medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

CONORM: Context-Aware Entity Normalization for Adverse Drug Event Detection

Anthony Yazdani, Hossein Rouhizadeh, Alban Bornet, Douglas Teodoro
University of Geneva, Geneva, Switzerland
anthony.yazdani@etu.unige.ch

Abstract

The increasing significance of Adverse Drug Events (ADEs) extracted from social media, such as Twitter data, has led to the development of various end-to-end resolution methodologies. Despite recent advancements, there remains a substantial gap in normalizing ADE entities coming from social media, particularly with informal and diverse expressions of symptoms, which is crucial for accurate ADE identification and reporting. To address this challenge, we introduce a novel end-to-end solution called CONORM: Context-Aware Entity Normalization. CONORM is a two-step pipeline. The first component is a transformer encoder fine-tuned for entity recognition. The second component is a context-aware entity normalization algorithm. This algorithm uses a dynamic context refining mechanism to adjust entity embeddings, aiming to align ADE mentions with their respective concepts in medical terminology. An integral feature of CONORM is its compatibility with vector databases, which enables efficient querying and scalable parallel processing. Upon evaluation with the SMM4H 2023 ADE normalization shared task dataset, CONORM achieved an F1-score of 50.20% overall and 39.40% for out-of-distribution samples. These results improve performance by 18.00% and 19.90% over the median shared task results, 7.60% and 10.20% over the best model in the shared task, and 5.00% and 3.10% over the existing state-of-the-art ADE mining algorithm. CONORM's ability to provide context-aware entity normalization paves the way for enhanced end-to-end ADE resolution methods. Our findings and methodologies shed light on the potential advancements in the broader realm of pharmacovigilance using social media data.

The model architectures are publicly available at https://github.com/anthonyyazdani/CONORM.

Introduction

Adverse Drug Events (ADEs) are defined as unintended complications or injuries resulting from the use of medication, accounting for a substantial portion of adverse events [1,2]. The economic burden they impose is equally staggering, costing billions of dollars in outpatient settings in the U.S. alone, with these costs showing an increasing trend [3,4]. The advent of machine learning and natural language processing technologies has presented unprecedented opportunities for mining valuable insights from large-scale data sources. One application that has received significant attention is the extraction of ADEs from various text corpora, particularly social media platforms like Twitter [5]. These platforms offer a wealth of user-generated data that can be exploited for real-time public health monitoring [6].

The importance of the normalization process lies in its ability to structure a wide variety of textual data into standardized concepts, thus enabling subsequent analysis for pharmacovigilance [7]. However, extracting meaningful health related content from social media is challenging due to its informal and unstructured nature. Factors such as slang, abbreviations, misspellings, and informal language are detrimental to automated information extraction [8].

While there have been traditional avenues for ADE detection such as clinical notes in electronic health records [9], these channels are not exhaustive. ADEs documented in electronic health records are poorly reported to official ADE registries, such as the FDA Adverse Event Reporting System (FAERS) [10,11]. This gap can be addressed by the exploration of alternative sources for capturing a more comprehensive view of ADEs, thus providing impetus for leveraging social media platforms.

In response to existing gaps and challenges, this paper presents CONORM: Context-Aware Entity Normalization, an end-to-end solution for ADE resolution. CONORM utilizes a fine-tuned BERT

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

architecture[12] for ADE entity recognition and leverages the power of Dynamic Context Refining (DCR) for context-aware entity normalization. The method offers a straightforward and easily implementable pipeline that achieves state-of-the-art performance. Additionally, it is designed to be compatible with vector databases, enabling advantages such as efficient querying and scalable parallel processing.

Related work

Machine learning and natural language processing methods are widely used in pharmacovigilance for extracting and normalizing ADEs from unstructured textual data such as social media content[5]. A leading approach in this area is DeepADEMiner[13]. This method has three main steps. First, it uses a classifier based on RoBERTa[14] to filter tweets that likely contain ADEs. Next, a span extractor that incorporates BERT, gated recurrent units[15], and conditional random fields[16] identifies ADE mentions. Finally, DeepADEMiner employs a normalization component, which is trained on an augmented dataset comprising 265,000+ unique instances. This dataset is generated by combining the SMM4H 2017 and 2019 datasets[17,18] with terminology from the Medical Dictionary for Regulatory Activities (MedDRA)[19] and the Unified Medical Language System (UMLS)[20].

Another approach utilizes an hybrid pipeline[21] that combines the DeepADEMiner span extractor model with a specialized model known as `mcn-en-smm4h`[18]. In this merged approach, the output file generated from the ADE extraction phase by DeepADEMiner is employed as the initial input for the ADE normalization task. The `mcn-en-smm4h` model itself is initialized from BioBERT[22] and pre-trained using SMM4H 2017 data, and then further fine-tuned to predict MedDRA terms for each ADE mention.

Another notable approach is our previously developed two-stage pipeline for ADE normalization, as presented in a separate study[23]. This approach ranked first in the SMM4H 2023 shared task 5 and demonstrated comparable performance to DeepADEMiner. A two-stage process was adopted for ADE normalization. The first stage involves fine-tuning BERTweet[24] for named entity recognition to identify ADE mentions in tweets. The second stage normalizes these recognized ADE mentions to corresponding MedDRA terminology using a zero-shot approach. This normalization employs reciprocal-rank fusion[25,26] of similarity rankings from sentence transformers[27]. These sentence transformers are pre-trained on diverse text corpora, and the reciprocal-rank fusion algorithm combines the similarity rankings from multiple sentence transformers to achieve more accurate normalization.

Benchmark dataset

Our study makes use of the SMM4H 2023 shared task 5 dataset[28], which is centered on the normalization of ADEs in English tweets. This dataset acts as a standard benchmark for ADE resolution to their corresponding MedDRA concept IDs. The task is designed to mine social media, specifically Twitter, for the early detection of ADEs and the discovery of novel instances. Diverging from previous iterations of the challenge[5], two primary evaluation metrics are employed. The first assesses the system's performance across all MedDRA IDs in the test set, while the second aims to evaluate the model's ability to generalize to ADEs that are out of the training distribution. The dataset is composed of approximately 18,000 labeled tweets for training and around 10,000 for testing. Each data entry presents the raw tweet, the ADE text span, and its associated MedDRA ID. To maintain the integrity of the evaluation, the test set labels are kept undisclosed, with participants submitting their predictions through CodaLab[29] for assessment.

CONORM-NER: Named entity recognition

Named Entity Recognition (NER) is an essential component of end-to-end entity resolution. In the broader CONORM pipeline, CONORM-NER serves as the sub-model specifically tasked with performing NER. In the context of our study, it plays a key role in identifying ADEs from tweet data, contributing to more accurate normalization, especially when a tweet contains multiple ADE mentions. This section details our approach for

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

NER. The subsections that follow describe the stages of data preprocessing, modeling, training and inference methods utilized in our study.

Data preprocessing

Data preprocessing serves as a pivotal stage in our study, designed to prepare the raw tweet data for effective NER. Initially, the raw tweets underwent a word-based tokenization process, which divided tweets into individual words and punctuations. This process utilized regular expressions to divide the text based on white spaces and punctuation marks. Subsequently, to ensure compatibility with the embedding layer of our chosen model, BERTweet, these pre-tokenized tweets were processed through the BERTweet wordpiece tokenizer[30]. This two-stage tokenization strategy is used to closely align our labeling scheme with the annotations in the dataset. Each tweet in the training set is labeled using the BIOX scheme. In this approach, "B" denotes the beginning of an ADE mention, "I" the inside, "O" the outside, and "X" is used for sub-words, irrespective of whether they are related to an ADE or not. Finally, the processed and tagged dataset was partitioned into training and validation sets, using a ratio of 90% for training and 10% for validation.

Modeling, training & inference

We utilize the pre-trained BERTweet-Large[31] model as the backbone for our token classification framework. Given an input sequence x, CONORM-NER produces a sequence of hidden states. The hidden states can be represented as:

$$h = \text{BERTweet}(x).$$

To reduce the risk of overfitting, we apply a 10% dropout layer to these hidden states, resulting in:

$$h_{drop} = D_{10\%}(h).$$

Sub-words present a unique challenge during the classification process. To address this, we introduce a dummy variable s, which is assigned a value of 1 for sub-words and 0 otherwise. Before forwarding these tokens to the classification layer, this dummy variables are concatenated to their embeddings. This can be formulated as:

$$h_{concat} = \text{concatenate}(h_{drop}, s),$$
$$y = \text{classification}(h_{concat}).$$

Expanding upon conventional training techniques, CONORM-NER integrates a two-phase learning strategy for rapid convergence. In the initial epoch, only task-specific parameters — namely the classification layer weights — are trained, leaving the pre-trained parameters unchanged. This strategy allows CONORM-NER to refine the randomly initialized layers before back-propagating errors to the pre-trained weights, which promotes faster convergence. From the second epoch onward, every parameter becomes accessible for updates.

Input sequences are kept within a maximum limit of 130 tokens. During training, we employ mini-batches of size 24. We utilized the AdamW optimizer[32] for optimization. In the first phase, the learning rate was set at 1×10^{-3}, while in the second phase, it was reduced to 3×10^{-5}. We adopted the cross-entropy loss function throughout. To stabilize training, both phases incorporate a linear learning rate schedule with a 10% warm-up and gradient clipping set at 1.0. We trained CONORM-NER over 20 epochs, selecting the weights that delivered the highest strict F1-score on the validation set, as evaluated by BRATEval[33].

We employ a two-step heuristic during the inference phase to transform token-based predictions into span-based mentions. In the first step, we remove all sub-word tokens predicted as "X". As a result, a sequence such as $[B, X, I, I, O, X, I]$ becomes $[B, I, I, O, I]$. In the subsequent step, we adjust the sequence to ensure continuous ADE predictions. Consequently, the aforementioned sequence is further transformed into $[B, I, I, I, I]$.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

CONORM-EN: Entity normalization

The challenge of extracting ADEs from textual data extends beyond recognizing these entities within the text. It is crucial to ensure that the recognized entities map to standardized terminologies in medical databases, enabling a unified understanding and further analysis. CONORM-EN is our dedicated sub-model for this entity normalization task. This section details our approach for entity normalization applied to identified ADEs in tweet data. The subsections that follow describe the stages of data preprocessing, modeling, training and inference methods utilized in our study.

Data preprocessing

Our approach for entity normalization employs two transformer tokenizers: a context and a isolated entity tokenizer. The two-tokenizer strategy stems from the inherent architecture of CONORM-EN. One sentence transformer is tasked with interpreting the broader context surrounding the identified entity, while the second focuses solely on the entity.

In our method, we introduce special tokens, denoted by $[S]$ for starting and $[E]$ for ending, to mark the edges of target entities within a context. These tokens play a important role in highlighting the target entity, steering the model to focus on one entity when multiple entities are present in the context. These special tokens are integrated into the context tokenizer's vocabulary and are allocated unique embeddings in the embedding matrix of the context sentence transformer.

Consider a tweet stating, "Experiencing relentless headaches after taking XYZ". Using CONORM-NER, an ADE is extracted, identified as "relentless headaches". Transitioning to CONORM-EN, the tweet undergoes a transformation to encapsulate the identified entity with the special tokens: "Experiencing [S] relentless headaches [E] after taking XYZ". The context tokenizer processes the entire transformed tweet, while the isolated entity tokenizer concentrates on the extracted entity in isolation. Specifically:

$$x_{context} = \text{tokenizer}_{\text{context}}(\text{"Experiencing [S] relentless headaches [E] after taking XYZ"}),$$
$$x_{isolated} = \text{tokenizer}_{\text{isolated}}(\text{"relentless headaches"})$$

Given the limited size of our dataset, which contains only 1,710 instances, we divided it into training and validation sets consisting of 1,539 instances (90%) and 171 instances (10%) respectively.

Modeling, training & inference

CONORM-EN constructs a representation for a target entity through a mechanism that we call Dynamic Context Refining (DCR). This process employs a late fusion that integrates contextual and isolated entity embeddings. This fusion is supported by a simplified self-attention mechanism with constant $\mathcal{O}(d)$ time and space complexity, where d is the embedding dimension.

Given a pair of token sequences from a context $x_{context}$, and an isolated entity $x_{isolated}$, we obtain there embeddings by mean pooling hidden states from two sentence transformers $S_{context}$ and $S_{isolated}$:

$$e_{\text{context}} = \frac{\sum S_{\text{context}}(x_{\text{context}})}{|x_{\text{context}}|},$$

$$e_{\text{isolated}} = \frac{\sum S_{\text{isolated}}(x_{\text{isolated}})}{|x_{\text{isolated}}|}.$$

Post this computation, DCR refines e_{isolated} by incorporating insights from e_{context}:

$$e'_{\text{isolated}} = \frac{\exp(a)}{\exp(a) + \exp(b)} \cdot e_{\text{isolated}} + \frac{\exp(b)}{\exp(a) + \exp(b)} \cdot e_{\text{context}}$$

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

where

$$a = \frac{e_{\text{isolated}}^T \cdot M_K \cdot M_Q \cdot e_{\text{isolated}}}{\sqrt{d}}, \quad b = \frac{e_{\text{context}}^T \cdot M_K \cdot M_Q \cdot e_{\text{isolated}}}{\sqrt{d}}.$$

In CONORM-EN, S_{context} and all M matrices are set as trainable parameters. Conversely, S_{isolated} remains frozen with its dropout layers deactivated. This configuration ensures that S_{isolated} operates in inference mode throughout, irrespective of the training status of the model. This setup allows CONORM-EN to produce context-free embeddings using the pre-trained S_{isolated}, even in the absence of surrounding context.

In the training regime of CONORM-EN where context is present, we employ balanced negative sampling to generate dissimilar pairs. Specifically, negative examples are produced by pairing entities with MedDRA terms that aren't explicitly marked as similar. With this approach, there are inherent uncertainties in the composition of negative pairs.

Given this backdrop, we utilize the cosine embedding loss rather than the mean squared error over cosine similarities. The loss function is as follows:

$$L(e'_{\text{isolated}}, m) = \begin{cases} 1 - \cos(e'_{\text{isolated}}, m) & \text{if positive pair,} \\ \max(0, \cos(e'_{\text{isolated}}, m)) & \text{if negative pair.} \end{cases}$$

For positive pairs, the loss pushes e'_{isolated} to have a cosine similarity close to 1 with it's corresponding MedDRA lowest level term m. For the negative pairs, the loss function provides a significant degree of freedom, contributing only for cosine similarities above 0. This is tailored to accommodate the unknown quality of our negative samples, which are created by randomly selecting lowest level terms having a dissimilar preferred term in MedDRA.

To determine the optimal pair of transformer encoders, we conducted a grid search on the validation set across a range of candidates. This included all available sentence transformers[27,34], BioClinicalBERT[35], general domain models[12,14], and BERTweet. Based on our findings, we initialized S_{context} using weights from `multi-qa-mpnet-base-dot-v1`[27,34] and employed `all-mpnet-base-v2`[27,34] for S_{isolated}. During the training phase, mini-batches of size 32 were utilized. The training leveraged the AdamW optimizer together with a one-cycle learning rate policy[36], setting both the learning rate and the upper learning rate boundary at 3×10^{-4}. To prevent exploding gradients, the gradient was clipped at a maximum magnitude of 1.0. We trained CONORM-EN for 100 epochs, choosing the model's weights that yielded the highest micro-averaged F1-score on the validation dataset.

In the inference phase, each ADE mention identified by CONORM-NER was embedded using CONORM-EN, resulting in the embeddings $\{e'_{\text{isolated},j}\}_{j=1}^{J}$, where J is the number of detected ADEs. For every detected ADE, we computed its cosine similarity with the mean-pooled hidden states of all MedDRA lowest-level terms derived from S_{isolated}. Formally, the cosine similarity for the jth ADE against each MedDRA term is:

$$c(e'_{\text{isolated},j}, m_k) = \frac{e'_{\text{isolated},j} \cdot m_k}{\|e'_{\text{isolated},j}\|_2 \cdot \|m_k\|_2},$$

where $\{m_k\}_{k=1}^{K}$ encompasses the embeddings for all K MedDRA lowest-level terms produced by mean-pooling hidden states of S_{isolated}.

Each ADE yields a set of similarity scores:

$$\text{scores}_j = \{\ c(e'_{\text{isolated},j}, m_1),\ c(e'_{\text{isolated},j}, m_2),\ \ldots,\ c(e'_{\text{isolated},j}, m_K)\ \}.$$

Subsequently, for each ADE, the MedDRA lowest-level term with the highest cosine similarity is identified. From this, the corresponding preferred term associated with the chosen lowest-level term is then selected as the final prediction for that specific ADE.

In the inference phase, CONORM-EN capitalizes on the simplicity of the dot product between normalized embeddings. This technique fits seamlessly into vector database architectures, leveraging their efficient

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

features such as locality-sensitive hashing and random projection. This integration offers significant benefits for real-time, large-scale ontology mapping.

Results

We compare CONORM's performance against various systems designed for ADE resolution. Our primary focus lies in comparing CONORM with four distinct entities: the top-performing system from the SMM4H 2023 ADE normalization shared task, the current state-of-the-art model known as DeepADEMiner, the mean, and median performances of all shared task participants. Table 1 and 2 present the overall and out-of-distribution performance metrics, respectively. In terms of overall performance, CONORM demonstrates a superior balance between precision and recall, achieving an F1-score of 50.20%. In out-of-distribution cases, CONORM maintains an F1-score of 39.40%, supported by its precision rate of 53.10%. Given that CONORM-NER employs the same NER algorithm as the top-performing system, it can be inferred that CONORM's enhanced performance arises from CONORM-EN's architecture, its training methodology, or a combination of both.

Table 1. Comparative table of the overall test set performance (F1-score, Precision, and Recall) for ADE resolution. The compared systems are the proposed CONORM, the top-performing system from the SMM4H 2023 ADE normalization shared task[23], DeepADEMiner[13], as well as the median and mean performance of all SMM4H 2023 ADE normalization shared task participants.

Metrics	CONORM	Top system	DeepADEMiner	Median	Mean
F1-score %	**50.20**	42.60	45.20	32.20	32.94
Precision %	**53.50**	44.90	40.70	24.90	29.30
Recall %	47.30	40.50	**50.80**	40.50	42.24

Table 2. Comparative table of the out-of-distribution (OOD) test set performance (F1-score, Precision, and Recall) for ADE resolution. The compared systems are the proposed CONORM, the top-performing system from the SMM4H 2023 ADE normalization shared task[23], DeepADEMiner[13], as well as the median and mean performance of all SMM4H 2023 ADE normalization shared task participants in OOD scenarios.

Metrics	CONORM	Top system	DeepADEMiner	Median	Mean
F1-score %	**39.40**	29.20	36.30	19.50	20.16
Precision %	**53.10**	24.90	33.50	12.80	15.08
Recall %	31.40	35.40	**39.50**	35.40	36.02

Conclusion

ADEs extracted from social media, particularly platforms like Twitter, have grown in significance. Our solution, CONORM: Context-Aware Entity Normalization, addresses the need for enhanced entity normalization techniques. It operates in two stages: the first employs a transformer encoder for entity recognition, while the second introduces a novel context-aware normalization algorithm. This normalization leverages a late fusion technique — dynamic context refining mechanism — mapping ADE mentions to the medical dictionary for regulatory activities terminology.

CONORM achieved an F1-score of 50.20% overall and an out-of-distribution score of 39.40%. When compared with the median performance in the shared task, CONORM showed improvements of 18.00% for overall and 19.90% for out-of-distribution ADEs. Against the top-performing system in the shared task, CONORM exhibited gains of 7.60% and 10.20%. Notably, CONORM surpassed the current state-of-the-art algorithm by 5.00% and 3.10% respectively, despite relying on a significantly smaller dataset (1,710 vs 265,000+ instances).

One aspect that stands out with CONORM is its compatibility with vector databases, which offers efficiency in querying and supports scalable parallel processing. However, there may be limitations that were not

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

addressed in this study. Future work can consider a broader application scope for CONORM, including other languages, more social media platforms, and clinical notes from electronic health records.

CONORM offers a promising approach to context-aware entity normalization for extracting ADEs from social media data. As the importance of ADEs from digital platforms continues to grow, systems like CONORM can play a crucial role in enhancing extraction and normalization methods for pharmacovigilance.

References

1. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. New England journal of medicine. 1991;324(6):377-84.

2. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events: implications for prevention. Jama. 1995;274(1):29-34.

3. Johnson J, Booman L. Drug-related morbidity and mortality. Journal of Managed Care Pharmacy. 1996;2(1):39-47.

4. Ernst FR, Grizzle AJ. Drug-related morbidity and mortality: updating the cost-of-illness model. Journal of the American Pharmaceutical Association (1996). 2001;41(2):192-9.

5. Gonzalez-Hernandez G, Weissenbacher D, editors. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. Gyeongju, Republic of Korea: Association for Computational Linguistics; 2022. Available from: https://aclanthology.org/2022.smm4h-1.0.

6. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. The Milbank Quarterly. 2014;92(1):7-33.

7. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. Drug safety. 2017;40:1075-89.

8. Zhou L, Zhang D, Yang CC, Wang Y. Harnessing social media for health information management. Electronic commerce research and applications. 2018;27:139-51.

9. Yazdani A, Proios D, Rouhizadeh H, Teodoro D. Efficient Joint Learning for Clinical Named Entity Recognition and Relation Extraction Using Fourier Networks:A Use Case in Adverse Drug Events. In: Proceedings of the 19th International Conference on Natural Language Processing (ICON). New Delhi, India: Association for Computational Linguistics; 2022. p. 212-23. Available from: https://aclanthology.org/2022.icon-main.27.

10. Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. Journal of the American Medical Informatics Association. 2010;17(6):671-4.

11. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool'shows that adverse events in hospitals may be ten times greater than previously measured. Health affairs. 2011;30(4):581-9.

12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86. Available from: https://aclanthology.org/N19-1423.

13. Magge A, Tutubalina E, Miftahutdinov Z, Alimova I, Dirkson A, Verberne S, et al. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. Journal of the American Medical Informatics Association. 2021;28(10):2184-92.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

14. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.

15. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1724-34. Available from: https://aclanthology.org/D14-1179.

16. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

17. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, et al. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task; 2019. p. 21-30.

18. Sarker A, Belousov M, Friedrichs J, Hakala K, Kiritchenko S, Mehryary F, et al. 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. 2018;25(10):1274-83.

19. MedDRA Terminology;. Available from: https://www.meddra.org/.

20. Unified Medical Language System;. Available from: https://www.nlm.nih.gov/research/umls.

21. Guellil I, Wu J, Wu H, Sun T, Alex B. Edinburgh_UCL_Health@ SMM4H'22: From Glove to Flair for handling imbalanced healthcare corpora related to Adverse Drug Events, Change in medication and self-reporting vaccination. In: Proceedings of COLING. International Conference on Computational Linguistics. vol. 2022. Europe PMC Funders; 2022. p. 148.

22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

23. Yazdani A, Rouhizadeh H, Alvarez DV, Teodoro D. DS4DH at# SMM4H 2023: Zero-Shot Adverse Drug Events Normalization using Sentence Transformers and Reciprocal-Rank Fusion. arXiv preprint arXiv:230812877. 2023.

24. Nguyen DQ, Vu T, Tuan Nguyen A. BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 9-14. Available from: https://aclanthology.org/2020.emnlp-demos.2.

25. Cormack GV, Clarke CL, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval; 2009. p. 758-9.

26. Teodoro D, Ferdowsi S, Borissov N, Kashani E, Vicente Alvarez D, Copara J, et al. Information retrieval in an infodemic: the case of COVID-19 publications. Journal of medical Internet research. 2021;23(9):e30161.

27. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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: Association for Computational Linguistics; 2019. p. 3982-92. Available from: https://aclanthology.org/D19-1410.

28. Klein A, Banda J, Guo Y, Flores Amaro J, Rodriguez-Esteban R, Sarker A, et al. Overview of the eighth Social Media Mining for Health Applications (SMM4H) Shared Tasks at the AMIA 2023 Annual Symposium. In: Proceedings of the Eighth Social Media Mining for Health Applications (SMM4H) Workshop and Shared Task; 2023. .

medRxiv preprint doi: https://doi.org/10.1101/2023.09.26.23296150; this version posted September 26, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

29. CodaLab;. Available from: https://codalab.org/.

30. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:160908144. 2016.

31. Vinai-Bertweet-large · hugging face;. Available from: https://huggingface.co/vinai/bertweet-large.

32. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. 2017.

33. GitHub - READ-BioMed/brateval;. [Accessed 20-09-2023]. https://github.com/READ-BioMed/brateval.

34. Sentence-Transformers documentation;. Available from: https://www.sbert.net/docs/pretrained_models.html.

35. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. NAACL HLT 2019. 2019:72.

36. Smith LN, Topin N. Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006. SPIE; 2019. p. 369-86.