CSE 6250: Big Data in Healthcare Project Report

Predicting Drug Recalls using drug reviews

Andrew Kogler, Arjun Chintapalli, Qin Peng, Gaurav Falia
Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA
andrew.kogler@gatech.edu, arjun.ch@gatech.edu, gfalia3@gatech.edu, qin.peng@gatech.edu

Video Presentation URL: https://youtu.be/CW-nGC5s1vI

***Abstract*—Manufacturers of consumer products across many industries all face the risk of recalls due to defects, unintended side-effects, among other reasons. These recalls impose high-cost implications that motivate the quality control processes to minimize the already low risk of a recall even further. This risk of recalls is even more pronounced in the drug manufacturing industry where recalls can be attributed to unintended side-effects and liability and magnitude of relevant lawsuits is enormous compared to that of say children toy manufacturers. Including organizationally driven internal quality control procedures, early detection and forecasting of product/drug recalls could incur huge cost savings.**

Our work intends to scan and parse drug review comments on a drug review site (drugs.com), for which we have a dataset, learn common review and textual features that are correlated with drugs that in the future underwent a recall, and attempt to utilize this in predicting the likelihood of recalls for other drugs shortly after their release where reviews are streaming input data. This specifically is a new area of study given the ubiquity of product review data, which is relatively unexplored outside of a few papers from no more than two years ago.

Keywords—NLP, FDA Recall

I. Introduction

Drug manufacturers and pharmaceutical companies produce numerous drugs every year. They must advance through many stages of review going from research & development, preclinical research, the four stages of clinical trials, FDA review, and then finally post-market review where the product is publicly available to consumers but monitored to ensure quality is maintained. Even in light of this drug recalls still occur given the sheer amount of drugs produced, and the high variability in consumer's bodies reactions to the drugs that can be severe and common enough to pose a greater market-wide health risk. This is understandable to a degree due to the complex nature of the system interactions between a patient's body and organs, the illness that the drug is targeted to alleviate, and the drug itself.

Even with significant quality control measures in place, another way of reducing risk, so far unexplored based on our understanding, is to look for early indicators that the drug is causing side-effects that may lead to recall right after the drug has been introduced to the market. The potential for side-effects to be observed only when being available mass market and having not been seen in clinical trials is significant, given the discrepant sample sizes of participants in clinical trials compared to that of the mass market.

A way of accomplishing this is to treat drug reviews on websites like drugs.com as streaming input and sources of feedback which could serve as statistical predictors of the likelihood that a drug could experience a recall from the FDA. Given the ubiquity of review sites in recent past, there is an increasing amount of valuable information that if isolated can address problems such as this.

The problem is many folded, in that the review text data is semi-structured, can contain mis-spellings, varying degrees of ambiguity, and can be even off-topic, in discussing parts of the patients experience that are at best tangentially related to the drug in question. Due to the lack of strict adherence to the topic in question, one cannot simply rely on sentiment analysis to predict the likelihood that a drug will be recalled, in that a effective drug could carry negative reviews simply for the reason of making the consumer tired, or hungry for example.

II. Background and Related Work

In tackling the specific problem of FDA recalls, the most similar work appears to be that of Elad Yom-Tov [1] in which Yom-Tov uses internet search queries via Bing from a time period in 2015, to see if those precede subsequent FDA recalls. This approach experiences around 20% recall in classifying drugs as likely to undergo a recall the day before, with an AUC of 0.791 implying a relatively high precision rate.

Slightly less similar work but still of general interest, Zhang et al [2] discusses the feasibility of using consumer comment board data to forecast likely vehicle part recalls. Their method uses "Smoke Words" filtration [3] to isolate comments that are highly likely to be discussing product defects as opposed to other negative experiences they might have had that are unrelated to the functioning of parts but relate to use of their vehicle in some other way. The group's best performing classifier, that being K Nearest Neighbors experienced recall approximating 0.5.

One of the primary causes of drug recalls are adverse drug reactions (ADR's), something that Huang et al [4] focused more specifically on. They used review data, ICD-10 codes associated with patients taking the drug in question, as well as some variables stemming from biomolecular network relationships and gene annotation information to augment the forecasting of these negative outcomes for patients. In this case, it appears Support Vector Machines produced the best results outperforming similarly trained Logistic Regression model variants.

Kastrin et al [5] produced related work on ADR prediction, specifically those driven by drug-drug interactions (DDI) which are a subset of all ADR's in general. In looking at pharmacological descriptors such as proteins and the chemical structure of the drug itself, the group of researchers were able to have moderately success in forecasting the likelihood of DDI's which make up a disproportionate number of drug recalls.

A quality survey of the problem of predicting ADR's using all methods established within the academic literature was conducted by Ho et al [6]. In it they list a wealth of data sources and datasets that have been investigated towards this end including SIDER, Drugbank, Pubchem, RxNorm, Matador, etc. which gives a good impression of the sheer scope of candidate datasets outside of what we're using, those simply being just the drugs.com UCI dataset, and the FDA historical enforcement reports recalls data.

Dandala et al [7] looked at the viability of using NLP techniques on medical data to predict long-term outcomes of patients via data from a longitudinal patient study. Although not the most similar work to our own, it offered good use-cases of textual analysis techniques that we've considered using in our own work, namely the usefulness in finding topic similarity features to find relevant reviews.

Lastly, Qiao et al [8] served a similar purpose by expanding on the use of latent-drichlet allocation (LDA) which is another means of performing topic modeling and relations which is another potential improvement we may incorporate if we have the time to do so.

III. Datasets

The datasets we used are publicly available, sourced from the U.S. Food & Drug Administration's (FDA) website and drugs.com via UC-Irvine who hosts a data repository commonly used for machine learning tasks within the academic community.

The FDA data specifically comes from the Recall Enterprise System [9]. We specifically filtered this database for recalls that were related to Drugs, and used all that were returned which included recalls going back to the beginning of 2016 until November 2018. This served as our universe of target labels as they ultimately serve as informal ground truth for what drugs are safe versus not.

The drugs.com data is managed by a research group at the University of California - Irvine [10][11]. This was assembled by researchers in the lab group, and is a common benchmark dataset to be used when performing research on sentiment analysis, rating prediction, etc. on customer review data.

To isolate the subset of data this is usable to attempt solving this problem, we filtered only for review data and FDA recall data where the product exists in both such, and where the review comes before the recall instatement date.

Performing this drug linkage across datasets, yielded us 777 pairs of drugs along with relevant reviews before the recall window. With respect to feature source records, this filtered 207,038 different review instances we can use to construct features from to then feed into our supervised algorithm to forecast likelihood of FDA recalls.

IV. General Approach

Our approach consists of applying common Natural Language Processing (NLP) techniques to the drugs.com candidate drug review text data and metadata, in addition to deriving features of the text itself. The textual features include the count of keywords of drugs that have been recalled via keyword clustering using K-Means, sentiment of the review, in as well as the subjective nature of the post itself. Our intuition is that reviews that exhibit strong negative sentiment are deeply indicative of experiences with drugs that may have caused side-effects to the reviewer. Also, our hypothesis is that reviews that are more personalized could be more highly correlated with drugs that would soon thereafter experience a recall.

Getting to the stages of applying the above techniques occurs after many steps of involving matching the drugs as they are physically described in the FDA dataset and how they're described in the drugs.com dataset where the drug name is a colloquial description as opposed to a drug unique-identifier that would make matching trivial.

With the above approach we were able to create a new dataset consisting of reviews only for drugs that have been recalled. We then further filtered these reviews by only considering recalled drugs that have more than 10 reviews.

We then split this dataset into a train and test dataset, reserving 80% for training, and 20% for test, maintaining the ratio of classes across the two sets.

Having found the target-label association between the drugs in the review dataset and the FDA recall dataset (target being 1 if the product experienced a recall, 0 otherwise), we then proceed to clean the review data removing stop words, replacing all HTML tags that are present, any hyperlinks, word tokens only composed of numbers, punctuation characters, etc.

We then take these reviews having already split them into the two groups based on target label, having cleaned the reviews following the above steps, and then run sentiment and subjectivity analysis of the reviews via the nltk.sentiment.vader library. We then have a tuple of two values, the first representing the sentiment score which is ranged from [-1.0, 1.0] where the sign of the score dovetails with negative and positive sentiment respectively, and the subjectivity score ranged from [0.0, 1.0] where a 1.0 reflects a completely subjective document, and 0.0 reflects the most objective document possible.

Figure 1: Data Pipeline

We intend to explore the performance of multiple models for prediction, the main models being Neural Networks, both convolutional, and recurrent, utilizing existing code from the last homework in which we implemented wrappers around the PyTorch library. We will focus significantly on hyper-parameter optimization using a mix of manual space-search, in addition to automated space search via Bayesian Optimization and/or GridSearch. Feature engineering will certainly play a role as it is iterative in nature and even if there is not total re-derivation of new metrics, there will be transformations to existing features to scale them, or lessen their role in model fitting.

V. Exploratory data analysis

Some familiarity with the two core datasets and how they can be interpreted and used towards solving our problem is crucial to developing a working solution of value.

With respect to our FDA recalls dataset universe, there are 6,821 different recall events which span years going from 2012-2018 with the latest recalls going up to October 22nd which was the latest produced when extracted. In this time-range many different drugs undergo recalls, with many drugs experiencing multiple recalls over time. For example, Bupivacaine underwent many recalls from various companies/distributors in every year from 2012-2016. The scale of these recalls varies quite highly as well, in addition to the severity. As one can imagine, sometimes small shipments are mislabeled causing a recall, other times an ingredient isn't named in the labeling incurring a total recall of all shipments that have been sent up until that point.

In looking at the companies that had recall sanctions brought against them, and specifically those who've experienced more than one, we saw a few companies that were present and over again in the recall universe during this time window. The pie-chart following this paragraph is composed of companies that fall in this category, with any explicitly named company having >= 2 recalls, and all others making up the light-blue section for which there are the majority of companies that had only one recall.

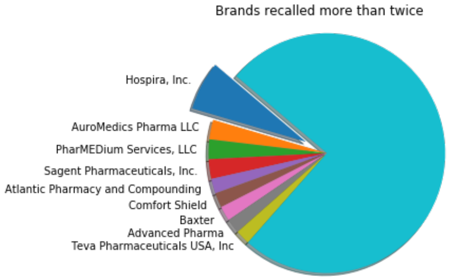


Figure 2: Pie Graph of Common Recalled Brands

We conducted a similar initial investigation into the drugs.com UCI dataset to gather a further understanding into how content could serve to predict subsequent recalls of drugs. One of the most pertinent things we wanted to know was the "sparsity" of feature data in that we would want a long historical view of patients experiences from the drug.

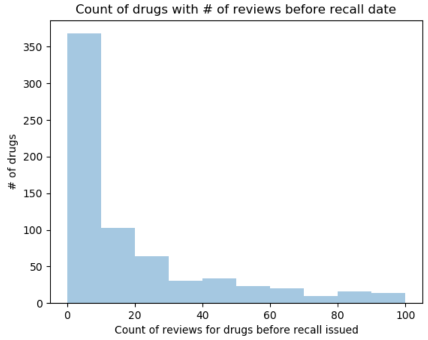


Figure 3: Histogram of Recalled Drugs binned by Review Count

In looking at the count of reviews before a recall instantiation per drug, and plotting it in a histogram we can come to understand how much of a problem feature source data and it's sparsity would pose. The single most common review count bucket was 0-10 reviews, which was true for over 350 drugs and their respective recalls. This counted for just under half of the labeled instances, or 383/777 (49.29%) with the remainder having ten or more reviews.

VI. Results

1. Development Phases

A. Phase 1:

Our very first iteration of the model started with a basic feature set which relied on the stock features already provided from UCI Drugs.com dataset, along with some naive numerical features computed based on the review text.

The features provided from drugs.com are the numerical rating of the patients experience with the drug, ranged from [1, 10] where 10 indicates the most ideal drug interaction, and 1 the worst. Another mappable feature is that of the useful count which is a count of unique users who found the review useful when surveying the drug for a particular diagnosis they have, which can serve as a proxy of importance or quality of the review itself.

The model types generated within this phase consisted of a decision-tree classifier, and a multi-layer perceptron with minimal parameter tuning/optimization. This effort was focused on getting benchmark performance statistics which we intended to improve upon subsequently.

B. Phase 2:

The second iteration of the models were constructed using more sophisticated features based on sentiment which were derived using NLTK VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment, along with TextBlob sentiment polarity and subjectivity analysis.. VADER is a lexicon-based and rules-based sentiment tool which has been specifically constructed to appropriately detect sentiment that is communicated via social media. As the review data is similar given it's textual and where users record reviews which are expressed in a similar fashion to the way they communicate via social media outlets, this is a valid extension and use-case of the technique.

VADER scoring is done via first computing a negative, neutral, and positive vector for each text review where the vector sums to 1.0 which resembles the total polarity components of the review. This is then aggregated to compute a normalized, weighted-compound score for the entire textual review, which is then thresholded to assign it to logical outcomes of positive, neutral, and negative. The ranges for negative are from [-1, -.05], neutral from (-.05, .05), and positive from [+.05, +1.0]. We retain all four features as part of our second iteration of models.

Next, the TextBlob analysis computes a slightly different sentiment polarity score, along with a separate measure which is the subjectivity score. The sentiment score computed by TextBlob is roughly equivalent to the composite sentiment score coming from VADER which is also ranged from [-1.0, +1.0]. The subjectivity score, which is a measure of how opinionated or personalized the review which is ranged from [0.0, +1.0] where 0.0 would be true for the most objective text body possible, where +1.0 would indicate the most subjective text feasible.

C. Phase 3:

In the third iteration, we examined and decided to leverage the Stanford CoreNLP open-source library for sentiment analysis. Similar to sentiment polarity from TextBlob and VADER analysis. The Stanford CoreNLP sentiment analyzer is actually querying a recurrent neural network on the backend that is the best current model available where annotations are sourced on a daily basis from movie reviews, restaurant reviews, social media posts, etc.

In this phase we also experimented merging the two sub-models together into an ensemble model as ensemble methods usually outperform individual models. We used weighted average soft-labeling so that we don't consider the two models equal where on most performance metrics the decision tree seems superior to the MLP.

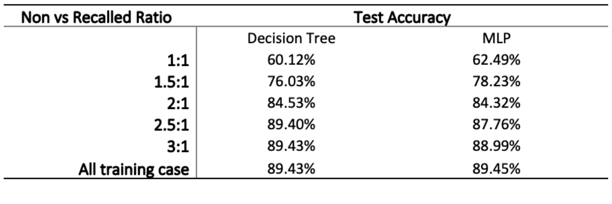
2. Undersampling

Since our dataset is unbalanced, with 192370 reviews (89.45%) for non-recalled cases and 22693(10.55%) for recalled cases. We needed to preprocess the data before training our model. There are several strategies for learning from unbalanced data, which can fall into five categories: undersampling, oversampling, cost-sensitive learning; ensemble learning; and combination. For simplicity, we used undersampling for our training set.

However, undersampling lowered the test accuracy in our benchmark study, which is shown in the below table. The results suggest that we should try some of the other methods for unbalanced data, while also using AUC as the evaluation metric.

In this phase, we also experimented more heavily with parameter tuning/optimization to see how well boosted the model performance could be. Parameters that were most modified were the number of layers in the multi-layer perceptron, which solver to use in the MLP, which criterion to use in the decision tree, and how many layers to use in the tree.

Table 1: Test accuracy with undersampling ratios



Non vs Recalled Ratio	Test Accuracy	
	Decision Tree	MLP
1:1	60.12%	62.49%
1.5:1	76.03%	78.23%
2:1	84.53%	84.32%
2.5:1	89.40%	87.76%
3:1	89.43%	88.99%
All training case	89.43%	89.45%

3. Benchmark models

In order to compare the performance of our sentiment feature model, we first tested models utilizing the predefined “rating”, “useful count” features available per review. We then constructed these raw text features for each review: "review_length", "review_word_count", "review_cleaned_word_count", "review_avg_word_length", "review_avg_cleaned_word_length".

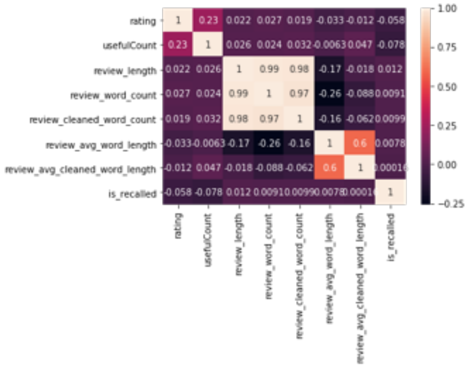


Figure 4: Similarity Matrix of Benchmark Features

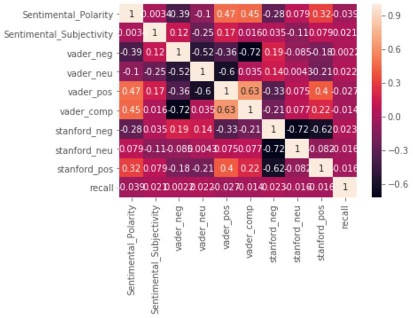


Figure 10: Sentiment feature PCA analysis

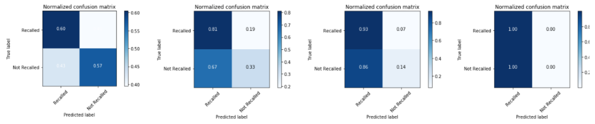
We plot above the similarity between these calculated features we found on our training set. This figure shows that most of these features are independent of each other. Of primary interest in these features are the “rating” and “useful count” features.

We then chose two popular but relatively simple models to test these feature sets: a multilayer perceptron, and a random forest model.

For both neural-based model, and tree-based model, we varied the undersampling rate and calculated the confusion matrix at each undersampling rate.

2.1 Random Forest

The less undersampling we did, the better accuracy we achieved. However, the best decision tree model we got is the model that classify all cases into the majority class, with an accuracy of 89.45%.

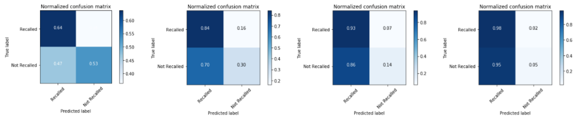


Non Vs Recalled: 1:1 **Non Vs Recalled: 1.5:1** **Non Vs Recalled: 2:1** **Non Vs Recalled: 2.5:1**

Figure 5: Benchmark Metrics for Decision Trees

2.2 Multilayer perceptron

Second, we trained a 4-layer fully-connected neural network model with Same undersampling ratios as the Decision Tree model. Again, the undersampling actually lower our model accuracy, and with lower undersampling rate, our model prone to predict all cases into non-recalled ones, with a highest 89.45% accuracy.



Non Vs Recalled: 1:1 **Non Vs Recalled: 1.5:1** **Non Vs Recalled: 2:1** **Non Vs Recalled: 2.5:1**

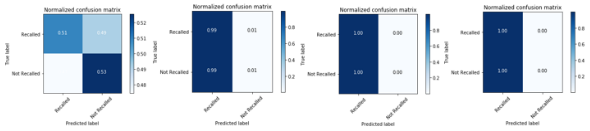
Figure 6: Benchmark Metrics for MLPs

4. Sentiment models

We then trained a multilayer perceptron, and a random forest model on our constructed sentiment features, subjectivity and polarity we calculated for each review. We then calculated the same metrics as our benchmark given the same random initialization.

3.1 Random Forest

Similar to before, the less undersampling we did, the better accuracy we achieved. However, the model performance is much worse then when using the ratings and raw review metrics.

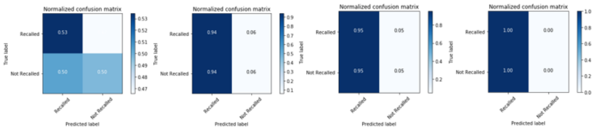


Non Vs Recalled: 1:1 **Non Vs Recalled: 1.5:1** **Non Vs Recalled: 2:1** **Non Vs Recalled: 2.5:1**

Figure 7: Sentiment Metrics for Decision Trees

3.2 Multilayer perceptron

Second, we trained a 4-layer fully-connected neural network model with same undersampling ratios as before. Again, the performance was much worse than the benchmark.

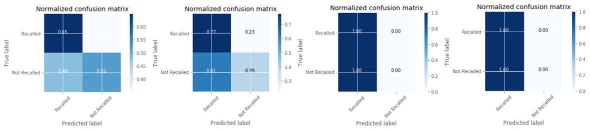


Non Vs Recalled: 1:1 **Non Vs Recalled: 1.5:1** **Non Vs Recalled: 2:1** **Non Vs Recalled: 2.5:1**

Figure 8: Sentiment Metrics for MLPs

3.3 Benchmark and Sentiment Random Forest

To see if we can achieve better results by combining the feature sets of both the sentiment and review metric features. We present these results for just the Random Forest case for brevity.



Non Vs Recalled: 1:1 **Non Vs Recalled: 1.5:1** **Non Vs Recalled: 2:1** **Non Vs Recalled: 2.5:1**

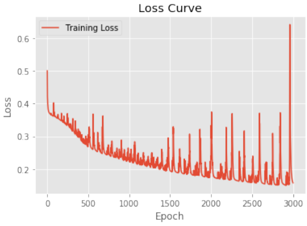
Figure 9: Combined Metrics for Decision Trees

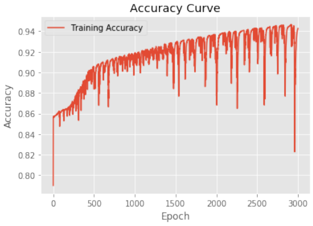
As shown above, adding the sentiment metrics to the raw review metrics did improve performance.

4. FINAL MODELS WITH ALL SENTIMENT FEATURES

4.1 MULTILAYER PERCEPTRON

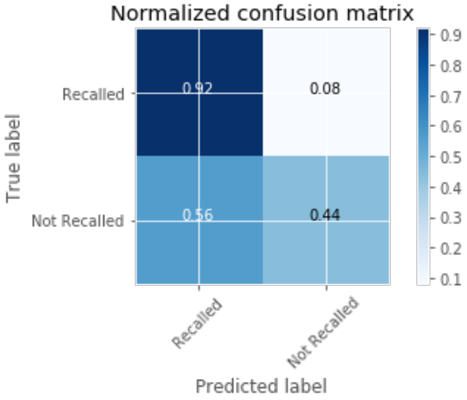
After including all features including those from previous stage along with those derived from Stanford CoreNLP, we have the following model performance plots to display. The learning and loss curves can be seen below:



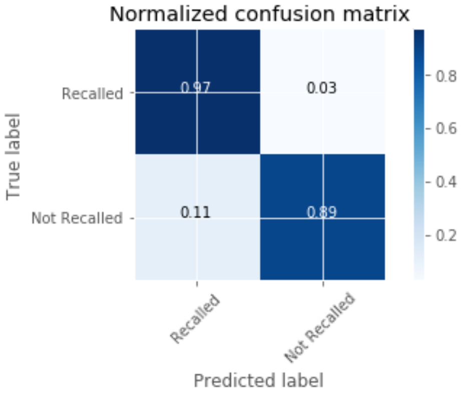


One can see the general trends are in improving directions but with significant spikes along the curve.

Lastly, the confusion matrix for this best MLP implementation can be seen below:



4.2 DECISION TREE



4.3 ENSEMBLE METHOD

The last model we experimented with was an ensemble learner that would aggregate the results of both the multi-layer perceptron and

In ours study, the best results (both accuracy and precision) are achieved by Random forest, which is a discriminative model using ensemble learning method. The Generative Models, the simple Tensorflow MLP model and Keras ensemble MLP model in general are inferior to the discriminative model. But with the ensemble method, the Keras ensemble MLP increased the precision dramatically from 0.39 to 0.81. The reason under that may be there are many unincluded factors which have huge impacts on the state of it being recalled. Due to missing important features, the generative models is not complete enough for complete generation.

5. FINAL RESULTS

Our best model performance metrics for each of the three different model types we experimented with can be seen below:

Model Name	Accuracy	Precision	Recall	F1-Score	ROC
Decision Tree	0.96	0.89	0.83	0.83	0.93
MLP	0.87	0.39	0.44	0.41	0.68

Ensemble	0.81	0.81	0.45	0.45	0.50

VII. DISCUSSION/FUTURE WORK

The initial run we accomplished with the sentiment features leaves a lot of room for improvement. We expect to build off our initial work by creating additional features from our subjectivity and polarity features. We plan on building a sentiment feature that gives sentiment from reviews only if they are relevant.

We also tried to create feature vectors based on TF-IDF statistics trained on the collection of recalled product reviews, but we dropped this method on concerns that this method will lead to bias towards recalled products. We will consider revisiting this approach after we find methods of mitigating this bias.

Finally, we will vary the neural network architecture to achieve better performance. We are planning on creating a RNN network to see if this will improve performance as well keep running experiments to find the optimal size of layers to improve performance.

We are also planning on using more metrics to evaluate our model performance in addition to our current use of confusion matrices and accuracy metrics, by also calculating AUC, precision and recall metrics as well.

After perfecting this sentiment feature based approach, this model can then be generalized to include other datasets such as Twitter data to expand the amount of training data available.

VIII. CONCLUSION

We were able to take a dataset of drug reviews that was originally used for NLP analysis on the review text to predict given review ratings as well as manually labelled ratings of side effects and effectiveness [11] to predict known drug recalls as documented by the FDA [9]. Although we have found that the predefined ratings and raw review text metrics is enough to predict recalls with high accuracy (89%) we hope to improve on this performance by using sentiment features from the review text.

Although our initial sentiment analysis produced poor results, the advantages of using a sentiment based approach are readily apparent. Using a sentiment based approach frees a data scientist from only using datasets with predefined features such as ratings. Thus the sentiment based approach can be generalized to solely text datasets such as say Twitter or search queries. Thus, our findings could then be applied to predict in real time the probability of a drug being recalled, thereby shortening the timespan a drug will be on market causing adverse side effects.

REFERENCES

[1] Yom-Tov, Elad. “Predicting Drug Recalls from Internet Search Engine Queries.” Arxiv.org, 2016, arxiv.org/abs/1611.08848.

[2] Zhang, Xuan, et al. “Predicting Vehicle Recalls with User-Generated Contents: A Text Mining Approach.” SpringerLink, Springer, 19 May 2015, link.springer.com/chapter/10.1007%2F978-3-319-18455-5_3.

[3] Abrahams, A., Jiao, J., Wang, A. and Fan, W. (2012). Vehicle defect discovery from social media. [online] Mba.vt.edu. Available at: https://mba.vt.edu/content/dam/mba_vt_edu/Abrahams-defect-research.pdf [Accessed 17 Nov. 2018].

[4] Huang, Liang-Chin et al. "Predicting Adverse Side Effects Of Drugs". Ncbi.Nlm.Nih.Gov, 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3287493/pdf/1471-2164-12-S5-S11.pdf. Accessed 17 Nov 2018.

[5] Kastrin, Andrej et al. "Predicting Potential Drug-Drug Interactions On Topological And Semantic Similarity Features Using Statistical Learning". Journals.Plos.Org, 2018, https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0196865&type=printable. Accessed 19 Nov 2018.

[6] Ho, Tu-Bao et al. "Data-Driven Approach To Detect And Predict Adverse Drug Reactions". Jaist.Ac.Jp, 2014, http://www.jaist.ac.jp/~bao/papers/ADRdraft.pdf. Accessed 17 Nov 2018.

[7] Dandala, Bharath et al. "Scoring Disease-Medication Associations Using Advanced NLP, Machine Learning, And Multiple Content Sources". Aclweb.Org, 2016, http://www.aclweb.org/anthology/W16-5114. Accessed 19 Nov 2018.

[8] Qiao, Zhilei et al. "A Domain Oriented LDA Model For Mining Product Defects From Online Customer Reviews". Core.Ac.Uk, 2017, https://core.ac.uk/download/pdf/77239685.pdf. Accessed 19 Nov 2018.

[9] United States Food & Drug Administration. “Recalls, Market Withdrawals, & Safety Alerts.” U S Food and Drug Administration Home Page, Center for Drug Evaluation and Research, 2018, www.fda.gov/Safety/Recalls/default.htm.

[10] Kallumadi, Surya, and Felix Gräßer. “Drug Review Dataset (Drugs.com) Data Set.” UCI Machine Learning Repository: Flags Data Set, 2018, archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+(Drugs.com).

[11] Gräßer, Felix et al. "Aspect-Based Sentiment Analysis Of Drug Reviews Applying Cross-Domain And Cross-Data Learning". Kdd.Cs.Ksu.Edu, 2018, http://kdd.cs.ksu.edu/Publications/Student/kallumadi2018aspect.pdf. Accessed 19 Nov 2018.