Appendix A: Correlation Matrix

Table A.1. Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
$adrPredictionsSVM_{i,t}$	-																				
$adrPredictionsBERT_{i,t}$	0.90*	-																			
$FAERS_{i,t}$	0.09*	0.12*	-																		
$professionals_{i,t}$	0.14*	0.17*	0.59*	-																	
$consumers_{i,t}$	0.05*	0.07*	0.95*	0.33*	-																
$substitute_{i,t}$	0.02*	0.05*	0.01	0.02*	0.00	-															
$clinicalTrials_{i,t}$	0.02*	0.03*	0.06*	0.09*	0.03*	-0.03*	-														
$news_{i,t}$	0.29*	0.31*	0.08*	0.09*	0.05*	0.02*	-0.02*	-													
$marketSize_{c,t}$	0.05*	0.04*	0.01	0.01	0.01	0.16*	0.07*	0.03*	-												
$numRecalls_{j,t}$	0.01*	0.00	0.02*	0.03*	0.02*	0.11*	0.11*	-0.05*	0.04*	-											
$numDrugs_{j,t}$	0.04*	0.04*	-0.01	-0.01	-0.00	0.12*	0.03*	0.01	0.07*	0.50*	-										
$competition_{j,t}$	0.04*	0.08*	0.04*	0.06*	0.03*	0.43*	0.02*	0.01	0.20*	0.13*	0.21*	-									
$complexity_{j,t}$	-0.01	-0.02*	0.02*	0.03*	0.02*	0.00	0.06*	-0.02*	-0.01	0.09*	0.11*	-0.05*	-								
$abuse_i$	0.06*	0.03*	-0.02*	-0.03*	-0.01*	0.02*	0.02*	0.05*	0.20*	0.04*	0.10*	-0.00	-0.03*	-							
$adverseReactions_i$	0.00	0.01	0.02*	0.03*	0.01*	-0.07*	0.22*	0.03*	0.01	0.05*	0.06*	-0.04*	0.09*	-0.02*	-						
$warnings_i$	0.02*	0.03*	0.00	0.01	-0.00	-0.07*	0.21*	0.03*	0.01	0.06*	0.07*	-0.06*	0.08*	-0.02*	0.91*	-					
$boxedWarnings_i$	0.04*	0.04*	0.10*	0.14*	0.06*	-0.05*	0.11*	0.01	0.06*	-0.01	0.00	0.03*	0.06*	0.19*	0.34*	0.34*	-				
$priorityReview_i$	0.02*	-0.02*	0.02*	0.04*	0.02*	-0.42*	0.04*	0.01*	-0.12*	-0.11*	-0.21*	-0.53*	0.04*	0.02*	0.09*	0.08*	0.09*	-			
$orphanDrug_i$	-0.03*	-0.04*	0.01	0.01	0.00	-0.16*	0.03*	-0.02*	-0.05*	-0.08*	-0.12*	-0.26*	0.04*	-0.05*	0.01*	0.02*	0.02*	0.47*	-		
$wideDistribution_i$	0.03*	0.02*	-0.00	-0.00	0.00	0.03*	0.02*	0.00	0.03*	0.10*	-0.05*	0.13*	0.01	0.05*	-0.00	-0.03*	-0.05*	-0.10*	-0.13*	-	
$approvalType_i$	-0.03*	-0.06*	-0.07*	-0.09*	-0.05*	-0.40*	-0.01	-0.03*	-0.17*	-0.01*	-0.05*	-0.69*	0.11*	-0.04*	0.14*	0.15*	-0.04*	0.46*	0.14*	0.00	-

Note. * $p < 0.05$

Appendix B: Text Classification for ADR Detection

B.1. Support Vector Machine (SVM)

We follow the procedure proposed by Sarker and Gonzalez (2015) to build an SVM classifier for ADR detection. First, we preprocess social media posts by lowercasing, stemming, and parsing all the tokens. Next, we extract the following features from the preprocessed tokens.

- *Unified Medical Language System (UMLS) semantic types and concept IDs (CUIs).* We use MetaMap toolbox[1] to identify terms that are related to the UMLS semantic types and the CUIs, which represent categories of medical concepts. To reflect the importance of a term, we compute the term frequency-inverse document frequency (TF-IDF) values for this set of features.
- *Syn-set expansion.* For each adjective, noun, or verb in a sentence, we use WordNet[2] to identify the synonyms of that term and add the synonymous terms as features. Similar to the previous feature set, we calculate the TF-IDF value for each derived synonym.
- *Change phrases.* To capture the polarity of the text, we use the lexicon proposed by Sarker et al. (2013) to derive the following four binary features: MORE-GOOD, MORE-BAD, LESS-GOOD, and LESS-BAD. For example, if a MORE-word and a GOOD-word both occur in a text, we denote the feature MORE-GOOD as one. The other three features are constructed in a similar manner.

[1] See http://metamap.nlm.nih.gov/

[2] See http://wordnet.princeton.edu/

- *ADR lexicon matches.* To incorporate domain-specific knowledge, we use the lexicon proposed by Leaman and Wojtulewicz (2010) to derive two features measuring ADR mentions in the text. The first feature is a binary feature indicating the presence of ADR mentions and the second feature is a numeric feature indicating the ratio of ADR mentions to the total number of words in the text segment.
- *Sentiword scores.* We use the lexicon proposed by Guerini et al. (2013) to assign each term with a sentiment score between -1 to 1 and then calculate the overall sentiment of a text by summing the individual scores divided by the length of the sentence in words.
- *Topic-based feature.* We use a topic modeling technique called Mallet[3] to generate two features: the topic terms and the sums of all the relevance scores of the terms.
- *Other features.* We include three simple features, including the length of the text segments in words, the presence of comparatives and superlatives, and the presence of modals.

After creating such a comprehensive set of features, we split a data set into a training set (80%) and a test set (20%) and then conduct a grid search with 10-fold cross-validation over the training set to find the best hyperparameters, such as the kernel function, the regularization parameter, and the weight assigned to each class. Next, we build an SVM classifier supervised by the training set and evaluate it on the test set. We report a standard set of performance measures for evaluation, such as accuracy, area under the ROC curve (AUC), precision, recall, and F_1 score.

B.2. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language representation model that pre-trains deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al. 2018). Due to the ability to learn contextual relations between words in a text, the BERT model has achieved state-of-the-art performance on multiple NLP tasks, such as the General Language Understanding Evaluation (GLEU) task sets, Stanford Question Answering Dataset (SQuAD), and Situations With Adversarial Generations (SWAG).

There are two steps to implement a BERT model: *pre-training* and *fine-tuning*. In the pre-training stage, the BERT model is trained on unlabeled data over two tasks: masked language model, the objective of which is to predict the original token of a masked token based only on its context; and next sentence prediction,

[3] See http://mallet.cs.umass.edu/index.php

the objective of which is to understand the relationship between two sentences. Devlin et al. (2018) provides two pre-trained models in their seminal work: $BERT_{BASE}$ (12 layers, 768 hidden vector size, 12 heads, and 110 million parameters) and $BERT_{LARGE}$ (24 layers, 1024 hidden vector size, 16 heads, and 340 million parameters). In the fine-tuning stage, the BERT model is first initialized with the pre-trained parameters, and then all parameters are fine-tuned using labeled data from the specific NLP task.

In this study, we consider the $BERT_{BASE}$ pre-trained by Devlin et al. (2018) as a starting point and utilize transfer learning to fine-tune the neural network with the annotated social media posts. More specifically, we split the data set into a training set (80%) and a test set (20%) and fine-tune the BERT model with the training set. Next, we evaluate the fine-tuned BERT model on the test set and report a standard set of performance measures.

Appendix C: Summary Statistics

Table C.1. Summary Statistics on Different Classes of Drug Recalls

Variable	No. of obs.	Mean	S. D.	Min.	Max.
Class I & II Recall					
$adrPredictionsSVM_{i,t}$	18,652	5.72	24.27	0	879
$adrPredictionsBERT_{i,t}$	18,652	5.07	18.36	0	562
$professionals_{i,t}$	18,652	21.42	49.00	0	3275
$consumers_{i,t}$	18,652	20.62	109.29	0	10525
Class III Recall					
$adrPredictionsSVM_{i,t}$	12,396	7.33	32.19	0	572
$adrPredictionsBERT_{i,t}$	12,396	5.54	21.02	0	392
$professionals_{i,t}$	12,396	15.04	33.64	0	1985
$consumers_{i,t}$	12,396	17.30	144.06	0	6762

Table C.2. Summary Statistics on ADR-related Tweets

Variable	Mean	S. D.	Definition
adrPredictionsSVM			
$adrPredictionsSVMTwitter_{i,t}$	4.63	24.49	Number of tweets that mention drug i's proprietary name and are predicted by SVM to be ADR-related at month t
$avgLogFollowers_{i,t}$	1.14	2.25	Average of logarithm-transformed number of followers for drug i at month t
$avgLogLikes_{i,t}$	0.01	0.10	Average of logarithm-transformed number of likes for drug i at month t
$avgLogRetweets_{i,t}$	0.01	0.09	Average of logarithm-transformed number of retweets for drug i at month t
adrPredictionsBERT			
$adrPredictionsBERTTwitter_{i,t}$	3.38	15.53	Number of tweets that mention drug i's proprietary name and are predicted by BERT to be ADR-related at month t
$avgLogFollowers_{i,t}$	1.30	2.52	Average of logarithm-transformed number of followers for drug i at month t
$avgLogLikes_{i,t}$	0.01	0.10	Average of logarithm-transformed number of likes for drug i at month t
$avgLogRetweets_{i,t}$	0.01	0.10	Average of logarithm-transformed number of retweets for drug i at month t

Note. Number of observations is 31,048.

References

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Guerini M, Gatti L, Turchi M (2013) Sentiment analysis: How to derive prior polarities from SentiWordNet. *arXiv preprint arXiv:1309.5843* .

Leaman R, Wojtulewicz L (2010) Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* 117–125.

Sarker A, Gonzalez G (2015) Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics* 53:196–207.

Sarker A, Molla D, Paris C (2013) Automatic prediction of evidence-based recommendations via sentence-level polarity classification. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 712–718.