Drug Review Sentiment Analysis

By: Myra Rust

Executive Summary

Sentiment analysis can assist companies in monitoring their brand and public opinion on their products. Machine learning can be utilized to gain insights on the sentiment of freestyle text, which makes it a perfect choice for pharmaceutical company's trying to gauge how their product is being received on social media platforms. This project builds a neural network that is able to classify freestyle text reviews on drugs into the categories of positive, neutral, and negative with an accuracy score of 83.3%. This is a great start and with further development could lead to an extremely successful drug review sentiment prediction model.

Introduction/Background

The pharmaceutical industry, which discovers, develops, manufactures, and distributes drugs, continues to be one of the largest industries in the United States. "In 2019, the United States remained the world's largest single pharmaceutical market, generating more than $490 billion of revenue." (Pharmapproach, 2020). With that amount of money at stake, it is important for companies to keep a close eye on public perception. By analyzing patient reviews of products on social media, companies are able to monitor brand and product sentiment and listen to the voice of the consumer.

The objective of this project is to create a model that is able to analyze text reviews of pharmaceuticals and determine the sentiment of that review based solely on text analysis. If the model is successful, this model can be deployed to monitor real-time web-based sentiment of drug performance on platforms such as Twitter or Meta (a.k.a. Facebook) and give pharmaceutical companies the heads up on any trends in public sentiment.

Preliminary Analysis

Data Source: The dataset contains reviews scraped from online pharmaceutical review sites and was retrieved from the UCI Machine Learning Library. (Gräßer et al., 2018).

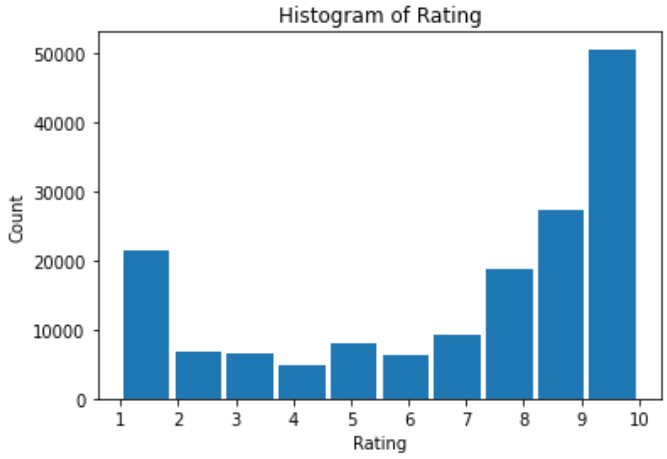
	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	November 27, 2016	37

Training dataset - 161,297 observations
Test dataset - 53,766 observations

The dataset contains 7 features:

- Unnamed:0 (numerical): Previous Index
- drugName (categorical): name of drug
- condition (categorical): name of condition
- review (text): patient review
- rating (numerical): 1-10 rating of drug effectiveness provided by patient (1 being worst, 10 being best)
- date (date): date of review entry
- usefulCount (numerical): number of users who found the review useful

It's important to note that the patient reviews being analyzed are not evidence based. They are simply opinion. The FDA has a drug review process that includes many steps such as clinical research and trials, which analyzes evidence. This project will focus solely on opinion mining, so let's take a look at the distribution of data based on rating.

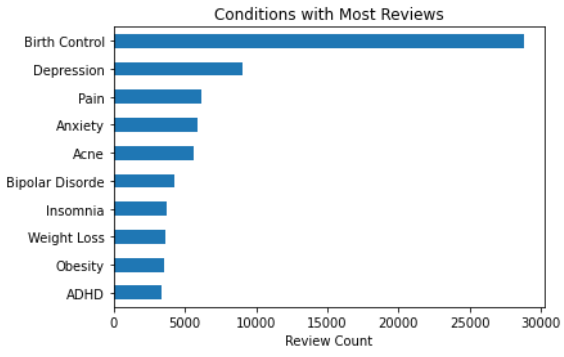


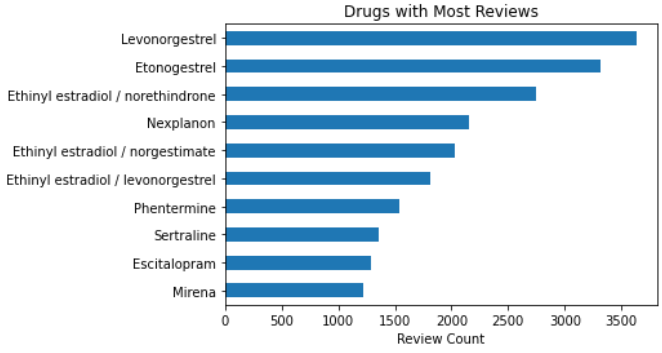
We can see that the data is heavily skewed left, showing that the majority of reviews are positive in nature. This class imbalance will have to be accounted for during modeling using assigned class weights when training the model.

Data Cleaning:

Observations containing parsing errors and NaN values were noted during preliminary analysis. Since there were only a small number of observations with these issues (1,799 in the training dataset and 566 in the test dataset) the records were simply removed from the datasets.

The following visualizations show different interesting statistics derived from the data, prior to modeling.





For a list of the top 10 drugs with the highest mean rating and top 10 conditions with the highest mean rating on drug effectiveness, refer to *Appendix A*: Additional EDA Data.

Feature Extraction & Selection

To prepare for modeling all features other than ‘review’ and ‘rating’ have been removed from the dataset.

Data Transformation:

The feature ‘rating’ is the target variable. We are trying to predict what rating would be given for the according text review. In order to better understand the sentiment of the target variable, rating has been broken down into three categories.

- Positive equal to a rating of 7 or higher. This person has a positive sentiment toward the product and is likely to be vocal in their opinion and recommend the product for use to others.
- Neutral equal to a rating of 5 or 6. This person has a passive sentiment toward the product and may not recommend the product to others.
- Negative equal to a rating of 4 or lower. This person has a negative sentiment toward the product and is likely to be vocal in their opinion and provide others with a negative review of the product.

Here is a glimpse of the data that will be fed into the model for analysis:

	review	target
0	"It has no side effect, I take it in combinati...	0
1	"My son is halfway through his fourth week of ...	0
2	"I used to take another oral contraceptive, wh...	1
3	"This is my first time using any form of birth...	0
4	"Suboxone has completely turned my life around...	0

At this time, training data was split into training and validation sets, with the test set remaining set aside for testing. Data was then turned into arrays, tokenized, and padding added to sequences to meet the requirements for model input.

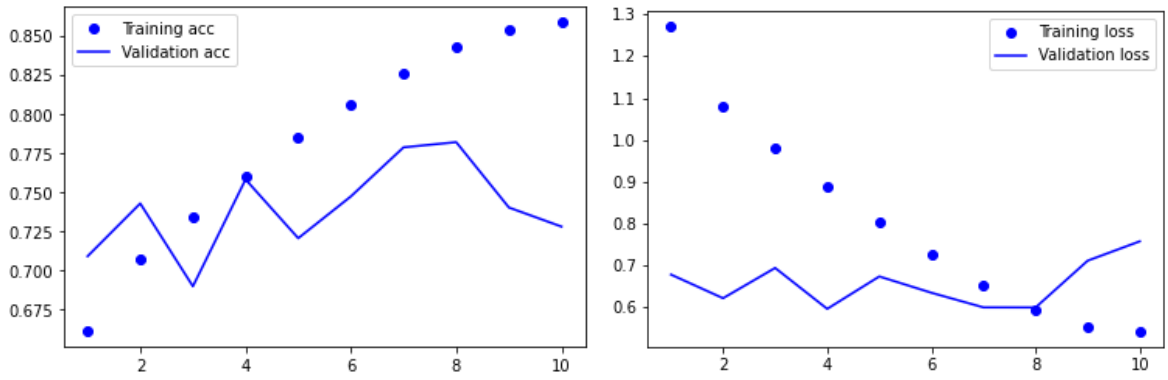
Model Selection and Evaluation

The objective of this project “to properly classify the sentiment of patient reviews” is a problem of single-label multiclass classification. Due to a large portion of the corpus being medical terminology or terminology specific to drug reviews, I decided to build a recurrent neural network that leveraged the Keras library and NLTK library for natural language processing.

The neural network contained one embedding layer, one LSTM layer, and one Dense layer. The LSTM layer deals with the vanishing gradient problem by saving information for later. This prevents older signals from gradually vanishing during processing. (Chollet, F., 2018). The Dense layer used the softmax activation function to allow results to be separated into multiclass labels.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 32)          320000
_________________________________________________________________
lstm (LSTM)                  (None, 32)                8320
_________________________________________________________________
dense (Dense)                (None, 3)                 99
=================================================================
Total params: 328,419
Trainable params: 328,419
Non-trainable params: 0
```

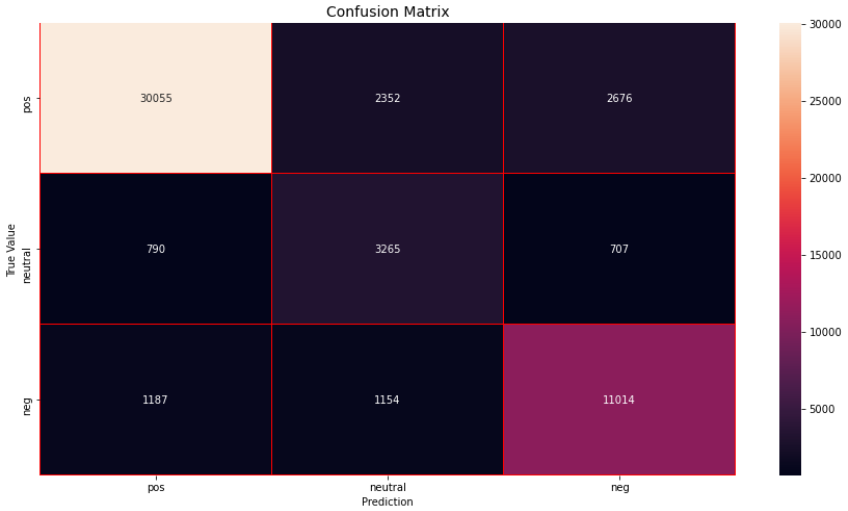
Because the target variable is a numeric value correlating to a category (0=pos, 1=neutral, 2=neg) and not one-hot encoded the sparse_categorical_crossentropy loss function was chosen. Adam proved to be the best option for the optimizer function. The model was compiled and fit to the training dataset using assigned class weights to account for the imbalanced classes. The model was then evaluated against the validation dataset. Below are the plots displaying the training and validation accuracy and loss results.



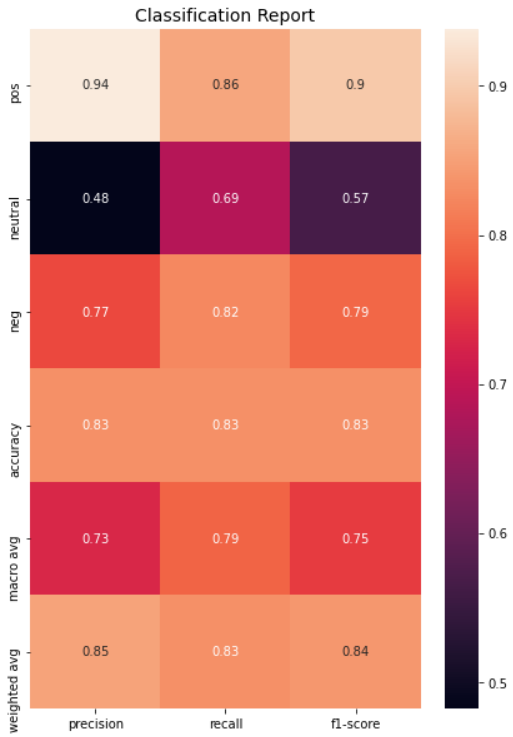
The model appears to achieve the best results around the 8-epoch mark after which accuracy declines and loss increases. The model design was changed to stop after 8 epochs and a new model was compiled and trained from scratch before using it to evaluate the test dataset.

Results

The resulting model achieved an accuracy of 83.3%, which is quite good considering the complexity of the text it was evaluating. We can view how the model scored the reviews in the confusion matrix located below.



For each of the classes in the target variable, the model correctly predicted the majority of reviews. No single area stands out as an area for concern because for each class, the incorrect predictions were equally dispersed between the remaining two classes. To get a better look at how the reviews were classified, let's take a look at the classification report.



The classification report contains a series of percentages relating to the models ability to predict each target class and to the model as a whole. Precision, also known as positive predictive power, is the percentage of predicted values that matched the true value for each target class. For example: The model predicted a total of 32,032 reviews as having positive sentiment. Of those 32,032 reviews, 30,055 were actually positive reviews resulting in a precision score of 94%. Recall, also known as the true positive rate, is the percentage of true values predicted correctly for each target class. For example: We know there were 35,083 positive reviews in the test dataset. Of those, 30,055 were accurately identified as being positive resulting in a recall score of 86%. Precision and recall scores are combined to provide the F1 Score which is used to measure model accuracy.

We can see in the classification report that the model struggled with neutral reviews across the board. It over predicted the existence of neutral reviews and did not properly classify actual neutral reviews very well. This doesn't surprise me as the very nature of a neutral review is ambiguous. A neutral review can contain all ambiguous words, positive words, negative words, both positive and negative words, or all three types. It really is a very hard sentiment to properly classify.

Conclusion

In conclusion, reviews evaluating the performance of pharmaceuticals to treat a particular condition contain very complex terminology and opinions can be difficult to classify correctly. The model performed well predicting 83.3% of reviews correctly. I would consider this project a success. With that being said, there may be potential to improve on this work in the future.

Future Work

Sentiment was measured by dividing the rating into three classes. 0-4 being negative, 5-6 being neutral, and 7-10 being positive. Due to the competitiveness of the market, really anything less than seven could be considered negative by some corporate standards. It may be worthwhile to eliminate the neutral class, which is the class that was the most problematic for the model and build a model that only classifies the review as either positive or negative using the breakdown of rating as 0-6 being negative and 7-10 being positive.

Another possibility for future work involves the creation of a domain specific sentiment lexicon as outlined in the article *SentiHealth: creating health-related sentiment lexicon using hybrid approach*. (Asghar et al., 2016). This method would take considerably more time and expertise to accomplish. However, should a medical sentiment lexicon become publicly available in the future, it could be incorporated into this work with little effort.

Acknowledgements

I would like to thank Professor Catherine Williams and all other professors that are part of Bellevue University's data science program. Without their direction, support, and guidance, this paper would not have been possible. Lastly and nearest to my heart, I would like to thank my daughter for sacrificing her time with me to allow me to pursue this degree for the betterment of myself.

References

Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R., & Kundi, F. M. (2016). SentiHealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, 5(1), 1139. https://doi.org/10.1186/s40064-016-2809-x.

Chollet, F. (2018). Deep Learning with Python. Manning Publications: Shelter Island, NY.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Gräßer, F., Kallumadi, S., Malberg, H., and Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125. DOI: https://dl.acm.org/doi/10.1145/3194658.3194677.

Pharmapproach. (2020, November 1). 15 Astonishing Statistics and Facts About the U.S. Pharmaceutical Industry. Pharmapproach.com. Retrieved from https://www.pharmapproach.com/15-astonishing-statistics-and-facts-about-u-s-pharmaceutical-industry/.

Appendix A: Additional EDA Data

Top 10 drugs with highest mean rating and >50 reviews

	drugName	rating
0	Stribild	9.466667
1	Cobicistat / elvitegravir / emtricitabine / te...	9.433962
2	Diethylpropion	9.305882
3	Chlorpheniramine / hydrocodone	9.275000
4	Campral	9.252747
5	Librium	9.230769
6	Carisoprodol	9.202899
7	Chlordiazepoxide	9.191011
8	Subutex	9.169492
9	Clobetasol	9.166667

Top 10 conditions with highest mean rating on drug effectiveness

	condition	rating
0	mance Anxiety	9.673797
1	Alcohol Withdrawal	9.244240
2	Dermatitis	9.096774
3	Cold Sores	9.068120
4	Hyperhidrosis	8.935007
5	Alcohol Dependence	8.894330
6	Barrett's Esophagus	8.864865
7	Erosive Esophagitis	8.842857
8	Herpes Simplex, Suppression	8.789157
9	Cluster Headaches	8.771084