Predicting adverse side effects of drugs using Twitter data

Yomna Genina

September 23, 2017

Abstract

Adverse drug reactions (ADR's) are among the leading causes of death in the United States [4]. Due to the shortcomings of ADR identification during the drug development process, in addition to a lack of post-market patient reporting - almost 90% of cases go unreported[2] - exploring alternative sources such as social media platforms with millions of posts per day is growing in popularity. This paper focuses on analyzing data-mined drug and side effect relationships extracted from Twitter by R. Eshleman and R. Singh [2016, BMC Bioinformatics]. Drug and side effect links from the SIDER database were taken as ground truth[5],[6]. This paper shows that machine learning models trained on Twitter data can predict already known adverse side effects found in SIDER with an accuracy of up to 85%, with a precision and recall of 86% and 90% respectively, leading to and F1 measure of 88% . In addition to that, it was shown that when compared to semantic features, topological features have a higher information gain, providing a stronger capability to distinguish adverse side effects from non-adverse ones. Nevertheless, semantic features increased predictive accuracy by about 3%, and it was shown that effect context features were overall more informative than drug context features.

1 Introduction

Adverse drug reactions (ADR's) are between the fourth and sixth leading causes of death [1]. Not all ADR's are identified during the drug development process, though Various attempts have been made to allow patients to report ADR's post-usage, such as MedWatch, 90% of cases still go unreported [2]. Recently, dispersed information in the form of clinical documents, electronic health records, and online review data has been publicly accessible, and so research was performed to gather all of the aforementioned publicly available information into one database that is easy to analyze. One such database is called SIDER, which contains information about 1430 drugs, 5868 side effects and 139756 adverse drug and side effect relationships. Nevertheless, this database might not be up-to-date, and so further research into extracting information from social media platforms is growing in popularity. One of the most popular social media platforms used for such purposes is Twitter which contains over 280 active users and 500 million tweets per day. This paper aims to analyze data-mined drug-effect relationships extracted from Twitter by R. Eshleman and R. Singh [2016, BMC Bioinformatics][1], to show that known ADR's on SIDER can be predicted by training a model on data extracted from Twitter. Five topological features were calculated for each drug-effect link, in addition to the 42 semantic features which come with the Twitter data, making a total of 47 features. As a result, this paper also aims to explore which features contain the highest capability of distinguishing between adverse and non-adverse side effects.

2 Data Summary

The Twitter data contains 263 drugs, 363 side effects, and 2766 drug-effect links of which 37% can be found on SIDER and hence labeled as *adverse*, while the rest labeled as *non-adverse*. Each drug-effect link has the following semantic features: one drug sentiment context, one effect sentiment context, 20 drug topics, and 20 effect topics (read more about them here[1]). Additionally, using the Twitter data to create a bipartite network with drugs and side effects as nodes and drug-effect links as edges, 5 topological features were calculated for each existing drug-effect link by using 5 different neighborhood-based similarity measures[7]:

[1] R. Eshleman and R. Singh, *Leveraging Graph Topology and Semantic Context for Pharmacovigilance through Twitter Streams, BMC Bioinformatics, Vol. 17, (Suppl 13):335, 2016.*

1. Adamic-Adar Index(d,se) = $\sum_{s \epsilon N(d) \cap N(se)} \frac{1}{\log_2(|N(s)|)}$
2. Jaccard Index(d,se) = $\frac{|N(d) \cap N(se)|}{|N(d) \cup N(se)|}$
3. Common Neighbours(d,se) = $|N(d) \cap N(se)|$
4. Resource Allocation(d,se) = $\sum_{s \epsilon N(d) \cap N(se)} \frac{1}{|N(s)|}$
5. Preferential Attachment(d,se) = $|N(d)| * |N(se)|$

, where N(d) and N(se) are the sets of 3^{rd} degree neighbours of a drug node and a side effect node respectively.

3 Method

3.1 Analyzing Feature Distributions

Afterwards, the topological feature distributions over adverse and non-adverse links were analyzed separately and compared:

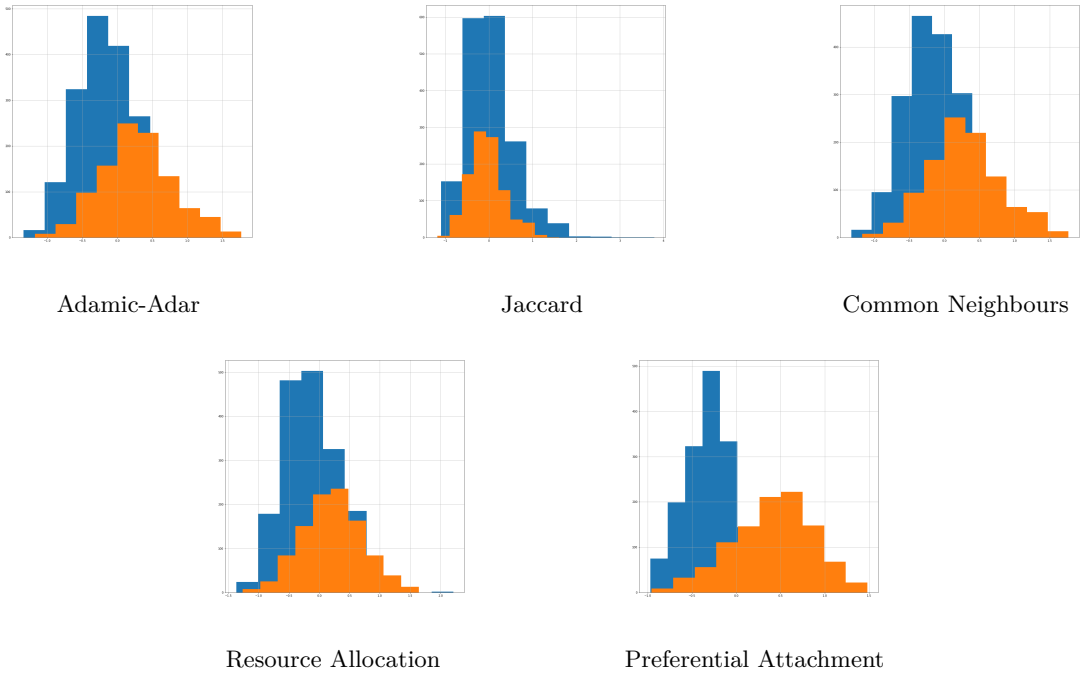
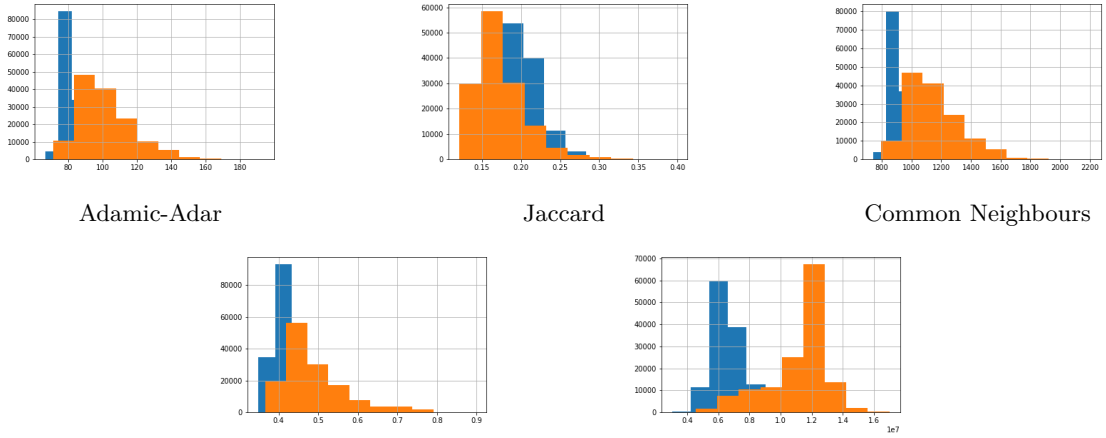


Figure 1: Showing Twitter's topological feature distributions over adverse (**orange**) and non-adverse links (**blue**)

As seen from figure 1, preferential attachment seems to have the highest partitioning capability between adverse and non-adverse classes. Next, adamic-adar, common neighbours, and resource allocation seem to give similar partitioning capabilities, followed by jaccard which isn't able to distinguish much the difference between adverse and non-adverse. As a result, we expect that feature importance will rank in the same way for our future machine learning model.

The SIDER data was augmented upon by inserting edges between drugs and side effects that weren't there at the start, and labeling them *non-adverse*. If the number of drugs is d, side effects se and the number of existing edges is $|e|$, then the number of missing edges is $d * se - |e|$. From the set of missing edges, a subset of length $|e|$ was chosen randomly, and then augmented into the SIDER data with a label of *non-adverse*. The 5 topological features were then calculated and their distributions grouped by class:



Adamic-Adar | Jaccard | Common Neighbours

Resource Allocation | Preferential Attachment

Figure 2: Showing SIDER's topological feature distributions over adverse (**orange**) and non-adverse links (**blue**)

Similar to Twitter's feature distributions in figure 1, figure 2 shows that preferential attachment showed the highest distinguishing power between adverse and non-adverse examples in SIDER.

3.2 Training a Random Forest Model

Firstly, the Twitter data was partitioned 75% and 25% for training and testing respectively. Within the training set, stratified 10-fold cross validation with 10 repeats to average across was used with a random forest model of 500 trees. All 47 features were used to fit this model, and feature importance was calculated and ranked:

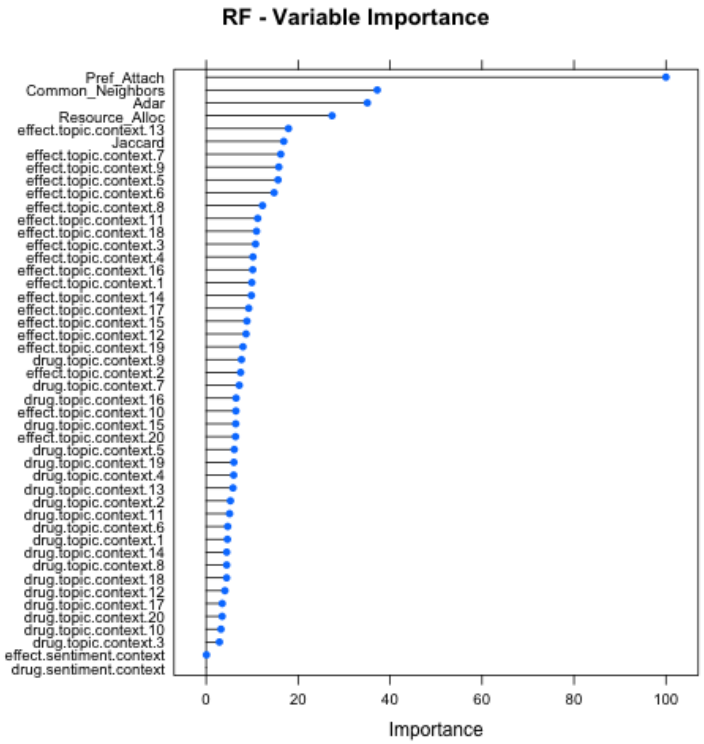


Figure 3: All 47 features ranked by how many times the model used them to branch out the decision trees

As expected from figure 1, figure 3 shows that preferential attachment ranked the highest in

importance, while jaccard the lowest amongst the topological features, and semantic features ranked lower than topological features overall. However, looking at semantic features alone, it can be seen that effect topic features were used more often than drug topic features.

3.3 Testing the RF Model

In order to find how many of the top features give the best generalization when tested on unseen data, 47 different RF models were trained such that the first model contains information about only the first top feature while the second contains the top 2 features and so on. All 47 models were the tested on the unseen test set and their F1 values, overall accuracies, and kappa accuracies compared:

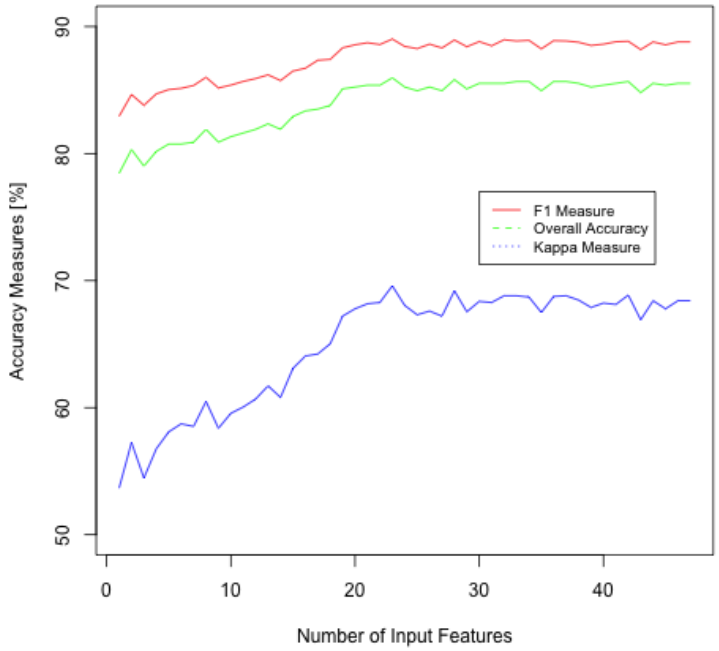


Figure 4: number of top features fed to the RF model vs. their F1, overall, and kappa accuracies on the test set

As shown in figure 4, we can cut the number of features fed into the RF model by half and attain a slightly higher predictive accuracy than when using all 47 features which causes a slight overfitting. In addition to that, we can see that using the top 4 topological features alone can achieve an F1 accuracy of 85%, however adding a subset of the semantic features increases the F1 value to 88%.

4 Conclusion

Twitter data can indeed be used to predict already known adverse side effects found on SIDER, which means that it can be used in the future to predict adverse side effects that are not yet on SIDER. In addition to that, topological features seemed to be more useful and have a high information gain for our Random Forest model, the best of which was preferential attachment. However, when semantic features were added on top of the topological features, accuracy increased by around 3%.

For the future, it could be useful to search for more discriminative topological features than just the 5 that were tested in this paper, as well as checking if modifying the number of neighbour degree when calculating the topological features would change our topological feature distributions between the adverse and non-adverse groups.

References

[1] Lazarou J, Pomeranz BH, Corey PN. *Incidence of Adverse Drug Reactions in Hospitalized PatientsA Meta-analysis of Prospective Studies.* JAMA. 1998;279(15):1200–1205. doi:10.1001/jama.279.15.1200

[2] Cieliebak, Mark, Dominic Egger, and Fatih Uzdilli. *Twitter can help you find Adverse Drug Reactions.*ERCIM NEWS

[3] R. Eshleman and R. Singh, *Leveraging Graph Topology and Semantic Context for Pharmacovigilance through Twitter Streams*, BMC Bioinformatics, Vol. 17, (Suppl 13):335, 2016.

[4] S. Katragadda, H. Karnati, M. Pusala, V. Raghavan and R. Benton, *Detecting adverse drug effects using link classification on twitter data*, 2015 IEEEpp. 675-679.

[5] Kuhn M, Letunic I, Jensen LJ, Bork P. *The SIDER database of drugs and side effects. Nucleic Acids Res.* 2015 Oct 19. doi: 10.1093/nar/gkv1075

[6] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. *A side effect resource to capture phenotypic effects of drugs* Mol Syst Biol. 2010;6:343. Epub 2010 Jan 19.

[7] G, Alanis-Lobato, *Mining protein interactomes to improve their reliability and support the advancement of network medicine*, Frontiers in Genetics, Vol. 6, 2015. doi: 10.3389/fgene.2015.00296