Normalization of Long-tail Adverse Drug Reactions in Social Media

Manousogiannis, E.; Mesbah, Sepideh; Bozzon, Alessandro; Sips, Robert-Jan; Szlávik, Zoltán; Baez Santamaria, Selene

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis

Citation (APA)
Manousogiannis, E., Mesbah, S., Bozzon, A., Sips, R.-J., Szlávik, Z., & Baez Santamaria, S. (2020). Normalization of Long-tail Adverse Drug Reactions in Social Media. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis* (pp. 49-58) https://www.aclweb.org/anthology/2020.louhi-1.6.pdf

Normalization of Long-tail Adverse Drug Reactions in Social Media

Emmanouil Manousogiannis[2], Sepideh Mesbah[1], Alessandro Bozzon[1],
Robert-Jan Sips[2], Zoltán Szlávik[2], and Selene Báez Santamaría[2]

[1]Delft University of Technology, Landbergstraat 15, 2628 CE Delft , the Netherlands
[2]MyTomorrows, Anthony Fokkerweg 61, 1059 CP Amsterdam, the Netherlands
manousogm@gmail.com, {s.mesbah, a.bozzon} @tudelft.nl
s.baezsantamaria@vu.nl,{r.sips, zoltan.szlavik} @mytomorrows.com

Abstract

The automatic mapping of Adverse Drug Reaction (ADR) reports from user-generated content to concepts in a controlled medical vocabulary provides valuable insights for monitoring public health. While state-of-the-art deep learning-based sequence classification techniques achieve impressive performance for medical concepts with large amounts of training data, they show their limit with long-tail concepts that have a low number of training samples. The above hinders their adaptability to the changes of layman's terminology and the constant emergence of new informal medical terms. Our objective in this paper is to tackle the problem of normalizing long-tail ADR mentions in user-generated content. In this paper, we exploit the implicit semantics of rare ADRs for which we have few training samples, in order to detect the most similar class for the given ADR. The evaluation results demonstrate that our proposed approach addresses the limitations of the existing techniques when the amount of training data is limited.

1 Introduction

Discovering adverse drug reactions (ADRs) is a critical component of drug safety. In addition to controlled clinical trials, continuous monitoring of adverse effects after market introduction provides valuable insights into ADRs. Studies have shown that traditional techniques (i.e., voluntary and mandatory reporting of ADRs by patients) of post-market ADRs are not able to fully characterize drugs' adverse effects (Harpaz et al., 2012; Chee et al., 2011; Ahmad, 2003; Sarker et al., 2015).

Social media could help to obtain more information on the occurrence of adverse effects in the real world, by monitoring the information discussed and shared by users for their personal experiences with pharmaceutical drugs. Such monitoring can aid in the monitoring of public health (Aramaki et al., 2011; Paul and Dredze, 2011), and provide new opportunities for the identification of adverse drug reactions (Lee et al., 2017b; Sarker and Gonzalez, 2015; Mesbah et al., 2019).

However, web users report ADRs using a different language style and terminology that depends on the user's medical proficiency, but also on the type of online medium (e.g. health forums vs micro-post social networks). Therefore, ADR reports from user generated content typically differ significantly from ADR statements in professional medical text. As exemplified in Table 1, lay people often use diverse dialects (Karisani and Agichtein, 2018) when describing medical concepts, and make abundant use of figures of speech (e.g. metaphors) and informal terminology. Additionally, social media text is usually informal and succinct, often due to limitations imposed by the communication platform, limiting thus the extent and semantic richness of the report (Baron, 2010).

A critical step in the practical use of user-generated text for ADR surveillance starts with the normalization of reported adverse events, meaning linking the user-reported event to a formal Knowledge Base such as the UMLS (Unified Medical Language System)[1] or its subsets like MedDRA (Medical Dictionary for Regulatory Activities)[2] .

There is a large body of work on ADR normalization in social media such as *Twitter* (Chowdhury et al., 2018; Nikfarjam et al., 2015; Manousogiannis et al., 2019), or forums like *Dailystrength* (Leaman et al., 2010; Nikfarjam and Gonzalez, 2011). Existing research on ADR normalization rely on different techniques such as rule-based,

[1] UMLS contains structured information about a large number of different medical entities like Diseases, Symptoms, Drugs etc, as well as the relations that those different entities have with each other (for instance a certain Disease is associated with certain Symptoms)

[2] https://www.meddra.org/

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 49–58
November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

Layman’s terminology	Medical concept in UMLS
’head spinning a little’	Dizziness
’lose 10 lbs’	Body Weight Decreased
’appetite on 10’	Increased Appetite
’terrible headache!!!!’	Headache

Table 1: Examples of layman’s text describing ADRs and their related medical concept in UMLS (Unified Medical Language System)

machine translation, supervised-learning (Aronson and Lang, 2010; Stewart et al., 2012; Combi et al., 2018; Soldaini, 2016; Leaman et al., 2013; Leaman and Lu, 2014) which all showed to have limited performance compared to deep learning- based techniques (Lee et al., 2017a; Limsopatham and Collier, 2016; Tutubalina et al., 2018; Han et al., 2017; Niu et al., 2018).

Original contribution. Driven by previous literature, we note that deep neural networks demonstrate a remarkable performance in many publicly available datasets with user-generated text. However, we show that this set-up does not fully correspond to a real-world setting, where: 1) The full collection of medical concepts in the controlled vocabulary is large and continuously increasing. 2) Many concepts have limited and insufficient training examples available. This is clearly shown in Figure 1, where we can see that more than 41% of the medical concepts (classes) present in SMM4H 2017 Twitter and CADEC dataset have just one training sample. As reported in SIDER (Side Effect Resource)[3], there are more than 5000 MedDRA codes mentioned as ADRs in medical documents, but around only 10% of them (e.g., 300-500) are contained in the publicly available training data for ADR normalization. This is unrealistic in comparison to mining for ADRs ’in the wild’, as the language is evolving in online and offline communication (Kershaw et al., 2016), and there is a constant emergence of new informal medical terms. This means ADR surveillance in the real world requires robustness for rare ADRs rather than performing well only in common classes. In this paper, we introduce a simple but competitive technique to adapt to the changes and emergence of informal medical layman’s terms with no training costs. Our intuition is that for rare ADRs for which we have few training samples, we can exploit its implicit semantics to detect the most similar class for the given ADR. In this case, we reduce the risk of overfitting. Based on this intuition, we present a technique that leverages pre-trained language representation models and revisits a simple Nearest-Neighbor (1-NN) approach to solve the real-world problem of normalizing rare ADR concepts.

We perform an extensive experimental evaluation of our presented approach and compare its effectiveness to the current state of the art on real-world data from social media. Our evaluation aims to provide the necessary insights into the strengths and weaknesses of the proposed approach from both quantitative and qualitative perspectives. Results show that our approach achieves superior performance to state-of-the-art deep learning methods for rare ADR concepts.

2 Related Work

Existing methods for normalizing ADRs in clinical text content fall into five categories: 1) *rule-based approaches* (Aronson and Lang, 2010; Stewart et al., 2012; Combi et al., 2018; Soldaini, 2016), ; 2) *deep learning approaches* (Lee et al., 2017a; Limsopatham and Collier, 2016; Tutubalina et al., 2018; Han et al., 2017; Niu et al., 2018) 3) *machine translation techniques* (Ghiasvand and Kate, 2014; Cossin et al., 2018; Lu et al., 2017); 4) other *supervised approaches* such as DNorm (Leaman et al., 2013; Leaman and Lu, 2014) and 5) *unsupervised techniques* (Tahmasebi et al., 2018).

Rule-based approaches mostly rely on string matching techniques, ignoring the semantics of each text mention, which results in a relatively poor performance (Limsopatham and Collier, 2016). The *unsupervised* techniques, despite the fact that they capture semantic similarity between text with minimal string similarity, barely outperform the rule-based techniques. The use of language in domains like social media or online forums is totally different than the official medical terminology used in Knowledge Bases and hence it is hard to find

[3] http://sideeffects.embl.de/

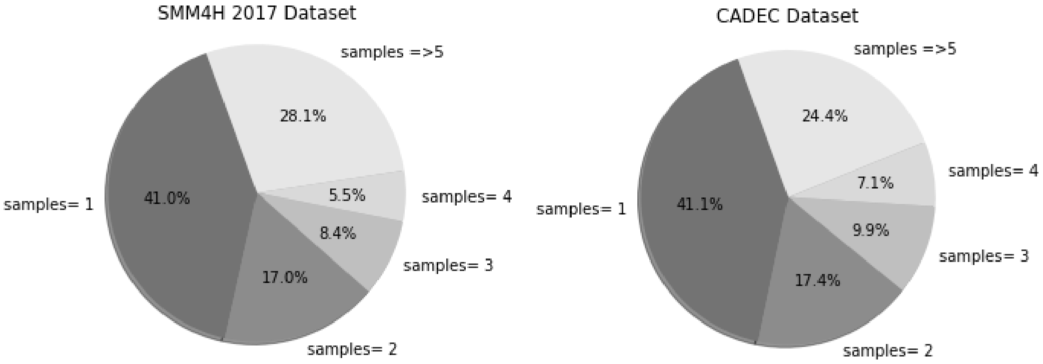

Figure 1: Available training samples per concept in SMM4H 2017 Twitter and CADEC Dataset

a common embedding space that is able to match semantically similar entities. The results published in (Limsopatham and Collier, 2016; Lee et al., 2017a; Tutubalina et al., 2018; Niu et al., 2018) show that supervised machine learning techniques such as *deep learning* and *machine translation* clearly outperform rule-based and unsupervised techniques (i.e., deep learning achieving the highest accuracy). In addition to the task specific deep learning techniques, BERT (Devlin et al., 2018) a bi-directional language representation model was recently fine-tuned in order to tackle the ADR normalization problem as a sequence classification task . This attempt demonstrated remarkable performance across different user-generated text datasets (Miftakhutdinov and Tutubalina, 2019). However, to perform properly, supervised techniques require large collection of labeled training data for each concept and are not suitable for normalizing concepts with none or few training samples. This is clearly shown in Figure 2, were we reproduced the state of the art RNN as presented in (Limsopatham and Collier, 2016) and also fine-tuned BERT, to visualize the performance of those models as a function of the available training samples that each concept (class) has on the SMM4H 2017 Twitter Dataset.

In social media posts some medical concepts like ADRs or symptoms, are way more common than others. For instance, it is common to find many different expressions referring to 'Headache' or 'Stomach Pain' caused by a drug use, rather than 'sleepwalking'. As a result, there is a high class imbalance between common and rare medical concepts that usually show up in social media. In Figure 1, we can clearly see that more than 40 % of the medical concepts (classes) present in this dataset (507) have just one training sample. In this context, it becomes obvious that there is a need for an alternative way of predicting concepts when we have insufficient training data.

Addressing the problem of scarcity in training data is a well-studied research topic. Few-shot learning (Wang and Yao, 2019) is a family of machine learning algorithms that are able to perform classification with only a few ‘shots’ (samples) from each class. To achieve that, most of those techniques leverage less complex models to avoid overfitting on the limited amount of training data. For instance, embedding-based few-shot learning models (Wang and Yao, 2019), try to classify unknown samples by creating a meaningful embedding (feature representation) of the labeled and unlabeled data and then classifying the unlabeled samples based on their most similar embedding from the training data. Recent studies (Wang et al., 2019) show that this family of Nearest Neighbor based approaches can achieve surprisingly promising results in a variety of tasks.

In Natural Language Processing, this family of techniques can profit from various word embedding models (Pennington et al., 2014; Mikolov et al., 2013; Bojanowski et al., 2017) to create a semantically meaningful representation of text. Those embedding models are proven to create vector representation of words that can capture their semantics, in such a way that semantically similar words will have similar vectors, in terms of cosine similarity. On the sentence and phrase level, several techniques can be used to derive a fixed-size vector representation for the whole phrase, regardless of the number of tokens the phrase includes. These techniques can vary from simple approaches like calculating the average of the individual token em-

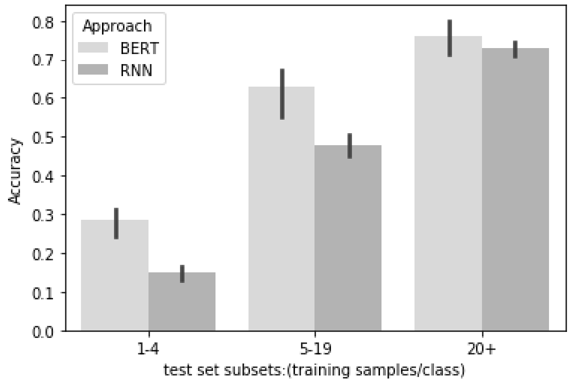

Figure 2: Accuracy of RNN and BERT as a function of the available training samples on SMM4H 2017 dataset.

beddings of a phrase, to more sophisticated techniques that use encoder-based models to derive a meaningful fixed-size vector for each multi-token phrase they receive as input. Recently, researchers in SBERT (Reimers and Gurevych, 2019) modified the BERT language representation model, leveraging Siamese and Triplet network structures, to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. The above was demonstrated as a primary drawback of the original BERT version. The network was fine-tuned with pairs of sentences from Natural Language Inference (NLI) and Semantic Textual Similarity (STS) datasets achieving state of the art performance in minimizing the embedding distance between semantically similar sentences compared to previous sentence embedding techniques.

Despite the aforementioned research in this domain, to our knowledge, none of the previous research in medical concept normalization in user-generated text proposed a solution that would address the problem of class imbalance and scarcity in the training data.

3 ADR Normalization

In the medical domain, getting more annotated data for rare concepts would be extremely expensive and time consuming. Therefore we need to focus our effort on predicting a class (medical concept) for which we have none or maybe few representatives. On a theoretical level, our basic hypothesis is to test whether creating fixed-size vector representations of ADRs from the training and test data is a valuable feature in order to normalize the unlabeled ADR mentions, based only on their vector similarities. In that case, we are expecting that a Nearest Neighbor (1-NN) approach can take advantage of its simplicity and demonstrate a better performance in medical concepts where training data is limited. As we mentioned before, this is a significant percentage of the data in a real world scenario.

Our method is composed of the following stages: 1) we use sentence embedding techniques to encode all ADR phrases in the training and test data into a fixed-size vector so that similar ADRs appear in close proximity; 2) we use cosine distance to measure the semantic similarity between the vector representation of an unlabeled ADR and each representation of the training samples; 3) we classify the unlabeled ADR with the label of its Nearest Neighbor from the training data. There are different techniques to encode words or phrases into a fixed-size vector representation. In this paper, we use two techniques, a simple averaging approach and the state-of-the-art sentence representation model (Reimers and Gurevych, 2019) described below:

- Simple average of word embeddings (avg-wordpiece): We create a fixed-size vector representation of an ADR phrase by using a pre-trained word embedding model, encoding each token in an ADR phrase and then averaging the individual token embeddings. As a pre-trained word embedding model, we use WordPiece embeddings used by BERT (Devlin et al., 2018) in order to establish a fair comparison with the BERT fine-tuned model.

- S-BERT encoder (SBERT): We also used Sentence BERT (Reimers and Gurevych, 2019), which is the state-of-the-art sentence representation model for semantic textual similarity tasks. Its Siamese network architecture is composed of two identical pre-trained BERT models. This model is then trained on pairs of similar and dissimilar sentences with the objective to reduce the distance between the semantically similar pairs and increase the distance between the dissimilar pairs. Further details can be found in the original implementation of this model.

SBERT model training details: To train the model for our task, we needed pairs of similar and dissimilar ADR phrases, which is a time-consuming and expensive process to obtain. For this reason, we used the two pre-trained models provided by the authors. The first model is trained on a combination of the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2017) datasets, consisting of approximately 1 million sentence pairs labeled as a contradiction, entailment, or neutral. The second one is trained on the aforementioned NLI data, and then further tuned on the STSb dataset (Tian et al., 2017), which consists of 8,628 sentence pairs from image captions and news. Those pairs are labeled with a number between 0 and 5, indicating the semantic relatedness of each sentence pair. Besides, we also experimented with medNLI (Romanov and Shivade, 2018) and BIOSSES STS (Soğancıoğlu et al., 2017) datasets, which are manually annotated sentence pairs on the clinical domain. Despite the different combinations of data tried, the best performing version was the provided pre-trained model on SNLI, MNLI, and STSb data, which will be further used in our experiments.

After encoding the phrases from the train and test data in the semantic vector space, the normalization of the test samples is done based on their nearest neighbor from the training data. As a similarity measure, we use cosine similarity, which is the most common similarity measure between word vector representations (Lofi, 2015). We finally assign the concept label of the Nearest Neighbor (i.e., the sample with the highest similarity measure) from the training data to the unlabeled ADR in the test set.

4 Experimental Settings

As an evaluation metric we use accuracy, which denotes the percentage of the correctly normalized ADR samples in the test data. The focus of our evaluation is on the variation of performance using the complete set of training data as well as the different fractions of available training data per class. Those subsets are selected based on how many training samples the true label of a test set sample has in the training data. In this way we can demonstrate the effectiveness of our technique in dealing with rare ADRs. Furthermore, we include the results of qualitative analysis, which will help us identify in more detail the strengths and the limitations of the proposed approach

4.1 Dataset Description

We evaluated our approach on SMM4H 2017 (Sarker et al., 2018) (i.e., the largest available Twitter dataset to our knowledge) and the Cadec dataset (Karimi et al., 2015). **Twitter SMM4H 2017** was published as part of the Social Media Mining for Health workshop in 2017. Unfortunately, the annotation data was only released for the training set and the development set. The annotations for the test set were not released in public. For this reason, we use the annotated part of the data for our experiments. The development set and train set were originally concatenated and then split into **5 equal folds**. However, out of the approximately 9500 mentions only **3629 ADR mentions** were unique. Those mentions were mapped to **507 medical concepts** from the MEDDRA Knowledge-Base. As there was a very high percentage of overlap between the training and test folds, we decided to remove all duplicates in order to avoid over optimistic estimations of our performance. After the duplicate removal, from the validation (development) and test set folds, the final test sets consisted of approximately 400 ADR mentions for testing and 3200 mentions for training in each one of the 5 different folds.

Dataset	#ADR mentions	#MEDDRA Codes
SMM4H 2017	3629	507
CADEC	3092	659

Table 2: Summary of Datasets used in the experimental procedure

Apart from SMM4H 2017 dataset, we leveraged **CADEC** dataset. This data does not only consist of concepts representing Adverse Drug Reactions. It also includes other medical concepts like drugs, diseases and symptoms. For this reason, we filtered all the annotated samples and excluded all data samples that represented different medical concepts, based on their assigned MedDRA code. The resulting dataset we used for evaluation consisted of **3092 samples** mapped to **659 distinct MedDRA codes**. Statistics on data used for training and testing are shown in Table 2.

4.2 Comparison Methods

In order to enable a direct comparison of our 1-NN approach, we reproduced two state of the art neural networks, **BERT** (Miftakhutdinov and Tutubalina, 2019) and the second is a **Recurrent Neural Network (RNN)** (Limsopatham and Collier, 2016) with a single GRU layer. In (Miftakhutdinov and Tutubalina, 2019), the authors fine-tuned BERT for ADR sequence classification, which demonstrated a remarkable performance and outperformed all task specific neural network architectures. In our experiments we are using the BERT-base version

Approach	SMM4H 2017	CADEC
RNN	0.561	0.459
1-NN-avgwordpiece	0.587	0.484
BERT	**0.667**	**0.571**
1-NN-SBERT	0.637	0.531

Table 3: 5 fold cross validation accuracy on SMM4H 2017 and CADEC datasets. Legend: 1-NN-avgwordpiece – using Nearest Neighbor (NN) with the avgwordpiece representation model; 1-NN-SBERT – using NN with SBERT representation model.

Approach	Rare classes Subset* CADEC (115 samples/fold)	Rare classes Subset* SMM4H 2017 (80 samples/fold)
BERT	0.295	0.285
1-NN-SBERT	**0.34**	**0.37**

Table 4: 5 fold cross validation accuracy of the two best performing models on a subset of the test data. The subset includes samples from classes that have less than 5 training samples in the training data.

of this model. The RNN model (Limsopatham and Collier, 2016), demonstrated superior performance compared to all other rule-based or ML based techniques in normalising medical concepts. In addition the authors made their implementation details public to the research community in order to ensure the accurate reproducibility of their approach.

5 Results And Discussion

The results of our experimental evaluation in the two aforementioned datasets are demonstrated in Table 3. Overall, BERT outperforms both the RNN and the 1-NN based models in SMM4H 2017 and CADEC datasets. However, 1-NN-SBERT (i.e., using NN with SBERT representation model) achieves a comparable performance to it, while outperforming the RNN architecture as well as the baseline 1-NN-avgwordpiece. Based on the above results, we could conclude that the BERT sequence classification technique is the big winner in the ADR normalization task, which is also in line with previous research findings (Miftakhutdinov and Tutubalina, 2019). However, as Figure 1 indicates, having insights on different subsets of the predicted medical concepts would give us a better picture of the strengths and the weaknesses of each technique.

For this reason, we divided all test sets into three different subsets, based on how many training samples the true label of test set sample has in the training data. The first test subset includes ADRs with a true label that has just 1 to 4 unique training samples in the corresponding training data. Accordingly, the second subset includes concepts with 5-19 samples in the training set and the third 20 or more of them. The bin limits were selected so that each subset has a significant number of samples.

The results of the accuracy in those subsets are shown in Figures 3 and 4. From these results, we can see that our proposed 1-NN-SBERT technique outperforms BERT and RNN in predicting classes with limited training samples (1-4). This represents the vast majority of the different medical concepts that are present in those two datasets, as seen in Figure 1. The exact accuracy metrics of the two best performing techniques in this subset are presented in Table 4. As expected, we can see that as the availability of training data increases in the other two bins, the sequence classification performance techniques are becoming more effective than the 1-NN based approach.

6 Qualitative analysis

We perform a qualitative analysis of our 1-NN-SBERT approach performance on the SMM4H 2017 dataset. The purpose of the qualitative analysis is to get insights about ’where’ and ’why’ this approach fails or succeeds and to highlight the underlying properties which are hard to digitize without losing any meaning. Table 5 shows examples of correctly normalized ADRs and their corresponding medical terms in the knowledge base, as well as different kinds of misclassified ADR samples. Based on our error analysis, we identified three types of errors where our approach produces the majority of the incorrectly normalized samples:

- 1-NN-SBERT fails to take into account some significant properties of textual data, like negation. The phrase ’never going to lose weight’ is erroneously normalized to Weight Decrease because its closest neighbor in the vector space is ’lose so much weight’.

- There exist inconsistencies or disagreement in the manual annotation of the test data. Nonetheless, our model selected semantically similar concepts in the classification procedure. The phrase ‘kills my sex drive’ is normalized to the medical concept ‘Loss of Li-

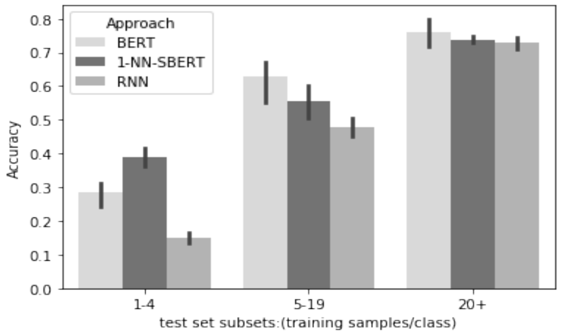

Figure 3: Accuracy of different techniques per available training samples on SMM4H 2017.

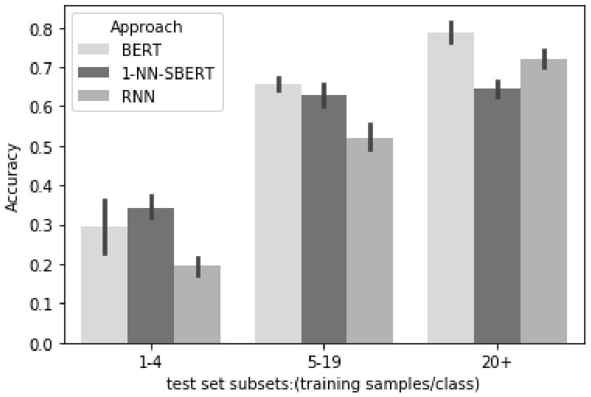

Figure 4: Accuracy of different techniques per available training samples on CADEC test set.

bido' while the ground truth is 'Libido decreased'. Despite having almost the same meaning, these medical concepts represent two different entities in MedDRA. This indicates that we could consider the medical concept normalization problem as a multi-label classification task.

- Our approach does not consider the context around an extracted ADR entity in order to achieve a more robust normalization performance. For instance, the phrase 'feel like I am having a heart attack' was once annotated as a 'Palpitations' adverse effect, while an almost identical phrase was annotated as 'Myocardial infarction'. Most likely, the annotations were based on the information provided by the rest of the text.

While the main focus of the paper was on normalizing rare concepts that have a few training samples available, we also started investigating an additional source of knowledge for the concepts for which no training data is available. For the medical concepts with no training data, we used all different synonymous terms (i.e., available in UMLS) associated with this concept as its representatives.

7 Conclusions and Future Work

In this work, we have presented a 1-NN-SBERT approach that is robust for normalizing rare classes of ADRs. Our technique cannot compete with a deep neural network in concepts where the training data is available on a larger scale. Yet, it easily outperforms deep learning techniques when dealing with rare classes of ADRs. This approach can easily scale as the number of classes increases and can adapt to the changes and emergence of new ADRs, with no training cost. Finally, our sentence embedding model can benefit from the concept of transfer learning, as it does not require task-specific training data to generate high-quality sentence and phrase representations. As future work our model can be improved in the direction of separating the common and the rare medical concepts at test time in a more effective and efficient way. Using multiple binary classifiers, or neural networks with multiple sigmoid functions instead of the final softmax layer have been used in similar domains where multi-label classification or open-set classification (Shu et al., 2017) is considered.

References

Syed Rizwanuddin Ahmad. 2003. Adverse drug event monitoring at the Food and Drug Administration: your report can make a difference. *Journal of General Internal Medicine* 18, 1 (2003), 57–60.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1568–1576.

Alan R. Aronson and François Michel Lang. 2010. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 3 (2010), 229–236. https://doi.org/10.1136/jamia.2009.002733 arXiv:arXiv:1502.05814v1

Naomi S Baron. 2010. *Always on: Language in an online and mobile world.* Oxford University Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors

ADR mention	Predicted Concept	Ground Truth
'Food doesn't look appetizing'	Decreased Appetite	Decreased Appetite
'Slept 10-14 hours '	Hypersomnia	Hypersomnia
'kills my sex drive'	Loss of libido	Libido decreased
'Never going to lose weight'	Weight decreased	Weight increased
'like I am having a heart attack'	Palpitations	Myocardial infarction
'fade into sleep'	Somnolence	Hypersomnia

Table 5: Examples of correctly and incorrectly normalized ADRs from our 1-NN-SBERT model.

with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

Brant W Chee, Richard Berlin, and Bruce Schatz. 2011. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 217.

Shaika Chowdhury, Chenwei Zhang, and Philip S Yu. 2018. Multi-Task Pharmacovigilance Mining from Social Media Posts. In *Proceedings of the 27th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 117–126.

C Combi, M Zorzi, G Pozzani, E Arzenton, and U Moretti. 2018. Normalizing Spontaneous Reports into MedDRA: some Experiments with MagiCoder. *IEEE Journal of Biomedical and Health Informatics* (2018). https://doi.org/10.1109/JBHI.2018.2861213

Sébastien Cossin, Vianney Jouhet, Fleur Mougin, Gayo Diallo, and Frantz Thiessard. 2018. IAM at CLEF eHealth 2018: Concept annotation and coding in French death certificates. *CEUR Workshop Proceedings* 2125 (2018). arXiv:1807.03674

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

Omid Ghiasvand and Rohit J Kate. 2014. UWM : Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns. SemEval (2014), 828–832.

Sifei Han, Tung Tran, Anthony Rios, and Ramakanth Kavuluru. 2017. Team UKNLP: Detecting ADRs, Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter. *Proceedings of the 2nd Workshop on Social Media Mining for Health Research and Applications* (2017), 49–53. http://ceur-ws.org/Vol-1996/paper9.pdf

Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics* 91, 6 (2012), 1010–1021.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics* 55 (2015), 73–81.

Payam Karisani and Eugene Agichtein. 2018. Did You Really Just Have a Heart Attack?: Towards Robust Detection of Personal Health Mentions in Social Media. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 137–146.

Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 553–562.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 22 (2013), 2909–2917. https://doi.org/10.1093/bioinformatics/btt474

Robert Leaman and Zhiyong Lu. 2014. Disease Named Entity Recognition and Normalization with DNorm. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '14)*. ACM, New York, NY, USA, 587. https://doi.org/10.1145/2649387.2660780

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*. Association for Computational Linguistics, 117–125.

Kathy Lee, Sadid A. Hasan, Oladimeji Farri, Alok Choudhary, and Ankit Agrawal. 2017a. Medical

Concept Normalization for Online User-Generated Texts. *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017* (aug 2017), 462–469. https://doi.org/10.1109/ICHI.2017.59

Kathy Lee, Ashequl Qadir, Sadid A Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. 2017b. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 705–714.

N Limsopatham and N Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, Vol. 2. 1014–1023. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85011838424{&}partnerID=40{&}md5=accab69c7d3d0bb0bf95a93ee0415605

Christoph Lofi. 2015. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. *Information and Media Technologies* 10, 3 (2015), 493–501.

C.J. Lu, D. Tormey, L. McCreedy, and A.C. Browne. 2017. *Enhanced lexsynonym acquisition for effective UMLS concept mapping*. Vol. 245. 501–505 pages. https://doi.org/10.3233/978-1-61499-830-3-501

Emmanouil Manousogiannis, Sepideh Mesbah, Alessandro Bozzon, Selene Baez, and Robert Jan Sips. 2019. Give it a shot: Few-shot learning to normalize ADR mentions in Social Media posts. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*. 114–116.

Sepideh Mesbah, Jie Yang, Robert-Jan Sips, Manuel Valle Torre, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. 2019. Training Data Augmentation for Detecting Adverse Drug Reactions in User-Generated Content. 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)*. Association for Computational Linguistics, Hong Kong, China, 2349–2359. https://doi.org/10.18653/v1/D19-1239

Zulfat Miftakhutdinov and Elena Tutubalina. 2019. Deep Neural Models for Medical Concept Normalization in User-Generated Texts. https://doi.org/10.18653/v1/P19-2055

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR* 2013 (01 2013).

Azadeh Nikfarjam and Graciela H Gonzalez. 2011. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 1019.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association* 22, 3 (2015), 671–681.

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2018. Multi-task Character-Level Attentional Networks for Medical Concept Normalization. *Neural Processing Letters* (2018), 1–18. https://doi.org/10.1007/s11063-018-9873-x arXiv:1005.4198

Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. *Icwsm* 20 (2011), 265–272.

Jeffrey Pennington, Richard Socher, and Christoper Manning. 2014. Glove: Global Vectors for Word Representation. *EMNLP* 14, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752* (2018).

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association* 25, 10 (2018), 1274–1283.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of Biomedical Informatics* 54 (2015), 202–212.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics* 53 (2015), 196–207.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: Deep Open Classification of Text Documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2911–2916. https://doi.org/10.18653/v1/D17-1314

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 33, 14 (2017), i49–i58.

Luca Soldaini. 2016. QuickUMLS : a fast , unsupervised approach for medical concept extraction. (2016).

S.A. Stewart, M.E. Von Maltzahn, and S.S.R. Abidi. 2012. Comparing metamap to mgrep as a tool for mapping free text to formal medical lexicons. In *CEUR Workshop Proceedings*, Vol. 895. 63–77.

Amir M Tahmasebi, Henghui Zhu, Gabriel Mankovich, Peter Prinsen, Prescott Klassen, Sam Pilato, Rob Van Ommering, Pritesh Patel, Martin L Gunn, and Paul Chang. 2018. Automatic Normalization of Anatomical Phrases in Radiology Reports Using Unsupervised Learning. (2018).

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. Ecnu at semeval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 191–197.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics* 84, June (2018), 93–102. https://doi.org/10.1016/j.jbi.2018.06.006

Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. 2019. SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning. *arXiv preprint arXiv:1911.04623* (2019).

Yaqing Wang and Quanming Yao. 2019. Few-shot Learning: A Survey.

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