Exploring Sentiment Analysis on Online Medical Drug Reviews

Arjun Soin
Stanford University
Stanford, CA
asoin@stanford.edu

Aditya Khandelwal
Stanford University
Stanford, CA
akhand@stanfod.edu

Abstract

In this paper, we describe several deep learning frameworks that aim to extract sentiment from online medical drug reviews. We use a dataset that agglomerates pharmaceutical drug reviews from a diverse range of websites, available at University of California Irvine's Machine Learning Repository (1). We describe baseline models that were used to compare our various deep learning neural network architectures with. Moreover, this paper goes over evaluation metrics, a brief description of the dataset and thoughts for future work in the area.

1 Introduction

Pharmaceutical product safety currently depends on clinical trials and specific test protocols. These studies are typically conducted under standardized conditions on a limited number of test subjects in a limited time span. As a consequence, the discrepancies in patient selection and treatment conditions can have significant impact on the effectiveness and potential risks of adverse drug reactions (ADRs). Post-marketing drug surveillance, i.e. pharmacovigilance, plays a major role concerning drug safety once a drug has been released. More recently, with the advent of e-commerce and modern drug delivery mechanisms, pharmacovigilance has taken the form of peer-to-peer drug recommendation systems that rely on user reviews to understand ADRs and suggest potential alternatives to a certain medical drug.

Moreover, patients use the same online peer-to-peer recommendation and review mechanisms to inform their decisions such as understanding the alternatives to one or more drugs, likely ADRs associated with pharmaceutical products, overall satisfaction and effectiveness of a drug and pricing. Since some of these decisions can have potentially life-altering consequences, from a Natural Language Understanding (NLU) perspective, it becomes extremely important to generate useful insights from such reviews that can be used by patients and big pharmacuetical companies alike.

Online reviews are generally accompanied by a rating on the *Likert Scale*, a way of measuring reliability of products using a 1-10 rating system. For our research, we are concerned with analyzing the sentiment of online drug review dataset and understanding the sentiment's relationship with the user rating on the Likert Scale. We assume that the rating accompanying a review is directly correlated to the positiveness of the sentiment of a review.

Our core hypothesis is that we can utilize recent advances in machine learning to understand the relationship between the sentiment of an online user review and compare it to several metrics such as user rating and usefulness of review to create a classifier that assigns scores to each review based on its likelihood of it being negative, positive or neutral.

1.1 Applications

As mentioned earlier, the applications of understanding sentiment in online drug reviews can be beneficial for both consumers and enterprise. Better pharmacovigilance leads to higher quality of drugs, more transparency in the side effects of drugs and greater decision making information to individual users of the drugs. Due to the recent groundbreaking research in transfer learning, our process and models can be easily applied to data from a diverse range of review websites to understand the sentiment for products even outside of the pharmaceutical industry.

1.2 Problem Statement

For our project, we wanted to implement a deep learning model in PyTorch and use it to test accuracy on trimodal sentiment classification of online

drug reviews. Our input consists of a corpus of textual reviews and we wanted to generate an output class - positive, neutral or negative for each review, as described above. Moreover, we wanted to tune our hyperparameters to yeild better results on the validation set in order to see and increased accuracy on the test set.

1.3 Assumptions

We assumed that the rating that accompanies a drug review is an accurate measure of the sentiment of the user writing the review. Therefore, without loss of generality, we decided to follow the following rubric:

1. Rating between 1 to 4: Negative
2. Rating between 4 to 7: Neutral
3. Rating over 8: Positive

Overall, we see that the assumption holds true when we randomly sampled a few reviews and tried to judge each review into one of the three categories by hand.

2 Related Work

2.1 Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning (2)

In this paper, Kallumadi et al. (2018) explore sentiment analysis techniques to examine online user reviews in the pharmaceutical field, specifically about drugs. The dataset used by the authors is created by scraping the internet for drug reviews. After acquiring the dataset, the authors perform sentiment analysis on the corpora. Furthermore, they present transfer learning as a viable option to tackle the challenge of missing annotated data, and cross-domain sentiment analysis. The authors use Cohen's Kappa as the metric of evaluation. They observe very high Kappa scores during in-domain analysis which implies that sentiment analysis is a good way to understand the ADRs associated with drugs that are used for a specific condition. However, lower Kappa scores are observed during their analysis of cross-domain and cross-data sentiment with the use of aspect-based transfer learning. The authors that aspect-based sentiment analysis requires more extensive, structured datasets to extract features with sufficient generalization capabilities.

2.2 Sentiment Analysis of Pharmaceutical Products Evaluation Based on Customer Review Mining (3)

In this paper, Mahboob and Ali (2018) use data-mining and sentiment analysis to understand the quality of pharmaceutical products, not necessarily just drugs, and the satisfaction of the people using these products. The authors use techniques of sentiment analysis in the context of pharmacological sciences to classify patient reviews and experiences as positive, negative or neutral. They employ a lexicon-based approach to measure the performance of a pharmaceutical products with grading functionalities. A pre-trained model called Sentistrength (which has implementations in both Python and R) was used to give a sentiment strength score to each string of text that the model was evaluated on. This score was compared to the annotated score given to each review by the authors, and precision was measured as the metric of evaluation. The authors conclude that certain medical categories, such as eye and skin treatments, yield better precision scores than other categories, such as dehydration and women's health.

2.3 Sentiment Analysis of User-Generated Content on Drug Review Websites(4)

This paper develops and presents an effective method for sentiment analysis of user-generated content on drug review websites. The aim is to leverage the unprecedented amount of user-generated content in the form of blogs, discussion forums, user review web sites, and social networking sites in domain that is more specific than conventional research avenues (movies, restaurants, product reviews). This is where the health and medical domain steps in - the authors claim that users are often looking for stories from patients like them on the Internet and that online community support can have a positive impact, giving rise to a need to further delve into an under-researched domain like drug reviews to inform the design and conception of online tools and applications. They also believe that the application of the proposed sentiment analysis approach will be useful not only for patients, but also for drug makers and clinicians to obtain valuable summaries of public opinion.

Towards this end, the authors develop a sentiment analysis algorithm at the clause-level

specificity since each sentence can well contain multiple clauses touching multiple aspects of a drug. The method then assigns sentiment as positive, negative or neutral of each clause from well-defined sentiment scores assigned to words. The authors then use MetaMap to map various health and medical terms, such as disease and drug names, to semantic types in the Unified Medical Language System (UMLS) Semantic Network to incorporate domain-specific knowledge into the algorithm. Drug review sentences are collected from the drug review website WebMD (www.webmd.com) to evaluate the developed algorithm. The target drugs are mainly diabetes, depression, ADHD (Attention Deficit Hyperactivity Disorder), slimming pills, and sleeping pills. For the algorithm development, they use the development dataset prepared from the drug review website DrugsExpert. Furthermore, they use the Stanford NLP library (de Marneffe, 2006) to process the grammatical relationships of words in a clause. 2,700 clause-based experimental results are gathered to confirm that the algorithm performs better than a baseline machine learning approach through Support Vector Machine (SVM). An important component of the robustness of this paper's analysis is an error analysis on 829 error clauses that were misclassified by the sentiment algorithm. Based on the nature of errors, they categorize the sources of errors into seven groups, with an inference problem (relating to modality and indirect expressions) representing the single largest type of error.

2.4 Patient opinion mining to analyze drugs satisfaction using supervised learning(5)

The paper "Patient opinion mining to analyze drugs satisfaction using supervised learning" is grounded in a similar motivation as Na and Kyaing (2015) in that it aims to break into untapped domains for opinion mining in the form of health and medical research on user-generated content. This paper, however, aims to predict drug satisfaction level among other patients who already experienced the effect of a drug. The authors aim to establish a neural network based opinion mining approach as the an effective for classification method on reviews of two different drugs. The paper confirms that neural network based mining approach outperforms the support vector machine in terms of precision, recall and f-score.

They construct two datasets by collecting the reviews of two popular drugs, cymbalta (treating depressive disorder) and depo-provera (a form of female hormone used to prevent pregnancy) from the web resource www.askapatient.com that allows people to query for previous patient experiences. The idea behind using these two drugs is that since they differ in their purposes, the medical terminology and general language asscociated with them is likely to be different.

SVM with a polynomial kernel and most default values from the Weka tool (an open source project with various machine learning algorithms) is used as baseline. From there, the authors proceed with a probabilistic neural network (PNN) which is basically a statistical bayesian classification algorithm. This approach is then contrasted with a Radial basis neural network (RBF) which is a type of feedforward network. It is found that the performance of the RBF neural network method outperforms the PNN for each performance measure used, while both neural networks together outperform the SVM approach. Despite showing that RBFN could be a solution to constantly improve classification performance, the one area for more consideration is how to deal with the largely indirect opinion formulation in the drug review domain. Patients usually express opinions through drug effectiveness or side effects. Future research can focus on more categorically identifying indirect opinion and contrasting it with more direct formulations, especially considering statistical approaches were found to have lower performance than neural network for the two drug types in this paper.

2.5 Sentiment Analysis Tool for Pharmaceutical Industry Healthcare (6)

In this paper, authors Grissette, Nfaoui and Bahir explore drug reviews to identify breakpoints in public opinion as a factor for pharmaceutical companies to gain a competitive edge in competing drug markets. The authors are primarily concerned with online marketing and customer service as opposed to understanding ADRs of drugs or improving drug quality itself. The authors are interested in gaining insights from unstructured data using sentiment analysis.

The authors have documented common ways of doing sentiment analysis on online medical

reviews of drugs. Some of them include pre-trained models such as Sentistrength, Chatterbox, AlchemyAPI, etc. These models are going to be helpful for our preliminary research on sentiment analysis for the dataset that we are using.

The authors suggest four ways of understanding insights from online reviews: subjectivity classification, sentiment classification, polarity determination, and opinion feature extraction and product aspects extraction. They propose an architecture that comprises of collecting data using Facebook and Twitter. They then perform a lexicon-based sentiment analysis based on previous studies. The tool developed by the authors incorporates statistical learning methods and machine learning approaches also known as hybrid approaches to identify the suitable sentiment polarity. This tool involves methods from Alchemy API Datumbox, two Machine Learning APIs with Python build distributions.

The research paper also presents a case study of Novartis International, a Swiss pharmaceutical company in Morocco. Using a bag of words representation of tweets related to Novartis drugs, the authors were successfully able to identify pain points amongst customers across drug categories. The proposed tool was able to classify sentiments and emotions conveyed by users in real time as positive, negative and neutral, regarding the pharmaceutical industry. The author concludes that this makes sentiment analysis possible to establish other polarities that show more clearly personal emotions.

2.6 Baseline model implemented by Penn State's Nittany Data Labs(7)

We also looked at a Kaggle kernel set up by Penn State's Nittany Data Labs conducting sentiment analysis on a UCI ML drug review dataset as one of our sources. They attempt to solve how sentiment plays into rating and usefulness of reviews as well as what machine learning models work best for predicting the sentiment or rating based on review. Moreover, what is specially interesting to us is that this project also aims to shed light on whether the problem suited for classification or regression and if one should sort the reviews into categories based on sentiment or predict the actual rating of the review. A related task comes from finding insight into what features or words are most important for predicting review rating.

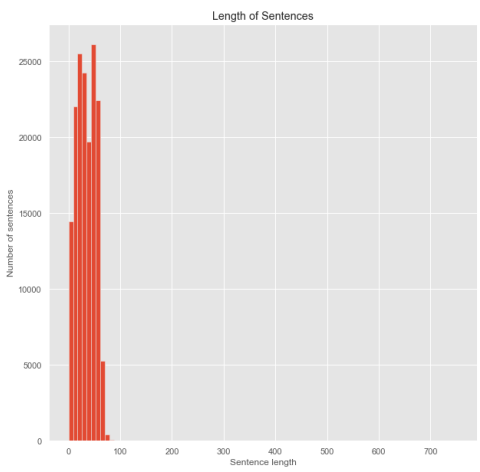
The dataset in question contains patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction. In treating the problem as a classification one, they first apply random forests for the purposes of sentiment analysis. The vectorization technique opted for is Term-Frequency Inverse Document-Frequency (TF-IDF) which works well for the chosen dataset despite only accounting for frequency and not relative position of word. The classification task is then extended using a neural network with Keras which helps scale with high-dimensional data. The authors find that the neural network gives the best overall accuracy with almost 90%, with related findings that softmax functions as the best activation for the output layer and a batch size of 128 and roughly 6 epochs to be optimal for their model. The keras approach is supplemented with feature analysis to derive importance of vectorized features by k-clustering by similarity, as keras has no built-in function to check for feature importance.

They also frame the problem in terms of a linear regression and predict score of the review with only 2.3 error. This is impressive considering how little preprocessing, feature engineering, hyperparameter tuning and model experimentation went into the model. It is concluded that the problem can be framed in terms of classification or regression depending on the preferred approach. Last but not least, exploring different NN architecture could have been beneficial, as certain types of neural nets are known to work very well for NLP problems.

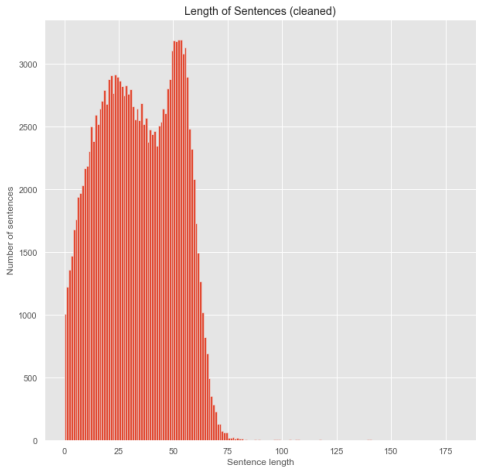
3 Data

We use a dataset agglomerated from pharmaceutical review sites available at University of California Irvine's Machine Learning Repository (1). The dataset provides patient reviews on specific drugs along with related conditions and a 20 star patient rating reflecting overall patient satisfaction. The dataset is further augmented by the number of likes each review received as a factor of its relative usefulness. The dataset is split into training (75%) and test (25%) sets. For each entry in the dataset, the following parameters are available to us:

1. Unique ID



(a) Data before outliers were removed



(b) Data after outliers were removed

Figure 1: Distribution of data in terms of length before and after outliers were removed

2. Name of the drug
3. Condition of the patient
4. Text review
5. User rating
6. Number of people who found the review useful
7. Date on which the review was posted

The dataset contains 215,063 combined reviews spanning 917 conditions. The average rating of all reviews is 6.99, which puts the total corpus in the neutral sentiment range. Several aspects of this dataset can be analyzed to understand the correlation between these different parameters. Importantly, we do not have information of how many people viewed the review since the dataset is not equally divided by condition.

4 Methods

We employ a variety of techniques to understand the nature of the data that we are working with, generate insights from the data that we use later in our work and preprocess data so that it is easier to use for training. Some of these methods are discussed in this section.

4.1 Data Anaysis

Initially, we decided to analyse the nature of the data we were working with. Basic data cleaning was undertaken, after which we generated a few important insights into the distribution of the ratings and usefullness statistic. We manipulated the dataset to represent confidence intervals for each condition in order to understand the condition-wise distribution of sentiment.

4.2 Removing & Replacing Outliers

First, we replaced all occurrences of contractions in reviews with the expanded forms of the word. For instance, the word 'isn't' was replaced by 'is not', and so on. Next, we removed all occurrences of special characters, emojis and punctuation so we only have to concern ourselves with textual data that derives its words from the English lexicon. Numerics were replaced by their english word translations. Finally, we decided to ignore long sentences (more than a 100 words) completely in order to keep our task easier, and not have to deal with multiple sentiments in the same review, since our analysis showed the existence of such outlier reviews.

4.3 Eliminating Medical Stopwords

In addition to using a corpus of regularly used English language stopwords, we created a list of commonly used medical drugnames in our dataset to expand the stopwords corpus. These stopwords were then categorically filtered out and eliminated from the training set in order to reduce the size of the data that we were training on and increase processing time significantly.

4.4 Tokenization & Stemming

We segmented reviews into tokens by splitting each review by word. Since we had already removed outliers and special characters from the reviews, tokenization helped us create feature vectors for our task and represent the information as a list of words instead of sentences or phrases. We then applied stemming to each word in a review to replace words with their root words in order to further reduce our feature space. We made the choice to use stemming instead of lemmatization, which is another way of producing root words for a word, because we were dealing with data that contained many medical terms that may not necessarily had common root words in English. However, stemming solves this issue by creating root words that may not exist in the grammar at all, if needed.

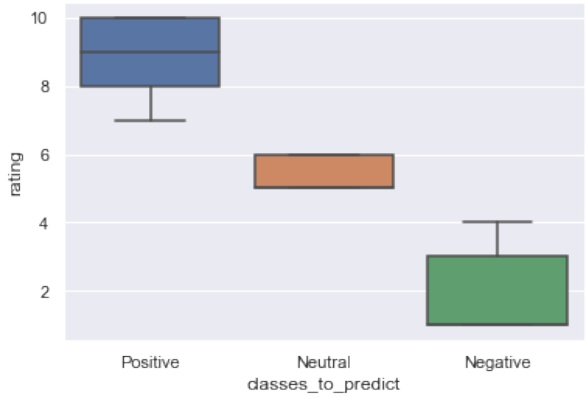


Figure 2: Division of training data based on sentiment

4.5 Sentiment Categorization

As described above, our assumption that the rating of a review is an accurate representation of the sentiment being represented by a patient meant that we followed the rubric mentioned above to categorize sentiment in our training and test sets into the three sentiment categories. Upon further inspection, we found that some of the reviews in the dataset were near-copies of each other for different medicines. While we expect this behavior from users, it was interesting to note that the same reviews had two different ratings. Since this was a big anomaly in our dataset, we removed such reviews altogether.

4.6 Train/Validation Split

Finally, before we started training, we split the given training set into two parts: 80% was dedicated to the training set and the remaining 20% was used as validation set. We tried to sample these datasets in such a way that the validation set had approximately the same sentiment distribution as the test set.

5 Experiments

5.1 Evaluation

We evaluated several models - both baseline and our original frameworks - on accuracy of classification of textual data into one of three classes mentioned before. Informally, accuracy is the fraction of predictions our models got right. Formally, accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{No. of Correct Predictions}}{\text{Total No. of Predictions}}$$

Our choice to optimize accuracy scores was influenced by the fact that if such models were applied in the real world, we would not care too much about false positives, but would want to increase correct predictions as much as possible.

5.2 Baseline Models

Inspired by some of the recent work in the field, our initial approach included applying several non-deep learning models to arrive at a baseline that we could use to compare and contrast our final model with. In this section, we describe and evaluate some of the models that we used.

1. **Multinomial Naive Bayes**
   This model uses simple conditional probability, namely Bayes' Rule, to classify each review into one of three sentiment categories. Since textual data has discrete features, such a word count, our Multinomial Naive Bayes Classifier was trained using tokenized review and the rating associated with each review. Upon testing, this classifier reported a 68.35% accuracy on the test set.

2. **Logistic Regression**
   This model is used when there are one or more independent variables that determine the prediction of a class. The goal is to fit the model to the textual data based on the way it is classified. In order to accomplish this task, we used four different ways of implementing logistic regression.

   (a) Naive implementation
   (b) Bag of Words

Model	Training Accuracy	Testing Accuracy	Parameters
Multinomial Naive Bayes	68.06%	68.35%	`alpha=1.0`
Logistic Regression (BoW)	76.66%	76.81%	`c=1.0`
Random Forest	66.27%	65.92%	`n_estimators=200`
Linear SVC	**79.13%**	**79.74%**	**`c=2.0`**

Table 1: Baseline Models Result Comparisons

(c) TFIDF + Bag of Words

(d) Word2Vec

Surprisingly, Logistic Regression with Bag of Words performed the best on our training and testing set. This model motivated us to implement more sophisticated baseline models since there was room for improvement from the 76.8% testing accuracy.

3. **Random Forest**
   A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. However, for our task, the random forest classifier yielded poor results with 65.92% testing accuracy.

4. **Linear SVC**
   The objective of a Linear Support Vector Classifier is to fit to the data, returning a best fit hyperplane that divides, or categorizes, data into many classes. This was the model we decided to use as our baseline since it provided a decent accuracy of over 79% on the testing data.

5.3 Gated Recurrent Neural Network

Next, we wanted to implement deep learning frameworks that have been shown to yield much better accuracy scores compared to our baseline scores. Recurrent Neural Networks are used in various analysis of natural language since the sentiment behind textual sequences is more important than the individual items (words) themselves, something that the recurrent nature of these neural networks is very adept at capturing.

Gated Recurrent Unit (GRU) is a type of RNN that solves the problem of vanishing gradients during back propagation. The vanishing gradient problem is when the gradient shrinks and become too small to operate on. If a gradient value becomes extremely small, it does not contribute in learning. GRUs solve this problem by impementing an update and reset gate that prevents gradients from becoming too small. A GRU Network is a kind of Recurrent Neural Network (RNN) that solves the problems of vanishing gradients in the backpropagation step by using update and reset gates.

Our implementation of a GRU yielded an accuracy of about 85% on the sentiment analysis of the test set. We tuned our hyperparameters using the validation set, which also had a similar accuracy score.

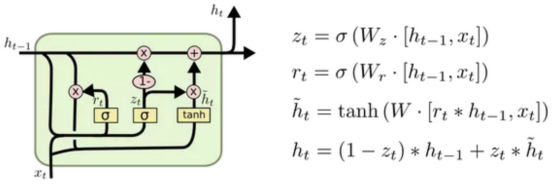


Figure 3: Structure of a GRU Model and Equations for Backprop (8)

5.4 1-Dimension Convolutional Neural Network

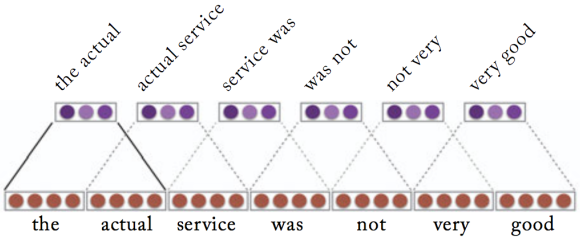


Figure 4: Structure of a 1-d CNN Model (9)

Traditionally, Convolutional Neural Networks (CNNs) are used in computer vision to compute feature maps. However, recent research has applied 1-dimension CNNs to textual data for classification. Given a sequence of words $w_i : n = w_1, \ldots, w_n$, where each word is associated with an embedding vector of dimension d, a 1-d convo-

Model	Training Accuracy	Validation Accuracy	Testing Accuracy	Parameters
Linear SVC	**79.13%**	NA	**79.74%**	**c=2.0**
GRU	83.28%	83.52%	84.64%	dropout=0.2
1-D CNN	96.93%	83.55%	85.09%	kernel_size=7
LSTM	**96.45%**	**88.20%**	**89.20%**	**input_shape=50**

Table 2: Deep Learning Models Result Comparisons

lution of width k is the result of moving a sliding-window of size k over the sentence, and applying the same convolution filter to each window in the sequence.

Our implementation with `kernel_size=7` recorded a testing accuracy of 85%, which was comparable to the GRU implementation. This model also lead to the highest training accuracy.

5.5 Long Short-Term Memory Neural Network

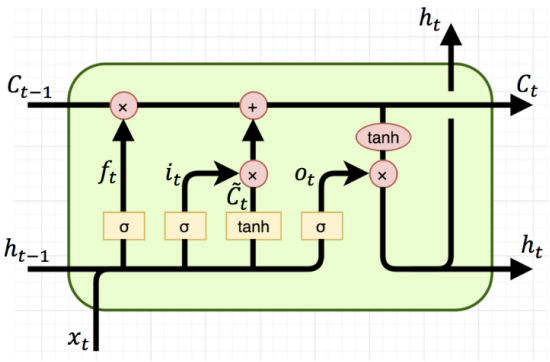


Figure 5: Structure of an LSTM Model (8)

An LSTM is similar to a GRU but consists of three gates: input, forget and output gates. The core concept of LSTMs are the cell state, and its the three gates. The cell state acts as a transport highway that transfers relative information all the way down the sequence chain. Our hypothesis was that an LSTM will be able to remember or forget the learnings from textual data more selectively and arrive at better sentiment predictions than a GRU.

In order to test this hypothesis, we implemented an LSTM with 50 features and produced a final testing accuracy of about 90%. This was higher than our baseline models and the other deep learning frameworks we implemented. Moreover, the model that we implemented ran fasted than the 1-d CNN, and can be used to characterize sentiment on medical drug reviews in real time.

6 Conclusion

In this paper, we present several deep learning models that achieve decent to excellent accuracy scores for sentiment analysis on drug review data. We experiment with various commonly used baseline models, and less traditional frameworks to provide a comprehensive understanding of the ways in which we can generate insights that can help patients and pharmaceutical companies.

Since we were bound by the constraints of time and computing resources, we decided to steer away from implementing deeper networks. Nevertheless, our experimental frameworks were still good enough - in terms of both accuracy, train time and run time to be applied in industry.

6.1 Future Considerations

A natural continuation of our work is to apply our models to other review databases and generate accuracy scores using transfer learning. Moreover, our results show that there is a lot of exciting work yet to be done in making better LSTM and CNN networks that can provide even better accuracy scores. We would also like to collect/scrape our own data by tapping into more resources (such as clinical trial surveys) in the future, and test with more complicated factors of sentiment, such as the number of people who liked a certain review.

7 Contributions

Both Arjun and Aditya contributed equally to the project. Arjun implemented the baseline and deep learning models and wrote the final paper, while Aditya assimilated the results from the models, analyzed the models quantitatively and qualitatively and wrote the final paper.

References

[1] Felix G. and Surya K. Dataset files are presented in the downloadable zip folder. On University of California, Irvine website https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29.

[2] Felix G., Surya K., Hagen M., and Sebastian Z. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross- Domain and Cross-Data Learning. In *2018 International Digital Health Conference, 2018 14th International Conference.* https://doi.org/10.1145/3194658.3194677, 2018.

[3] Mahboob K, Ali F. Sentiment Analysis of Pharmaceutical Products Evaluation Based on Customer Review Mining. In *2018 Journal of Computational Science Systems Biology* 11: 190- 194. doi:10.4172/jcsb.1000271, 2018.

[4] Na, J. C. and Kyaing, W. Y. M. Sentiment Analysis of User-Generated Content on Drug Review Websites. In *Journal of Information Science Theory and Practice. 2015. Mar, 3(1): 6-23. Korea Institute of Science and Technology Information.* DOI:10.1633/JISTaP.2015.3.1.1

[5] Gopalakrishnan, V and Ramaswamy, C. Patient opinion mining to analyze drugs satisfaction using supervised learning. In *J. Appl. Res. Technol. 15, 311319 (2017).* https://doi.org/10.1016/j.jart.2017.02.005.

[6] Hanane Grissette, EL Habib Nfaoui, Adil Bahir. Sentiment Analysis tool for Pharmaceutical Industry Healthcare. In *Transactions on Machine Learning and Artificial Intelligence Vol.5 No.4*. 10.14738/tmlai.54.3339, 2017

[7] Oakes I, Ashtekar N, Wright W, Dalsania S and Li M. Team NDL: Algorithms and Illnesses. In *Kaggle kernel: Penn State's Nittany Data Labs* 2019.

[8] Kaushik M. GRU's and LSTM's. In Towards Data Science Blog https://towardsdatascience.com/grus-and-lstm-s-741709a9b9b1 2019.

[9] Yoav G. Neural Network Methods for Natural Language Morgan In *Claypool Publishers* 2017.