Negation detection for robust Adverse Drug Event extraction from social media texts*

Simone Scaboro[1], Beatrice Portelli[1], Emmanuele Chersoni[2], Enrico Santus[3] and Giuseppe Serra[1]

[1]*University of Udine, Italy*
[2]*The Hong Kong Polytechnic University, Hong Kong*
[3]*CSAIL MIT, Cambridge (MA), United States of America*

Abstract

Adverse Drug Event (ADE) extraction from user-generated content has gained popularity as a tool to aid researchers and pharmaceutical companies to monitor side effect of drugs in the wild. Automatic models can rapidly examine large collections of social media texts. However it is currently unknown if such models are robust in face of linguistic phenomena such as *negation* and *speculation*, which are pervasive across language varieties. We evaluate three state-of-the-art systems, showing their fragility against negation, and then we introduce two possible strategies to increase the robustness of these models: (i) a pipeline approach, using a specific component for negation detection; (ii) an augmentation of the dataset with artificially negated samples to further train the models. We show that both strategies bring significant increases in performance.

Keywords

Bio-medical data, Social media, Annotated corpora creation, Negation detection, Adverse drug events

1. Introduction

As more users keep reporting their personal experience with drugs on social media, blogs and health forums, automatic Adverse Drug Event (ADE) detection in social media texts is becoming a fundamental tool in the field of pharmacovigilance [2, 3]. It is common for Internet users to report their personal experiences with drugs on forums and microblogging platforms, but also messaging pharmaceutical companies directly on social media, via chatbots or emails. This is why both researchers and the industry are looking for ways to make use of this great amount of unprocessed and potentially informative data. User-generated texts, and social media texts in particular, are inherently noisy (containing colloquial language, slang and metaphors, non-standard syntactic constructions etc.) and require specialized data cleaning and handling

* This is an extended abstract of [1]

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, IT
scaboro.simone@spes.uniud.it (S. Scaboro); portelli.beatrice@spes.uniud.it (B. Portelli); emmanuele.chersoni@polyu.edu.hk (E. Chersoni); esantus@gmail.com (E. Santus); giuseppe.serra@uniud.it (G. Serra)
0000-0003-2533-1298 (S. Scaboro); 0000-0001-8887-616X (B. Portelli); 0000-0001-8742-0451 (E. Chersoni); 0000-0002-7327-2731 (E. Santus); 0000-0002-4269-4501 (G. Serra)

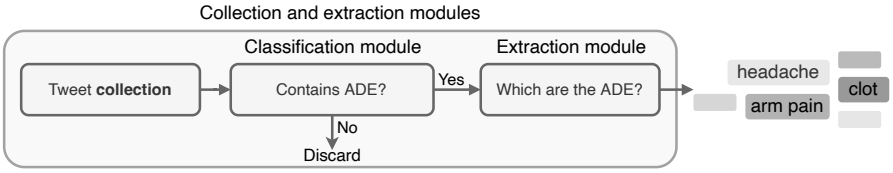

Figure 1: High level structure of a Tweet classification and ADE extraction pipeline system. Our objective is to increase the robustness of the extraction module against challenging linguistic phenomena.

techniques. The task becomes even more complicated if the final objective is to map them to a formal medical dictionary or ontology.

In the last decade, the Natural Language Processing (NLP) community dedicated a consistent effort in developing robust methods for mining biomedical information from user-generated texts, also leading to the creation of several dedicated shared tasks series on ADE detection (SMM4H – Social Media Mining for Health) [4, 5, 6, 7, 8]. Although these models have seen great advancements in the last years, it is still unknown how robust they are in face of some pervasive linguistic phenomena such as negation and speculation. However, general investigations on machine comprehension and question answering tasks confirmed that such phenomena often pose a serious challenge [9]. The distinction between certain, hypothesized and negated and speculated events is of key importance in biomedical NLP tasks [10, 11]. In the same way, it is essential to know whether the causal link between a drug and an ADE is being stated or negated in pharmacovigilance.

Detecting the scope of negation and speculation has been object of NLP research for at least one decade, via both rule-based and machine learning approaches. An early, popular system was introduced by Chapman et al. [12], whose NegEx algorithm exploited regular expressions to identify negations in clinical documents in English. The latest advancements are represented by BERT-based models [13, 14], also with the aid of multitask learning architectures [15].

As of today, the research in biomedical NLP mostly focused on scope detection of negations and speculations *per se* and on more formal types of texts (e.g. clinical notes, articles). Given the growing demand to process and analyze large collections of user-generated content from social media, we choose to focus on ADE detection on Twitter posts. They are characterized by a noisier and more informal writing style. The goal is to enable to systems to be more successful at distinguishing between factual and non-factual information.

In this paper, introduce an extended dataset to analyze the performance of ADE extraction in presence of asserted and negated Adverse Events. We show that the latest state-of-the-art ADE detection systems cannot recognize and handle negations correctly and introduce two strategies to increase the robustness of existing systems: (i) adding a negation detection module in a pipeline fashion to exclude the negated ADEs predicted by the models; (ii) augmenting the training set with artificially negated samples.

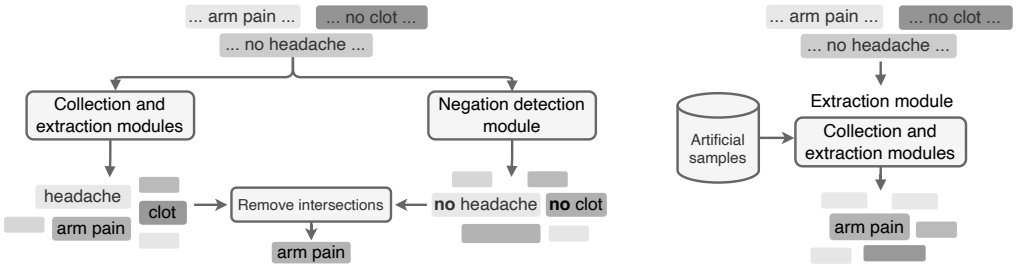

Figure 2: Structure of the pipeline system (left) and the system with data augmentation (right).

2. Proposed Strategies

Following the latest advancements in the SMM4H Shared Tasks, we choose three Transformer-based models that showed high performance on the ADE extraction dataset of SMM4H [16, 17], and are currently at the top of the corresponding leaderboard: BERT [18], SpanBERT [19] and PubMedBERT [20]. The models are fine-tuned for token classification, predicting an IOB label for each token in the sentence to detect the boundaries of ADE mentions.

We analyze two possible strategies to increase the robustness of the baseline models: (i) adding a negation (or speculation) detection module in a pipeline fashion to exclude some incorrect adverse events predicted by the models; (ii) augmenting the training set with artificially created samples. Figure 2 illustrates the two approaches.

2.1. Specialized negation detection modules

We propose a simple pipeline to enhance the robustness of the base models against negation by combining them with a negation detection module. Let us consider a text t, a ADE extraction base model $\mathscr{B}$ and a negation detection module $\mathscr{N}$. Given t, $\mathscr{B}$ outputs a set of substrings of t that are labeled as ADE mentions: $\mathscr{B}(t) = \{b_1, \dots, b_m\}$. Similarly, $\mathscr{N}$ takes a text and outputs a set of substrings, which are considered to be entities within a negation scope: $\mathscr{N}(t) = \{n_1, \dots, n_t\}$.

A combined *pipeline model* is obtained by discarding all ADE spans $b_i \in \mathscr{B}(t)$ that overlap one of the negation spans $n_j \in \mathscr{N}(t)$: $\mathscr{B}\mathscr{N}(t) = \{b_i \in \mathscr{B}(t) \mid \forall j(n_j \in \mathscr{N}(t) \wedge b_i \cap n_j = \emptyset)\}$

Modules used We introduce two negation detection modules: NegEx, a Python implementation [21] of the NegEx algorithm, based on simple regular expressions, which evaluates whether named entities are negated; BERTneg, a BERT model (bert-base-uncased) that we finetuned for token classification. We trained BERTneg on BioScope [22], which contains medical texts annotated for the presence of negation and speculation cues and their related scopes. We selected 3190 sentences (2801 with a negation scope) and finetuned the model for scope detection (10 epochs, learning rate $1e-4$).

2.2. Data Augmentation

While there are several datasets for ADE detection on social media texts [23, 24], the largest collection is the one released yearly for the SMM4H Workshop and Shared Task.

However, most datasets are made of samples that either *do* or *do not* contain an ADE (useful to train the *Classification module* in Figure 1). Because of this, they include a small number of negated ADEs by construction: no particular attention is given to these samples when curating the data and, even when they are present, they are labelled as **noADE** samples. This makes it harder to study this phenomenon.

We augment the SMM4H19$_E$ dataset (the training set for the ADE *extraction* Task of SMM4H19 [7]) in two ways: (i) recovery of real samples; (ii) generating negated versions of real samples. Both activities were carried out by four volunteer annotators with a high level of proficiency in English.

Recovery of real samples We look for real samples that negate the presence of an ADE using SMM4H19$_C$ and SMM4H20$_C$, the datasets for the binary classification tasks in [7] and [8]. These are meant to be used as test samples, to check the robustness of the model.

Generation of negated samples We manually create negated versions for the **ADE** tweets in the test split of SMM4H19$_E$. These are meant to be used as additional training samples, to teach the model how to distinguish asserted and negated adverse events. The result of this procedure is a new set of tweets denying the presence of an ADE. As an example, the original tweet "fluoxetine, got me going crazy" was transformed into "fluoxetine, didn't get me going crazy".

3. Data Partitioning

We split the available data in a train and a test set, both containing the three categories of tweets: **ADE**, **noADE** and **negADE**. Given the small amount of real **negADE** tweets, we use all of them in the test set to evaluate the performance only on real tweets. Conversely, the training set only contains the manually generated **negADE** samples.

4. Experiments

All the reported results are the average over 5 runs. For the Transformer models we used the same hyperparameters reported by Portelli et al. [16]. As metrics, we consider the number of false positive predictions (FP) and the relaxed precision (P), recall (R) and F1 score as defined in the SMM4H shared tasks [7]: the scores take into account "partial" matches, in which it is sufficient for a prediction to partially overlap with the gold annotation. We report the number of FP both on the whole test set and on individual partitions (**ADE**, **noADE** and **negADE** samples). For brevity, here we report the results for just one of the baseline models (PubMedBERT, Table 1). Results for the other baseline models behave similarly and can be found in [1].

As a preliminary step for all experiments, the two negation detection models are trained and used to predict the negation scopes for all the test samples once. This allows us to compute the predictions of any pipeline model.

Exp 0 (row 1) To provide a measure of the initial robustness of the base models and their general performance, we train them on the **ADE** and **noADE** samples only. The base models have a high number of FP, especially in the **negADE** category. This strongly suggests that they are not robust against this phenomenon.

Table 1
P, R, F1 score and number of False Positives (FP) for all the tested models.

		P	R	F1	FP	ADE	noADE	negADE	
1	$\mathscr{B}$ **(base model)**	53.24	67.41	59.47	144.2	37.0	40.4	67.4	1
2	$\mathscr{B}$**+NegEx**	59.21	59.76	59.47	93.4	32.0	37.6	23.8	2
3	$\mathscr{B}$**+BERTneg**	57.69	62.64	60.04	106.2	30.6	39.0	36.6	3
4	$\mathscr{B}$**+negSamp**	63.28	63.33	63.20	84.2	30.6	34.6	19.0	4
5	$\mathscr{B}$**+NegEx +negSamp**	63.24	58.29	60.58	76.6	29.4	34.6	12.6	5
6	$\mathscr{B}$**+BERTneg +negSamp**	64.74	60.98	62.72	74.2	27.2	34.2	12.8	6

Exp 1 (rows 2–3) We test the efficacy of the pipeline negation detection method, applying NegEx and BERTneg to the base models. When combined with NegEx (row 2), the FP decreases by almost 50 points, showing that the regular expression module removes a great number of unwanted predictions. BERTneg decreases the number of FP too, but only by 38 points, being less aggressive than NegEx. However, if we look at P and R in the first three rows, we see that the negation detection modules increase P at the cost of large drops in R: some correct predictions of the base models get discarded (i.e., ADEs that contain a negation such as "After taking this drug *I cannot sleep anymore*").
Exp 2 (row 4) We add to the training set all **negADE** generated samples and train the base models on them to test the effect of augmenting the dataset. This lowers the number of FP predictions for all models as much as using NegEx (compare row 2 and 4), especially on the **negADE** set. We still observe a drop in R, but less severe than in Exp 1 (less true positives are being discarded). The increase in P is also more noticeable, leading to an overall increase in F1.
Exp 3 (rows 5–6) To investigate whether the two methods are complementary in their action, we combine the two strategies, applying the pipeline architecture to the models trained on the augmented dataset. They are in some way complementary, as shown by the further decrease in FP in all categories. However, combining the two approaches might not be the best strategy, as it leads to a further decrease in R.

Observations The results show that introducing a small number of new samples (even if artificial) is the best way to directly increase the model knowledge about the phenomenon. However, this solution could be expensive in absence of annotated data. For this reason, the pipeline models might be a viable alternative, as they maintain the F1 score while still decreasing the number of FP.

5. Conclusions

In this paper, we evaluate the impact of negations on state-of-the-art ADE detection models. We introduce and compare two strategies to tackle the problem: using a negation detection module and adding **negSamp** samples in the training set. Both of them bring significant increases in performance. Future work should focus on more refined techniques to accurately model the semantic properties of the samples, also by jointly handling negation and speculation phenomena. This might be an essential requirement for dealing with the noisiness and variety of social media texts.

References

[1] 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. URL: https://aclanthology.org/2021.wnut-1.26.

[2] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, C. Paris, Text and Data Mining Techniques in Adverse Drug Reaction Detection, ACM Computing Surveys (CSUR) 47 (2015) 1–39.

[3] A. Sarker, G. Gonzalez, Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-corpus Training, Journal of Biomedical Informatics 53 (2015) 196–207.

[4] M. Paul, A. Sarker, J. Brownstein, A. Nikfarjam, M. Scotch, K. Smith, G. Gonzalez, Social Media Mining for Public Health Monitoring and Surveillance, in: Biocomputing 2016, 2016, pp. 468–479.

[5] A. Sarker, G. Gonzalez-Hernandez, Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017, Training 1 (2017) 1239.

[6] D. Weissenbacher, A. Sarker, M. Paul, G. Gonzalez, Overview of the Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018, in: Proceedings of the EMNLP Workshop on Social Media Mining for Health Applications, 2018.

[7] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. Paul, G. Gonzalez, Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019, in: Proceedings of the ACL Social Media Mining for Health Applications Workshop & Shared Task, 2019.

[8] A. Klein, I. Alimova, I. Flores, A. Magge, Z. Miftahutdinov, A.-L. Minard, K. O'connor, A. Sarker, E. Tutubalina, D. Weissenbacher, et al., Overview of the Fifth Social Media Mining for Health Applications Shared Tasks at Coling 2020, in: Proceedings of the COLING Workshop on Social Media Mining for Health Applications, 2020.

[9] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, in: Proceedings of ACL, 2020.

[10] E. Velldal, L. Øvrelid, J. Read, S. Oepen, Speculation and Negation: Rules, Rankers, and the Role of Syntax, Computational Linguistics 38 (2012) 369–410.

[11] N. P. Cruz Díaz, Detecting Negated and Uncertain Information in Biomedical and Review Texts, in: Proceedings of the RANLP Student Research Workshop, 2013.

[12] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan, A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries, Journal of Biomedical Informatics 34 (2001) 301–310.

[13] A. Khandelwal, S. Sawant, Negbert: A Transfer Learning Approach for Negation Detection and Scope Resolution, in: Proceedings of LREC, 2020.

[14] B. K. Britto, A. Khandelwal, Resolving the Scope of Speculation and Negation using Transformer-Based Architectures, arXiv preprint arXiv:2001.02885 (2020).

[15] A. Khandelwal, B. K. Britto, Multitask Learning of Negation and Speculation using Transformers, in: Proceedings of the EMNLP International Workshop on Health Text Mining and Information Analysis, 2020.

[16] B. Portelli, E. Lenzi, E. Chersoni, G. Serra, E. Santus, BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection,

in: Proceedings of EACL, 2021.
[17] B. Portelli, D. Passabì, E. Lenzi, G. Serra, E. Santus, E. Chersoni, Improving Adverse Drug Event Extraction with SpanBERT on Different Text Typologies, arXiv preprint arXiv:2105.08882 (2021).
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL, 2019.
[19] 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 (2019) 64–77.
[20] 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).
[21] J. Pizarro, L. Reteig, L. Murray, jenojp/negspacy: Minor Bug Fix, Improve Chunk Prefix Functionality (Version v0.1.9)., 2020. URL: https://doi.org/10.5281/zenodo.3702544. doi:10.5281/zenodo.3702544.
[22] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik, The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes, BMC Bioinformatics 9 (2008) 1–9.
[23] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, Cadec: A Corpus of Adverse Drug Event Annotations, Journal of Biomedical Informatics 55 (2015) 73–81.
[24] N. Alvaro, Y. Miyao, N. Collier, TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations, JMIR Public Health Surveillance 3 (2017) e24. URL: https://doi.org/10.2196/publichealth.6396. doi:10.2196/publichealth.6396.