Boosting Adverse Drug Event Normalization on Social Media: General-Purpose Model Initialization and Biomedical Semantic Text Similarity Benefit Zero-Shot Linking in Informal Contexts

François REMY
University of Ghent
francois.remy
@ugent.be

Simone Scaboro
University of Udine
scaboro.simone
@spes.uniud.it

Beatrice Portelli
University of Udine
portelli.beatrice
@spes.uniud.it

Abstract

Biomedical entity linking, also known as biomedical concept normalization, has recently witnessed the rise to prominence of zero-shot contrastive models. However, the pre-training material used for these models has, until now, largely consisted of specialist biomedical content such as MIMIC-III clinical notes (Johnson et al., 2016) and PubMed papers (Sayers et al., 2021; Gao et al., 2020). While the resulting in-domain models have shown promising results for many biomedical tasks, adverse drug event normalization on social media texts has so far remained challenging for them (Portelli et al., 2022). In this paper, we propose a new approach for adverse drug event normalization on social media relying on general-purpose model initialization via BioLORD (Remy et al., 2022) and a semantic-text-similarity fine-tuning named STS. Our experimental results on several social media datasets demonstrate the effectiveness of our proposed approach, by achieving state-of-the-art performance. Based on its strong performance across all the tested datasets, we believe this work could emerge as a turning point for the task of adverse drug event normalization on social media and has the potential to serve as a benchmark for future research in the field.

1 Introduction

Adverse drug events (ADEs) are unexpected and possibly undocumented negative effects related to the correct use of a drug, and they have the potential to result in serious harm to patients. ADEs can also increase hospitalization costs, reduce patient satisfaction, and erode trust in the health care system. For these reasons, ADEs are a major concern for patients, healthcare providers, and regulators.

However, detecting and reporting emerging ADEs (a process known as pharmacovigilance) is not an easy task (Pappa and Stergioulas, 2019). Most of the information about ADEs is buried in unstructured text sources, such as medical case reports, social media posts, or online reviews (Audeh et al., 2020). The two latter sources often contain informal language, abbreviations, slang, or misspellings, that make machine learning models unable to accurately extract and normalize ADEs present within them, a process known as biomedical concept normalization. Models trained exclusively on clinical data are particularly likely to be affected (see Figure 1). This is a real concern, as mapping ADEs to standardized ontologies, such as MedDRA (Brown et al., 1999) or SNOMED CT (IHTSDO, 2008), is an important step to facilitate the analysis and comparison of ADE data across different sources and domains (Adel et al., 2019).

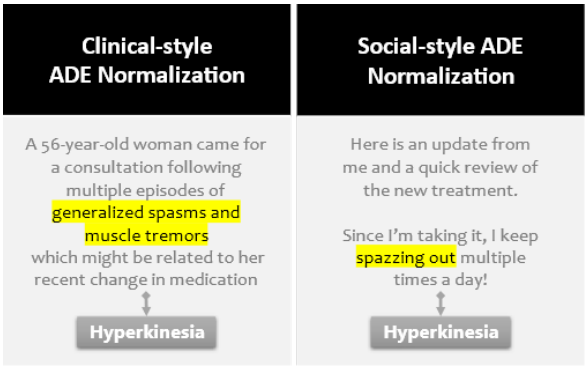

Figure 1: Normalization of concepts in the clinical domain made large progresses, but social media content remains more challenging due to informal language.

In a short time span, between the years 2020 and 2022, the field of biomedical concept normalization has seen significant advancements with the introduction of self-supervised contrastive models. Originally introduced by Chen et al. (2020) for computer vision, these models are trained to produce identical latent representations for multiple views of a same concept, yet contrasted representations for each concept. In the biomedical domain, these views are usually constructed based on biomedical ontologies, by pairing canonical names of a concept with some of their known synonyms.

Proceedings of the 11th International Workshop on Natural Language Processing for Social Media, pages 47–52
November 1, 2023.

State of the art biomedical entity normalization is now dominated by models relying on this technique such as BioSyn (Sung et al., 2020), CODER (Yuan et al., 2022), and SapBERT (Liu et al., 2021). What makes these models extremely versatile is that it is possible to encode a new set of target concepts at inference time, which means that using the same model is possible irrespective of the target ontology, enabling smooth system updates.

2 Our contributions

2.1 General-Purpose Initialization

All models cited thus far were initialized from language models pre-trained on text from the biomedical domain, as this is thought to improve entity normalization performance somewhat in the clinical domain. In this paper, we propose a new approach for adverse drug event normalization on social media by employing BioLORD, a general-purpose model initialization approach pioneered by Remy et al. (2022). We hypothesize that its pre-training on general texts will help tremendously in understanding the informal language used on social media, while previous state of the art models struggled at that specific task, due to the domain shift between clinical and social media languages.

2.2 MedSTS Finetuning

In addition, after noticing that semantic-text-similarity finetuning helps achieving better performance, we improve this new approach even further by incorporating two distinct semantic-text-similarity (STS) fine-tuning phases to the training, both before and after the BioLORD pre-training. We release the improved BioLORD-STS model as part of this paper, and show that it achieves a performance far exceeding the current state-of-the-art.

3 Methodology

In this paper, we set out to show that general-purpose models fine-tuned on biomedical definitions perform better, for the task of ADE normalization in the social media, than state-of-the-art models trained exclusively on biomedical corpora.

Our hypothesis is based on the following observations derived from the extensive ablation studies performed in the BioLORD pre-training paper (Remy et al., 2022): applying the BioLORD pre-training strategy on a general-purpose model can help the model learn to generalize across different writing styles, including non-clinical ones that are rare in biomedical corpora; meanwhile, possessing a strong biomedical knowledge at initialization time did not appear essential to achieve good performance when using the BioLORD pre-training.

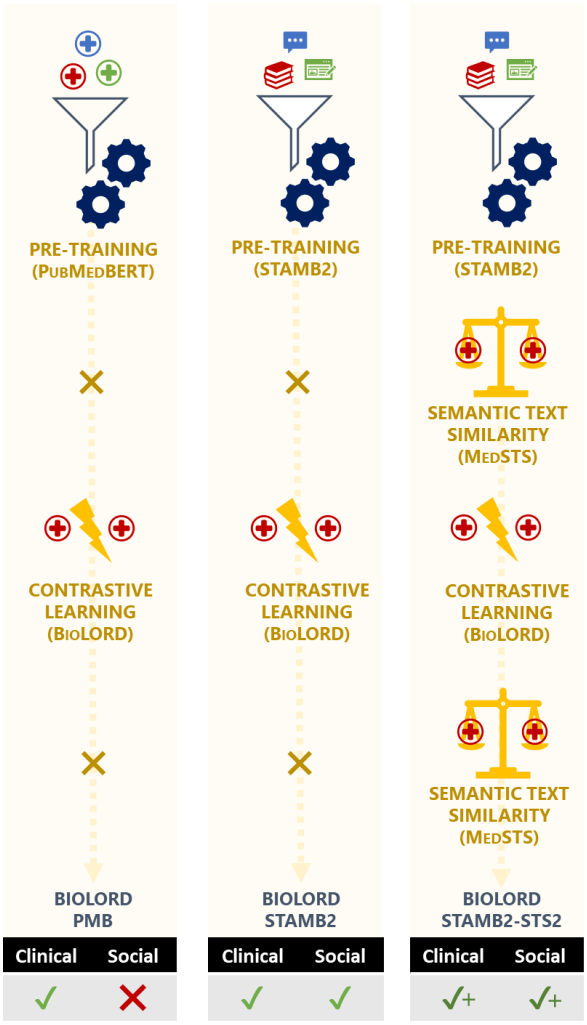

Figure 2: Schema of the pre-training and fine-tuning steps of the four candidate models: BioLORD-PMB, BioLORD-STAMB2, and BioLORD-STAMB2-STS2.

In a final experiment, we also verify that biomedical text similarity is a useful pre-training step to apply before training BioLORD-type models.

We benchmark all models on four social media datasets: CADEC (Karimi et al., 2015), PsyTAR (Zolnoori et al., 2019), SMM4H (Weissenbacher et al., 2019), and TwiMed (Alvaro et al., 2017), which are further described in Section 3.3.

Through this extensive benchmarking, we aim to demonstrate that our proposed approach significantly outperforms the previous state of the art in informal contexts, which cover a wide range of social media activities (such as: forum messages, online reviews, and tweets).

3.1 Candidate Models

To disentangle the impact of the base model initialization from the impact of the BioLORD pre-training strategy, we consider two different base models: STAMB2[1] (Reimers and Gurevych, 2019), the same general-purpose model used in the BioLORD paper, and PubMedBERT (Gu et al., 2020), a robust domain-specific model pre-trained on medical texts. The resulting models are named **BioLORD-STAMB2** and **BioLORD-PMB**.

We also analyze the effect of fine-tuning the models on a medical semantic text similarity task (STS) in addition to the BioLORD pre-training. We do so by fine-tuning some of the models on the MedSTS task (Wang et al., 2020), using the same hyperparameters described in the BioLORD paper. The base STAMB2 model is fine-tuned for STS before applying the BioLORD pre-training. This model then undergoes a second stage of STS fine-tuning, resulting in **BioLORD-STAMB2-STS2**.

Figure 2 illustrates the differences between the proposed models.

3.2 Baseline Models

We choose two BERT-based models trained with contrastive learning strategies as baselines: **CODER** (Yuan et al., 2022) and **SapBERT** (Liu et al., 2021). They are among the best dataset-agnostic models for medical term embeddings at the time of writing. Both of them are trained on the UMLS ontology (Bodenreider, 2004) and were tested on several term normalization datasets, showing promising results. SapBERT was the first large-scale contrastive model to leverage UMLS and is based on PubMedBERT. It is trained by using UMLS synonyms to create contrastive pairs. CODER, on the other hand, leverages both term-term pairs and term-relation-term triples.

3.3 Datasets

We evaluate all the candidate and baseline models using four medical entity normalization datasets containing ADEs. All of them contain informal texts coming from different social media platforms. One of the datasets also contains a subset of formal samples (TwiMed-PM). We include this subset in our experiments to verify that all the tested models perform well on ADE normalization in the clinical domain too.

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

The **TwiMed** dataset (Alvaro et al., 2017) provides a comparable corpus of texts from PubMed (abstracts) and Twitter (posts), allowing researchers in the area of pharmacovigilance to better understand the similarities and differences between the language used to describe disease and drug-related symptoms on PubMed (**TwiMed-PM**, clinical domain) and Twitter (**TwiMed-TW**, social media domain). Both sets of data contain 1000 samples.

The CSIRO Adverse Drug Event Corpus (**CADEC**) dataset (Karimi et al., 2015) is a corpus of user-generated reviews of drugs that has been annotated with adverse drug events (ADEs) and their normalization. It contains 1250 posts from a medical forum, which were annotated by a team of experts from the University of Arizona.

The Psychiatric Treatment Adverse Reactions (**PsyTAR**) dataset (Zolnoori et al., 2019) contains patients' expression of effectiveness and adverse drug events associated with psychiatric medications, originating from a sample of 891 drugs reviews posted by patients on an online healthcare forum.

The Social Media Mining for Health Applications (**SMM4H**) dataset (Gonzalez-Hernandez et al., 2020) is a dataset for Adverse Drug Event (ADE) normalization. It was used in the SMM4H 2020 shared task on ADE normalization. The aim of the subtask was to recognize ADE mentions from tweets and normalize them to their preferred term in the MedDRA ontology. The dataset includes 1212 tweets containing ADEs.

For each evaluated dataset, we perform zero-shot entity normalization using a setup identical to Portelli et al. (2022) with four test splits.

4 Results

We report the zero-shot evaluation results of the various models in Table 1.

Looking at the results on TwiMed-PM, the samples coming from the clinical domain, we observe that almost all of the models have a similar performance (between 69.99 and 70.60), showing that all models (general-purpose or in-domain) can reach a good performance on formal datasets.

The gap in performance of CODER and SapBERT between TwiMed-PM and all the social-media datasets highlights the existence of a significant difference in language distribution between texts in the clinical domain, and the less formal texts found in online reviews or social media.

	TwiMed-PM	CADEC	PsyTAR	SMM4H	TwiMed-TW
CODER	65.31 ± 1.85	35.29 ± 1.27	52.40 ± 0.71	33.14 ± 1.28	42.80 ± 2.06
SAPBERT	70.05 ± 1.56	40.42 ± 1.27	64.82 ± 1.36	43.37 ± 1.07	48.29 ± 2.85
BIOLORD-PMB	69.99 ± 1.87	58.23 ± 0.36	60.22 ± 0.84	41.80 ± 2.24	47.14 ± 2.21
BIOLORD-STAMB2	70.44 ± 1.19	58.69 ± 0.97	64.70 ± 0.76	46.51 ± 2.08	48.46 ± 1.53
BIOLORD-STAMB2-STS2	70.60 ± 1.19	60.28 ± 0.80	65.49 ± 0.74	47.33 ± 1.42	50.57 ± 1.72

Table 1: Accuracy@1 of the evaluated models on all the datasets. Datasets are ordered according to the formality of their language, from more formal (TwiMed-TW) to more informal (SMM4H and TwiMed-TW).

If we focus on the models trained using only the BioLORD pre-training, we can see that they perform better than the two state-of-the-art baseline alternatives across all the datasets. In particular, BioLORD-STAMB2 significantly outperforms SapBERT on CADEC (58.69 vs 40.42) and SMM4H (46.51 vs 43.37). We also observe that BioLORD-STAMB2, the general-domain model, outperforms BioLORD-PMB, the domain-specific variant, proving that the findings of the original BioLORD paper extend to the social media domain.

Our results also highlight that the newly-introduced BioLORD-STAMB2-STS2 manages to move the needle even further (with an average accuracy gain of 1 point with respect to BioLORD-STAMB2), indicating that priming general-purpose models (STAMB2) for biomedical text understanding (STS) before and after the BioLORD pre-training enables to achieve better performance.

On the CADEC dataset in particular, our BioLORD family of models achieves zero-shot accuracy@1 of above 60% for Preferred Term classification. To the best of our knowledge, this is by far the best zero-shot performance ever reported for this dataset.

This seems to confirm that ADE normalization will continue to move towards self-supervised contrastive models, as these models perform well, are very versatile, and can be used to map concepts to any new updated ontology at test time without requiring any retraining. In a field where such models are expected to continue to thrive, the improvements proposed in this paper should be particularly of interest to other researchers.

5 Conclusion

In this paper, we confirmed that BioLORD is an effective pre-training strategy for biomedical entity normalization. We were additionally able to show that applying BioLORD on general-purpose models like STAMB2 provides additional benefits, and that these benefits are more important for sources originating from social media than from clinical notes. Finally, we report that STS-fine-tuning of models both before and after undergoing the BioLORD pre-training can bring additional benefits even in the ADE normalization task, especially in the case when the source originates from social media documents.

Limitations

This paper did not investigate the impact of the proposed pre-training strategies on ADE identification, the task of finding ADE mentions in a text.

We also did not investigate the impact of fine-tuning models on the task, although we have performed some preliminary experiments on this, which seem to confirm the conclusions for zero-shot models apply to fine-tuned models as well.

Ethics Statement

The authors do not foresee that their work would raised any particular ethical concern.

References

Ebtsam Adel, Shaker El-Sappagh, Sherif Barakat, and Mohammed Elmogy. 2019. Chapter 13 - ontology-based electronic health record semantic interoperability: A survey. In Nilanjan Dey, Amira S. Ashour, Simon James Fong, and Surekha Borra, editors, *U-Healthcare Monitoring Systems*, Advances in Ubiquitous Sensing Applications for Healthcare, pages 315–352. Academic Press.

Nestor Alvaro, Yusuke Miyao, and Nigel Collier. 2017. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill.*, 3(2):e24.

Bissan Audeh, Florelle Bellet, Marie-Noëlle Beyens, Agnès Lillo-Le Louët, and Cédric Bousquet. 2020. Use of social media for pharmacovigilance activities: Key findings and recommendations from the Vigi4Med project. *Drug Saf.*, 43(9):835–851.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32:D267–70.

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety*, 20(2):109–117.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Graciela Gonzalez-Hernandez, Ari Z. Klein, Ivan Flores, Davy Weissenbacher, Arjun Magge, Karen O'Connor, Abeed Sarker, Anne-Lyse Minard, Elena Tutubalina, Zulfat Miftahutdinov, and Ilseyar Alimova, editors. 2020. *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, Barcelona, Spain (Online).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

International Health Terminology Standards Development Organisation IHTSDO. 2008. Snomed: Clinical terms. `https://www.snomed.org/`. Accessed: 2023-04-27.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

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

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Dimitra Pappa and Lampros K. Stergioulas. 2019. Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions. *International Journal of Data Science and Analytics*, 8(2):113–135.

Beatrice Portelli, Simone Scaboro, Enrico Santus, Hooman Sedghamiz, Emmanuele Chersoni, and Giuseppe Serra. 2022. Generalizing over long tail concepts for medical term normalization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8580–8591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

François Remy, Kris Demuynck, and Thomas Demeester. 2022. BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, Aron Marchler-Bauer, Melissa Landrum, Stacy Lathrop, Zhiyong Lu, Thomas L Madden, Nuala O'Leary, Lon Phan, Sanjida H Rangwala, Valerie A Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W Trawick, Kim D Pruitt, and Stephen T Sherry. 2021. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 49(D1):D10–D17.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and

Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D. Shah, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, and Hamideh Moayyed. 2019. The psytar dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24:103838.