervision [9].

In this work, we explore a range of techniques to boost the performance of our transformer model, optimized to detect drug mentions from Twitter data. However, the different approaches applied to increase the size of the data did not result in the expected increase of model performance. On the other hand, a knowledge driven baseline, a string match approach using known facts, quickly showed promising results on the development set when compared to our transformer model. As our knowledge driven approach and transformer complement each other, we combined both methods to get the best of two worlds. In the following, we present the given data and challenge, followed by the different techniques we tested, and our final system which has been used for the challenge.

II. Medication Mentions in Tweets: Task and Data

This work has been carried out in context of Track 3 - "Automatic extraction of medication names in tweets" of the BioCreative VII Challenge. The task addresses the detection of drug/medication mentions in user tweets. The data are based on the UPennHLP Twitter Pregnancy Corpus [10]–[12]. The dataset includes complete Twitter timelines of women who announced their pregnancy, collected between 2014-2017. To generate the data of this challenge, a subset of approximately 400 timelines was collected and annotated. The resulting data include about 89k tweets for training, 39k for validation and 54k for the test set. Interestingly, only a very small fraction of tweets did in fact include a medication mention, namely 218 in the training, 93 in the validation, and a currently unknown number of positives in the test set.

In addition to the dataset of the challenge, the organizers also provided another resource, namely the data of the SMM4H'18 [13] shared tasks. Similarly, this dataset also includes tweets with mentions of drugs and phrases ambiguous with drug names. However, this dataset does not target pregnant women, and it includes a balanced set of positive and negative tweets.

Finally, the authors provided a baseline model [14], which included some additional resource/dictionary files, namely (a) a file including formated RXNorm drug mentions, (b) some

drug variants, and (c) a list of generic drug mentions, such as “drugs for depression” or "anxiety meds".

III. Preprocessing, Methods and Initialization

Within the first step we preprocess all tweets using ekphrasis [14] for normalization (URL, email, time, numbers, dates) and tokenization. Finally, the tweet is converted to conll format.

A. Extending and Modifying the Dataset

As the BioCreative dataset is strongly imbalanced towards negative data, and as the number of positive training examples is relatively small, we intended to change this ratio and increase the number of training examples. A very first strategy regarding this was the downsampling of the negative instances. Here, we tested random downsampling with different strengths (50%, 75%, 90%). At the same time, we tried to increase the number of positives, using the following three strategies:

1) **SMM4H**: Although the annotations of SMM4H partially cover different entities, we enrich the original BioCreative data to increase the number of positives and to change the ratio of positives and negatives.
2) **EDA**: To generate additional examples we apply EDA [16] for data augmentation. To do so, we slightly amended the library and used synonym replacement (sr=0.15), as well as random character deletion (rcd=0.05) and random character swap (rcs=0.15). Using this augmentation technique we created for each tweet five new ones. With the modifications on character level we intended to create realistic tweets including typos. These modifications also affected medication names.
3) **Back-translation & Drug Substitution (BT)**: For every document of the training set, we chose those containing a drug, replaced the drug with a random drug from a drugs list containing 3617 medication names collected from the internet, translated[1] the message to German and afterwards back to English. If the drug is still present in the back-translated version, we add the new document to the training data. Back-translation is a well known technique and was already applied by e.g. [17]–[19] amongst others. Using this technique, we added about 120 new documents to each training fold.

B. BERT

We decided to use a transformer model, specifically a BERT [20] variant, topped with a token classification head as our base model. Among a selection of different BERT models, BioRedditBERT-uncased (BRB) [21] turned out to be the best performing candidate on the given validation set. It is initialized with BioBERT-Base [22] v1.0 + PubMed 200K + PMC 270K and then retrained using 300 million tokens and a vocabulary size of approximately 780,000 words obtained from 68 health-themed subreddits [21].

The Huggingface framework (plus Dataset, Trainer, Metric APIs)[2] was used to implement the fine-tuning pipeline consisting of loading the data into the correct format, applying tokenization followed by labeling, fine-tuning the classification head and evaluation. A dataset consisting of sentences represented as a list of tokens with corresponding labels is used for training the model.

An extra tokenization step is applied on the token level for words not included in the model's vocabulary, resulting in new sub-tokens. If a drug is tokenized, different labeling strategies can be applied for labeling the resulting sub-tokens. The strategy with the best results from the experimentation phase was labeling all resulting sub-tokens as a drug, instead of labeling the first subtoken only and disregarding the rest, as suggested in [20]. The employed labeling strategy is not the usual approach but works well for this challenge. We assume this is due to the following: First of all, drugs are more likely to be tokenized into new sub-tokens as they are usually not included in the model's vocabulary. Second, having more tokens labeled as a drug might improve the issue of the imbalanced quota of positive to negative drug mentions in the dataset. After dealing with the sub-tokens, a truncation is applied to the sentences according to the transformer's input length, resulting in a consolidated dataset length-wise.

The next stage comprises fine-tuning the randomly initialized weights of the classification head to be used for drug detections in tweets. However, the resulting predictions still require a couple of conversion steps, which will be described in the following.

Post-processing:
The post-processing consists of three successive stages. We start with merging all sub-tokens resulting from the transformer tokenizer with their corresponding labels into whole tokens again. The next stage identifies the spans of the predicted drugs following the IOB tagging schema as a first conversion method. Furthermore, a custom version of the IOB schema is used as a second span identification method that merges all beginning drug tokens not separated by white spaces. The second method resulted in increasing the strict F1 score among all experiments. Finally, a remapping of the identified drug spans to the original sentence is done using a string matching function.

C. String Match using Background Knowledge

In order to have a meaningful baseline, we applied string matching (SM) using some known medications. Specifically

[1] we apply a simpler version of the original back-translation [8] and simply access several translation APIs with the help of translatepy (https://github.com/Animenosekai/translate)

[2] https://huggingface.co/

we used the following three sources: 1) The medication annotations of the BioCreative (training) data, 2) the medication annotations of the SMM4H dataset, and 3) the generic mentions provided by the organizers. In the case of BioCreative and SMM4H we extracted the given annotations and mapped them back to unlabelled text. We refer to this data as 'background knowledge'. Originally, we also intended to use RXNorm for the mapping. However, a first test using the formatted RXNorm file achieved a very low precision. As the string match was originally intended to be a baseline, and the usage of RXNorm appeared to require much hand engineering, we did not consider RXNorm for the rest of the work.

D. Preliminary Results before the Final Submission

In order to examine the efficiency of our different approaches we split the training data into a stratified set of five folds to train the models and to find the best parameters and configuration. The preliminary evaluation has been then carried out on the official development set. The results are presented in Table 1.

TABLE I. PRELIMINARY RESULTS OF OUR CROSS-VALIDATION ON STRATIFIED DATASET WITH BERT, BIOBERT (BIOB), BIOREDDITBERT (BRB), AS WELL AS STRING MATCH (SM).

		Overlapping			**Strict**		
Model	**Mode**	**PRE**	**REC**	**F1**	**PRE**	**REC**	**F1**
SM	1	0.461	0.676	0.548	0.442	0.648	0.525
SM	1+2	0.130	0.849	0.226	0.116	0.762	0.201
SM	1+3	0.484	0.736	0.584	0.453	0.695	0.549
BERT		**0.829**	0.671	0.729	0.736	0.606	0.652
BioB		0.791	0.734	0.750	0.732	0.684	0.696
BRB		0.813	0.772	0.788	**0.745**	0.716	0.726
BRB	RD50%	0.800	0.805	**0.798**	0.737	0.754	**0.742**
BRB	SMM4H	0.605	**0.888**	0.719	0.580	**0.855**	0.691
BRB	EDA	0.732	0.796	0.756	0.678	0.749	0.706
BRB	BT	0.824	0.768	0.792	0.730	0.690	0.706

The preliminary results show at first glance that the different transformer models easily outperform the string match baseline (1=only patterns of training; 2=SMM4H patterns; 3=generic mentions file). Also, we can see that the additional SMM4H data leads to an increase of recall, but a very low precision. Regarding the transformer models, the table shows that BioRedditBERT (BRB) outperforms BioBert (BioB), which outperforms BERT. Therefore, all other experiments have been carried out using BioRedditBERT. Regarding downsampling, a 50% random downsampling strategy led to the best performance when compared to the other two downsampling strategies. Next, we examined the extension of the dataset using additional SMM4H data, as well as EDA and BT (back-translation), also in combination with different downsampling strategies. Those modifications never outperformed the standard BRB model. Thus, we decided to focus on the BRB model with the 50% random downsampling for the challenge.

E. Further Analysis

An analysis of the string matching method revealed that some patterns used for the matching resulted in a large number of false positives. Removing (filtering) some false positives, led to much better results (see Table 2), due to a strong increase in precision, and only a minimal drop of recall.

Next, we analysed the predictions of the best transformer model and the (filtered) string match (1+3). The string match is obviously restricted to already "known" entities in the background knowledge, but is able to be very precise. In contrast, the transformer model has the advantage to detect new (unknown) medication mentions by taking the context into account. On the other hand, it also makes clear mistakes if the surrounding context is ambiguous. For instance, in case of "I wasn't sure if my Sensodyne toothpaste was actually working...and then I stopped using it and owe. Okay, it works.", the transformer model labelled "Sensodyne" as medication.

F. System Architecture

As both approaches, the transformer and the string matching, have some advantages, we decided to set up a combined approach. In order to reduce the above mentioned errors, we apply a very simple filtering step to the results of the transformer: We keep only those annotations which also occur within the SMM4H background knowledge set. Note, a more sophisticated filtering with additional sources, as well as a check for partial matches would probably further increase the performance. However, for our purpose this was a quick and easy way to restrict to known medications only.

In addition to that, we slightly update the background knowledge, as some patterns resulted in a larger number of false positives. To do so, we examined within a cross-validation on the training set which patterns have a negative influence on the overall performance and removed them from the final list of terms.

TABLE II. PRELIMINARY CROSS-VALIDATION RESULTS OF OUR COMBINED MODEL ON STRATIFIED DATASET

		Overlapping			**Strict**		
Model	**Mode**	**PRE**	**REC**	**F1**	**PRE**	**REC**	**F1**
BRB	RD50%	0.884	0.778	0.828	0.811	0.733	0.770
BRB	filtering	**0.946**	0.667	0.782	**0.946**	0.667	0.782
String Match	1+2	0.940	0.736	0.825	0.880	0.695	0.777
combined model		0.870	0.813	0.841	0.820	0.781	0.800
combined model	**filtering**	0.933	0.783	**0.851**	0.876	0.743	**0.804**

The results of our approach, including the filtering, are tested on the validation set. The results are presented in Table 2. Note, we mainly focussed and optimized towards overlapping, not the strict matching. First, different to Table 1, string matching is further optimized and achieves a higher precision, with only a small drop in recall. Moreover, the table shows that filtering the results of the transformer strongly boosts the precision, but decreases the recall. Finally, we can see that combining the transformer results with the string match leads to further improvements. Most interestingly, the combination with the filtered transformer model results in an overlapping precision of 0.933 and a F1 of 0.851.

IV. The Challenge

Within the challenge, each participant was allowed to submit up to three submissions. Our first submission (run 1) builds upon the most reliable setup on the validation set before the challenge. This was the BioRedditBERT-uncased transformer model (50% random downsampling), and a string match using the reduced background knowledge, consisting of the BioCreative training data and the generic mentions. The results of the transformer model were then filtered using SMM4H data and then combined using the largest match.

For the second submission (run 2) we trained our BERT model on the complete training **and validation set**, using the parameters of submission one. In this way we hoped to increase performance by increasing the number of training examples. The rest of the setup is equivalent to submission one.

TABLE III. Reported results of our submissions on the BioCreative test data

	Overlapping			Strict		
Run	**PRE**	**REC**	**F1**	**PRE**	**REC**	**F1**
1	**0.841**	**0.721**	**0.777**	**0.786**	**0.673**	**0.725**
2	0.835	0.721	0.774	0.780	0.673	0.723
3	0.829	0.723	0.773	0.767	0.673	0.717

The third submission (run 3) is equivalent to submission one, but we used a reduced background knowledge, consisting of the BioCreative training **and validation data** and the generic mentions.

The results of our models on the BioCreative test data are presented in Table 3. The table shows that the first submission achieved the best results of all submissions. However the performance of all submissions is very similar, mainly due to a slight decrease in precision. This is surprising, as Submission 2 and Submission 3 intended to increase the recall, by including additional information from the validation set - either by using more training data or by extending the background knowledge set.

Moreover, the results show a strong drop in performance (7 points overlap and 8 points strict) in comparison to the results on the validation set. This is surprising as we assumed a very similar data to the validation set (same users, but different tweets of the history). Reasons for this might be (a) overfitting, (b) differences in the test set which caused the drop in performance, or (c) errors from our side preparing the submission. However, a detailed analysis, particularly an analysis of the single classifiers (transformer and string match) would be necessary to identify the reasons for the performance drop.

V. Conclusion

This work presented our contribution to Track 3 of the BioCreative VII Challenge, which targeted the automatic extraction of medication names in tweets. Domain specific language and the reduced context, due to the length of tweets in general, make the task difficult. However, the main challenge of Track 3 is the low number of positive examples, and at the same time, the large number of tweets in general (unbalanced ratio of positives and negatives).

Originally, we intended to tackle the problem of limited data by increasing the size of positive training examples. We started using additional data of SMM4H and augmented data, but this did not lead to significant improvements. Instead, our simple string match using known facts (entities which have been already labelled) appeared to be more reliable compared to our machine learning model. Thus, for this challenge we decided to apply a transformer model, combined with a string match using background knowledge. Generally our model is able to provide a good precision (more on validation than on test data), while it lacks the ability to achieve a very high recall. An improvement of the transformer model, or/and an improvement of the filtering step to increase recall might result in overall better performance.

Considering the large number of overall tweets and considering that a large number of entities can be reliably detected (using background knowledge), we assume that for the given scenario a semi automatic approach might be very efficient: Assuming that we make about 200 predictions in a dataset of 50k tweets, and assuming that we can predict half of those with a high precision, we might need to manually examine only about 100 tweets (0.2%). We believe that this manual effort is reasonable, considering the increasing precision and the number of overall tweets which can be processed within a short time. However, this is something which could be examined in future work.

Acknowledgment

This research was supported by the German Federal Ministry of Education and Research (BMBF) through the project BIFOLD (01IS18025E).

References

1. A. Jimeno-Yepes, A. MacKinlay, B. Han, and Q. Chen, "Identifying Diseases, Drugs, and Symptoms in Twitter," *Stud Health Technol Inform*, vol. 216, pp. 643–647, 2015.
2. A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association*, p. ocu041, Mar. 2015, doi: 10.1093/jamia/ocu041.
3. K. O'Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez, "Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions," *AMIA Annu Symp Proc.*, pp. 924–933, 2014.
4. R. Roller, P. Thomas, and S. Schmeier, "Football and Beer - a Social Media Analysis on Twitter in Context of the FIFA Football World Cup 2018," *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pp. 1–4, Oct. 2018, doi: 10.18653/v1/W18-5901.
5. L. Seiffe, O. Marten, M. Mikhailov, S. Schmeier, S. Möller, and R. Roller, "From Witch's Shot to Music Making Bones - Resources for Medical Laymen to Technical Language and Vice Versa," *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 6185–6192, May 2020.
6. A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. Accessed: Mar. 15, 2019. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
7. J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6381–6387. doi: 10.18653/v1/D19-1670.
8. R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016, pp. 86–96. doi: 10.18653/v1/P16-1009.
9. R. Roller and M. Stevenson, "Making the most of limited training data using distant supervision," *BioNLP 2015*, pp. 12–20, Jul. 2015.
10. S. Golder *et al.*, "Pharmacoepidemiologic Evaluation of Birth Defects from Health-Related Postings in Social Media During Pregnancy," *Drug Saf*, vol. 42, no. 3, pp. 389–400, Mar. 2019, doi: 10.1007/s40264-018-0731-6.
11. A. Sarker, P. Chandrashekar, A. Magge, H. Cai, A. Klein, and G. Gonzalez, "Discovering Cohorts of Pregnant Women From Social Media for Safety Surveillance and Analysis," *Journal of Medical Internet Research*, vol. 19, no. 10, p. e8164, Oct. 2017, doi: 10.2196/jmir.8164.
12. D. Weissenbacher, A. Sarker, A. Klein, K. O'Connor, A. Magge, and G. Gonzalez-Hernandez, "Deep neural networks ensemble for detecting medication mentions in tweets," *J Am Med Inform Assoc*, vol. 26, no. 12, pp. 1618–1626, Dec. 2019, doi: 10.1093/jamia/ocz156.
13. D. Weissenbacher, A. Sarker, M. J. Paul, and G. Gonzalez-Hernandez, "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, 2018, pp. 13–16. doi: 10.18653/v1/W18-5904.
14. D. Weissenbacher, S. Rawal, A. Magge, and G. Gonzalez-Hernandez, "Addressing Extreme Imbalance for Detecting Medications Mentioned in Twitter User Timelines," in *Artificial Intelligence in Medicine*, Cham, 2021, pp. 93–102. doi: 10.1007/978-3-030-77211-6_10.
15. C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, Aug. 2017, pp. 747–754. doi: 10.18653/v1/S17-2126.
16. J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6381–6387. doi: 10.18653/v1/D19-1670.
17. G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised Machine Translation Using Monolingual Corpora Only," presented at the International Conference on Learning Representations, Feb. 2018. Accessed: Oct. 07, 2021. [Online]. Available: https://openreview.net/forum?id=rkYTTf-AZ
18. F. A. Laureano De Leon, H. Tayyar Madabushi, and M. Lee, "UoB at ProfNER 2021: Data Augmentation for Classification Using Machine Translation," in *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, Mexico City, Mexico, 2021, pp. 115–117. doi: 10.18653/v1/2021.smm4h-1.23.
19. Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," *34th Conference on Neural Information Processing Systems (NeurIPS 2020),* p. 13, 2020.
20. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," presented at the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
21. M. Basaldella, F. Liu, E. Shareghi, and N. Collier, "COMETA: A Corpus for Medical Entity Linking in the Social Media," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020, pp. 3122–3137. doi: 10.18653/v1/2020.emnlp-main.253.
22. J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.