

Detection of Adverse Drug Events from Social Media Texts – Research Project Overview

Simone Scaboro¹, Beatrice Portelli¹ and Giuseppe Serra¹

¹*Department of Mathematics, Computer Science and Physics, University of Udine, Udine UD 33100, IT*

Abstract

This paper presents an overview of the current research project on the detection of Adverse Drug Events from social media texts led by the Artificial Intelligence Laboratory of Udine (AILAB Udine).

In the latest years, patients started reporting Adverse Drug Events (ADEs) on social media and similar online outlets, making it necessary to monitor them for pharmacovigilance purposes. For this reason, systems for the automatic extraction of ADEs are becoming an important research topic in the Natural Language Processing community.

In this paper we present our research project, focused on the Detection, Extraction and Normalization of ADEs, detailing its objectives, achievements and future directions.

Keywords

Adverse Drug Events, Social Media, Deep Learning

1. Introduction

Each year, dozens of new drugs and medical compounds are developed, tested and released in the market. As an example, in 2021, the Food and Drug Administration (FDA) approved the use of 50 new drugs [1], while the European Medicines Agency (EMA) recommended 92 medicines for marketing authorization [2]. These regulators are in charge of supervising the process that leads to the release of these new medicines, authorizing only the products that are deemed safe. For this reason, new drugs are approved only after rigorous medical trials, which aim to prove their therapeutic efficacy and their safety. In particular, regulatory authorities are interested in documenting possible Adverse Drug Events (ADE), that are unexpected reactions or effects related to the correct use of a medicine.

However, unexpected ADEs might emerge once the new medication is used by a larger population. This is why pharmaceutical and governance agencies have looked into new ways to detect these events in large populations, creating Pharmacovigilance programs.

Traditionally, the process to collect ADEs relies on formal reporting methods, based on the communication between patients, healthcare providers, pharmaceutical companies, and local PV authorities. ADEs can also be extracted (either manually or automatically) from formal medical documents, such as Electronic Health Records (EHR) (see [3] for a recent overview).


HC@AIxIA 2022: 1st AIxIA Workshop on Artificial Intelligence For Healthcare, November 30, 2022, Udine, IT

✉ scaboro.simone@spes.uniud.it (S. Scaboro); portelli.beatrice@spes.uniud.it (B. Portelli); giuseppe.serra@uniud.it (G. Serra)

🌐 <http://conceptbase.sourceforge.net/mjf/> (G. Serra)

🆔 0000-0003-2533-1298 (S. Scaboro); 0000-0001-8887-616X (B. Portelli); 0000-0002-4269-4501 (G. Serra)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Recently, a lot of effort has been put into applying AI models to automatically extract useful information. In particular, user-generated content such as social media posts present a great opportunity: an increasing number of patients discuss their health on forums and micro-blogging platforms, including details on their physical and mental health, as well as feedback on medications and medical procedures. AI systems could be used to analyze this health-related chatter on social media, extract useful information, and perform automatic tasks including ADE detection [4].

However, social media texts introduce a series of new challenges, such as dealing with highly informal speech and the presence of layman terms, typos, and linguistic phenomena like humour, irony, speculations and negations that could affect the meaning of the message. These characteristics make the texts difficult to understand for current natural language understanding techniques, which often rely on models pretrained on large general-domain corpora based on Wikipedia, English literature or other resources. These models have an hard time adapting to the informal language of social media, as they are not used to encounter typos and they also need to learn new contextual meanings for the words used in metaphors and slang. To tackle these problems, a growing branch of the Natural Language Processing (NLP) community has focused on the extraction of ADE mentions from social media texts, encouraging research in this field through conferences and dedicated challenges. One of the most significant examples is the Social Media Mining for Health Applications (SMM4H) Shared Task, which is co-located with top-tier international conferences such as ACL and COLING.

Our international research team led by the Artificial Intelligence Laboratory of Udine (AILAB Udine)¹ is working on an ongoing research project focused on deep learning methods for automatic handling of ADE mentions in social media texts. In this paper we present a summary of our activities and the future directions of our work.

2. Current Research Project

Our research project stemmed from the growing importance of social media in the field of digital pharmacovigilance. The aim of pharmacovigilance systems is to identify Adverse Drug Events, document them and analyze them. Traditionally, this task has been broken down into three sub-tasks (see Figure 1):

1. **ADE Detection:** the binary classification of pieces of text as either containing or not containing a possible ADE. This initial step is especially needed when working with social media posts, due to the large volume of input data.
2. **ADE Extraction:** the extraction of the exact ADE mention(s) from a given texts. During this step, the model is given a full text and is expected to return only the excerpt which is relevant to describe the ADE.
3. **ADE Normalization:** the ADE extracted at the previous step, which is usually written in informal or colloquial language, is mapped to a formal term belonging to a medical ontology. This is the last step of process is necessary to identify precisely what kind of reaction was caused by the drug, and allows medical professionals to perform further analyses on the aggregated data.

¹<http://ailab.uniud.it/>

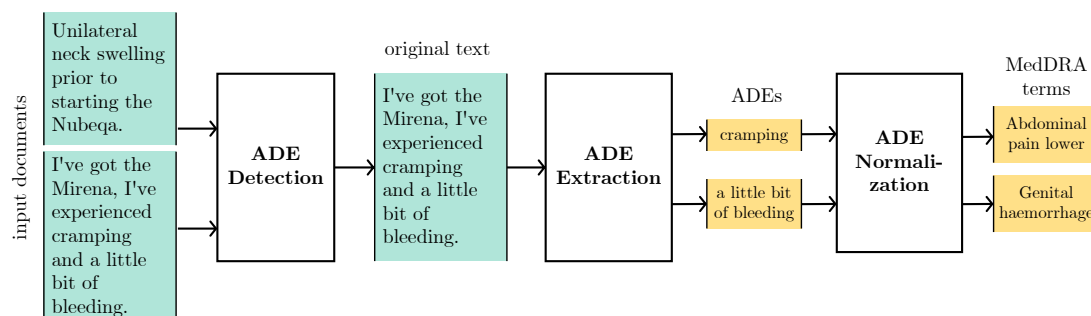


Figure 1: Example of a pipeline applying the three sub-tasks (ADE Detection, Extraction and Normalization) to two input texts.

Among the three tasks, Extraction and Normalization are the ones on which the community has focused more in the latest years, as they present multiple challenges connected to the peculiarities of informal social media language. Additionally, texts mined from different sources, such as internet forums, tweets, and reddit threads, might have different linguistic characteristics, which make it harder to build a unique system that works for all of them.

Our research group briefly tackled the task of ADE Detection, to then focus mainly on the ADE Extraction, its limits and its applications. Finally, we recently tackled the problem of ADE Normalization, with interesting findings on how to improve the generalization capabilities of models on large ontologies.

2.1. Materials and Methods

Most of the resources and datasets used in our research are publicly available. The datasets used in our works are widely used benchmarks for ADE-related tasks on social media texts. The following is a list of the main resources used in the projects:

- CADEC [5]. Public dataset containing 1,250 posts from the health-related forum “AskaPatient”. Each post is annotated for the presence of ADEs, and each ADE is mapped to the corresponding medical term in various medical ontologies (MedDRA and SNOMED CT).
- SMM4H’19 [6], SMM4H’20 [7] and SMM4H’22 [8]. Public datasets for the following shared tasks: SMM4H 2019 - Task 1–3, SMM4H 2020 - Task 3, SMM4H 2022 - Task 1. They contains 2,000 to 23,000 tweets which mention at least one drug name. A portion of the tweet contain one or more ADEs, which are also mapped to a term (Preferred Term or Lowest Level Term) in the MedDRA ontology.

The texts contained in the SMM4H datasets are short, highly informal and rich of typos and/layman terms. Differently, the forum posts contained in CADEC are longer (up to 5 times the length of the tweets), more descriptive and sometimes contain more precise medical terms, as they were posted on an health-related forum. Using datasets with different textual styles allows us to test and develop models in different scenarios, working towards the creation of models that can deal with varying levels of informality within texts.

Most of the systems we developed built upon Transformer-based models which use BERT variants [9]. These large language models proved to be extremely effective in several NLP tasks, including ADE Detection/Extraction from social media texts. Moreover, over the years researchers developed several BERT variants pre-trained on medical texts which hold useful in-domain knowledge for the these tasks (e.g., PubMedBERT [10]).

2.2. ADE Detection

As previously mentioned, our group briefly touched on the topic of ADE Detection, mainly to test the capabilities of current BERT-based models on the task, and to have an in-house system ready to use for future tasks.

We developed a simple BERT-based model the binary classification of texts as containing or not containing an ADE. It consists in a bert-base-uncased model with a binary classification head. If the text contains more than one sentence, it is split in single sentences and each one is evaluated for the presence of ADEs. If at least one of the sentences is labeled as positive, the whole sample is marked as positive for the presence of ADEs.

In Table 1, we compared the performance of this model with the best systems reported in literature on two datasets: SMM4H’19 and CADEC. The model reaches a good performance on SMM4H’19, surpassing in Precision the two best models which took part in the shared task. On CADEC, our system largely outperforms previously reported results in Recall and F1-score. Since most of the texts in CADEC are composed by several sentences, with an average of 5 sentences per sample, we believe that splitting them and evaluating them separately is what contributed the most to the increase in performance.

In following works, we empirically replaced the BERT module with other BERT variants pre-trained on medical texts (e.g., PubMedeBERT) to get small performance boosts in specific tasks and datasets.

Table 1

Comparison of our BERT-based classifier with the best performing systems on SMM4H’19 and CADEC for ADE Detection. Performances are reported in terms of Precision (P), Recall (R) and F1-score (F1), the harmonic mean of P and R.

Dataset	Model	F1	P	R
SMM4H’19	Chen et al. [11]	64.57	60.79	68.85
	Ellendorff et al. [12]	60.48	64.78	56.71
	Our (BERT-based classifier)	61.93	65.67	58.68
CADEC	Alimova and Tutubalina [13]	80.3	84.4	77.3
	Alimova and Tutubalina [14]	81.5	–	–
	Our (BERT-based classifier)	84.3	78.5	90.9

2.3. ADE Extraction

We addressed the task of ADE extraction as a tagging task (or Named Entity Recognition task): given a text, our goal is to identify whether it contains an ADE and its precise span. To this

end, we applied for the first time SpanBERT [15], a BERT variant specialized in multi-token text spans, such as those describing ADE. In our experiments [16], we use a SpanBERT model to generate one embedding for each word in the input text. The embeddings of each word are then fed to a linear layer to output a 3-class BIO label, using the Begin-Inside-Outside tagging scheme employed for NER tasks. The B label marks the beginning of an entity, while the I label is used to mark the following tokens in a multi-token entity. Figure 2 (left) illustrates this architecture. We also experimented the combination of SpanBERT with a Conditional Random Field (CRF) [17] module to further de-noise its output. The CRF is applied to the output of the linear layer, producing another set of BIO labels, as represented in Figure 2 (right).

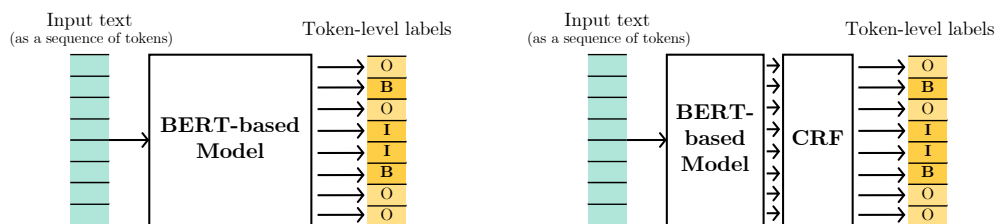


Figure 2: Architectures used for the ADE Extraction task using BERT-based models and Conditional Random Fields (CRF). The input of the model is a text, seen as a sequence of tokens. The output is a series of BIO labels (Begin-Inside-Outside) that mark the presence of ADE entities.

The performance of the system was evaluated on SMM4H’19 and CADEC. The results we of our experiments show that this approach obtains competitive performance compared to other models that took part in the SMM4H’19 task, and also other BERT variants. Moreover SpanBERT+CRF outperforms all the other approaches on both datasets.

The system we developed in currently at the first place on the public leaderboard of the SMM4H’19 ADE Extraction task.²

2.3.1. Application: A Web Platform to Monitor Adverse Reactions to COVID-19 Vaccines on Twitter.

The development of our ADE Extraction system coincided with the start of the COVID-19 vaccination campaigns at the end of 2020. The diffusion of COVID-19 vaccines through massive vaccination campaigns brought a lively debate on social media, with thousands of public discussions about possible ADEs. This was the perfect environment to test the ability of an automatic system for ADE Extraction, which could also provide insightful information to general users. In this context, our international research team developed a web platform³ for monitoring English tweets about COVID-19 vaccines [18].

The objective of the platform is to collect and process tweets, providing visual representations of the information extracted. This is possible thanks to a set of modules (shown in Figure 3) that interact with each other. These modules include the ADE Extraction module described in the previous section. This extracted information is then aggregated and visualized through

²<https://competitions.codalab.org/competitions/20798>, Sub-Task 2, Post-evaluation

³<http://ailab.uniud.it/covid-vaccines/>

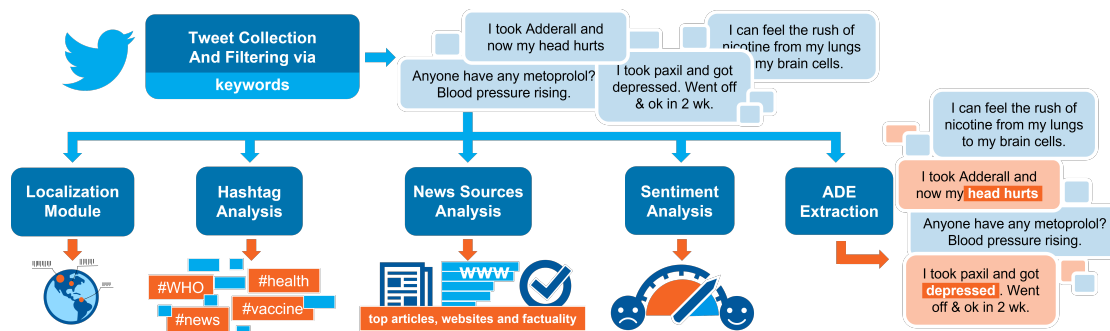


Figure 3: Schema of the modules implemented on the web platform developed to analyze tweets about COVID-19 vaccines.

an interactive word-cloud that can be filtered by date and vaccine name, permitting the user to investigate of trends across time and vaccines. In fact, thanks to this tool, is possible to monitor the tweets that mention a specific drug and obtain an overview of the possible side effects reported by the users, their frequency, and the original texts from which the ADEs were extracted. Despite the promising performance of the model on benchmark datasets, it is important to highlight that the data produced by these current automatic systems still need an expert's supervision to be validated, and should not be used without human revision. However, they are a powerful tool to redirect the experts' attention to potentially interesting phenomena.

As previously mentioned, the portal is also enriched with other modules that give different information on the population of users, such as their geographical location, an important piece of information which could be used to identify patterns in ADE reports. The system also analyzes the URLs that the users share together with information about the vaccines, which is important to take into consideration in the fight against misinformation.

The web portal proved to be useful to perform some analyses on social media chatter surrounding the AstraZeneca vaccines, and it is still actively processing data.

2.3.2. Focus: The Effect of Negations and Speculation on ADE Extraction

Analyzing the incorrect outputs of our ADE Extraction system and other baselines systems we discovered that a great number of False Positives (incorrectly extracted ADEs) were caused by the presence of particular linguistic phenomena, such as irony, negations and speculations. These phenomena are pervasive in social media texts, and they seriously hamper the ability of an automated system to discriminate between factual and nonfactual statements. Therefore, we studied methods to develop more robust models for ADE Extraction in face of negations and speculations [19, 20].

In our research, we took into consideration some systems for ADE detection on social media texts and introduced SNAX, a benchmark dataset to test their performance against negations and speculations on ADE Extraction from social media texts. The dataset is composed by a set of samples (tweets) belonging to four categories: tweets containing ADEs, tweets not containing ADEs, tweets that explicitly negate the presence of an ADE, and tweets that speculate an ADE.

We then introduced two possible strategies to increase the robustness of these models:

- data augmentation: the use of artificially negated and/or speculated samples during the training of the model;
- negation/speculation detection module: the use of a negation and/or speculation module as part of the ADE Extraction pipeline;

All the baseline ADE Extraction models that we analyzed are not robust against negations and speculations. The two strategies that we introduced successfully lower the number of False Positive predictions of the baseline models. However, they have the drawback of lowering the Recall of the models, which is especially noticeable when using the second strategy. Combining both strategies leads to the best results in terms of FP reduction but exacerbates the drop in Recall, so it is not recommended in practical scenarios.

With these works, we proved the lack of robustness of the current state-of-the-art models for the ADE Extraction task, providing a challenging setting to test future models. Furthermore, we introduced two strategies to address the problem, highlighting the improvements they achieved and also their weaknesses. In conclusion, we believe that these kinds of studies, which explore the lack of robustness against different natural language phenomena, should be the backbone of the analysis of models that deal with informal language.

2.4. ADE Normalization

Currently, our research project is focused on ADE Normalization, that is mapping the extracted mentions to large formal medical ontologies (e.g., MedDRA, containing over 24K unique Preferred Terms).

We [21] recently took part in the SMM4H'22 Shared Task for ADE Normalization, using a system based on GPT-2, a text generation model. It ranked third in the task [8], reaching 76% normalization accuracy.

However, one of the main challenges of this task is the high cardinality of the output space, and long tail distribution of the labels in available datasets. This is because current datasets for this task contain at most 5,000 samples, covering only 200-500 of the possible 24K output classes, and the label distribution is highly skewed towards few very frequent terms. For this reason, the models usually perform well on examples that are seen in the training set, but are unable to generalize on rare or unseen ones.

Therefore, we focused our research on ways to overcome the long tail distribution of the training terms and increase the generalization capabilities of ADE Normalization models. We proposed "Ontology Pretraining" (OP) [22], a pre-training strategy based on the hierarchical nature of the MedDRA medical ontology, that improves the generalization capabilities of large language models. This pretraining step injects domain knowledge about all the output labels of the ontology before performing a classical fine-tuning using just the few labels present in the actual training dataset. The OP strategy uses the Lowest Level Terms (LLT) present in MedDRA and teaches the ADE Normalization model to map them to MedDRA Preferred Terms (PT), our actual output labels. LLTs sub-categories of PTs, and are more informal names, which make them stylistically closer to ADEs. Therefore, a model trained with OP will have an advantage when presented with ADEs as input.

We performed several experiments on CADEC, SMM4H'20 and a proprietary dataset, using several ADE Normalization models. The experiments show that using the OP strategy before classical fine-tuning drastically improves the generalization abilities of all models without damaging their performance on seen concepts. The models trained using OP also showed promising results for zero-shot cross-dataset normalization, which is interesting in the perspective of developing a model that can work well across different text typologies.

3. Conclusions and Future Directions

Our research group tackled the tasks of ADE Detection, Extraction and Normalization on social media texts. The results consisted in the development of effective deep learning models for the tasks, some of which were also applied to create an online web platform for ADE monitoring.

The activities also highlighted several limitations of current models, which still require further exploration. For example, the performance of ADE Detection models is still scarce on highly informal texts, such as tweets. However, these are kind of texts for which the community needs a robust ADE Detection model, as they could become the input of real-time digital pharmacovigilance systems. Moreover, ADE Extraction models still lack the ability to deal with some pervasive linguistic phenomena (e.g., irony and humor), and the proposed solutions of negation/speculation detection could be improved to reduce drops in recall. Right now, ADE Extraction models cannot be blindly applied to real-time streams of social media posts, as they lack the ability to distinguish personal ADE reports from news pieces and second-hand reports (i.e., recounting what happened to another person). Even more importantly, they are not able to distinguish real ADE reports from maliciously-constructed posts that spread misinformation. Therefore, there is a need to look into ways to carefully integrate them in online pharmacovigilance activities, with the aid of other systems and expert human supervision.

As regards the future directions of this project, one of the top priorities is extending the current systems to other languages. SMM4H recently introduced small datasets in Spanish, French and Russian, but there is still a lack in resources and models for languages other than English. Medical ontologies, on the other hand, come with several official translations, so it would be interesting to explore the idea of leveraging them to solve other tasks. We also plan to focus more on term normalization, incorporating different ontologies (e.g., SNOMED CT and UMLS). This will help in two ways: by adding new complementary knowledge to train the AI systems, and by creating more bridges between the outputs of the NLP systems and the formal medical world. Finally, it would be interesting to move towards to extraction and normalization of other medical entities in informal texts, such as drugs and diseases.

References

- [1] B. G. de la Torre, F. Albericio, The pharmaceutical industry in 2021. an analysis of fda drug approvals from the perspective of molecules, *Molecules* 27 (2022). doi:10.3390/molecules27031075.

- [2] European Medicines Agency, Human medicines highlights 2021, https://www.ema.europa.eu/documents/report/human-medicines-highlights-2021_en.pdf, 2022. Accessed: 2022-10-07.
- [3] C. Feng, D. Le, A. B. McCoy, Using Electronic Health Records to Identify Adverse Drug Events in Ambulatory Care: A Systematic Review, *Applied Clinical Informatics* 10 (2019) 123–128.
- [4] H. Yang, C. C. Yang, Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis, *ACM Trans. Intell. Syst. Technol.* 6 (2015). URL: <https://doi.org/10.1145/2700482>. doi:10.1145/2700482.
- [5] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, Cadec: A Corpus of Adverse Drug Event Annotations, *J. Biomed. Inform.* 55 (2015) 73–81.
- [6] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O’Connor, M. Paul, G. Gonzalez, Overview of the Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019, in: *Proceedings of the ACL Workshop on Social Media Mining for Health Applications*, 2019.
- [7] G. Gonzalez-Hernandez, A. Z. Klein, I. Flores, D. Weissenbacher, A. Magge, K. O’Connor, A. Sarker, A.-L. Minard, E. Tutubalina, Z. Miftahutdinov, I. Alimova, *Proceedings of the COLING Social Media Mining for Health Applications Workshop & Shared Task*, 2020. URL: <https://aclanthology.org/2020.smm4h-1.0>.
- [8] D. Weissenbacher, J. Banda, V. Davydova, D. Estrada Zavala, L. Gasco Sánchez, Y. Ge, Y. Guo, A. Klein, M. Krallinger, M. Leddin, A. Magge, R. Rodriguez-Esteban, A. Sarker, L. Schmidt, E. Tutubalina, G. Gonzalez-Hernandez, Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022, in: *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 221–241. URL: <https://aclanthology.org/2022.smm4h-1.54>.
- [9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of NAACL*, 2019.
- [10] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, *arXiv preprint arXiv:2007.15779* (2020).
- [11] S. Chen, Y. Huang, X. Huang, H. Qin, J. Yan, B. Tang, HITSZ-ICRC: A report for SMM4H shared task 2019-automatic classification and extraction of adverse effect mentions in tweets, in: *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 47–51. doi:10.18653/v1/W19-3206.
- [12] T. Ellendorff, L. Furrer, N. Colic, N. Aepli, F. Rinaldi, Approaching SMM4H with merged models and multi-task learning, in: *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 58–61. doi:10.18653/v1/W19-3208.
- [13] I. Alimova, E. Tutubalina, Automated detection of adverse drug reactions from social media posts with machine learning, in: W. M. van der Aalst, D. I. Ignatov, M. Khachay, S. O. Kuznetsov, V. Lempitsky, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, A. V. Savchenko, S. Wasserman (Eds.), *Analysis of Images, Social Networks and Texts*, Springer International Publishing, Cham, 2018, pp. 3–15.

- [14] I. Alimova, E. Tutubalina, Detecting adverse drug reactions from biomedical texts with neural networks, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 415–421. doi:10.18653/v1/P19-2058.
- [15] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving Pre-training by Representing and Predicting Spans, Transactions of the Association for Computational Linguistics 8 (2020) 64–77.
- [16] B. Portelli, D. Passabì, E. Lenzi, G. Serra, E. Santus, E. Chersoni, Improving Adverse Drug Event Extraction with SpanBERT on Different Text Typologies, Springer International Publishing, Cham, 2022, pp. 87–99. doi:10.1007/978-3-030-93080-6_8.
- [17] J. Lafferty, A. McCallum, F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: Proc. of ICML 2001, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289.
- [18] B. Portelli, S. Scaboro, R. Tonino, E. Chersoni, E. Santus, G. Serra, Monitoring user opinions and side effects on covid-19 vaccines in the twittersphere: Infodemiology study of tweets, J Med Internet Res 24 (2022) e35115. doi:10.2196/35115.
- [19] S. Scaboro, B. Portelli, E. Chersoni, E. Santus, G. Serra, NADE: A benchmark for robust adverse drug events extraction in face of negations, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, Online, 2021, pp. 230–237. doi:10.18653/v1/2021.wnut-1.26.
- [20] S. Scaboro, B. Portelli, E. Chersoni, E. Santus, G. Serra, Increasing adverse drug events extraction robustness on social media: case study on negation and speculation, 2022. doi:10.48550/ARXIV.2209.02812.
- [21] B. Portelli, S. Scaboro, E. Chersoni, E. Santus, G. Serra, AILAB-Udine@SMM4H’22: Limits of transformers and BERT ensembles, in: Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 130–134. URL: <https://aclanthology.org/2022.smm4h-1.36>.
- [22] B. Portelli, S. Scaboro, E. Santus, H. Sedghamiz, E. Chersoni, G. Serra, Generalizing over Long Tail Concepts for Medical Term Normalization, in: Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022.