Conversation-Aware Filtering of Online Patient Forum Messages

Anne Dirkson
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA,
Leiden, the Netherlands

Suzan Verberne
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA
Leiden, the Netherlands

Wessel Kraaij
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA
Leiden, the Netherlands

{a.r.dirkson, s.verberne, w.kraaij}@liacs.leidenuniv.nl

Abstract

Previous approaches to NLP tasks on online patient forums have been limited to single posts as units, thereby neglecting the overarching conversational structure. In this paper we explore the benefit of exploiting conversational context for filtering posts relevant to a specific medical topic. We experiment with two approaches to add conversational context to a BERT model: a sequential CRF layer and manually engineered features. Although neither approach can outperform the F_1 score of the BERT baseline, we find that adding a sequential layer improves precision for all target classes whereas adding a non-sequential layer with manually engineered features leads to a higher recall for two out of three target classes. Thus, depending on the end goal, conversation-aware modelling may be beneficial for identifying relevant messages. We hope our findings encourage other researchers in this domain to move beyond studying messages in isolation towards more discourse-based data collection and classification. We release our code for the purpose of follow-up research.[1]

1 Introduction

In the past decade, social media has emerged as a source of valuable knowledge in the health domain (Gonzalez-Hernandez et al., 2017), for instance during the COVID-19 pandemic (Sarker et al., 2020) (Klein et al., 2020). In order to use social media to answer a medical question, it is necessary to identify posts on the forum that are relevant to the question at hand e.g. posts mentioning adverse drug responses (ADRs) (Li et al., 2020), personal experiences (Dirkson et al., 2019), medication abuse (Sarker et al., 2016) or medical misinformation (Kinsora et al., 2017). This filtering step is often the first step of the analysis pipeline. In this paper, we will refer to this specific type of filtering as relevance classification.

Previous automatic methods for medical relevance classification generally consider posts as units without context, thereby ignoring any information that can be gained from conversational context. One example of such an approach is the recent shared task on ADR relevance classification (Weissenbacher et al., 2019). Yet, including the conversational context may prove beneficial to relevance classification, as responses in a thread often relate to previous responses. For example, responses to a question or comment about a specific side effect are likely to also concern this side effect. To test this hypothesis, we investigate how positive labels are distributed across and within conversational threads.

At present, only one study into medical relevance classification has included some engineered features to capture aspects of the conversational structure (Kinsora et al., 2017). However, as this study includes only two discourse-based features, the effect of including manually engineered features that capture conversational structure is still largely unknown for relevance classification tasks.

Furthermore, including the relation between posts on a discourse level may also be able to improve classifier performance. Each post serves a conversational function in a dialogue, e.g. a question, explanation or statement (Austin, 1962). These functions are called *dialogue acts* (Stolcke et al., 2000). We have not found any study that included dialogue acts as features for medical relevance classification.

[1] Our code is available at: https://github.com/AnneDirkson/ConversationAwareFiltering

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 11–18
Barcelona, Spain (Online), December 12, 2020.

As an alternative to using manually engineered features, conversational threads can also be modelled with a sequential model. This has proven beneficial in other fields such as rumor classification in social media discussions (Zubiaga et al., 2018). As of yet, the use of sequential models for medical relevance classification has also not been explored.

We address the following research questions in this paper:

RQ1 To what extent can the addition of a sequential model on top of state-of-the-art non-sequential models improve medical relevance classification of social media data?

RQ2 To what extent can the addition of manually engineered features for conversational structure and discourse improve medical relevance classification?

We use two different datasets for answering our questions. In our current research, we are particularly interested in discovering ADRs in online discussions. We have collected and annotated a dataset about this topic. Since this dataset is new, no other results have been published for it. We therefore use one other dataset for evaluating our methods: the medical misinformation dataset by Kinsora et al. (2017). We use a BERT-based model as baseline. BERT models constitute the current state of the art for most NLP tasks (Devlin et al., 2019) including ADR relevance classification (Weissenbacher et al., 2019).

In the following section, we will elaborate on related work. Hereafter, we describe our methodology and data in Section 3 and 4 respectively. Finally, we present and discuss our results in Section 5 and 6.

2 Related Work

The use of conversational structure for improving the performance of classifiers of social media posts is prevalent in the field of rumor classification (Zubiaga et al., 2018) and related fields like disagreement detection (Rosenthal and McKeown, 2015). Conversational structure has previously been exploited through (a) manually engineered features or (b) sequential classifiers.

The most commonly employed engineered features to model the conversational structure are the similarity to the previous message and to the thread in general (Zubiaga et al., 2018). In addition to these features, the current state-of-the-art model on a leading shared task for rumor stance classification (RumourEval-2019) uses the label of the previous message and the distance to the start of the thread (Li et al., 2019). In the health domain, the only study that employs manually engineered features for conversational structure is Kinsora et al. (2017). Specifically, they use the running count of positive labels and the distance to the previous positive label. In this study, we will employ the above features as well as expand upon them with additional discourse-related features.

Other studies have used sequential classifiers to model the discursive nature of social media, although according to Zubiaga et al. (2018) this is "still in its infancy" (p. 276). Their comparison of various classifiers for rumor stance classification revealed that sequential classifiers outperform non-sequential classifiers overall. This is probably due to their ability to leverage information about sequential structure and preceding labels. Furthermore, Zubiaga et al. (2018) found that sequential classifiers did not benefit from contextual features representing thread context (e.g. similarity to the source tweet) whereas non-sequential classifiers did. They speculate that sequential classifiers take the surrounding context into account implicitly. To see if this also holds true for relevance classification in medical social media, we will compare the addition of conversation-aware features to both sequential and non-sequential models.

3 Methods

3.1 Models

CRF As a sequential model we use Conditional Random Fields (CRF). We train the models using the implementation in sklearn-crfsuite. L1 and L2 regularization parameters were tuned for each fold.

Linear SVM As a non-sequential counterpart, we use the sklearn implementation of Linear Support Vector Machines. The hyper-parameter C is tuned per fold with a grid of 10^{-3} to 10^{3} in steps of $\times 10$.

Feature type	Name	Description	Explanation (if applicable)
Local	+Emb	Sentence Vectors	We use Universal Sentence Encoder (USE) (Cer et al., 2018) to encode sentences into 512 dimensional vectors based on pre-trained embeddings so their cosine similarity (normalized between 0 and 1) approximates their semantic similarity.[2]
	+BERTpred	distilBERT predictions	The raw confidence scores for each label
Relational	+PrevSim	Similarity to previous message	Similarity is calculated using the USE sentence vectors
	+ThreadSim	Thread similarity	Similarity to USE sentence vector of all other posts in the thread combined into one vector
Positional	+Dist	Absolute distance from start of thread	
Label distribution	+PrevLbl	Label of previous post	We use the true labels for training and the predicted labels for testing for all label distribution features.
	+CountPos	Absolute running count of preceding positive labels in thread	
	+CountNeg	Absolute running count of preceding negative labels in thread	
	+RelPos	Percentage of preceding positive labels	
	+DistPos	Distance from previous positive label	
	+DistNeg	Distance from previous negative label	
Discourse	+DA	Dialogue act of post	Dialogue acts are calculated using the Dialogue Act tagger as trained by Tortoreto et al. (2019)
	+PrevDA	Dialogue act of previous post	

Table 1: Manually engineered features to model conversational structure

DistilBERT As BERT model, we opt for DistilBERT (distilbert-base-uncased), which is a lighter, more computationally efficient variant of BERT (Sanh et al., 2019). We use the Huggingface implementation (Wolf et al., 2019) with the wrapper ktrain (Maiya, 2020) to train our models. The initialization seed is set to 1. We use the default learning rate of 5×10^{-5} and tune the number of epochs (3 or 4) per fold.

Ensemble models To investigate the benefit of adding a sequential model on top of the DistilBERT model, we experiment with a blending-based ensemble method: we input the raw confidence scores from DistilBERT for each label as features in a CRF model (i.e. CRF + BERTpred). We create an equivalent non-sequential baseline by using the same approach with an SVM (i.e. SVM + BERTpred).

3.2 Feature analysis

To explore the benefit of manually engineered features that capture thread context, we use step-wise greedy forward feature selection using the features in Table 1. For each step-wise iteration, we select the best feature to add to the model until the F_1 score no longer improves. We use 10-fold cross-validation in which per fold features are selected on the development data (10%) and tested on a held-out test set (10%). For a fair comparison, we keep folds and hyper-parameters the same as for the respective base model. Since the label distribution features could leak information, we omit these gold annotated features for evaluation. Instead, we perform an initial run without these features and use the resulting predictions to calculate them for the final evaluation.

3.3 Model comparison

We used 10-fold cross validation in all experiments. Instead of splitting per message, we split on whole discussion threads to ensure possible dependencies between posts do not bias the outcome. Statistical comparisons of model performance are done using Wilcoxon signed rank tests across the 10 folds. To avoid the multiple testing problem, we only compare the three best models – namely those with the highest F_1 score, precision and recall – to the BERT baseline.

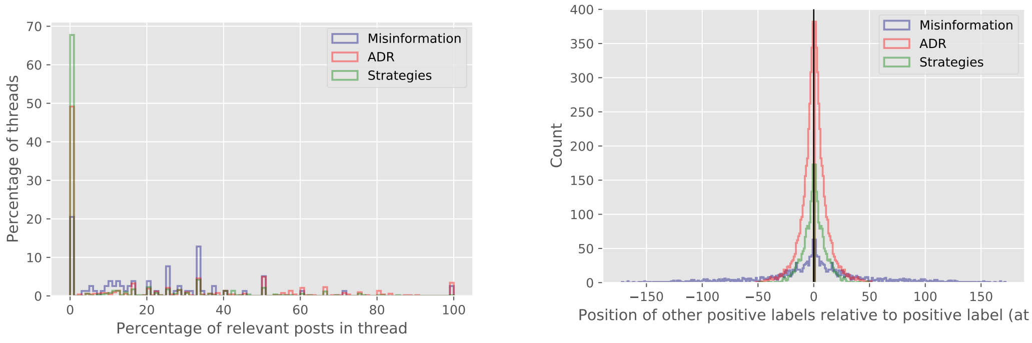
4 Data

Data collection At present, there is only one publicly available medical relevance classification data set that includes the conversational structure: the Medical Misinformation Data set (Kinsora et al., 2017).

[2] We opt for USE instead of BERT embeddings, as cosine similarity cannot be applied directly to BERT embeddings

Data set	Target	#Posts	#Discussions	Median length	% Positive
Medical Misinformation Dataset (Kinsora et al., 2017)	Misinformation	1,566	78	8.0	15.0 %
ADR Discussions (In-house)	Adverse Drug Response (ADR) & Coping Strategies	4,195	527	6	22.9 % & 12.3%

Table 2: Statistics on the data sets. The ADR Discussions data set has two target classes.



(a) Distribution of target posts *across* threads

(b) The relative position of target posts to each other *within* threads

Figure 1: Distribution of the target class (i.e. positively labelled posts)

It is based on MedHelp data and annotated for the presence of misinformation. We collected a second data set from a Facebook group of Gastro Intestinal Stromal Tumor (GIST) patients. We selected 527 discussions based on their likelihood to contain an ADR: We selected the threads that contained (1) at least one drug name according to a match with RxNorm (U.S. National Library of Medicine, 2020) and (2) a high percentage of posts in which authors shared experiences. The latter criterion was included since sharing that you had an ADR is an example of experience sharing. To estimate this, we used a previously developed classifier (Dirkson et al., 2019). According to our classifier, at least 80% of the posts within each selected thread is a personal experience. Due to privacy issues and ownership of the data by the GIST International patient organization, we are not able to share this data set at present. See Table 2 for more details on the data sets.

Data annotation Following a pilot annotation round, the data was annotated by the first author and three patients for the presence of ADRs and coping strategies for dealing with ADRs (hereafter also called: Strategies) using an annotation guideline.[3] The pair-wise inter-annotator agreement was substantial for ADR (mean κ =0.71) and moderate for Coping Strategies (mean κ =0.54).

5 Results

5.1 Distribution of the target class in the discussion threads

As visualized in Figure 1a, the target class is not distributed equally across the discussion threads for any of the data sets; There appear to be many threads with few or no target posts. According to z-tests, the distribution is significantly different from normal. An inspection of the relative position of target posts *within* discussion threads reveals that target posts also cluster together (see Figure 1b). The probability that the post after a target post is also a target post is 27% for Misinformation and 40% and 34% for ADRs and Coping Strategies respectively. These probabilities are higher than is to be expected based on the percentage of positively labelled posts (see Table 2). Thus, it appears that the conversational structure is indeed related to the probability of a post being relevant and consequently incorporating conversational structure or discourse may be able to improve performance of relevance classifiers.

[3] Available at: `https://github.com/AnneDirkson/ConversationAwareFiltering`

	Misinformation		
	F_1	P	R
BERT	0.366 ± 0.155	0.386 ± 0.154	0.396 ± 0.235
SVM+Emb	**0.478 ± 0.083**	0.492 ± 0.109	0.482 ± 0.111
+ *Features*	0.392 ± 0.089	0.457 ± 0.169	0.405 ± 0.156
CRF+Emb	0.424 ± 0.155	**0.565 ± 0.148**	0.352 ± 0.162
+*Features*	0.457 ± 0.137	0.557 ± 0.155	0.420 ± 0.167
SVM + BERTpred	0.443 ± 0.078	0.449 ± 0.082	0.479 ± 0.151
+*Features*	0.454 ± 0.070	0.449 ± 0.081	**0.492 ± 0.140**
CRF + BERTpred	0.434 ± 0.079	0.453 ± 0.100	0.447 ± 0.138
+*Features*	0.428 ± 0.078	0.435 ± 0.092	0.446 ± 0.126

(a) Misinformation data set

	ADR				Strategies		
	F_1	P	R		F_1	P	R
BERT	**0.714 ± 0.034**	0.715 ± 0.038	0.718 ± 0.062	**BERT**	**0.581 ± 0.060**	0.622 ± 0.087	**0.563± 0.111**
SVM+Emb	0.640 ± 0.054	0.673 ± 0.055	0.613 ± 0.069	**SVM+Emb**	0.517 ± 0.101	**0.660 ± 0.111**	0.434 ± 0.111
+*Features*	0.610 ± 0.068	0.621 ± 0.087	0.624 ± 0.128	+*Features*	0.502 ± 0.108	0.603 ± 0.137	0.453 ± 0.128
CRF+Emb	0.654 ± 0.059	0.710 ± 0.036	0.611 ± 0.086	**CRF+Emb**	0.441 ± 0.134	0.597 ± 0.120	0.373 ± 0.151
+*Features*	0.638 ± 0.067	0.695 ± 0.037	0.601 ± 0.110	+*Features*	0.512 ± 0.106	0.609 ± 0.110	0.462 ± 0.143
SVM + BERTpred	**0.714 ± 0.035**	0.724 ± 0.043	0.707 ± 0.056	**SVM+Bertpred**	0.578 ± 0.059	0.632 ± 0.091	0.545 ± 0.089
+*Features*	0.677 ± 0.121	0.673 ± 0.164	**0.738 ± 0.103**	+*Features*	0.561 ± 0.095	0.601 ± 0.146	0.552 ± 0.087
CRF+ BERTpred	**0.714 ± 0.038**	**0.728* ± 0.040**	0.704 ± 0.062	**CRF + BERTpred**	**0.581 ± 0.065**	0.629 ± 0.087	0.558 ± 0.115
+*Features*	0.713 ± 0.039	0.726 ± 0.040	0.705 ± 0.060	+*Features*	0.573 ± 0.058	0.635 ± 0.090	0.539 ± 0.100

(b) ADR and Strategies data set

Table 3: Evaluation results of mean model performance over 10 folds. Features are selected through step-wise greedy feature selection. **<0.01 *<0.05

5.2 Model comparison

The results of model evaluation are presented in Table 3. It appears that neither the addition of a sequential layer nor manual features can improve upon the F_1 score of the BERT model. Misinformation detection appears to be the exception to this; any additional layer, sequential or not, outperforms the BERT baseline model. The highest F_1 is attained by an SVM model based on USE sentence vectors (+Emb), which were specifically designed for representing whole sentences. Perhaps sentence vectors perform better than BERT embeddings when the BERT model performs poorly (F_1= 0.366). Additional research will be necessary to substantiate this.

Despite a lack of improvement in the F_1 score for the detection of ADR and Strategies, an additional layer does seem to offer flexibility in tailoring the model towards a higher recall or precision. On the one hand, recall can be improved for two target classes by adding a non-sequential SVM layer with manual features to the BERT model. On the other hand, precision can be improved through the addition of a sequential CRF layer on top of BERT predictions for all target classes. Adding manually engineered features in addition to the sequential layer only improves the precision further for the detection of coping strategies. Our findings are thereby in line with Zubiaga et al. (2018). They speculated that sequential classifiers may take the surrounding context into account implicitly and therefore do not benefit from features representing thread context.

The only significant increase according to Wilcoxon signed rank tests is in the precision for ADR detection. This may be related to the high variance between folds. Further research is necessary to validate these results and advance our understanding of how conversation-aware modelling can be best be used for relevance classification. We believe that this first study shows that this is a promising direction.

5.3 Analysis of selected features

There is large variation in which features are selected per fold. Manual inspection of the selected features shows that features relating to the distribution of labels in the thread are chosen most often, especially the running count of negative and positive labels in the thread (CountNeg, CountPos), and the label of

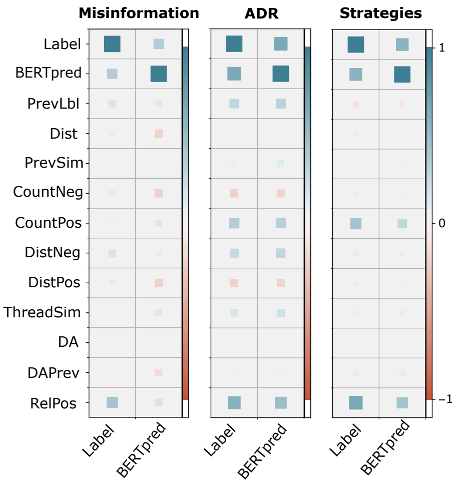


Figure 2: Correlation matrix of ground truth labels and BERT predictions with the manually engineered features. The size and colour of the squares corresponds to the strength of the correlation

the previous post (PrevLbl) (see Table 1). Features of this type may therefore be the most promising for future work. The number of features that is chosen is more consistent; On average, 1 or 2 of the 11 features are chosen.

To further explore why certain features are chosen, we compute the correlations between the target label and the manually engineered features and between the BERT predictions and the manually engineered features (see Figure 2). We find, firstly, that features relating to the label distribution indeed appear to correlate most strongly with the ground truth labels. Secondly, the correlation between these features and the BERT predictions is often equal to or stronger than the respective correlation to the ground truth. This might indicate that this variance is already captured by the BERT model and therefore manually engineered features have little to add to the baseline model.

6 Discussion

We find that the distribution of target posts across discussion threads is skewed and that within a conversational thread posts cluster together. Thus, our hypothesis that the probability of a target post occurring is related to the conversational structure appears valid.

In answer to **RQ1**, we find that a sequential CRF layer on top of a BERT model improves precision slightly, although only significantly so for ADR detection. In answer to **RQ2**, we find that the addition of manually engineered features representing thread context often does not aid performance. The one consistent exception is when combined with a non-sequential SVM layer on top of a BERT model. This combination can improve recall for all target classes, although not significantly. An additional layer on top of a BERT model that is able to capture the thread context appears to offer flexibility in tailoring the model towards a higher recall or precision. In future work, we plan to investigate the benefit of including conversational context for other tasks such as concept normalization of ADR.

For all the data sets included in this study, a pre-selection of discussion threads was made prior to annotation to ensure a higher proportion of target posts. We expect that both sequential models and manually engineered features of thread context may prove more beneficial when such a pre-selection does not take place and the target class is even more imbalanced. Thus, our results may be an underestimation of the benefit of conversational context for finding ‘needles in the haystack’.

Finally, our findings call into question the practice of splitting data into folds without taking the discussion context into account. In this study, we split the folds per discussion thread and we recommend others to consider doing so when dealing with multiple posts from the same thread, as neglecting to do so when there are dependencies between posts may bias model performance. This is especially important when threads contain duplicate posts.

Acknowledgements

This work was funded by the SIDN fonds. We would like to thank our annotators for their time and effort and our anonymous reviewers for their feedback.

References

John Langshaw Austin. 1962. *How to do things with words*. Oxford university press.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2019. Narrative Detection in Online Patient Communities. In A. Jorge, R. Campos, A. Jatowt, and S. Bhatia, editors, *Proceedings of the Text2StoryIR'19 Workshop*. CEUR-WS.

Graciela Gonzalez-Hernandez, Abeed Sarker, Karen O 'Connor, and Guergana Savova. 2017. Capturing the Patient's Perspective : a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of medical informatics*, pages 214–217.

Alexander Kinsora, Kate Barron, Qiaozhu Mei, and Vinod Vydiswaran. 2017. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In *IEEE International Conference on Healthcare Informatics*.

Ari Klein, Arjun Magge, Karen O'Connor, Haitao Cai, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. A chronological and geographical analysis of personal reports of covid-19 on twitter. *medRxiv (preprint)*.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. eventAI at SemEval-2019 Task 7: Rumor Detection on Social Media by Exploiting Content, User Credibility and Propagation Information. In *Proceedings ofthe 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 855–859.

Zhiheng Li, Zhihao Yang, Ling Luo, Yang Xiang, and Hongfei Lin. 2020. Exploiting adversarial transfer learning for adverse drug reaction detection from texts. *Journal of Biomedical Informatics*, page 103431.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv*, arXiv:2004.10703 [cs.LG].

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic, September. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

Abeed Sarker, Karen O'connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Safety*, 39.

Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. 2020. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *Journal of the American Medical Informatics Association*, 27(8):1310–1315, 07.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Giuliano Tortoreto, Evgeny A Stepanov, Alessandra Cervone, Mateusz Dubiel, and Giuseppe Riccardi. 2019. Affective Behaviour Analysis of On-line User Interactions: Are On-line Support Groups more Therapeutic than Twitter? In *Proceedings ofthe 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 79–88.

U.S. National Library of Medicine. 2020. RxNorm.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the Fourth Social Media Mining for Health (#SMM4H) Shared Task at ACL 2019. In *Proceedings ofthe 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Transformers: State-of-the-art Natural Language Processing. *ArXiv*.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-Aware Rumour Stance Classification in Social Media Using Sequential Classifiers. *Information Processing & Management*, 54(2):273–390.