NADE: A Benchmark for Robust Adverse Drug Events Extraction in Face of Negations

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

[1] University of Udine, Italy
[2] The Hong Kong Polytechnic University, Hong Kong
[3] DSIG - Bayer Pharmaceuticals, New Jersey, USA

{scaboro.simone,portelli.beatrice}@spes.uniud.it,
emmanuele.chersoni@polyu.edu.hk,
esantus@gmail.com, giuseppe.serra@uniud.it

Abstract

Adverse Drug Event (ADE) extraction models can rapidly examine large collections of social media texts, detecting mentions of drug-related adverse reactions and trigger medical investigations. However, despite the recent advances in NLP, it is currently unknown if such models are robust in face of *negation*, which is pervasive across language varieties.

In this paper 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: a pipeline approach, relying on a specific component for negation detection; an augmentation of an ADE extraction dataset to artificially create negated samples and further train the models.

We show that both strategies bring significant increases in performance, lowering the number of spurious entities predicted by the models. Our dataset and code will be publicly released to encourage research on the topic.

1 Introduction

Exploring social media texts is becoming more and more important in the field of pharmacovigilance (Karimi et al., 2015b; Sarker and Gonzalez, 2015), since it is common for Internet users to report their personal experiences with drugs on forums and microblogging platforms. Given the inherent noisiness of social media texts (colloquial language, slang and metaphors, non-standard syntactic constructions etc.), the Natural Language Processing (NLP) community dedicated a consistent effort in developing robust methods for mining biomedical information from social media outlets. This also led to the creation of several dedicated shared tasks series on ADE detection (SMM4H – Social Media Mining for Health) (Paul et al., 2016; Sarker and Gonzalez-Hernandez, 2017; Weissenbacher et al., 2018, 2019; Klein et al., 2020).

Although these models have seen great advancements in the last years, also thanks to the introduction of pre-trained Transformers-based architectures (Vaswani et al., 2017; Devlin et al., 2019), it is still unknown how robust they are in face of some pervasive linguistic phenomena such as negation. However, general investigations on machine comprehension and question answering tasks confirmed that such phenomena often pose a serious challenge (Ribeiro et al., 2020). Managing to efficiently handle the scope of negations and speculations in clinical notes is a key problem in biomedical NLP (Velldal et al., 2012; Cruz Díaz, 2013), and similarly, for digital pharmacovigilance it is essential to recognize whether the association between a drug and an ADE is actually being stated or negated because the consequences of extracting misleading information about the possible side effects of drugs can be extremely serious.

In this paper, we analyze the performance of some of the latest state-of-the-art ADE detection systems on NADE: a new dataset derived from SMM4H data, which contains a relevant amount of annotated samples with negated ADEs. We also introduce and analyze two strategies to increase the robustness of the models: adding a negation detection module in a pipeline fashion to exclude the negated ADEs predicted by the models; augmenting the training set with artificially negated samples. As a further contribution, our dataset and scripts will be made publicly available for researchers to test the robustness of their ADE extraction systems against negation.[1]

[1] https://github.com/AilabUdineGit/NADE-dataset

2 Related Work

Detecting negation scopes is a traditional topic of NLP research. An early, popular system was introduced by Chapman et al. (2001), whose NegEx algorithm exploited regular expressions to identify negations in clinical documents in English. Later, machine learning approaches became more popular after the publication of a common gold standard, i.e. the BioScope corpus (Vincze et al., 2008). Several proposals consisted of a two-steps methodology: a first classifier to detect which token in a sentence is a negation/speculation cue, and a second classifier to determine which tokens in the sentence are within the scope of the cue word (Morante et al., 2008; Cruz Díaz et al., 2012; Attardi et al., 2015; Zou et al., 2015).

More recently, approaches based on neural networks (CNN, Qian et al. 2016; BiLSTM, Fancellu et al. 2016, 2017; Dalloux et al. 2019) have been introduced in the literature, showing some crosslinguistic and crossdomain transferability. Moreover, BERT-based models have been proposed to handle this phenomenon (Khandelwal and Sawant, 2020; Britto and Khandelwal, 2020), also with the aid of multitask learning architectures (Khandelwal and Britto, 2020).

To our knowledge, the research in biomedical NLP mostly focused on scope detection *per se* and on more formal types of texts (e.g. clinical notes, articles). In our research we focus instead on the specific task of ADE detection and on the impact of negation on the performance of ADE systems for noisy social media texts (e.g. tweets, blog posts), with the goal of making them able to distinguish between factual and non-factual information.

3 NADE Dataset

While there are several datasets for ADE detection on social media texts (Karimi et al., 2015a; Alvaro et al., 2017), the biggest collection of tweets tagged for ADE mentions is the one released yearly for the SMM4H Workshop and Shared Task. We used the following resources:

a) **SMM4H19**$_{ext}$. The training set for the ADE *extraction* Task of SMM4H19 (Weissenbacher et al., 2019), consisting of 2276 tweets that mention at least one drug name. 1300 of them contain ADEs, and annotations of their position in the text (**ADE** class). The other 976 are control samples with no ADE mentions (**noADE**). As the blind test set is not publicly available, we rely on the training set only and use the splits by Portelli et al. (2021a), which balance positive and negative samples;

b) **SMM4H19**$_{cls}$ and **SMM4H20**$_{cls}$ (Weissenbacher et al., 2019; Klein et al., 2020). The training sets for SMM4H *classification* Tasks, containing tweets labeled as **ADE** or **noADE** (1:9 ratio).

The community focused on the ADE extraction task, so most datasets are made of samples that either do or do not contain an ADE. 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 when they are present they are treated as **noADE** samples and not explicitly labelled. This leads to their class being misrepresented and makes it harder to study this phenomenon.

3.1 Data Augmentation

In order to perform our analysis we created a new set of samples containing negated ADEs (**negADE**) in two ways: looking for real samples negating the presence of an ADE in SMM4H19$_{cls}$ and SMM4H20$_{cls}$ (**negADE_r**); manually creating negated versions for the **ADE** tweets in the test split of SMM4H19$_{ext}$ (**negADE_g**). The original test split includes 260 tweets, 7 of which were discarded during the generation process, leading to 253 samples. Further details in Appendix A.

Recovery of Real Samples

SMM4H19$_{cls}$ and SMM4H20$_{cls}$ contain a total of 24857 unique **noADE** tweets, so there is an high chance of encountering negated ADEs. The samples have no other annotation apart from their binary labels and analyzing all of them manually would be extremely time-consuming. We performed a preliminary filtering, keeping only the tweets containing a negation cue. Then we manually analyzed the filtered tweets, assessing whether the negation refers to an ADE and if the message actually negates the presence of the ADE. In the following examples: the first tweet is valid (it negates the ADE); the second one contains a negation, but does not negate the ADE.

1. This **#HUMIRA shot** has me feeling like a normal human... **No pain no inflammation** no nothinggggh

2. But I'm **not on adderall and I am feasting**.

The tweets were evaluated by four volunteer annotators with a high level of proficiency in English, and we only kept the samples for which they were in agreement. As a result we obtained **negADE_r**, a set of real tweets containing negated ADEs, that

allows us to create a test set containing up to 16% negated samples.

Generation of Artificial Samples

The generation process for the **negADE_g** samples was carried out by the same four volunteers as before. Each one of them was given part of the 260 tweets from the SMM4H19$_{ext}$ test set, and was asked to alter them (as conservatively as possible) to generate a new version of the tweet that negates the presence of the ADE. Each volunteer was asked to review the tweets generated by the other participants and to propose modifications, in case the tweets looked ambiguous or unnatural. If no agreement was found about the edits, the augmented tweets were discarded. The result of this procedure is **negADE_g**, a new set of tweets denying the presence of an ADE. Here is an example of an original tweet and its negated version (highlighting the cue word added to negate the ADE):

Original: fluoxetine, got me going crazy.

Negated: fluoxetine, **didn't** get me going crazy.

3.2 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** (Table 1). Given the small amount of **negADE_r** 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_g** samples.

	ADE	noADE	negADE	Total
Train	842 45%	785 42%	253 13%	1880
Test	200 43%	195 42%	73 16%	468

Table 1: Distribution of samples in NADE.

4 Analyzed models

4.1 ADE Extraction Models

We choose three Transformer-based models that showed high performance on SMM4H19$_{ext}$ (Portelli et al., 2021a,b), and are currently at the top of the SMM4H19$_{ext}$ ADE extraction leaderboard: BERT (Devlin et al., 2019), SpanBERT (Joshi et al., 2019) and PubMedBERT (Gu et al., 2020). 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.

4.2 Negation Detection Models

We introduce two negation detection modules: NegEx, a Python implementation (Pizarro et al., 2020) 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 the BioScope dataset, which consists in medical texts annotated for the presence of negation and speculation cues and their related scopes. We selected 3190 sentences (2801 of which with a negation scope) and finetuned the model for scope detection (10 epochs, learning rate 0.0001).

4.3 Pipeline Models

Let us consider a text t, a ADE extraction base model $\mathcal{B}$ and a negation detection module $\mathcal{N}$. Given t, $\mathcal{B}$ outputs a set of substrings of t that are labeled as ADE mentions: $\mathcal{B}(t) = \{b_1, \ldots, b_m\}$. Similarly, $\mathcal{N}$ takes a text and outputs a set of substrings, which are considered to be entities within a negation scope: $\mathcal{N}(t) = \{n_1, \ldots, n_t\}$.

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

5 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. (2021a).

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 (Weissenbacher et al., 2019): the scores take into account "partial" matches, in which it is sufficient for a prediction to partially overlap with the gold annotation.

As a preliminary step, 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 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 (842+785 samples). We then test the efficacy of the pipeline negation detection method, applying NegEx and BERTneg to the base models.

Exp 2: to test the effect of augmenting the training data with artificial samples (i.e., the second

		BERT				PubMedBERT				SpanBERT			
		FP	ADE	noADE	negADE	FP	ADE	noADE	negADE	FP	ADE	noADE	negADE
1	$\mathcal{B}$ **(base model)**	161.2	46.2	42.6	73.4	144.2	37.0	40.4	67.4	245.6	66.2	79.8	100.8
2	$\mathcal{B}$**+NegEx**	106.4	41.6	40.4	24.4	93.4	32.0	37.6	23.8	170.8	58.4	74.0	38.6
3	$\mathcal{B}$**+BERTneg**	120.2	40.0	41.0	39.2	106.2	30.6	39.0	36.6	184.0	56.2	74.8	53.2
4	$\mathcal{B}$**+50**	146.6	40.2	38.4	69.0	126.2	34.4	35.4	57.0	183.2	46.2	58.6	79.2
5	$\mathcal{B}$**+100**	166.8	50.4	48.4	69.0	108.0	33.0	34.6	40.4	230.0	58.6	77.4	95.4
6	$\mathcal{B}$**+150**	125.8	43.0	41.4	42.4	100.8	29.8	41.4	29.6	156.8	47.6	50.8	59.2
7	$\mathcal{B}$**+200**	105.4	41.0	36.0	29.0	79.0	29.2	27.6	22.2	134.2	38.6	44.4	51.6
8	$\mathcal{B}$**+253**	101.6	42.6	35.8	23.2	84.2	30.6	34.6	19.0	179.8	55.2	60.6	65.2
9	$\mathcal{B}$**+NegEx +253**	91.6	42.2	35.8	13.6	76.6	29.4	34.6	12.6	136.4	52.0	57.0	27.6
10	$\mathcal{B}$**+BERTneg +253**	93.6	41.0	35.8	16.6	74.2	27.2	34.2	12.8	145.4	51.4	57.4	36.8

Table 2: False Positives for: the base models; the pipeline models; base models trained with an increasing number of **negADE_g** samples; pipeline models trained with all **negADE_g** samples.

		BERT			PubMedBERT			SpanBERT		
		P	R	F1	P	R	F1	P	R	F1
1	$\mathcal{B}$ **(base model)**	50.15	65.6	56.78	53.24	67.41	59.47	43.65	73.28	54.61
2	$\mathcal{B}$**+NegEx**	55.59	58.03	56.73	59.21	59.76	59.47	48.65	65.04	55.55
3	$\mathcal{B}$**+BERTneg**	54.37	60.7	57.30	57.69	62.64	60.04	47.82	67.39	55.85
⋮	⋮	⋮			⋮			⋮		
8	$\mathcal{B}$**+253**	58.85	63.86	61.21	63.28	63.33	63.20	48.52	68.76	56.85
9	$\mathcal{B}$**+NegEx +253**	58.65	58.03	58.29	63.24	58.29	60.58	51.63	61.29	55.98
10	$\mathcal{B}$**+BERTneg +253**	59.22	60.17	59.65	64.74	60.98	62.72	51.31	64.10	56.94

Table 3: Precision, Recall and F1 score for: the base models; the pipeline models; base models trained with an increasing number of **negADE_g** samples; pipeline models trained with all **negADE_g** samples.

negation detection method), we add to the training set an increasing number of **negADE_g** samples (**+50** to **+253**, in steps of 50 samples). During preliminary experiments, we added 100 **noADE** samples from SMM4H19$_{cls}$ to the training set. The performance of all the models *did not vary* in this case, showing that the results of Exp 2 are caused by the nature of the samples, and not simply by the increased size of training set All the base models are then fine-tuned on the resulting dataset.

Exp 3: 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.

5.1 Results

Table 2 and 3 contain a summary of the most relevant metrics for the tested models. We report the number of FP both on the whole test set and on individual partitions (**ADE**, **noADE** and **negADE_r** samples).

Exp 1 (rows 1–3): all base models (row 1) have a high number of FP, especially in the **negADE** category. This strongly suggests that they are not robust against this phenomenon. When combined with NegEx (row 2), the FP decreases by 34%, showing that the regular expression module removes a great number of unwanted predictions. BERTneg decreases the number of FP, too, but only by 25%. This is due to the difference between the brute-force behaviour of NegEx and the contextual prediction of a deep machine learning model. We can notice that the two pipeline models slightly reduce the number of FP also in the **ADE** category (e.g. from 66.2 to 56.2 for SpanBERT).

However, if we look at P and R in the first three rows, we can see that the negation detection modules bring an increase in P at the cost of large drops in R. Some correct predictions of the base models get discarded, in particular the ADEs that contain a negation (e.g., "After taking this drug *I cannot sleep anymore*"). As this effect is undesirable, we investigated the use of the **negADE_g** samples to mitigate it.

Exp 2 (rows 4–8): adding **negADE_g** samples to the training set (from **+50** to **+253**) lowers the number of FP predictions for all models. Using all the available samples brings down the number of FP as much as using NegEx on the initial model (compare +NegEx and **+253**). This reduction is given by the decrease of FP in the **negADE** set, while the number of FP in the **ADE** and **noADE** categories remains roughly stable.

Comparing row **+253** with row 1 shows, as for Exp 1, an increase in P and a drop in R. However, the drop in R is less severe than before (5 points

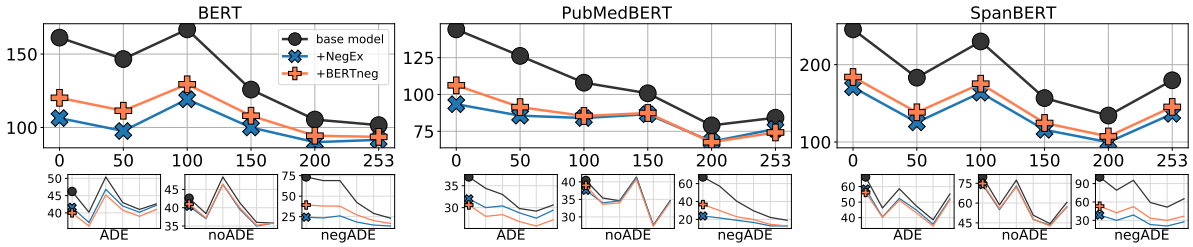

Figure 1: Top: total FP for the three base models and their pipeline versions with the increasing of **negADE** samples in the training set (x-axis: from 50 to 253). Bottom: number of FP for all the models by sample category.

at most), meaning that less true positives are being discarded. Also, P increases twice as much, leading to an overall increase in F1.

Exp 3 (rows 9–10): the effect of augmented dataset and negation modules are complementary, as shown by the further decrease in FP. However, combining the two approaches does not seem to be a winning strategy, as it leads to a further decrease in R without the benefit of increasing P.

The same behaviour can be observed for all the base models, despite the different initial performance (with SpanBERT having a generally higher R and PubMedBERT an higher P).

Figure 1 offers another visualization of the effect that adding **negADE_g** samples has on the number of False Positives. The number of **negADE** samples in the training set increases from left to right in each plot (from 0 to 253). The top row shows how the total number of predicted FPs decreases (for all base and pipeline models) when adding the generated negated samples. The plots in the bottom row show how the number of FP varies for the three categories of samples separately (**ADE**, **noADE** and **negADE**, bottom row). As observed in Table 2, the decrease is most significant in the **negADE** partition.

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 a large quantity of negated data. For this reason, the pipeline models might be a viable alternative, as they maintain the F1 score while still decreasing the number of false positives.

6 Conclusions

In this paper, we evaluate the impact of negations on state-of-the-art ADE detection models. We introduced NADE, a new dataset specifically aimed at studying this phenomenon. The dataset proves to be a challenging setting and the experiments show that current methods lack mechanisms to deal with negations. We introduce and compare two strategies to tackle the problem: using a negation detection module and adding **negADE_g** samples in the training set. Both of them bring significant increases in performance.

Both the dataset and the code are made publicly available for the community to test the robustness of their systems against negation.

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. The main short term directions are increasing the quality and quantity of real **negADE** samples (possibly via crowdsourcing), and creating a model that is able to discard **negADE** (keeping an high precision level), without sacrificing recall.

7 Acknowledgments

We would like to thank the three anonymous reviewers for their insightful feedback.

References

Nestor Alvaro, Yusuke Miyao, and Nigel Collier. 2017. TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. *JMIR Public Health Surveillance*, 3(2):e24.

Giuseppe Attardi, Vittoria Cozza, and Daniele Sartiano. 2015. Detecting the Scope of Negations in Clinical Notes. In *Proceedings of CLiC.it*.

Benita Kathleen Britto and Aditya Khandelwal. 2020. Resolving the Scope of Speculation and Negation using Transformer-Based Architectures. *arXiv preprint arXiv:2001.02885*.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings

and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Noa P Cruz Díaz. 2013. Detecting Negated and Uncertain Information in Biomedical and Review Texts. In *Proceedings of the RANLP Student Research Workshop*.

Noa P Cruz Díaz, Manuel J Mana López, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. 2012. A Machine-learning Approach to Negation and Speculation Detection in Clinical Texts. *Journal of the American Society for Information Science and Technology*, 63(7):1398–1410.

Clément Dalloux, Vincent Claveau, and Natalia Grabar. 2019. Speculation and Negation Detection in French Biomedical Corpora. In *Proceedings of RANLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural Networks for Negation Scope Detection. In *Proceedings of ACL*.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting Negation Scope Is Easy, Except When It Isn't. In *Proceedings of EACL*.

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. *arXiv preprint arXiv:2007.15779*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chenchen Wang. 2015a. CADEC: A Corpus of Adverse Drug Event Annotations. *Journal of Biomedical Informatics*, 55:73–81.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015b. Text and Data Mining Techniques in Adverse Drug Reaction Detection. *ACM Computing Surveys (CSUR)*, 47(4):1–39.

Aditya Khandelwal and Benita Kathleen Britto. 2020. Multitask Learning of Negation and Speculation using Transformers. In *Proceedings of the EMNLP International Workshop on Health Text Mining and Information Analysis*.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution. In *Proceedings of LREC*.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, et al. 2020. 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*.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the Scope of Negation in Biomedical Texts. In *Proceedings of EMNLP*.

Michael Paul, Abeed Sarker, John Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen Smith, and Graciela Gonzalez. 2016. Social Media Mining for Public Health Monitoring and Surveillance. In *Biocomputing*, pages 468–479.

Jeno Pizarro, Leon Reteig, and Luke Murray. 2020. jenojp/negspacy: Minor Bug Fix, Improve Chunk Prefix Functionality (Version v0.1.9).

Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. 2021a. BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection. In *Proceedings of EACL*.

Beatrice Portelli, Daniele Passabì, Edoardo Lenzi, Giuseppe Serra, Enrico Santus, and Emmanuele Chersoni. 2021b. Improving Adverse Drug Event Extraction with SpanBERT on Different Text Typologies. *arXiv preprint arXiv:2105.08882*.

Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo. 2016. Speculation and Negation Scope Detection via Convolutional Neural Networks. In *Proceedings of EMNLP*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of ACL*.

Abeed Sarker and Graciela Gonzalez. 2015. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-corpus Training. *Journal of Biomedical Informatics*, 53:196–207.

Abeed Sarker and Graciela Gonzalez-Hernandez. 2017. Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. *Training*, 1(10,822):1239.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of NIPS*.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and Negation: Rules, Rankers, and the Role of Syntax. *Computational Linguistics*, 38(2):369–410.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(11):1–9.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. 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 (# SMM4H) Workshop & Shared Task*.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. 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*.

Bowei Zou, Qiaoming Zhu, and Guodong Zhou. 2015. Negation and Speculation Identification in Chinese Language. In *Proceedings of ACL-IJCNLP*.

A Data Augmentation Process

The data augmentation process was carried out by four volunteers with a high level of proficiency in English. More specifically, the volunteers were: two graduate students (Master's degree in Computer Science and Artificial Intelligence) and two Ph.D. in Natural Language Processing. All of them have a minimum English level of C1.[2]

A.1 Recovery of Real Samples

The **noADE** samples from the SMM4H19$_{cls}$ and SMM4H20$_{cls}$ binary classification datasets have been filtered using the negation cues from BioScope (e.g. *none*, *missing*, *no longer*, *etc.*). Thanks to this first filtering, only the remaining 3897 samples have been analyzed by the volunteers. A tweet was kept only if it negated the presence of an ADE.

In the following examples, 1 and 2 are valid tweets, while 3 and 4 contain negations but do not negate the ADE:

1. This **#HUMIRA shot** has me feeling like a normal human... **No pain no inflammation** no nothingggggh #RAproblems
2. @UKingsbrook That's correct! **Metoprolol** is **NOT known to cause hypokalemia**.
3. I've seen so much **Tamiflu** these past couple of days I'm **not even surprised I'm shivering and experiencing aches** right now. *sigh
4. But I'm **not on adderall and I am feasting**.

A.2 Generation of Artificial Samples

According to the split provided by Portelli et al. (2021), we extracted and modified only the **ADE** samples from the test set (260 samples). The volunteers were instructed to modify the samples with as little edits as possible, while still generating a plausible tweet. They were encouraged to add one or more negation cue words to negate the ADE reported in the tweet. In the case it was not possible to negate the meaning of the tweet just by adding cue words, they were allowed to perform more edits in the sentence and use longer expressions.

Each annotator was asked to review the edits done by the others, and asked to point out which samples seemed unrealistic or failed to negate the ADE. During the augmentation process, if no agreement was found about the edits, the tweet modified tweet was discarded. At the end of the process 7 tweets were removed from the final dataset. Due to the generation process we implemented, we could not directly measure the inter-annotator agreement, which could, however, be inferred by the number of discarded samples.

B Full Results

Table 4 reports the metrics for all the models (average over 5 runs). The table includes also the results for the pipeline models trained with an increasing number of **negADE_g** samples, which show the same trend as the base models. The most relevant combinations (discussed in the main part of the paper) are highlighted in color (base models in gray, base models trained with 253 **negADE_g** samples in orange, pipeline models in blue).

[2] https://www.efset.org/cefr/c1/

False Positives

negADE_g samples	0	50	100	150	200	253
BERT	161.20	146.60	166.80	125.80	105.40	101.60
+NegEx	106.40	97.60	119.20	100.00	**90.00**	91.60
+BERTneg	120.20	111.40	129.40	108.00	94.40	93.60
PubMedBERT	144.20	126.20	108.00	100.80	79.00	84.20
+NegEx	93.40	85.60	84.00	86.60	68.00	76.60
+BERTneg	106.20	91.40	85.40	87.40	**67.60**	74.20
SpanBERT	245.60	183.20	230.00	156.80	134.20	179.80
+NegEx	170.80	125.40	163.80	115.80	**99.80**	136.40
+BERTneg	184.00	138.40	175.40	124.60	107.20	145.40

Precision

negADE_g samples	0	50	100	150	200	253
BERT	50.15	51.47	49.17	54.24	57.82	58.85
+NegEx	55.59	56.79	53.52	56.53	58.52	58.65
+BERTneg	54.37	55.34	52.78	55.88	58.53	**59.22**
PubMedBERT	53.24	56.43	59.25	59.46	62.94	63.28
+NegEx	59.21	61.60	61.98	60.29	63.84	63.24
+BERTneg	57.69	61.24	62.44	61.12	**64.84**	64.74
SpanBERT	43.65	53.32	45.46	48.24	52.80	48.52
+NegEx	48.65	**60.44**	49.95	52.83	55.95	51.63
+BERTneg	47.82	56.60	49.25	51.43	55.54	51.31

F1 score

negADE_g samples	0	50	100	150	200	253
BERT	56.78	56.59	56.18	58.41	59.45	**61.21**
+NegEx	56.73	56.48	56.10	57.03	57.19	58.29
+BERTneg	57.30	56.73	56.61	57.65	58.06	59.65
PubMedBERT	59.47	61.37	62.72	61.59	61.74	**63.20**
+NegEx	59.47	60.91	61.26	59.22	59.26	60.58
+BERTneg	60.04	62.15	62.70	61.10	61.36	62.72
SpanBERT	54.61	45.76	56.20	49.47	54.75	56.85
+NegEx	55.55	46.54	56.69	48.85	53.44	55.98
+BERTneg	55.85	46.61	**57.17**	49.61	54.49	56.94

Recall

negADE_g samples	0	50	100	150	200	253
BERT	**65.60**	63.04	65.57	63.41	61.85	63.86
+NegEx	58.03	56.32	58.97	57.66	56.43	58.03
+BERTneg	60.70	58.38	61.09	59.66	58.16	60.17
PubMedBERT	67.41	**67.44**	66.67	64.03	60.67	63.33
+NegEx	59.76	60.33	60.59	58.32	55.45	58.29
+BERTneg	62.64	63.19	62.99	61.19	58.36	60.98
SpanBERT	73.28	59.73	**74.19**	59.22	60.32	68.76
+NegEx	65.04	53.01	65.95	52.86	53.91	61.29
+BERTneg	67.39	54.98	68.48	55.14	56.27	64.10

Table 4: All metrics for: the original base models (gray); the base models trained with different quantities of **negADE_g** samples (the results of training with 253 **negADE_g** samples in orange); the two pipeline models (in blue); the combination of the two methods. For each model, in bold is highlighted the best value for the specific evaluation metric.