PROJECT PART 5

TEAM 10:

Ravi Tarun Prasad Nimmalapudi
Krishna Koushik Parimi
Naveen Kumar Pallanti

SUMMARY

The use of supervised machine learning algorithms to analyze drug evaluations and apply sentiment analysis to the pharmaceutical industry is a major advancement in the use of computational approaches to interpret patient experiences and opinions. Natural language processing (NLP) has a branch called sentiment analysis that is used to systematically classify textual input into sentiment-based categories like positive, negative, or neutral. This method is very helpful in the healthcare industry for analyzing patient input regarding drugs, which is crucial for determining the effectiveness of drugs, improving patient care, and influencing healthcare choices.
Together, the three research papers under review examine various approaches and machine learning algorithms for analyzing drug reviews. This shows the viability and significance of sentiment analysis in gaining a more profound understanding of patients' attitudes towards drug use, which can help pharmaceutical companies and healthcare providers better understand and improve patient outcomes.

Paper 1 - "*Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms.*"
Published - 21 March 2020, by Sairam Vinay Vijayaraghavan and Debraj Basu.

Objective:
The paper aims to conduct sentiment analysis on drug reviews using supervised machine learning algorithms, focusing on the textual content of reviews and their associated numerical ratings. The authors explore the impact of different words used in reviews for drugs treating similar conditions on the ratings.

Approach:
In preparing the data, the authors first cleaned and processed the text information and then converted it into numeric formats using Term Frequency Inverse Document Frequency (TFIDF) and Count Vectors (CV). For the machine learning aspect, they applied supervised classification algorithms, exploring embeddings like TFIDF and CV. The algorithms used included Artificial Neural Networks (ANN), Recurrent Neural Networks (LSTM and GRU), Support Vector Machines (SVM), Random Forests, and Logistic Regression. The study focused on prevalent conditions such as "Birth Control," "Depression," and "Pain" in the dataset. Notably, Count Vectorizer showed better representation than TFIDF, likely because it directly captured word counts. Deep learning models, especially neural networks like ANN, outperformed traditional machine learning methods. The optimal results were achieved by combining deep learning algorithms with Count Vectorizer encoding for each studied condition.

Accuracy Percentages for Best Models:
For "Birth Control" with ANN – TFIDF and Count Vectorizer accuracies up to 93.4114% and up to 93.85%
For "Depression" with ANN - TFIDF and Count Vectorizer accuracies up to 90.1% and best 92.10%.
For "Pain" with ANN - TFIDF and Count Vectorizer accuracies up to 89.6667% and best 91.286%.

Paper 2 - "*Drug Review Sentiment Analysis*"
Published - 2020 by Sumit Mishra

Objective:
The primary objective of the research paper "Drug Review Sentiment Analysis" by Sumit Mishra is to perform sentiment analysis on patient reviews of specific drugs. The goal is to classify these reviews as either positive or negative based on the sentiment expressed in the text. Sumit Mishra aims to develop a model that accurately predicts the sentiment of drug reviews, providing insights into the overall satisfaction of patients with medications.

Approach:
Sumit Mishra's approach involves merging and preprocessing datasets from the UCI Machine Learning Repository, conducting exploratory data analysis (EDA) for insights, and engineering features for sentiment analysis of drug reviews. The dataset, sourced from online pharmaceutical review sites, includes categorical and numerical data. Mishra applies three classification models (LightGBM, XGBoost, and CatBoost) after splitting the data into training and testing sets. The emphasis is on meticulous feature engineering, detailed EDA, and evaluating model performance.

Accuracy Percentages for Best Models:
The accuracy percentages for the best models in Sumit Mishra's research paper on "Drug Review Sentiment Analysis" are: LightGBM- 88.79% and CatBoost: 88.37%.

Paper 3 - **"***Drug Sentiment Analysis using Machine Learning Classifiers."*
Published - January 2022 by Mohammed Nazim Uddin and colleagues.

Objective:
The study aimed to assess the effectiveness of drugs through sentiment analysis of user reviews, utilizing machine learning classifiers as a sub-dimension of natural language processing. The focus was on user reviews that rated drugs based on their effectiveness, leading to both binary and multiclass classification of drugs' effectiveness.

Approach:
In their January 2022 study, "Drug Sentiment Analysis using Machine Learning Classifiers," authors Mohammed Nazim Uddin, Md. Ferdous Bin Hafiz, Sohrab Hossain, and Shah Mohammad Mominul Islam aimed to assess drug effectiveness through sentiment analysis of user reviews. They utilized a dataset from the UCI machine learning repository, focusing on binary and multiclass classification of drug effectiveness. Employing tokenization and lemmatization for preprocessing, the authors used Naive Bayes, Random Forest (achieving the highest accuracy at 94.06%), Support Vector Classifier, and Multilayer Perceptron for binary classification. Linear SVC was chosen for multiclass classification, particularly excelling in categorizing moderately effective drugs with an AUC of 0.82. The study highlights the machine learning algorithms' potential in aiding consumers and manufacturers to understand drug effectiveness based on user sentiment.

Accuracy Percentages for Best Models:
Random Forest has the highest accuracy 94.06% among binary classification models in drug sentiment analysis. Linear SVC, the best-performing model for multiclass classification, demonstrated promising results, particularly in categorizing moderately effective drugs with an AUC of 0.82.

REPORT

DRUG REVIEW ANALYSIS

The Drug Review Analysis project is a comprehensive examination of patient-provided reviews on various pharmaceuticals, with a primary focus on assessing drug efficacy, patient satisfaction, and the prevalence of side effects. Patients are increasingly using internet venues to discuss their experiences with drugs in this era of abundant information. Beyond the controlled setting of scientific trials, these reviews provide priceless insights into the actual efficacy of medications. Our initiative intends to close the gap between clinical data and patient-reported outcomes by methodically analyzing these patient testimonies. This will provide a comprehensive understanding of drug effectiveness and safety from the patient's point of view.

METHODOLOGY

Data Collection – We have sourced our dataset from UCI Machine Learning Repository.
Dataset Link -

Exploring Data – Initially we had our data split into training and testing sets respectively.
Train Set - drugsComTrain_raw.tsv – 161297 records.
Test Set - drugsComTest_raw.tsv – 53766 records.

The dataset comprises a total of 7 columns, reflecting a blend of both numerical and textual data.

Textual Data –

- 'drugName': The name of the drug/medication reviewed.
- 'condition': The medical condition the drug was prescribed for.
- 'review': The patient's written review detailing their experience with the drug.

Numerical Data –

- 'rating': A numerical score given by the patient.
- 'date': The date when the review was posted.
- 'usefulCount': A count of how many times other users found the review helpful.
- 'uniqueId': Unique ID for each review.

We will be featuring 'drugName', 'condition', 'review', 'rating' and 'usefulCount' columns for our other analysis as these probably are the best sources of information and align with the goals of our analysis approach.

Data Cleaning –
We find out that we have some missing values in the column 'condition'.
We just ignored them as we could not rely on empirical methods to impute those missing values based on existing assumptions to ensure data integrity.
As mentioned earlier, we initially had a very clean and organized dataset except for these missing values.

We further explored our data so as to know the nature it is comprehending.
From further exploration, we got to know the distribution of ratings that we had almost 51,000 records having a rating '10' and 21,000 records having a rating '1'.
Similarly, we highlighted the most effective drugs according to patient reviews and also plotted the top 10 conditions having the most reviews.

Regression Approach – We started it with initial examination of the numerical columns in our dataset, including 'rating' and 'usefulCount'. This analysis did not reveal any significant correlation.

Sentiment Analysis - Then to enrich our analysis, we extracted sentiment scores from the textual review data. This process involved applying natural language processing (NLP) techniques to convert the qualitative feedback into a quantitative sentiment score, reflecting the overall positivity or negativity of each review. Despite this enhancement, our correlation analysis between the newly derived sentiment scores and other numerical variables still failed to uncover meaningful relationships.

Development of a Predictive Model - The primary objective of this model is to predict ratings based on the available review text and other drug information.
About the Model – We designed our predictive model using transformers, leveraging the 'Hugging Face' library. This approach allowed us to harness the power of advanced NLP techniques for deep understanding and generation of (predicted) ratings from review texts.

After analyzing the pair plots plotted between the remaining numeric columns we couldn't find any significant correlations therefore, we proceeded to quantitatively assess the relationships by calculating Pearson's Correlation Coefficients.
As expected, we found a strong correlation of '0.81' between our model's 'predictedRating' and the actual 'rating'.

Then further exploring our analysis, we tried to assess the linear relationship between the predicted ratings from our built model and their actual ratings which would even evaluate our model's performance and also visualized this relationship using a scatter plot.
We achieved an **R Squared Score of '0.77'** which nearly indicates a **good fit.**
We also believe that training our prediction model for more epochs would give us a better fit.

Clustering – We first started off by preparing the data for our analysis, this included computing the length of each review in terms of word count and stored in separate columns then we performed dimensionality reduction using Principal Component Analysis (PCA) by reducing the dimensionality of the data to just 2 dimensions so that it becomes feasible to visualize the data points in a scatter plot.

Furtherly, we have employed various clustering methods to segment our dataset effectively. Specifically, we implemented Agglomerative Clustering, K-Means, Mini-Batch K-Means and Mean Shift clustering algorithms. To evaluate the performance and the quality of clusters formed by these methods, we calculated the silhouette scores for each.

Clustering Method	**Silhouette Score**
Agglomerative Clustering	0.4160178320097562
K-Means Clustering	0.45949033717341753
Mini-Batch K-Means	0.3369146167419203
Mean Shift Clustering	0.5552562045906827

Mean Shift Clustering performs the best of the methods evaluated above obtaining the highest silhouette score of 0.555 with 4 clusters.

Classification – In a similar manner we start off by preparing the data for classification by encoding categorical features 'condition' and 'drugName' using LabelEncoder for both training and testing datasets. By encoding this data, machine learning algorithms are able to understand and make use of the connection between the analyzed sentiment and the medical state.

Then we started our classification modelling by dropping unnecessary columns. Then through a written function we started interpreting and visualizing the performances of various classification models.

Then we chose 'Logistic Regression', 'Decision Trees' and 'KNN' classifiers as they are computationally efficient and can handle relatively large datasets.
After performing performance evaluation of our classification methods below are the accuracies of different classifiers:

Classifier	**Accuracy**
Logistic Regression	0.69
Decision Tree	0.69
KNN	0.58

Considering the performance of the considered classifiers, we thought it was an indication that the dataset's features have limited predictive power to the target variable.
Therefore, we introduced our own transformer model which can analyze and understand textual data more effectively.

About the model – We implemented a transformer-based sentiment analysis model for analyzing drug reviews. This model is trained on our dataset of drug reviews labeled with sentiment (positive or negative) based on the rating. The trained model is used to predict sentiments for new reviews, and the results are labeled accordingly.
The transformer model which was created especially to handle natural language tasks performed exceptionally well at extracting rich semantic information from the drug review texts, while other classifiers, such as Logistic Regression, Decision Tree, and KNN, found it difficult to capture the complex patterns and relationships within the textual data.
With an accuracy score of 0.93, this transformer model outperformed all these classifiers.

Classifier	**Accuracy**
Logistic Regression	0.69
Decision Tree	0.69
KNN	0.58
Transformer Model	0.93

Our Drug Review Analysis research has revealed insightful patterns and linkages among patient evaluations through painstaking analysis and the application of cutting-edge machine learning techniques. These discoveries open the door for future advancements in personalized medicine and healthcare analytics, as well as improving our understanding of medication efficacy and patient happiness.