Extracting patient-reported outcomes and side effects from online drug reviews for real-world evidence generation

Moritz Blum[1], Matthias Hartung[1], and Philipp Cimiano[1,2]

[1] Semalytix GmbH, Bielefeld, Germany
{moritz.blum,hartung,cimiano}@semalytix.com
http://www.semalytix.com
[2] Bielefeld University, Faculty of Technology
cimiano@techfak.uni-bielefeld.de

Abstract. We present a modular and hierarchical deep learning architecture based on transformer models to extract outcomes and adverse drug reactions (ADR) from online drug reviews written by patients. The method is suited to generate real-world data that can complement the evidence gathered in randomized controlled clinical trials (RCTs). We provide results for four diseases (Diabetes Type 2, Obesity, Breast Cancer and Psoriasis), showing that the model generalizes to related diseases (e.g., from Diabetes to Obesity), while struggling to generalize across very different disease types (e.g., Diabetes to Breast Cancer). We show that with a very limited amount of data samples for five additional diseases (Migraine, Muscle spasm, Depression, Parkinson's Disease, Crohn's Disease), our model can generalize also to new diseases with limited amounts of training. We present three use cases showing how our method can support comparative effectiveness research, pharmacovigilance and exploration of main drivers of non-adherence to therapies.

Keywords: Drug Reviews · Patient-reported Outcomes · Pharmacovigilance · Natural Language Processing

1 Introduction

In recent years, there is an increasing interest in complementing clinical evidence generated from randomized controlled trials (RCTs) with additional evidence derived from real-world data, i.e., data derived from a number of sources that are associated with outcomes in a heterogeneous patient population in real-world settings. The real-world evidence (RWE) [17] derived from such additional data can support the assessment of the comparative effectiveness of different treatments and thus provide additional data for health technology assessment (HTA) and benefit-risk analysis in particular. Real-world evidence can also provide crucial insights to focus drug development on patients' needs (see FDA guidelines on patient-focused drug development[3]). Social media has been identified as one promising source of evidence to reveal insights about the subjective experience of patients related to outcomes, but also adverse drug reactions (ADR), and about how drugs improve their quality of life [4, 16].

In this paper we present a new approach to deriving patient-focused real-world evidence regarding outcomes and ADRs from online drug reviews written by patients. While previous work has mainly focused on approximating outcomes by extracting sentiment at the very coarse-grained level of drugs or interventions, our approach allows for a deeper analysis of drug reviews beyond sentiment by extracting outcomes including information about individual variables that explain observed outcomes, ADRs and their severity, as well as the duration of the treatment and information about whether the patient has stopped taking the drug. The model we use is a modular hierarchical deep learning architecture composed of several transformer-based components that are trained to extract the relevant information. As dataset, we rely on the public corpus of online drug reviews provided by Gräßer et al. [8]. We evaluate the architecture on four diseases (Diabetes Type 2, Obesity, Breast Cancer and Psoriasis) for which we have annotated 300 drug reviews each (1.200 in total) and we test the generalizability on five further diseases (Migraine, Muscle spasm, Depression, Parkinson's Disease, Crohn's Disease) for which we have less data (100 samples per disease). We show that our models reach an exact match F_1 score of 0.53 for extracting text passages describing outcomes, and 0.73 for extracting ADR descriptions. In a more lenient evaluation considering overlap in terms of tokens we reach even F_1 scores of 0.66 and 0.76, respectively. Beyond these descriptions, we can extract the specific outcome measure / variable as well as magnitude of improvement / severity of ADR with F_1 scores between 0.37 and 0.47. In addition, on treatment duration and sentiment F_1 scores of up to 0.79 are reached. We further show how our approach can be used to generate relevant evidence in a number of use cases, showing how it can support comparative effectiveness research by comparing outcomes across treatments, as well as pharmacovigilance, showing distribution and severity of ADRs. Finally, our approach allows to investigate the main ADRs negatively affecting adherence.

2 Dataset

As dataset of online patient drug reviews we rely on the public dataset provided by Gräßer et al. [8] that is free for research purposes. It contains more than 215k reviews from patients summarizing their experiences with a specific treatment. These reviews have on average a length of around six sentences. A review that we use as running example is the following:

"I have been taking Invokana for about 5 months and I feel great!! My a1c went from 9.8 to 7.4 in 3 months. The only side effects I have noticed is some

[3] https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical

weight loss. But the weight loss stopped. It does make me really thirsty, small price to pay for controlled blood sugar levels."

We manually annotated reviews with our targets of interest as follows:

- **sentiment:** The annotation describes the overall sentiment, *(positive, neutral, negative)*. The example review above has a positive sentiment.
- **improvement:** We annotate whether the patient reports an *improvement*, *no improvement* or even *worsening* of their condition; this represents thus a 3-class classification task. In the example review, the patient describes an improvement w.r.t the core Diabetes symptoms.
- **outcomes:** In terms of outcomes, we mark the strings describing the outcome (e.g. some weight loss, a1c went from 9.8 to 7.4, and controlled blood sugar levels). For each outcome, we annotate the variable/measurement related to the outcome (a1c, weight, blood sugar), direction *(increase, decrease, no change, got normal, not applicable)* as well as whether the outcome is positive, negative or unknown *(+, -, *)*.
- **side-effects:** We annotated the side-effects reported as a connected sequence of tokens in addition to the patient-reported severity (minor, normal, severe). In the above case, the patient reports severe thirst.
- **duration of treatment:** We mark the duration of the treatment as mentioned by the patient, that is *5 months* in our example.
- **treatment stop:** We annotate whether the patient reports having stopped the treatment (binary classification). In our example the patient does not report about a stopped treatment.

Overall, our dataset comprises of 1.700 samples: 300 samples annotated for Diabetes Types 2, Obesity, Breast Cancer and Psoriasis each, and 100 samples for five further diseases (Migraine, Muscle Spasm, Depression, Parkinson's disease, Crohns's Disease) to test the generizability of the model. For each disease we held back 50 randomly selected samples for testing purposes. Our annotations will be made available for further research upon acceptance of the paper.

3 Methods

We propose a complex, modular and hierarchical deep learning architecture to extract all the information of interest from online drug reviews. We rely on pre-trained transformers ([18, 5, 13]) which are based on attention mechanisms to better capture long-distance relationships in texts beyond what is typically possible with recurrent architectures. Devlin et al. have shown [5] that their pre-trained BERT model fine-tuned with one additional output layer can reach state of the art performance in various NLP tasks. We describe the modules for extracting the different target variables in what follows:

Sentiment & Improvement: In order to extract sentiment and the improvement experience by the patient, we rely on a pre-trained transformer (Roberta, [13]) with a fully-connected linear classification layer on top that is fine-trained for the tasks.

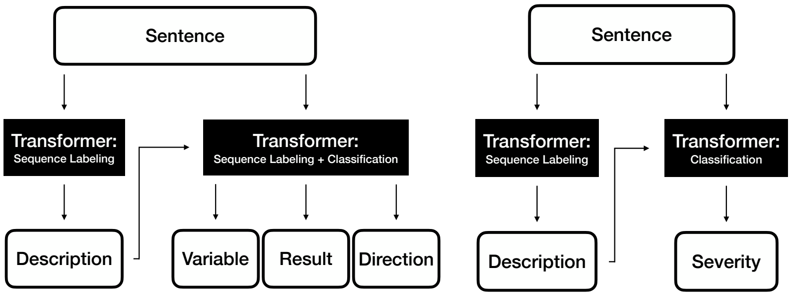
Outcomes: Outcome extraction involves the prediction of a 4-tuple *(outcome description, variable, result, direction)*. We split the problem into two parts: 1) recognizing the outcome description (modelled as a sequence labelling task) and 2) extracting the three other variables from the outcome description, thus having a hierarchical architecture that is common for this type of tasks [11]. Fig. 1a shows the architecture schematically. For sequence labeling we use the pre-trained transformer [13] and add a linear layer for IOB sequence labelling on top. The second part of the architecture has a linear layer as output that can handle sequence labeling and two classifications at once. The sequence labeling is responsible for extracting the variable whereas the first classification handles the result and the second the direction. The output has the triple format: ($\mathbb{R}^{3\times m}$, $\mathbb{R}^3$, $\mathbb{R}^5$), where m is the maximum sequence length (default 150 tokens). Experiments have shown that for the outcome and side effect extraction, feeding the text split into sentences rather than as a whole into the model yields better results.
Side Effects: To extract tuples of side effect descriptions and their severity, a two-part architecture similar to the one for outcome extraction is used. First, the description string is extracted, using a sequence labeling transformer as presented before. The output is then handed over to a second model for estimating the severity for each side effect. This second transformer has three linear output heads and gets the raw input together with the side effect description to perform the severity classification. Fig. 1b shows this architecture.
Duration of Treatment: For extracting the medication duration, a sequence labelling transformer is used as explained before.
Treatment Stop: Classification architecture as presented but with a two-class output layer.

To evaluate the performance of the sequence labeling models on our data we use the exact match precision, recall and F_1 and, furthermore, defined a score that is more meaningful and better suited for our setting. The score is called $F_{1,BOT}$ and considers the predictions and ground truth as bag-of-tokens on which the micro-averaged F_1 is computed.

4 Results

All models are trained on a batch size of 32 for 20 epochs by backpropagating the gradient of the *cross-entropy loss*. We made a 80%/20% train/validation split and the best models are chosen according to their performance on the validation set. Results are reported on 50 unseen samples for each indication.

Simple extraction results: For the sentiment and the improvement classifiers, micro-averaged F_1 scores of 0.79 are achieved for both tasks. A F_1 score of 0.93 is reached for the treatment stop classification. For these classification tasks precision and recall are quite balanced. A $F_{1,BOT}$ score of 0.53 is reached for detecting the outcome description and a score of 0.73 is reached for the side effect description. For these tasks, precision and recall show a larger but still acceptable gap. Tab. 1 shows the results.



(a) outcome extraction

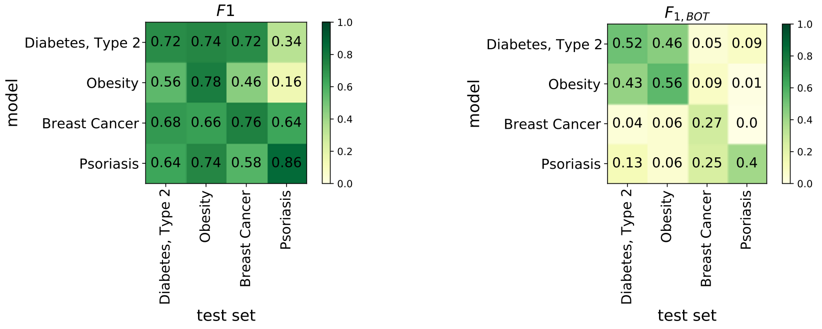
(b) side effect extraction

Fig. 1: system-architectures: prediction of precise medication information from drug review sentences

Tuple extraction results The advanced extractions involve the extraction of structured tuples for outcomes and ADRs. This involves the extraction of the measurement/variable and direction for outcomes, and the extraction of a severity for ADRs. For this, we feed the corresponding descriptions into a further model that is trained by summing up the loss of the different outputs and computing the gradient based on this. An output triple is considered correct if it is contained in the set of outcome triples for the considered sentence, irrespective of the position where the outcome was found. We see that the model reaches a very high precision in extracting the outcome variable (0.71), which is a positive regarding result. However, it misses some variables as shown by the lower recall (0.35). Regarding the extraction of the whole tuples for outcomes and side effect, the F-measure is 0.37 and 0.39, respectively. To annotate the side effect with their severity, a similar mechanism is used. The severity classification transformer is trained by putting in the side effect description along with the enclosing sentence and the side effects' severity as target. The scores are again evaluated based on the extractions for each sentence, this time on the side effect description and severity tuples. The trained model was able to achieve a F_1 score of 0.39. These results are shown in Tab. 1.

Table 1: Results in terms of F_1, Precision and Recall for our different target variables

	$F1, exact\, match$	precision	recall	$F_{1,BOT}$
sentiment	0.79	0.79	0.79	-
improvement	0.79	0.79	0.79	-
outcome description	0.53	0.55	0.51	0.66
side effect description	0.73	0.60	0.77	0.76
duration of treatment	0.57	0.52	0.63	0.66
treatment stopp	0.93	0.92	0.93	-
outcome variable	0.47	0.71	0.35	-
outcome ($variable, result, direction$)	0.37	0.56	0.28	-
side effect ($description, severity$)	0.39	0.32	0.49	-



(a) sentiment classification

(b) outcome descriptions extraction

Fig. 2: Generalization capabilities of models trained on data about a certain disease and tested against unseen diseases

Generalization to new diseases: In order to test the generalization capabilities of our approach, we train one model for each specific indication and test its performance on the other (unseen) diseases. We show results for this cross-disease transfer setting for the case of predicting the sentiment and the outcome description in Figures 2a and 2b, respectively. We see that the sentiment classification can be generally transferred well across diseases, reaching transfer F-Measures of between 0.56 and 0.70 for the group Diabetes, Obesity and Breast Cancer. It seems however, that the transfer works not so well from Diabetes/Obesity to Psoriasis. This requires further investigation. Regarding the transfer of the models for predicting outcomes, we see that the transfer is very limited, as expected, as the extraction of outcomes is specific for a particular indication. Due to the closeness of indications, it is thus as expected to see better resulting when transferring from Diabetes to Obesity than across less related diseases.

We further tested transferrability of our model trained on the four diseases considered above to five new diseases (Migrane, Muscle Spasm, depression, Parkisons's Disease and Crohn's Disease). While the zero-shot transfer setting did not work well (having $F_{1,BOT}$ scores of around 0.10 only), fine-tuning the model with only 50 samples from each of these diseases yielded reasonable $F_{1,BOT}$-Measures of 0.56 (Migraine), 0.44 (Muscle spasm), 0.40 (Depression), 0.48 (Parkinson's Disease), 0.50 (Crohn's Disease), respectively. This shows that our model can be transferred to new diseases with a minimal additional annotation effort.

5 Use Cases

This section briefly discusses some use cases supported by the information extracted from the online drug reviews.

Comparative effectiveness in terms of patient-reported outcomes: Using the outcomes extracted by our machine learning models, the reported effectiveness of different drugs can be investigated in comparison to each other. Fig. 3a shows

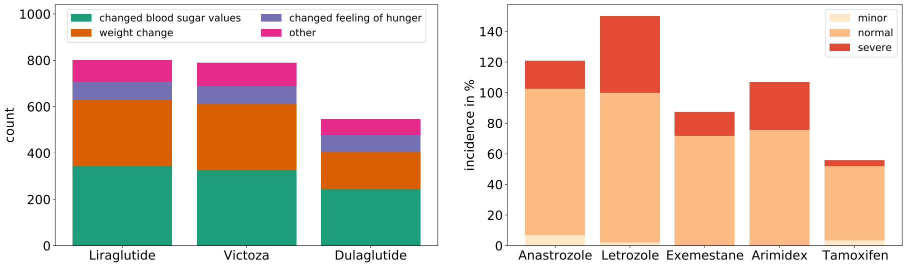


Fig. 3: a) Reported outcomes for the 3 most frequently reviewed drugs treating Diabetes Type 2, b) Mentions of joint pain for the 5 most frequently reviewed drugs treating Breast Cancer

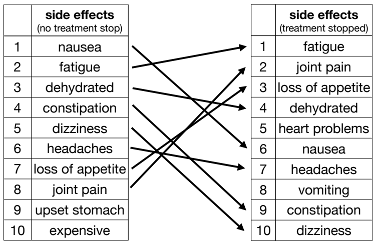
the extracted outcomes for the three most frequently reviewed drugs for Diabetes Type 2. The following table shows the percentage of Diabetes patients that report a given outcome (reduction of blood sugar levels, weight reduction and reduction of HbA1c) in comparison across the three drugs Liraglutide, Victoza, Dulaglutide:

success rate	Liraglutide	Victoza	Dulaglutide
changed blood sugar values	0.73	0.73	0.79 (*)
weight change	0.87	0.88	0.90
HbA1c change	0.95	0.95	0.96

The analyses show that according to a t-test ($\alpha = 0.1$), Dulaglutide has a significantly higher amount of outcomes compared to Liraglutide and Victoza in terms of improving patients' blood sugar levels.

Pharmacovigilance use case: The most frequently named ADRs for Breast Cancer drugs in a total of 272 reviews are joint pain (262 times), hot flashes (89 times) and fatigue (78 times). Fig. 3b shows the incidence in percent that joint pain was mentioned as an ADR in comparison across the three most reviewed Breast Cancer drugs. The analysis might help to understand which drug is most suited for a patient that is particularly sensitive to joint pain.

Challenges to adherence: Given that our classifiers predict whether patients have decided to stop treatment, we can investigate which are the most frequent ADRs mentioned in the context of treatment stops and thus represent a challenge to adherence. Fig. 4 for example shows the top-10 side effects mentioned in reviews about Obesity in which patients mention a stopped treatment in comparison to uninterrupted treatments. We see that nausea, constipation and dizziness are quite frequent side effects that patients are tolerating, since they do not lead to a treatment stop with the same relative frequency. Joint pain, loss of appetite and fatigue are overall less frequent, but are frequently named in context of a treatment stop. Such data might provide valuable insights to new product commercialization and strategic marketing in pharma.



	side effects (no treatment stop)
1	nausea
2	fatigue
3	dehydrated
4	constipation
5	dizziness
6	headaches
7	loss of appetite
8	joint pain
9	upset stomach
10	expensive

	side effects (treatment stopped)
1	fatigue
2	joint pain
3	loss of appetite
4	dehydrated
5	heart problems
6	nausea
7	headaches
8	vomiting
9	constipation
10	dizziness

Fig. 4: Side effects for obesity ranked from frequent to infrequent: mentioned in reviews of continued treatment (left) vs. in context of a treatment stop (right)

6 Discussion and Related Work

It has been shown in previous research that patient-reported outcomes (PROs) can provide relevant insights into patients' treatment experiences. PROs extracted from traditional data sources like medical studies have been compared to PROs extracted from social media sites, finding that the latter are able to confirm known and highlight novel or rare ADRs and thus support hypotheses generation and validation. These studies have also shown that social media provides more detailed information on patient experiences than other sources [7].

Existing approaches for extracting medical information from the web are already considering drug reviews [8, 14], medical forum messages [3, 12], social media posts like e.g. Twitter [2] or search queries [19]. Work so far on extracting patient experiences from online sources and social media includes basis tasks consisting in classifying entire reports or sentences into sentiment and polarity categories [8, 15]. Beyond a mere classification, other approaches have focused on sentiment analysis [8], ADRs or medication outcomes [2], and aspect related sentiment or polarity analysis [10, 20, 15]. Our goal has been to develop an approach that allows for a more in-depth extraction of patient-reported outcomes beyond sentiment or polarity only, attempting to extract the specific variable/measure as well the direction and level of improvement to the patient. Our approach does clearly go beyond pure classification approaches involving sentiment/polarity classification or detection of ADRs / outcomes without further detail. In this sense our approach allows for a deeper understanding of patient experiences and supports the aggregations that we have demonstrated in our use cases. From a methodological perspective, most related works rely on hand-crafted rules [14] or basic machine learning models such as SVMs or multilayer perceptrons [3], whereas state of the art neural network architectures like LSTMs or transformers are rarely used in this domain [9, 18, 6].

Regarding our own results, we have (very) good F-Measures of between 0.53 and 0.93 on detecting sentiment, subjective improvement, duration of treatment as well as outcome and ADR descriptions. The extraction of the structured tuples for outcomes and ADRs yields however lower results; the main reason is that the evaluation underestimates the performance of our approach, requiring to extract each component of the tuple as annotated in the gold standard. Arguably, even extractions that are not 100% correct w.r.t. to this strict evaluation are useful. Yet, the extraction of the specific measure/variable and direction is key to aggregate evidence across reviews to support our three use cases. Our results differ across diseases, showing differences in the way patients report outcomes. Our data showed that Diabetes patients are able to report their outcomes and track their health status precisely by reporting their self measured blood sugar. In contrast, Breast Cancer patients write more about side effects than outcomes in their online reviews.

7 Conclusion

We have presented a modular and hierarchical deep learning architecture for extracting patient-reported outcomes and ADRs from online drug reviews written by patients. While our results can be definitely improved, our research shows that it is possible to extract such PROs and ADRs with reasonable performance (F-Measures of 0.66 for extracting text passages describing outcomes, and 0.76 for extracting ADR descriptions). We have also shown that we can apply the models to new diseases given a small amount of annotated examples (50 samples). The model relies on explicit mentions of improvements / reductions / increases by patients. It would be interesting to equip our system with background knowledge about which values (e.g. Hb1AC) are 'normal' or 'pathological' so that outcomes can be detected when the patient only mentions values but does not mention an improvement specifically. It would also be interesting to include data from other sources to allow cross-linking of user-generated content with structured databases like MetaMap or UMLS Methathesaurus CHV (consumer Health Vocabulary) [1, 12]. Finally, future work should investigate whether our models and results can be transferred to less structured data sources such as social media or forums.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the Annual AMIA Symposium. p. 17 (2001)
2. Bian, J., Topaloglu, U., Yu, F.: Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the Intl. Workshop on Smart health and wellbeing. pp. 25–32 (2012)
3. Chee, B.W., Berlin, R., Schatz, B.: Predicting adverse drug events from personal health messages. In: Proc. of the Annual AIMA Symposium. p. 217 (2011)
4. Cook, N.S., Kostikas, K., Gruenberger, J.B., Shah, B., Pathak, P., Kaur, V.P., Mudumby, A., Sharma, R., Gutzwiller, F.S.: Patients' perspectives on copd: findings from a social media listening study. ERJ Open Research **5**(1) (2019)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR,abs/1810.04805 (2018)

6. Fan, B., Fan, W., Smith, C., et al.: Adverse drug event detection and extraction from open data: A deep learning approach. Information Processing & Management **57**(1), 102131 (2020)
7. Golder, S., Norman, G., Loke, Y.K.: Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. British journal of clinical pharmacology **80**(4), 878–888 (2015)
8. Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S.: Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In: Proceedings of the 2018 International Conference on Digital Health. pp. 121–125 (2018)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Kaiser, C., Bodendorf, F.: Mining patient experiences on web 2.0-a case study in the pharmaceutical industry. In: Proc. of the Annual SRII Global Conference. pp. 139–145. IEEE (2012)
11. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. Tech. rep., Stanford InfoLab (1997)
12. Liu, X., Chen, H.: A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports. Journal of biomedical informatics **58**, 268–279 (2015)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. CoRR,abs/1907.11692 (2019)
14. Na, J.C., Kyaing, W.Y.M.: Sentiment analysis of user-generated content on drug review websites. Journal of Information Science Theory and Practice **3**(1), 6–23 (2015)
15. Niu, Y., Zhu, X., Li, J., Hirst, G.: Analysis of polarity information in medical text. In: Proc. of the annual AMIA Symposium. p. 570 (2005)
16. Patalano, F., Gutzwiller, F.S., Shah, B., Kumari, C., Cook, N.S.: Gathering structured patient insight to drive the pro strategy in copd: Patient-centric drug development from theory to practice. Advances in Therapy **37**(1), 17–26 (2020)
17. Sherman, R.E., Anderson, S.A., Dal Pan, G.J., Gray, G.W., Gross, T., Hunter, N.L., LaVange, L., Marinac-Dabic, D., Marks, P.W., Robb, M.A., Shuren, J., Temple, R., Woodcock, J., Yue, L.Q., Califf, R.M.: Real-world evidence — what is it and what can it tell us? New England Journal of Medicine **375**(23), 2293–2297 (2016)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proc. of Advances in neural information processing systems **30**, 5998–6008 (2017)
19. White, R.W., Tatonetti, N.P., Shah, N.H., Altman, R.B., Horvitz, E.: Web-scale pharmacovigilance: listening to signals from the crowd. Journal of the American Medical Informatics Association **20**(3), 404–408 (2013)
20. Xia, L., Gentile, A.L., Munro, J., Iria, J.: Improving patient opinion mining through multi-step classification. In: Proc. of the Int. Conf. on Text, Speech and Dialogue. pp. 70–76. Springer (2009)