SENTIMENT ANALYSIS OF DRUG REVIEW DATA

ACHINTYA RAYA POLAVARAPU AND NATAVAN SAFIKHANOVA

Abstract. Patient reviews of drugs abound on the Internet today. There is a wealth of insights hidden in these unstructured reviews that can benefit healthcare stakeholders, if only we can tap into them. This project explores how natural language processing, machine learning and deep learning techniques could be leveraged to analyze a dataset of drug reviews.

1. Motivation

Medication errors pose a significant threat to patient safety and can have devastating consequences. In a systematic review [1, 2] of studies published between 1990 and 2011 to gain insight into this issue, the search identified 45 studies from 10 Middle Eastern countries that examined medication errors. These studies suggest medication error rates in the region may be high. Prescribing error rates ranged from 7.1% to 90.5% of medications, while administration error rates spanned 9.4% to 80%. The most common prescribing errors were incorrect dosage, frequency, and strength. Researchers pointed to lack of knowledge about drugs among doctors and nurses as contributing factors. However, most studies were of poor quality and did not assess the clinical impact of the errors. Medication errors are a global problem. In the U.S. alone, up to 9 million patients are impacted by medication errors each year. Between $7,000$ to $9,000$ Americans die annually from medication errors[6]. Out-patient medication errors injure over $500,000$ people per year in the U.S.

Unsafe medication practices and errors are a leading cause of harm in healthcare worldwide. Globally, medication errors cost an estimated 42 billion USD per year. Errors can happen at any point in the medication process. One promising way to help reduce medication errors could be providing patients and doctors with more details about specific drugs. The many reviews of medications that patients post on various websites offer a huge amount of information that could potentially be used for this purpose. Using tech tools like machine learning, natural language processing, and

Date: April 15, 2024.

sentiment analysis, we could analyze patients' drug reviews to give personalized[4, 5, 3].

2. Problem Statement

Can machine learning be used to analyze a huge number of reviews patients post about medications on Drugs.com and gain useful insights to improve how drugs are developed, regulated, marketed, and prescribed? This project aims to explore ways to tap into a subset of the dataset of over 8 million reviews of thousands of prescription and over-the-counter drugs using machine learning and natural language processing techniques. This shall pose to be a challenge since there has not been much work done applying sentiment analysis to medication reviews since these reviews tend to be much more complex to analyze. They contain medical terms like disease names, reactions, and chemical names used in making the drug, but expressed in many different ways.

Being able to access what real patients say about the medicines they take is an exciting new frontier for improving health and healthcare. However, the manner in which people write the reviews are just too unstructured for people to systematically understand patients' experiences with different treatments and turn those insights into practical knowledge. By designing machine learning models to detect patterns in these reviews, identify factors linked to good or bad experiences, and group together patients discussing similar conditions and outcomes, this project will explore how to address major challenges in obtaining useful opinions from patient reviews to benefit pharmaceutical companies, doctors, and patients themselves.

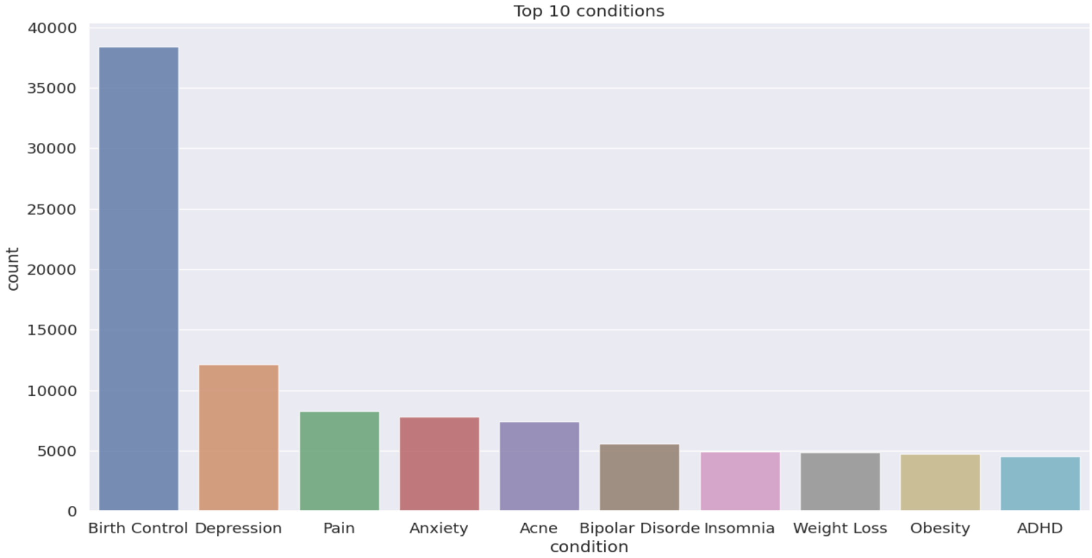
3. Study Objective

The goals of this project are:
Create a base to automate the recommendation system and use machine learning and natural language processing and deep learning models to make sense of reviews real patients wrote on Drugs.com about their experiences with different medications for drug development.

Find patterns in these reviews to figure out what makes patients rate their experiences with some drugs higher or lower. This could reveal what drives good or bad reviews of different medicines so companies know how to make treatments better. Check reviews for side effects that get mentioned frequently so drug makers and doctors are aware of common issues, safety concerns, or the need for more warnings. Group together reviews from patients dealing with similar medical conditions and experiences with drugs to pinpoint chances to tailor treatments and results to patients' needs. This might support more customized care focused on what really matters to patients. Summarize what these reviews say about how well drugs work, their side effects, and safety in an easy-to-understand way for regular people and professionals. This could encourage patients and doctors to make fully informed choices about treatment options and take advantage of knowing what other patients think.

4. Information on Data Set

The dataset used in this project is Drug Review Dataset (Drugs.com) taken from the UCI ML repository. The Drug Review Dataset contains over 8 million reviews of prescription and over-the-counter medications from the Drugs.com website. Reviews span thousands of drugs from 1999 to 2019. Drug Name, Drug ID, Date, Useful Count, Rating, Condition, Side Effects and Review Text. However, this proved to be too big of a data set for the computation powers available to us. Hence, we take a subset of this data by only considering the top 10 conditions and taking just 3000 reviews to begin with. Below, is the visual of the data set about the top 10 medical conditions.



5. Study Methodology

The first step in the sentiment analysis after data cleaning is Text Cleaning where an important step of it was the stop-words removal. Stopwords are common words like 'the', 'a', 'is'that do not add much meaning. Removing them reduced noise and focused the analysis on important words. Removed punctuation: Punctuation marks do not provide semantic information. Removing them resulted in a consistent representation of text. Lemmatization: Lemmatization reduced inflected words to their root form. This helped group together related words, improving analysis. Standardization: Applying techniques like removing extra spaces, converting to lowercase, handling abbreviations etc. standardized the text, enabling consistent treatment of words. Reduced sparsity: The cleaned, standardized text contains a smaller set of words that are meaningful for analysis. This made patterns in the text easier to detect with higher word counts for important features. Improved performance: Cleaned data required less storage space and enabled faster NLP analysis due to fewer distinct words and standardized format. Machine learning models also tend to perform better with clean, dense data. Facilitated word associations: By grouping inflected words and removing unimportant terms, helped revealing semantic themes in the text.

The next step involves summarization. Here the key topics represent a high-level summary of the main themes in a large collection of data. By analyzing the words in each row, we get a quick sense of the major areas of discussion. For instance side effects, cost, effectiveness and administration. In sentiment analysis - If topics represent different aspects such as side effects or advantages, sentiment analysis specifically on reviews allocated to each drug to determine the overall positive or negative sentiment towards those aspects. We create a new binary column called ‘sentiment’in both the training and testing datasets. The sentiment is assigned a value of 1 (positive) if the user rating is greater than 5, and 0 (negative) otherwise. This process is accomplished using the apply() function with a lambda function.

Machine learning and deep learning models with a brief overview of their algorithms:

Logistic Regression - - A statistical model that uses a logistic function to model a binary dependent variable. It uses maximum likelihood estimation to fit a sigmoid curve to the data. The logistic function, also called the sigmoid function, is a mathematical function that maps values to probabilities. It’s an S-shaped curve that can take any real valued number and map it into a value between 0 and 1, representing a probability.

Support Vector Machines(SVM) - In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the optimal hyper-plane that differentiates the two classes very well. So it is a discriminative classifier that finds a hyperplane to separate classes. It maximizes the margin between classes which leads to good generalization. Kernels is used to handle non-linear boundaries. The core idea behind SVMs is that we try to find the optimal hyperplane that separates the data points from different classes with the maximum margin.

Decision Tree - Non-parametric model that splits the feature space into regions based on decision rules. It uses a greedy algorithm to find the most significant splits that result in the most homogeneous branches. The hierarchical structure of a decision tree leads us to the final outcome by traversing through the nodes of the tree. Each node consists of an attribute or feature which is further split into more nodes as we move down the tree. At each node of the tree, the algorithm selects the predictor variable that best separates the observations into different classes based on the Gini index.

Random Forest - An ensemble method that combine multiple decision trees. Random Forest works by creating many different samples of training data using a technique called bootstrap sampling. For each sample, it grows a decision tree. However, when constructing each tree, it only considers a random subset of features when splitting nodes. This helps make each tree unique. After generating many diverse decision trees, it aggregates their predictions for new data. For classification, it takes a vote from all the trees and the most common prediction is the final prediction. Some key parts of how Random Forest works:

Bootstrap sampling: Random Forest creates many resamples of the training data by sampling with replacement. Each sample is used to train one decision tree. This makes each tree's training data slightly different and less correlated with the others.

Restricting feature use: At each node in growing a tree, only a subset of features are considered for the best split. This randomly chosen subset is different for each tree, which makes them even more different from each other.

Aggregating predictions: The predictions from all the individual trees are combined, either by voting for classification or averaging for regression. This ensemble approach typically outperforms any of the individual trees.

Different techniques used for representing text data in machine learning models:
Normalization: Converted to lowercase, handle abbreviations. Create vocabulary: Builded a list of the remaining words after preprocessing. These are the words we want to represent. Generate co-occurrence matrix: Counted how frequently each word co-occurs with every other word in a fixed window size. This captured the context each word appears in.

TF-IDF: It stands for Term Frequency-Inverse Document Frequency. Raw word counts can be weighted using Term Frequency-Inverse Document Frequency to give less weight to words that occur very frequently across all reviews. TF-IDF for word i in document j is:

$$TF - IDF_{ij} = (\text{Number of times word i appears in review j})\times$$
$$\log\left(\frac{\text{Total number of reviews}}{\text{Number of reviews with word i in it}}\right)$$

. This gives more weight to words that appear frequently in a text, but rarely in the corpus. For sentiment analysis, TF-IDF help emphasize words that are meaningful for a text's sentiment. L2 regularization penalizes weights in a model with large magnitudes to prevent overfitting. When applied to a document-word matrix X, it modifies the objective function to be:

$$J(w) = \frac{1}{2}||Xw - y||^2 + \frac{\lambda}{2}||w||^2$$

where w are the weights, y are the labels, and λ is the regularization strength. For sentiment analysis, X is a TF-IDF weighted text-word matrix, w the weights connecting words to sentiment predictions, and y vectors indicating positive/negative labels. L2 regularization helps constrain the w values connecting each word to the sentiment predictions, preventing the model from becoming overly complex. But because of the complexity of the data, L2 regularisation did not give better accuracy on TF-IDF. There is a scope of better hypertunning parameters on TF-IDF for better accuracy. Benefits of these methods on the data were that they captured semantic meaning, handled previously unseen words based on context, provided vector representations that enabled mathematical operations on words, grouped together related words in the vector space and improved performance on the accuracy of the sentiment analysis.

Summary: Below is the chart showing a summary of the above steps in a diagrammatic form.

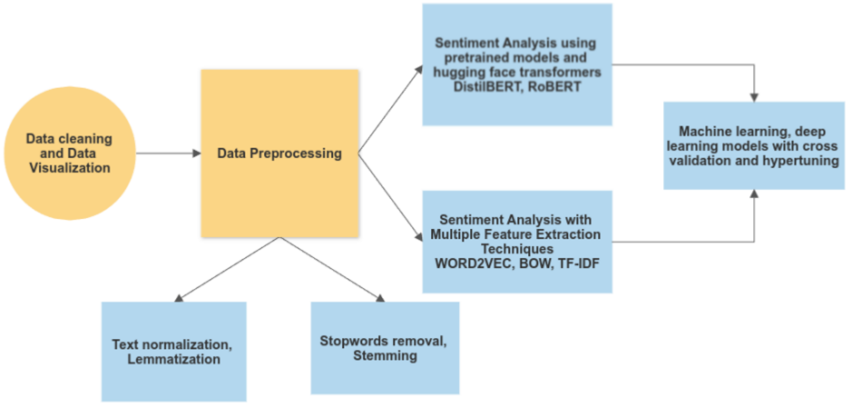
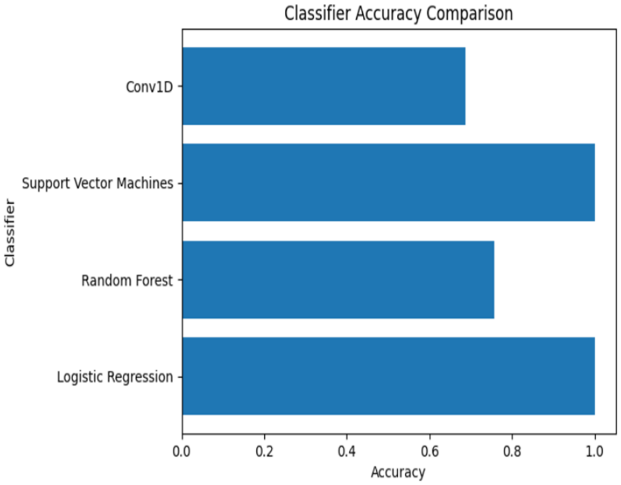

Data cleaning and Data Visualization
Data Preprocessing
Sentiment Analysis using pretrained models and hugging face transformers DistilBERT, RoBERT
Sentiment Analysis with Multiple Feature Extraction Techniques WORD2VEC, BOW, TF-IDF
Machine learning, deep learning models with cross validation and hypertuning
Text normalization, Lemmatization
Stopwords removal, Stemming

FIGURE 3.2: summary of the methodology

6. Results

In our study, we applied sentiment analysis to a dataset containing patient reviews of various drugs used for different medical conditions. We utilized the pre-trained DistilBERT model from the Hugging Face Transformers library to perform sentiment analysis on the reviews. We applied our analysis to Logistic Regression, Random Forest, Support Vector Machines and Conv1D. We notice that Support Vector Machines and Logistic Regression give an accuracy of 1.00. On the other hand, Conv1D gives an accuracy of about 65% while Random forrest gives an accuracy of 75%. Below is the accuracies obtained for the various models mentioned.



Conclusion and Criticism:

We note again that the results above were obtained for a dataset of 3000 entries. Hence in the above graph, we must take the high accuracy of Logistic Regression and Support Vector Machines with a grain of salt. Since, they obtain an accuracy of 100%, we can assume this is because, these models did not could not analyse the limited data. However, we are confident that given more computational capacity, we could have obtained a degree of accuracy that would allow these models to be useful in the sentiment analysis of drug reviews.

References

[1] Aseeri M, Banasser G, Baduhduh O, Baksh S, Ghalibi N. Evaluation of Medication Error Incident Reports at a Tertiary Care Hospital. Pharmacy (Basel). 2020 Apr 19;8(2):69. doi: 10.3390/pharmacy8020069. PMID: 32325852; PMCID: PMC7356747.

[2] Alomar, M.J., Tayem, Y.I., Qawasmeh, R.I., Al-Azzam, S.I., AbuRuz, S., 2021. A descriptive study of medication errors in a tertiary care hospital in saudi arabia. Saudi Pharmaceutical Journal 29, 72–80. URL: https://www.ncbi.nlm.nih. gov/pmc/articles/PMC7796638/.

[3] Goel, V., Gupta, A.K., Kumar, N., 2018. Sentiment analysis of multilingual twitter data using natural language processing, in: 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), IEEE. pp. 208–212. doi:10.1109/CSNT.2018.8820254.

[4] Habimana, O., Li, Y., Li, R., Gu, X., Yu, G., 2020. Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences 63, 1–36

[5] Li, J., Xu, H., He, X., Deng, J., Sun, X., 2016. Tweet modeling with lstm recurrent neural networks for hashtag recommendation, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1570–1577. doi:10.1109/IJCNN. 2016.7727385. arXiv:2104.01113v2 [cs.IR] 5 Apr 2021

[6] Shimada, K., Takada, H., Mitsuyama, S., et al., 2005. Drug-recommendation system for patients with infectious diseases. AMIA Annual Symposium Proceedings 2005, 1112

7. Appendix

The following is the link to the GitHub repository containing the code for the above project as well as the data used for the project:
`https://github.com/aytnihca/SentimentAnalysis`.

Email address: polavara@ualberta.ca
Email address: safikhan@ualberta.ca

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, T6G 2G1, Canada.